Introduction to Applied Statistical Methods

Larry Winner
University of Florida
Department of Statistics

August 25, 2020

Contents

Chapter 1

Introduction

1.1 Basic Concepts of Statistical Analysis

Statistical tools and methods are used to describe data and make inferences regarding states of nature in a wide variety of areas of study. From simple graphs and numeric summaries provided in mainstream press to highly complex models used to describe measurements across a wide range of individuals or sampling units, we see reports making use of statistical tools and methods constantly. We will go through many of the commonly used methods in these notes.

After a brief introduction to **descriptive statistics**, making use of numeric and graphical summaries of variables, we will spend the remainder of the notes on **inferential statistics** that make use of information from a sample to make statements regarding a larger population of units. When conducting a study, researchers typically use the following strategy.

1. Define the problem/research question of interest, including what to measure and all relevant conditions or groups to study.
2. Collect the data by means of a controlled experiment, observational study, or sample survey.
3. Summarize the data numerically in tabular form and/or graphically.
4. Analyze, interpret, and communicate the study's findings.

Many methods exist for the final part, data analysis, that we describe in detail in these notes. Many factors lead to the choice of the statistical methods to use for the analysis, including: data type(s), sampling method, and distributional assumptions regarding the measurements.

Populations will be thought of as the universe of units, while **samples** will refer to subsamples of the populations that are observed and measured. In practice, we observe the sample with the goal of making **inferences** regarding the corresponding population. Consider the following examples.

- A study compared 3 electronic reader models, each at 4 illumination levels in a sample of 60 subjects, measuring the times to read a document. The goal was to compare the effects of the models and illumination levels in the general population [13].
- Many studies have been conducted involving extrasensory perception (ESP). In a typical study, there are 4 choices of what target the sender is viewing and the receiver must identify which target was being viewed. Researchers wish to determine whether the true proportion of successful trials exceeds 1/4 from a sample of trials [48].
- Studies are conducted to measure general consistency within and between evaluators when assessing common items (e.g. fingerprints, x-rays, foods/beverages) based on sampled judges and targets [18].

Note that populations can be "fixed", a well defined and identified population of units (e.g. all National Hockey League players for the 2014-2015 season) or "conceptual" (e.g. all people with a particular condition currently or in the near future). In our work, we will often make use of taking random samples from fixed populations to understand the properties of statistical procedures as they are applied to different samples from a given population.

1.2 Data Collection

Once a research question has been made, then data is collected to attempt to answer the question. Three common methods of collecting data are: controlled experiments, observational studies, and sample surveys.

In a **Controlled Experiment**, a sample of experimental units is obtained, and randomized to the various treatments or conditions to be compared. There are many ways that these can be conducted, and we will describe many variations of them throughout this course and its sequel. Some elements of controlled experiments are given here.

Factors Variable(s) that are controlled by the experimenter (e.g. new drug vs placebo, 4 doses of a pesticide, 3 packages for food product)

Responses Measurements/Outcomes obtained during the experiment (e.g. change in blood pressure, weeds killed, consumer ratings for the product)

Treatments Conditions that are generated by the factor(s). When only 1 factor, these are the levels. With 2 or more factors, these are combinations of levels.

Experimental Unit Entity that is randomized to the Treatments. These can be individual items (patients in clinical trial, plants in botanical experiment) or groups of items (classrooms of students in an education experiment, pens of animals in a feed study).

Replications Treatments are assigned to more than one experimental unit, allowing for experimental error (variation) to be measured.

Measurement Unit Entity on which measurements are obtained. These can be experimental units when individuals are randomized, or subunits within the experimental units (students in a classroom, pigs in a pen).

Controlled experiments can be conducted in laboratories/hospitals/greenhouses, but can also be conducted in the "real world" where they are often referred to as "field studies" or "natural experiments."

There are many different treatment designs that are commonly applied. Some classes of designs are given below.

Single Factor Designs In these designs, there is a single factor to be studied with various levels.

Multi Factor Designs More than one factor is varied. Treatments correspond to combinations of factor levels.

Completely Randomized Designs Experimental units are randomly assigned to treatments with no restriction on randomization.

Randomized Block Designs Experimental units are grouped into homogeneous blocks, with treatments assigned so that each block receives each treatment.

Latin Square Designs Two or more blocking factors are available.

Repeated Measure Designs Units can be assigned to each treatment or be measured at multiple occasions on the same treatment.

Note that in designs with 2 or more factors, researchers are often interested in whether the effects of the levels of one factor depend on the levels of the other factor(s). When the effects do depend on the levels of the other factor, this is referred to as an **interaction**.

Example 1.1: Electronic Reader Reading Task Times by Model and Illumination

An experiment was conducted to compare reading times for a long duration reading task (Chang, Chou, and Shieh (2013) [13]). There were two factors: e-reader model with 3 levels (Sony PRS 700, Amazon Kindle DX, iRex 1000s) and 4 illumination levels (200 lx, 500, 1000, 1500). Thus there were 12 treatments (combinations of e-reader and illumination level). There were a total of 60 subjects, who were randomly assigned so that 5 subjects were assigned to each treatment (each subject read only 1 reader under only 1 illumination level). The response was the time to read the document in seconds.

∇

In many settings, it is not possible or ethical to assign units to treatments. For instance, when comparing quality of products of various brands, you can take samples from the various brands, but not assign "raw materials" at random to the brands. Studies comparing residents of various parts of a country can only take samples of residents from the areas, not assign people to them. In studies of the effects of smoking or drinking, it is unethical to assign subjects to the conditions. In all of these cases, we refer to these as **Observational Studies**. Typically the method of analysis is the same for controlled experiments and observational studies, however the ability to imply "cause and effect" is more difficult in observational studies than controlled experiments. Researchers in such studies must try and control for any potential alternative explanations of the association. For an interesting discussion of various aspects of observational studies, including: external validity (generalizing results beyond the original study), causation, reliability of measurement, and inclusion of covariates, involving study of interruption and multitasking, see Walter, Dunsmuir, and Westbrook (2015) [51].

In many research areas, data are collected through **Sample Surveys**. In particular, they are often used in Public Opinion, by Government Bureaus, Business, and Recreational Services. Unless surveys are based on some sort of sampling based method, they are generally not reliable for making inferences regarding a population.

It should be noted that certain problems tend to arise with surveys. The primary problem is **nonresponse**. If the individuals who do not respond tend to be different from those who do respond, then any estimates of population based quantities will be biased. Also, when the questions are "sensitive" such as illegal behavior, there will tend to be **response bias**. **Recall bias** occurs when some sampled elements are more likely to recall a previous experience than others. This can effect observed associations in retrospective surveys. Needless to say wording of questions can have a large impact on responses.

Some commonly used sampling methods are as follow.

Simple Random Sampling All possible samples of size n from a population of size N are equally likely. This needs a frame listing all elements of the population and a random number generator.

Stratified Random Sampling Elements of the population are classified by group (strata) and simple random samples are taken within each group.

Cluster Sampling Elements of the population are classified by cluster (possibly physical location) and a random sample of clusters is taken. Elements within the sampled clusters are the sampled units.

Systematic Sampling When elements of the population are in a sequence, a random starting point is selected, and every k^{th} subsequent element is sampled.

Note that these techniques are often applied in combination in many government/business/political surveys. Also, these techniques generalize to taking samples of individuals or elements from any population to be observed and measured. For instance, in quality control, items may be sampled and tested from an assembly line by systematic sampling.

All methods covered in this course are based on simple random sampling. Some adjustments for estimates and standard errors are used for the other sampling plans. For a detailed and accessible coverage of sampling, see e.g. Scheaffer, Mendenhall, and Ott (1990) [46].

1.3 Variable Types

In most settings, researchers have one or more "output" variable(s) and one or more "input" variable(s). For instance, a study comparing salaries among males and females would have the output variable be salary and possible input variables: gender (1 if female, 0 if male), experience (years), and education (years). The output variables are often referred to as **dependent variables**, **responses**, or **end points**. The input variables are often referred to as **independent variables**, **predictors**, or **explanatory variables**.

Variables are measured on different scales, and the data analysis methods are determined by variable types. Variables can be **categorical** or **numeric**. Categorical variables can be **nominal** or **ordinal**, while numeric variables can be **discrete** or **continuous**.

Subject	Age	Gender	Dysphonia	Subject	Age	Gender	Dysphonia	Subject	Age	Gender	Dysphonia
1	10	M	3	11	45	F	3	21	57	F	2
2	19	M	1	12	47	F	3	22	59	F	2
3	27	F	1	13	48	M	1	23	60	F	3
4	32	M	1	14	49	F	2	24	60	M	1
5	37	F	2	15	50	F	3	25	62	F	2
6	37	M	0	16	51	F	3	26	62	M	3
7	39	F	3	17	51	M	0	27	64	F	3
8	42	F	2	18	51	M	0	28	70	M	3
9	44	F	2	19	53	F	1	29	77	F	3
10	45	F	2	20	57	F	3	30	89	F	2

Table 1.1: Age, Gender, and Dysphonia Grade for 30 Subjects - VALI Study

Examples of nominal variables include gender, hair color, and automobile make. These are categories with no inherent ordering. Ordinal variables are categorical, but with an inherent ordering, such as: strongly disagree, disagree, neutral, agree, strongly agree. Discrete variables can take on only a finite or countably infinite set of values, these can be counts of number of occurrences of an event in a series of trials or in a fixed time or space, or the number facing up on a roll of a dice. Continuous variables can take on any value along a continuum, such as temperature, time, or blood pressure. When discrete variables take on many values, they are often treated as continuous, and continuous variables are often reported as discrete values.

Example 1.2: Consistency of Ratings Based on a Rating Scale for Videostroboscopy

A study was conducted to measure inter-rater and intra-rater reliability of the Voice-Vibratory Assessment with Laryngeal Imaging (VALI) rating form for assessing videostroboscopy and high-speed videoendoscopic (HSV) recordings (Poburka, Patel, and Bless (2017) [43]). Table 1.1 contains information on the 30 subjects in the study. These include: subject ID, Age (continuous, reported as a discrete variable), gender (nominal), and an overall dysphonia grade (ordinal, with 0=normal, 1=mild, 2=moderate, 3=severe).

∇

Chapter 2

Describing Data

Once data have been collected, it is typically described via graphical and numeric means. The methods used to describe the data will depend on its type (nominal, ordinal, or numeric). We also need to distinguish whether the data corresponds to a sample or a population. In this chapter, we focus purely on describing a set of measurements, not making inferences. First we consider graphical and numeric descriptions of a single variable. Then we consider pairs of variables.

2.1 Graphical Description of a Single Variable

Depending on the type of measurement, common plots are **pie charts**, **bar charts**, **histograms**, **box plots**, and **density plots**.

Pie charts can be used to describe any variable type. Continuous numeric variables must be collapsed into "bins" or "buckets." The size of the sectors of the pie represent the relative frequency of each category.

Bar charts are used to describe nominal or ordinal data. The variable levels are arrayed on the bottom (or left side) of the plot and bars above (or beside) the levels represent the frequency or relative frequency of the number of observations belonging to the various categories.

Histograms are used for numeric variables, where the heights of the bars above the bins represent the frequency or relative frequency of the various bins.

Box plots are used on numeric variables. They identify particular percentiles of a distribution and are useful in detecting outlying observations and spread in the distribution.

Density plots are used for numeric variables. They represent smoothed histograms describing the proportion of measurements within some distance of each point on the continuum.

Example 2.1: Charlotte, NC Traffic Stops - 2016

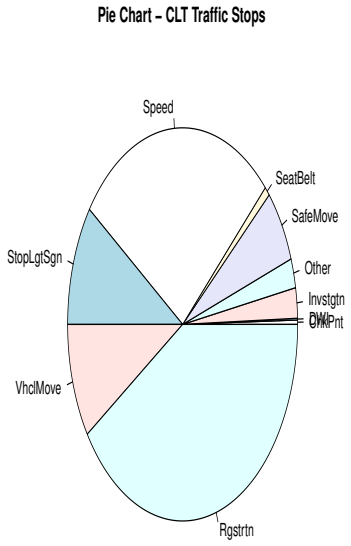

Figure 2.1: Pie Chart for Charlotte, NC traffic stops by Reason for Stop

Data for a population of 79884 traffic stops in Charlotte, North Carolina in 2016 were obtained from Data.gov. There were 10 possible reasons for the traffic stops (including a category 'Other'). A pie chart (Figure 2.1) and a bar chart (Figure 2.2) are displayed. Note that the pie chart does a very poor job with the categories "DWI" and "Check Point." Pie charts should generally be avoided. It is clear that Registration and Speed violations are the most often occurring reasons.

R Commands and Output

```
### Output

> (table.RsnStop <- table(RsnStop))
RsnStop
    ChkPnt        DWI   Invstgtn      Other   SafeMove   SeatBelt      Speed
       286        114       1992       1926       4827        631      22222
StopLgtSgn   VhclMove    Rgstrtn
      7946       7535      32405
```

∇

Example 2.2: Body Mass Index for National Hockey League Players - 2013/2014 Season

Body mass index (BMI) is a measure of body fat that is based on the the work of Adolphe Quetelet, a renowned Belgian researcher in astronomy and statistics and other areas, particularly social sciences. In terms

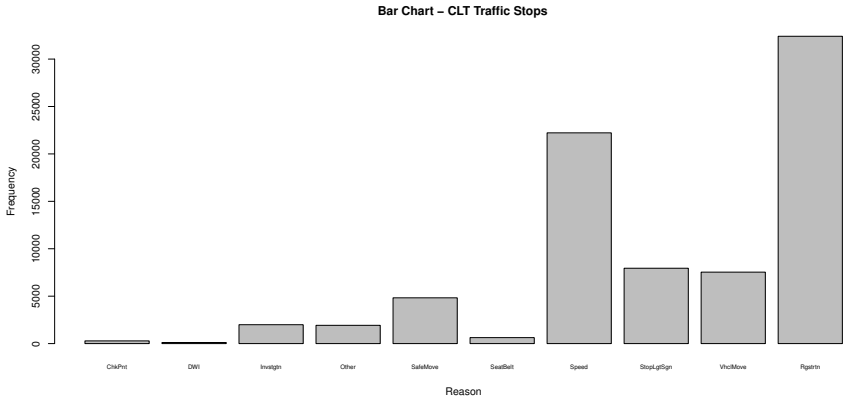

Figure 2.2: Bar Chart for Charlotte, NC traffic stops by Reason for Stop

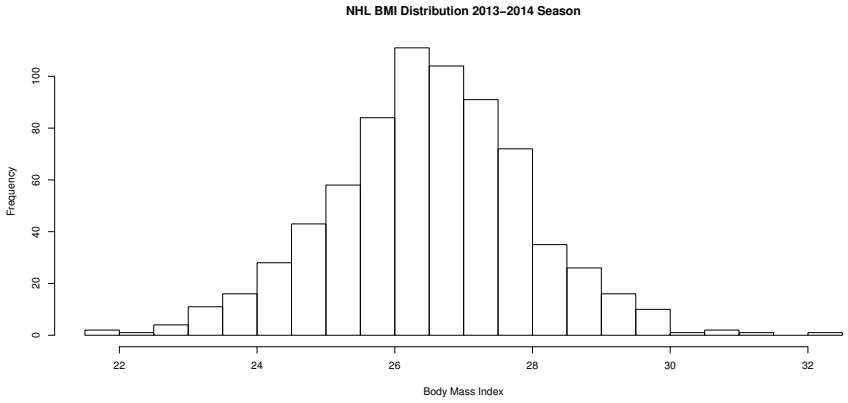

Figure 2.3: Body Mass Index for 2013/2014 season National Hockey League Players

of metric units, BMI is mass(kg)/height(m)^2; in the American system, BMI is 703*mass(lbs)/height(in)^2. Data for all National Hockey League (NHL) players are obtained, reported in pounds (lbs) and inches, discretely. A histogram is given in Figure 2.3. The histogram is approximately symmetric and mound-shaped, centered between 26 and 28.

∇

Example 2.3: Female and Male Speeds at Washington, DC Rock and Roll Marathon - 2015

The 2015 Rock and Roll Marathon in Washington, D.C. was completed by 1045 female and 1454 male participants. Each participant's time to complete the marathon was converted to a speed (miles per hour). Histograms and kernel density plots for females and males are given in Figure 2.4, and side-by-side box plots are given in Figure 2.5. For both genders, there tend to be more cases at lower speeds with a few extreme cases with higher speeds. These distributions are **right-skewed**. The box-plot identifies from bottom to top the following elements.

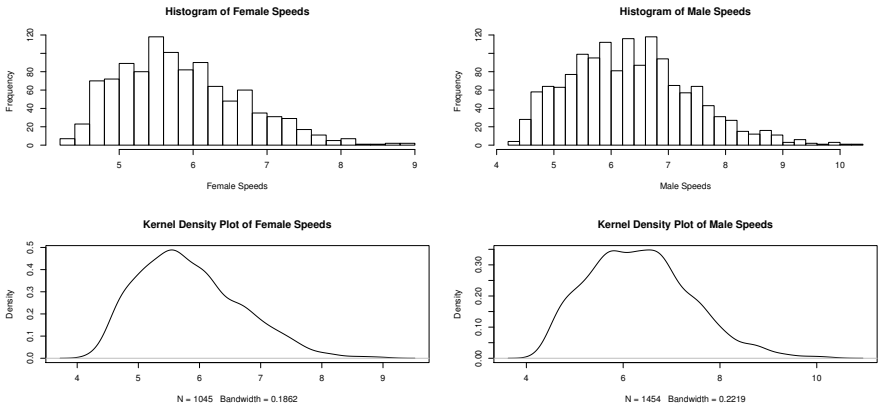

Figure 2.4: Histograms and density plots of Rock and Roll marathon speeds by gender

1. Minimum: Bottom of line at bottom of plot
2. Range for slowest 25% of participants: Line below box
3. 25th percentile: Bottom line of box
4. Range for the 25th to 50th percent of participants: Between bottom of box and second horizontal line
5. Median (50th percentile): Second horizontal line
6. Range for the 50th to 75th percent of participants: Between second horizontal line and top of box
7. 75th percentile: Top line of the box
8. Range for 75th to 100th percent of participants: Line extends to either the Maximum speed or 1.5 times the distance between 75th and 25th percentiles (height of the box), whichever is lowest. Circles represent outlying measurements (very fast runners).

A smooth version of a boxplot, which does not separate the measurements into quantiles is a **violin plot**. For the marathon data, one is displayed in Figure 2.6.

R Output

```
### Output
> ## Obtain mean and standard deviation by gender
> tapply(mph,Gender,mean)
       F        M
5.839839 6.336979
> tapply(mph,Gender,sd)
        F         M
0.8310405 1.0576868
```

∇

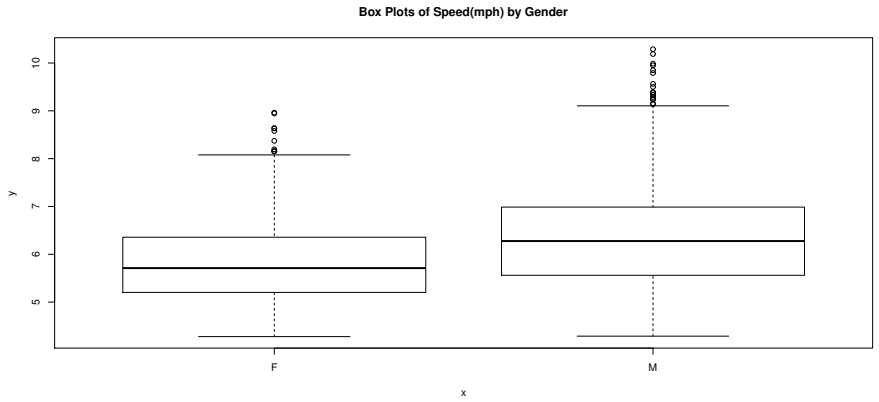

Figure 2.5: Side-by-side box plots of Rock and Roll marathon speeds by gender

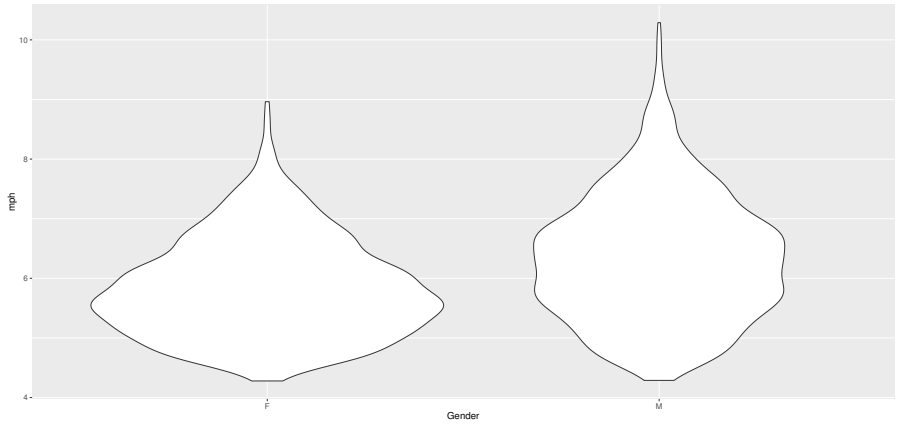

Figure 2.6: Side-by-side violin plots of Rock and Roll marathon speeds by gender

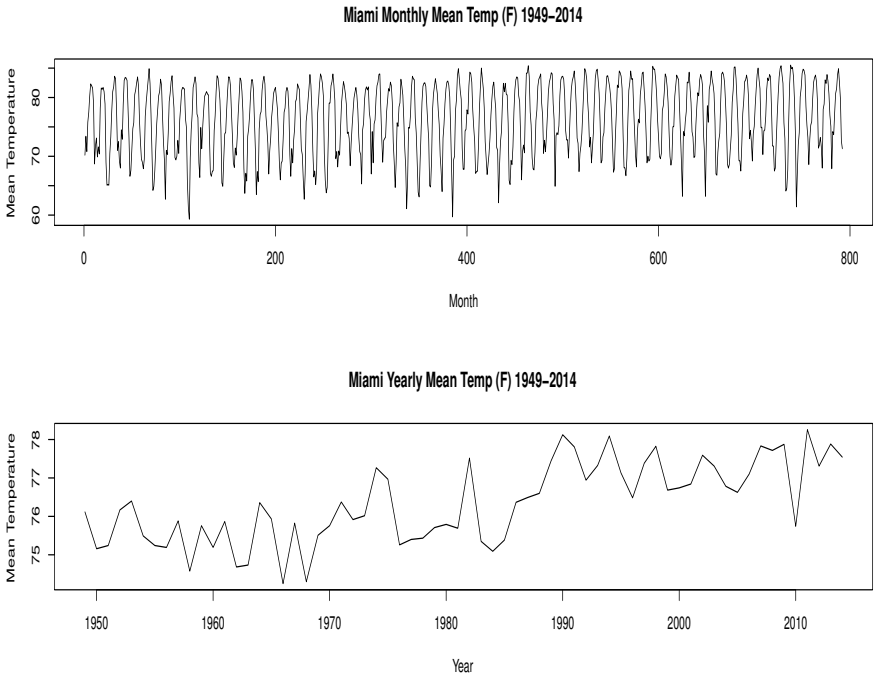

Figure 2.7: Monthly Mean Temperature in Miami, FL (January 1949 - December 2014)

Time series plots are widely used in many areas including economics, finance, climatology, and biology. These graphs include one or more characteristics being observed in a sequential time order. These plots can be based on virtually any level of sampling interval. They can be used to detect trend and cyclical patterns over time. Figure 2.7 shows the the monthly and annual mean temperature in Miami for the years 1949 through 2014. Clearly there is a cyclical pattern occurring within years, and after a flat early annual series, there certainly appears to be evidence of an increasing trend over approximately the second half of the series (after about 1970).

R Output

```
### Output (condensed)
> (yearMeanTemp <- aggregate(meantemp ~ year, mw1, mean))
   year meantemp
1  1949 76.11667
2  1950 75.15833
3  1951 75.24167
...
64 2012 77.30833
65 2013 77.88333
66 2014 77.54167
```

Data maps are very popular as more and more spatial datasets are available. Figure 2.8 displays Bigfoot sightings for the 50 United States.

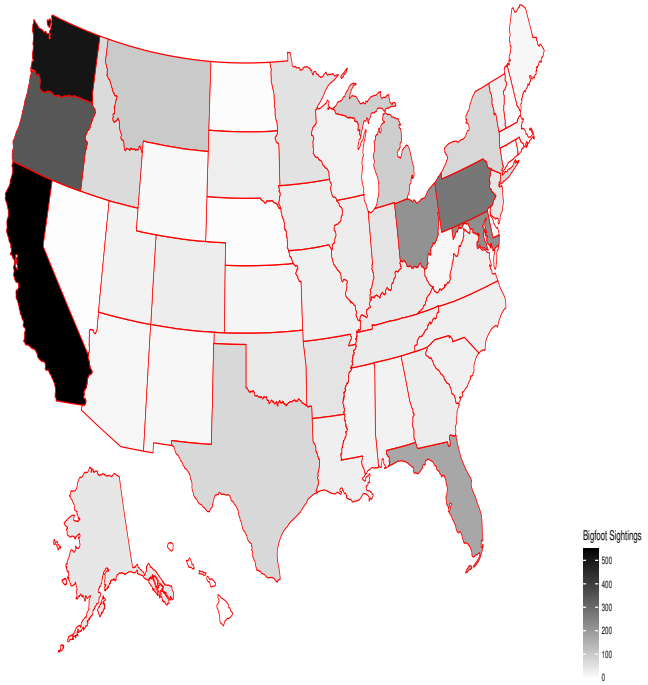

Figure 2.8: Bigfoot sightings by US state

2.2 Numerical Descriptive Measures of a Single Variable

Numerical descriptive measures describe a set of measurements in quantitative terms. When describing a **population** of measurements, they are referred to as **parameters**; when describing a **sample** of data, they are referred to as **statistics**.

In terms of nominal and ordinal data, **proportions** are generally the numeric measures of interest. These are simply the fraction of measurements falling into the various possible levels (and must sum to 1). For ordinal variables, the **cumulative proportions** are also of interest, representing the fraction of measurements falling in or below the various categories.

Example 2.4: CLT Traffic Stops and the VALI Laryngeal Study

For the Charlotte traffic stops, there were 10 categories for the reason for the stop. These reasons are treated as nominal, as there is no inherent ordering of the levels.

R Commands and Output are given below. The **table** function counts the number of cases (traffic stops) that are of each category, and dividing by their sum turns them into proportions.

R Output

```
### R Commands/Output (using previous dataset)
(table.RsnStop <- table(RsnStop))
round(table.RsnStop / sum(table.RsnStop), 5)

> (table.RsnStop <- table(RsnStop))
RsnStop
    1     2     3     4     5     6     7     8     9    10
  286   114  1992  1926  4827   631 22222  7946  7535 32405
> round(table.RsnStop / sum(table.RsnStop), 5)
RsnStop
      1       2       3       4       5       6       7       8       9      10
0.00358 0.00143 0.02494 0.02411 0.06043 0.00790 0.27818 0.09947 0.09432 0.40565
```

For the VALI study, the ordinal dysphonia rating had levels: 0, 1, 2, 3. There were 3, 6, 9, and 12 cases for those categories (total of 30 subjects). The proportions for the categories are:

$$0 : 3/30 = .10 \quad 1 : 6/30 = .20 \quad 2 : 9/30 = .30 \quad 3 : 12/30 = .40$$

The cumulative proportions (at or below that score) are:

$$0 : .10 \quad 1 : .10 + .20 = .30 \quad 2 : .30 + .30 = .60 \quad 3 : .60 + .40 = 1.00$$

In these examples, the traffic stop data can be thought of as a population (all traffic stops in Charlotte, N.C. in 2016), and the VALI dysphonia data is most certainly a sample.

∇

2.2.1 Measures of Central Tendency

There are two commonly reported measures of central tendency, or location for a set of measurements. The **mean** is the sum of all measurements divided by the number of measurements, and is reported often as "per capita" in economic reports. The mean is the "balance point" of a set of measurements in a physical sense. The **median** is the point where half of the measurements fall at or below it, and half of the measurements fall at or above it. It is also the 50th percentile of the set of measurements. Many economic reports state median values. A third, less reported measure is the **mode** which really is only appropriate for discrete variables, and is the value that occurs most often. For a histogram of discretely measured data, the mode is the level with the highest bar.

Note that the mean is affected by outlying measurements, as it is the sum of all measurements, evenly distributed among all of the measurements. The median is more "robust" as it is not affected by the actual values of individual measurements, only the center of them. The formulas for the population mean μ, based on a population of N items and the sample mean $\overline{y}$ for a sample of n items are given below.

$$\text{Population Mean: } \mu = \frac{\sum_{i=1}^{N} y_i}{N} \qquad \text{Sample Mean: } \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

To obtain the median, measurements are ordered from smallest to largest, and the middle observation (odd population/sample size) or the average of the middle two observations (even population/sample size) are identified.

Example 2.5: NHL BMI's and Rock and Roll Marathon Speeds

Using the **mean** and **median** functions in R, we obtain the population means for NHL BMI's and marathon speeds by gender for the Rock and Roll marathon.

R Output

```
### Output

> cbind(head(bmi.nhl.sort), tail(bmi.nhl.sort))
         [,1]     [,2]
[1,] 21.56757 29.98314
[2,] 21.75521 30.12259
[3,] 22.14871 30.51215
[4,] 22.64680 30.82813
[5,] 22.75987 31.39688
[6,] 22.75987 32.00386
> round(bmi.cent.out, 4)
       N      sum    mean  median
[1,] 717 19000.61 26.5002 26.5159
>
> ### Use built-in mean and median functions
> mean(bmi.nhl)
[1] 26.50015
> median(bmi.nhl)
[1] 26.51586
```

Note that the mean (26.50) and median (26.52) are very close, as is expected for an (approximately) symmetric distribution.

For the marathon speeds, we use the **tapply** function in R that will compute functions separately for different groups (gender).

R Output

```
> tapply(mph,Gender,mean)
       F        M
5.839839 6.336979
> tapply(mph,Gender,median)
       F        M
5.711109 6.276599
```

These distributions are skewed-right, with a few very fast runners in each gender. This causes the means (F=5.84, M=6.37) to be larger than the medians (F=5.71, M=6.28).

∇

Outliers are observations that lie "far" away from the others. These may be data that have been entered erroneously or just individual cases that are quite different from others. As stated above, means can be affected by outliers, while medians generally are not. A measure of the mean that is not affected by outliers is the **trimmed mean**. This is the mean of observations in the "middle" of the measurements. For instance, 90% trimmed mean is the mean of the middle 90% of the ordered measurements (removing the smallest 5% and largest 5%).

2.2.2 Measures of Variability

Along with the "location" of a set of measurements, researchers are also interested in their variability (aka dispersion). The **range** is the distance between the largest and smallest measurements (note that this differs from the standard meaning which would just give the lowest and highest values). The **interquartile range** (IQR) is the distance between the 75th percentile (3/4 of measurements lie below it) and the 25th percentile (1/4 of the measurements lie below it). That is, the IQR measures the range for the middle half of the ordered measurements.

Measures that are more widely used in making inferences are the **variance** and its square root, the **standard deviation**. In terms of measurements, the variance is approximately the average squared distance of the individual measurements from the mean (for a population, it is the average). The formulas for the population and sample variance are given below. Note that unless stated otherwise specifically, software packages are reporting the sample version.

$$\text{Population Variance: } \sigma^2 = \frac{\sum_{i=1}^{N}(y_i-\mu)^2}{N} \qquad \text{Sample Variance: } s^2 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n-1}$$

The reason for dividing by $n-1$ in the sample variance is to make the estimator an unbiased estimator for the population variance. That is, when computed across all possible samples, the "average" of the sample variance will be the population variance. The standard deviation is the positive square root of the variance and is in the same units as the measurements. The population standard deviation is denoted as σ, the sample standard deviation is denoted as s. For many (but certainly not all) distributions, approximately 2/3 of the measurements lie within one standard deviation of the mean and approximately 19/20 lie within two standard deviations of the mean.

Example 2.6: NHL BMI's and Rock and Roll Marathon Speeds

We compute the range, interquartile range, variance, and standard deviations for the NHL BMI's and the Rock and Roll mathon speeds by gender. Since we treat each of these as a population, we will make a slight adjustment to R's "built-in" functions **var** and **sd**, which compute the sample versions by default.

R Output

```
### Output
> var(bmi.nhl)              # Sample Variance with "var" function
[1] 2.116228
> (N-1)*var(bmi.nhl)/N      # Pop variance with "var" function
[1] 2.113277
```

```
> sd(bmi.nhl)               # Sample Std Dev with "sd" function
[1] 1.454726
> sqrt((N-1)/N)*sd(bmi.nhl)  # Population Std Dev with "sd" function
[1] 1.453711
> round(bmi.var.out1, 3)
       min    max  range    LQ     UQ   IQR
    21.568 32.004 10.436 25.62 27.439 1.819
> round(bmi.var.out2, 3)
     mean sum(dev^2) sigma^2   s^2 sigma     s P(mu+/-1sigma) P(mu+/-2sigma)
[1,] 26.5   1515.219   2.113 2.116 1.454 1.455          0.706          0.946
```

For the marathon speeds, we will simply use the **var** and **sd** functions in R, applied separately to Females and Males. As both population sizes exceed 1000, the adjustment for population variances and standard deviations would be very small.

R Output

```
### Output
> round(rr.var.out, 3)
           N  mean sigma^2 sigma P(mu+/-1sigma) P(mu+/-2sigma)
Females 1045 5.840   0.691 0.831          0.662          0.964
Males   1454 6.337   1.119 1.058          0.665          0.964
```

Male speeds tend to be higher and more variable than Female speeds. All three distributions have approximately 2/3 of individuals lying with one standard deviation of the mean, and approximately 95% lying within two standard deviations from the mean.

∇

Two other measures of variation are given here. The **median absolute deviation** (MAD) is the median absolute deviation to the sample (population) median. When data are from a normal (Gaussian) distribution, this should be approximately 0.6745σ. The other is the **coefficient of variation** (CV), which is the ratio of the standard deviation to the mean (and is sometimes reported as a percentage). The coefficient of variation is often reported as a measure of the accuracy of laboratory equipment.

Example 2.7: NHL BMI's and Rock and Roll Marathon Speeds - MAD and CV

Here MAD and CV are computed for the three datasets. Note that the MAD for the NHL BMI's, when divided by 0.6745 is 1.364, while $\sigma = 1.454$, so they are similar, as expected as the BMI distribution is well approximated by a normal distribution. The CV is .055, so that the magnitude of the standard deviation is 5.5% of the mean.

The output for the Rock and Roll marathon speeds is given as well, by gender. The MAD's divided by 0.6745 are almost identical to the population standard deviations. The CV's are between 14 and 17 percent, reflecting that the spread of the distributions relative to the mean are higher than the NHL body mass indices.

R Output

```
### Output
> round(mad.cv.out, 3)
     median  MAD MAD/0.6745 sigma   mu    CV
[1,] 26.516 0.92      1.364 1.454 26.5 0.055

> round(mad.cv.out, 3)
        median   MAD MAD/0.6745 sigma    mu    CV
Females  5.711 0.564      0.836 0.831 5.840 0.142
Males    6.277 0.714      1.059 1.057 6.337 0.167
```

∇

2.2.3 Higher Order Moments

Two other measures are occasionally reported: **skewness** and **kurtosis**. Skewness is used to measure the symmetry of the distribution, and kurtosis measures the heaviness of the tails of the distribution. Positive values for skewness correspond to right-skewed distributions, while negative values correspond to left-skewed distributions. Negative values of kurtosis imply a distribution has fewer extreme values (lighter tails) than a normal distribution, while positive values imply more extreme values (heavier tails) than a normal distribution. These measures are reported in many fields, and are especially important in financial modeling. For a set of measurements, the skewness and kurtosis are computed as follow.

Population Skewness: $\frac{\mu_3}{\sigma^3}$ $\mu_3 = \frac{\sum_{i=1}^{N}(y_i-\mu)^3}{N}$ Sample Skewness: $\frac{m_3}{s^3}$ $m_3 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^3}{n}$

Population Kurtosis: $\frac{\mu_4}{\sigma^4}-3$ $\mu_4 = \frac{\sum_{i=1}^{N}(y_i-\mu)^4}{N}$ Sample Kurtosis: $\frac{m_4}{s^4}-3$ where $m_4 = \frac{\sum_{i=1}^{n}(y_i-\overline{y})^4}{n}$

Under normality, the standard errors for the sample skewness and sample kurtosis are given below, and depend only on the sample size.

$$\hat{SE}\left\{\frac{m_3}{s^3}\right\} = \sqrt{\frac{6n(n-1)}{(n-2)(n-1)(n+3)}} \qquad \hat{SE}\left\{\frac{m_4}{s^4}-3\right\} = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n-1)(n+3)(n+5)}}$$

Example 2.8: NHL BMI's and Rock and Roll Marathon Speeds

Skewness and kurtosis and their standard errors (temporarily treating them as samples) are computed for the three datasets here.

R Output

```
### NHL BMI Output
```

```
> round(skew.kurt.out, 4)
        mu3 Skewness SE{Skew}     mu4 Kurtosis SE{Kurt}
[1,] -0.1098  -0.0357   0.0914 15.5403   0.4797   0.0068

### Rock and Roll Marathon Output

> round(skew.kurt.out, 4)
           mu3 Skewness SE{Skew}    mu4 Kurtosis SE{Kurt}
Females 0.3617   0.6302   0.0757 1.4835   0.1103   0.0047
Males   0.5792   0.4895   0.0642 3.8334   0.0631   0.0034
```

Skewness is very close to 0 for the NHL BMI data, as expected from the histogram. The skewnesses for the Female and Male marathon speeds are positive, and well away from 0, again consistent with their histograms. The kurtosis for the NHL BMI data is greater than 0, corresponding to heavier tails than a normal distribution; the measures for marathon speeds are closer to 0.

∇

2.3 Describing More than One Variable

So far, we have looked at cases one variable at a time, although the marathon speed data set has two variables: speed and gender. Now we consider describing relationships when two variables are observed on each sampling/experimental unit. These can be extended to more than two variables, but can be harder to visualize. We consider graphical techniques as well as numerical measures. Keep in mind that variable types (nominal, ordinal, and numeric) will dictate which method(s) is (are) appropriate.

When both variables are categorical (nominal or ordinal), two methods of plotting them are **stacked bar graphs** and **cluster bar graphs**. For the stacked bar graph, one variable is on the horizontal axis (one slot for each level) and the other variable is displayed within the bars with subcategories for each of its levels. In a cluster (grouped) bar graph, one variable forms "major groupings," while the second variable is plotted "side-by-side" within the groupings. Both methods are based on results of a **contingency table** also known as a **crosstabulation**. These are tables where rows are the levels of one categorical variable, columns are levels of another variable, and numbers within the table are counts of the number of units falling in that cell (combination of variable levels). Often these are converted into proportions either overall (cell probabilities sum to 1), or within rows or columns marginally. A contingency table is typically of the form in Table 2.1.

Example 2.9: Thumb Styles of Blues Guitarists by Region and Period

A study reported hand and thumb styles of Blues guitarists as well as the region they were from and when they were born (Cohen (1996) [16]). The regions are 1=East, 2=Delta, and 3=Texas. The thumb styles are 1=Alternating, 2=Utility, and 3=Dead. The birth period was labeled post1906 with 0=Born before 1906, 1=born after 1906. First, the association between region (row) and thumb style (column) is considered, then birth period is added. The crosstabulations are given below in the R code. Figure 2.9 gives the Stacked and Cluster Bar Graphs.

R Output

		Column				
		1	2	$\cdots$	c	Total
Row	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
	Total	$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..}$

Table 2.1: Contingency Table for Row Variable with r levels, and Column variable with c columns

```
### Output

> (reg_ts <- table(region, thumbSty))
       thumbSty
region  Alternating Utility Dead
  East           20       8    7
  Delta           9      19   19
  Texas           1       2    8
> ## Obtain Row (1) and Column (2) Marginal Totals
> margin.table(reg_ts,1)
region
 East Delta Texas
   35    47    11
> margin.table(reg_ts,2)
thumbSty
Alternating     Utility        Dead
         30          29          34
> ## Obtain Proportions across all Cells
> reg_ts/sum(reg_ts)
       thumbSty
region  Alternating    Utility       Dead
  East   0.21505376 0.08602151 0.07526882
  Delta  0.09677419 0.20430108 0.20430108
  Texas  0.01075269 0.02150538 0.08602151
> ## Obtain Row Proportions (Thumb Style w/in Region)
> prop.table(reg_ts,1)
       thumbSty
region  Alternating    Utility       Dead
  East   0.57142857 0.22857143 0.20000000
  Delta  0.19148936 0.40425532 0.40425532
  Texas  0.09090909 0.18181818 0.72727273
> ## Obtain Column Proportions (Region w/in Thumb Style)
> prop.table(reg_ts,2)
       thumbSty
region  Alternating    Utility       Dead
  East   0.66666667 0.27586207 0.20588235
  Delta  0.30000000 0.65517241 0.55882353
  Texas  0.03333333 0.06896552 0.23529412
```

If there are three or more categorical variables, then tables of higher order dimensions and **mosaic plots** can be constructed. Here we consider the three variables: Post1906, thumb style, and region. The mosaic plot is constructed within the **vcd** (visualizing categorical data) package and is shown in Figure 2.10.

R Output

```
### Output
```

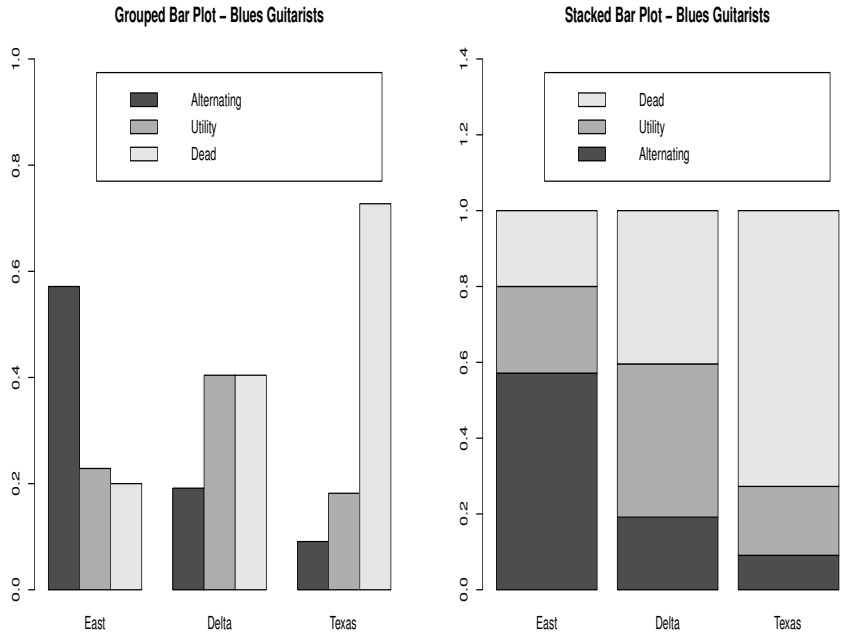

Figure 2.9: Stacked and Cluster (Grouped) Bar Charts - Blues Guitarists - Region and Thumb Style

```
> table(post1906,region,thumbSty)
, , thumbSty = Alternating

        region
post1906 East Delta Texas
       0    5     7     1
       1   15     2     0

, , thumbSty = Utility

        region
post1906 East Delta Texas
       0    1     8     0
       1    7    11     2

, , thumbSty = Dead

        region
post1906 East Delta Texas
       0    4    11     5
       1    3     8     3
```

∇

When the independent variable is categorical (nominal or ordinal) and the response (dependent variable) is numeric, we can construct side-by-side histograms and density plots (see Figure 2.4), box plots (see

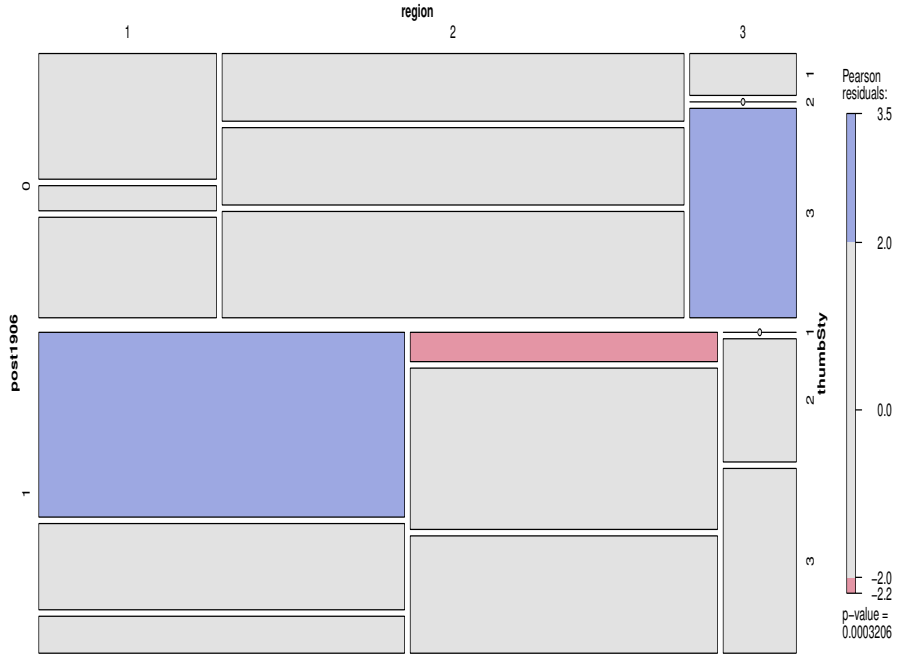

Figure 2.10: Mosaic plot for Blues Guitarist Data - Birth Period is on Left Axis, Region on Top Axis, Thumb Style on Right Axis

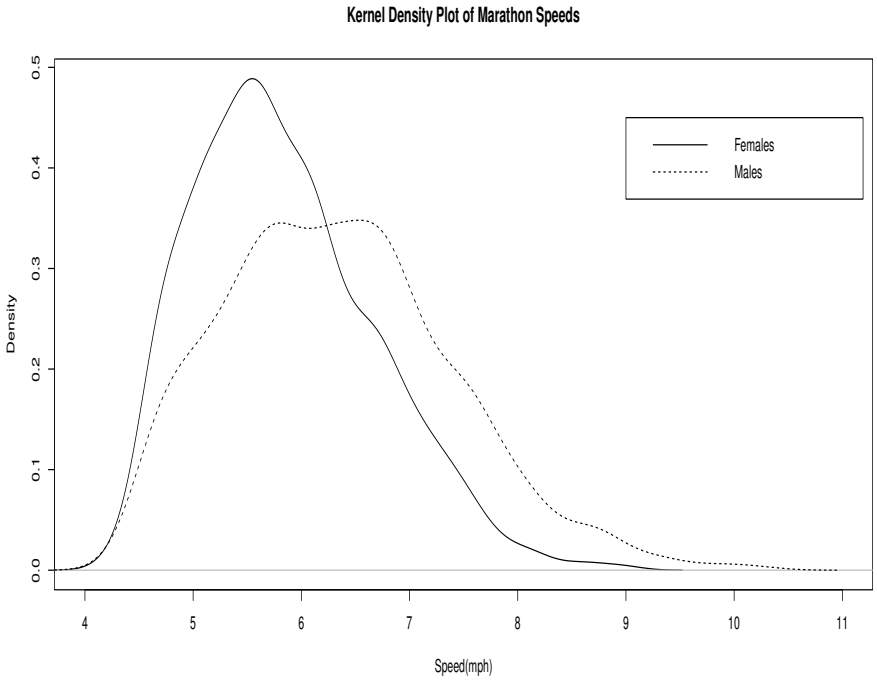

Figure 2.11: Density plot for Females and Males - Rock and Roll Marathon Speeds

Figure 2.5), or violin plots (see Figure 2.6). Histograms and densities can also be placed into single plots with different colors or patterns.

Example 2.10: Rock and Roll Marathon Speeds by Gender

A density plot using basic plotting functions in R is displayed in Figure 2.11, and a combined histogram using the **ggplot2** package is given in Figure 2.12.

∇

When two variables (labeled x and y) are both numeric, one numeric descriptive measure that is widely reported is the **correlation** between the two variables. Technically, this is called the Pearson product moment coefficient of correlation. This measure is only for the **linear**, or "straight line" relation between the two variables. Unlike in Regression (described later), the variables are not necessarily (but can be) identified as an independent and or dependent variable. The formula for this measure (population and sample) are given below.

$$\text{Population Correlation: } \rho = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sigma_x\sigma_y} = \frac{\sum_{i=1}^{N}(x_i-\mu_x)(y_i-\mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu_x)^2\sum_{i=1}^{N}(y_i-\mu_y)^2}}$$

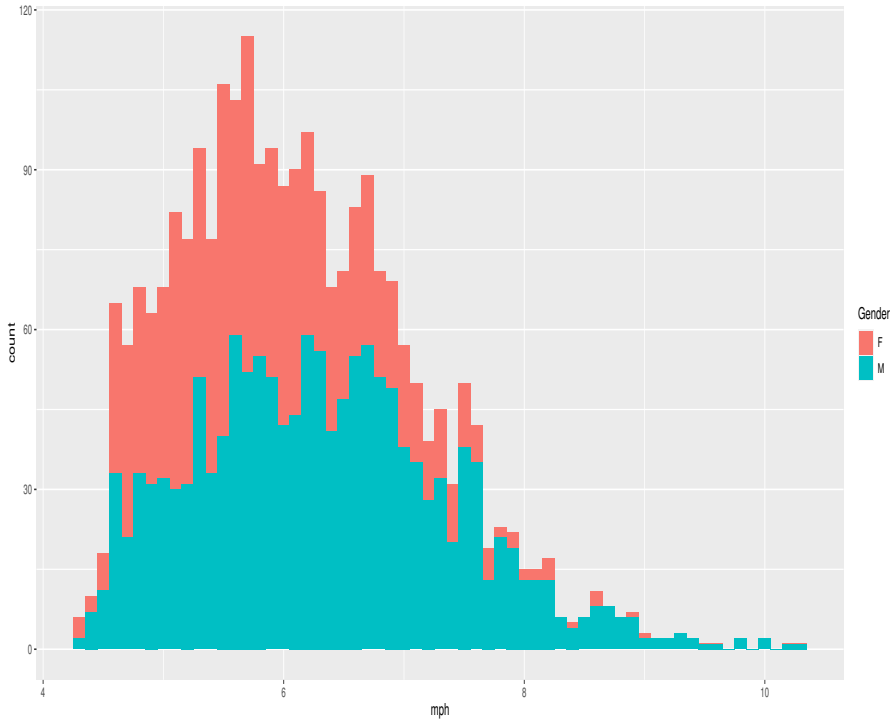

Figure 2.12: Combined Female/Male Histogram for Rock and Roll marathon speeds

$$\text{Sample Correlation: } r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{s_x s_y} = \frac{\sum_{i=1}^{N}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\overline{x})^2\sum_{i=1}^{N}(y_i-\overline{y})^2}}$$

A **scatterplot** is a plot where each case's x and y pairs are plotted in two dimensions. When one variable is the dependent variable, it is labeled y, and plotted on the vertical axis and the independent variable is labeled x, plotted on the horizontal axis. We are interested in any pattern (linear or possibly nonlinear, or none at all) between the variables.

Example 2.11: Software Project Development - Size and Effort of Projects

A pair of studies considered the size (number of function points) and the effort needed for completion (hours) for 17 software development projects (Jeffery and Stathis (1996) [28] and Jorgensen, Indahl, and Sjoberg (2003) [29]). The data are given in Table 2.2. Note that Project 17 is much larger than the others and was not used in the Jorgensen paper. We consider data with and without that case, and also data based on natural logarithms of size (x) and effort (y). For the full dataset, based on the original scale, we obtain a correlation of $r = .9752$, see calculations in Table 2.2, based on an Excel spreadsheet. Also, for the full dataset, based on natural logarithms of size and effort (which often helps meet model assumptions when data are skewed with extreme case(s), as here), we find the correlation to be $r = .8791$. This was obtained using the **correl** built-in function in Excel. Plots of the four cases (original/log scale and with/without Project 17) are given in Figure 2.13, along with the "least squares regression line", which minimizes the error sum of squares (SSE), obtained as follows.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \qquad \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \qquad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \qquad SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\left(y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right)^2$$

The plots were obtained in R, and the correlations for the 4 cases were obtained using the **cor** function. The **abline** command after each **plot** command adds the least squares regression line described above.

R Output

```
### Text Output

> cor(sizeProj,effortProj)
[1] 0.9752405
> cor(sizeProj[1:16],effortProj[1:16])
[1] 0.9261634
> cor(log(sizeProj),log(effortProj))
[1] 0.8791134
> cor(log(sizeProj[1:16]),log(effortProj[1:16]))
[1] 0.8131933
```

Note that the extreme Size of Project 17 had the impact of pulling the regression line toward its Effort level and tended to increase the correlation. That project has high "leverage" on the calculated regression line.

∇

We often are interested in relationships among more than two numeric variables. Scatterplot and correlation matrices can be constructed to demonstrate the bivariate association of all pairs of variables.

Example 2.12: Compressive Strength and Microfabric Properties of Amphibolites

A study (Ali, Guang, and Ibrahim (2014) [5]) reported the relationship between Uniaxial Compression Strentgh (UCS) and 8 predictor variables including: percent hornblende (hb), grain size (gs), and grain area (ga). A simple scatterplot matrix of plots of all pairs of these four variables is given in Figure 2.14. The correlation matrix is given along with R code below. Note that this can be extended to all pairs of variables, the plot just gets very difficult to focus on particular pairs of variables.

R Output

```
### Text Output

> cor(rs1[,c(2,6,7,8)])
           UCS         hb         gs         ga
UCS  1.0000000  0.6935996 -0.8535317 -0.8537215
hb   0.6935996  1.0000000 -0.7200409 -0.6641698
gs  -0.8535317 -0.7200409  1.0000000  0.9845240
ga  -0.8537215 -0.6641698  0.9845240  1.0000000
```

projID	size(x)	effort(y)	$x-\overline{x}$	$y-\overline{y}$	$(x-\overline{x})(y-\overline{y})$	x^*=ln(x)	y^*=ln(y)
1	1164	3777	612.76	1683.71	1031715.54	7.06	8.24
2	1834	4389	1282.76	2295.71	2944850.48	7.51	8.39
3	388	1647	-163.24	-446.29	72850.95	5.96	7.41
4	336	1318	-215.24	-775.29	166870.66	5.82	7.18
5	116	529	-435.24	-1564.29	680836.01	4.75	6.27
6	182	691	-369.24	-1402.29	517776.48	5.20	6.54
7	65	291	-486.24	-1802.29	876339.01	4.17	5.67
8	160	448	-391.24	-1645.29	643697.13	5.08	6.10
9	185	262	-366.24	-1831.29	670684.54	5.22	5.57
10	168	415	-383.24	-1678.29	643181.54	5.12	6.03
11	422	2070	-129.24	-23.29	3010.42	6.05	7.64
12	296	1947	-255.24	-146.29	37339.42	5.69	7.57
13	129	1500	-422.24	-593.29	250509.72	4.86	7.31
14	143	1114	-408.24	-979.29	399782.42	4.96	7.02
15	38	362	-513.24	-1731.29	888561.25	3.64	5.89
16	89	921	-462.24	-1172.29	541875.72	4.49	6.83
17	3656	13905	3104.76	11811.71	36672567.54	8.20	9.54
Mean	551.24	2093.29		Sum/(n-1)	2940153.05		
SD	923.09	3265.97		Correlation	0.9752	Correlation	0.8791

Table 2.2: Software Projects Sizes and Effort Levels and Correlation Calculations

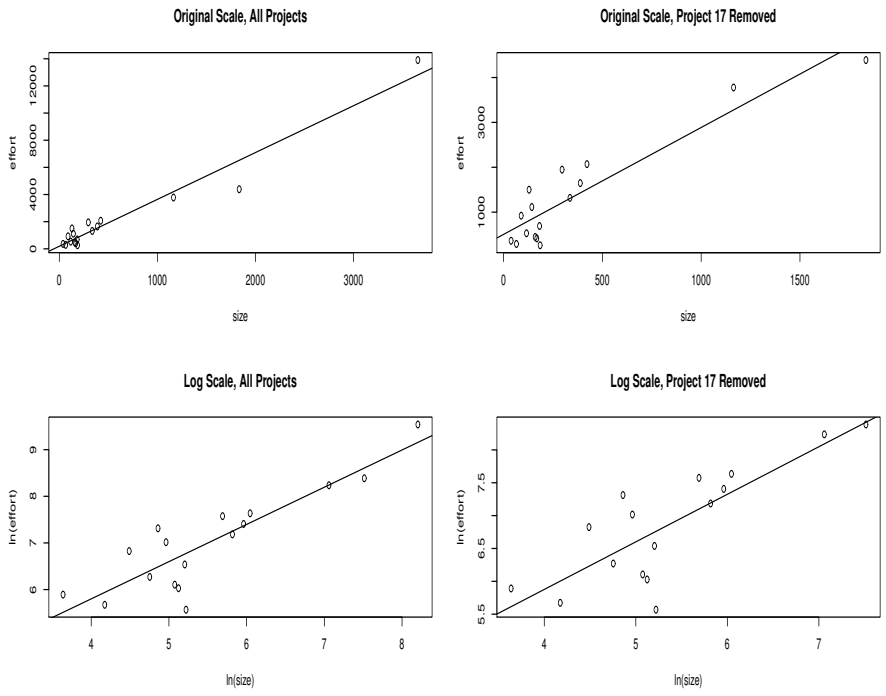

Figure 2.13: Plots of Effort (y) versus Size (x) for Original/log scales and with/without Project 17

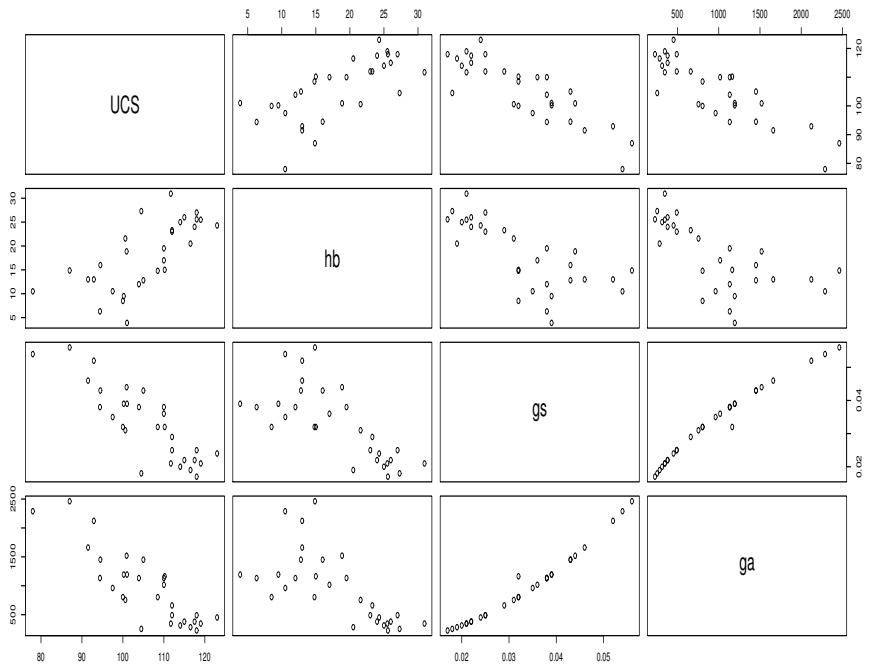

Figure 2.14: Bivariate Plots of Uniaxial Compression Strength (UCS), Percent Hornblende (hb), Grain Size (gs), and Grain Area (ga)

∇

When data are highly skewed, as in the software development example, individual cases have the ability to have a large impact on the correlation coefficient. An alternative measure that is widely used is the Spearman Rank Correlation Coefficient (aka Spearman's rho). This coefficient is computed by ranking the x and y values from 1 (smallest) to n or N (largest), and applying the formula for Pearson's coefficient to the ranks. This way, extreme x or y values do not have as large of an impact on the coefficient. Also, in many situations, the natural measurements are the rankings or ordering themselves.

Example 2.13: NASCAR Start and Finish Positions 1975-2003

A study of NASCAR races for the years 1975-2003, considered the correlation between starting and finishing positions among drivers for the 898 races during those seasons (Winner (2006) [52]). As the data were orderings, it was natural to compute the correlation using Spearman's rank correlation. The summary of the correlations is given below, and a density plot and histogram are given in Figure 2.15.

R Output

```
### Output
> length(spearman)
```

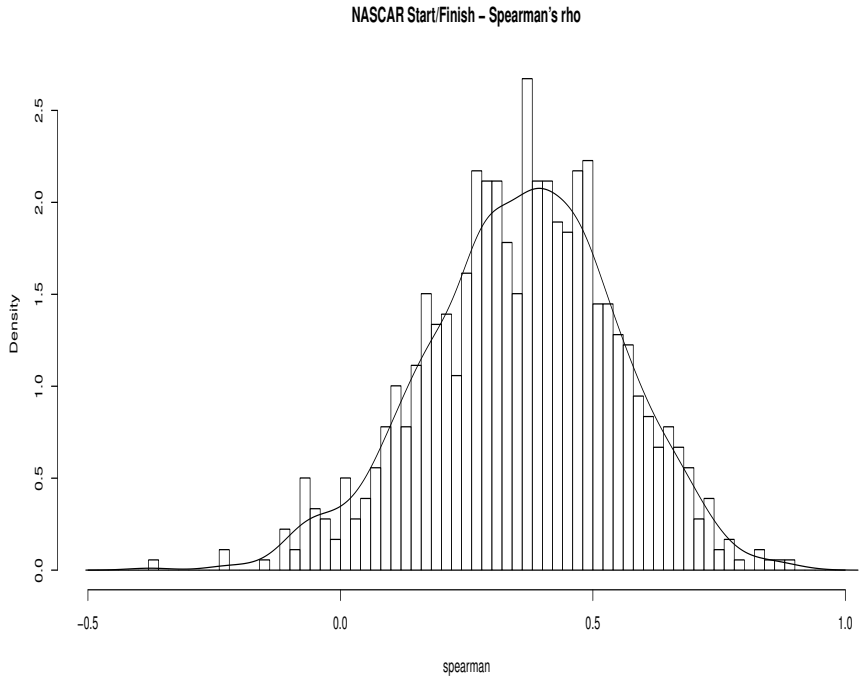

Figure 2.15: NASCAR Races 1975-2003 - Spearman's rank correlation coefficient for start/finish positions

```
[1] 898
> summary(spearman)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-0.3768  0.2399  0.3690  0.3590  0.4869  0.8977
```

∇

Many series (particularly when measured over time) display **spurious correlations**, particularly when both variables tend to increase or decrease together with no **causal** reason that the two (or more) variables move in tandem. For instance, the correlation between annual U.S. internet users (per 100 people) and electrical power consumption (kWh per capita) for the years 1994-2010 is .7821 (data source: The World Bank). Presumably increasing internet usage isn't leading to large increases in electrical consumption, or vice versa.

2.4 R Code for Chapter 2

```
### Chapter 2

### Example 2.1
```

```
## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

head(clt2016)    ## Print first 6 observations

## Assign RsnStop to be a factor (categorical) variable
## Assign labels to levels to the Catgories of Reasons for Stop
RsnStop <- factor(RsnStop)
levels(RsnStop) <- c("ChkPnt","DWI","Invstgtn","Other",
   "SafeMove","SeatBelt","Speed","StopLgtSgn","VhclMove","Rgstrtn")

## Obtain and print frequency table for Reasons for Stop
(table.RsnStop <- table(RsnStop))

## Figure 2.1 - Pie chart based on Table and Labels from above
pie(table.RsnStop, main="Pie Chart - CLT Traffic Stops")

## Figure 2.2 - Bar chart based on Table and Labels from above (cex shrinks size of levels)
barplot(table.RsnStop,
   main="Bar Chart - CLT Traffic Stops", xlab="Reason", ylab="Frequency",
   cex.names=0.6)

rm(list=ls(all=TRUE))

### Example 2.2

### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

### Obtain histogram - Figure 2.3
hist(bmi.nhl, breaks=30, xlab="Body Mass Index",
main="NHL BMI Distribution 2013-2014 Season")

rm(list=ls(all=TRUE))

### Examples 2.3 and 2.5 (Rock and Roll Marathon)

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

## Obtain mean and standard deviation by gender
tapply(mph,Gender,mean)
tapply(mph,Gender,median)

tapply(mph,Gender,var)
tapply(mph,Gender,sd)

## Obtain the densities (for plotting) of mph by gender
d.F <- density(mph[Gender=="F"])
d.M <- density(mph[Gender=="M"])

## Figure 2.4
## Set up a 2x2 grid for plots
par(mfrow=c(2,2))
## Histograms for Female and Male mph
hist(mph[Gender=="F"],breaks=25,main="Histogram of Female Speeds",
```

```
   xlab="Female Speeds")
hist(mph[Gender=="M"],breaks=25,main="Histogram of Male Speeds",
   xlab="Male Speeds")
## Density Plots for Female and Male mph
plot(d.F,
  main="Kernel Density Plot of Female Speeds")
plot(d.M,
  main="Kernel Density Plot of Male Speeds")

## Figure 2.5
## Reset Plot to 1 per page and obtain side-by-side boxplots
## Gender is a factor vatiable (on the x-axis)
par(mfrow=c(1,1))
plot(Gender, mph, main="Box Plots of Speed(mph) by Gender")

## Figure 2.6
## Obtain a "violin plot" - a "smoothed density" version of boxplot
# install.packages("ggplot2")
require(ggplot2)
ggplot(rr.mar, aes(y=mph, x=Gender)) + geom_violin()

rm(list=ls(all=TRUE))

### Miami Weather Plots

## Read data and set up data frame
mw1 <- read.csv("http://www.stat.ufl.edu/~winner/data/miami_weather.csv")
attach(mw1); names(mw1)

## Obtain mean temperature by year
(yearMeanTemp <- aggregate(meanTemp ~ Year, mw1, mean))

## Figure 2.7
## Stack Monthly and Annual plots
par(mfrow=c(2,1))
## Monthly Plot gives only "y", not "x", this is a line plot
## type="l" draws lines meeting points
plot(meanTemp, type="l", main="Miami Monthly Mean Temp (F) 1949-2014",
  xlab="Month", ylab="Mean Temperature")

## Plot "x"=Year (first column of yearMeanTemp) and
##  "y"=mean temp (second column of yearMeanTemp0
plot(yearMeanTemp[,1], yearMeanTemp[,2],
  type="l", main="Miami Yearly Mean Temp (F) 1949-2014",
  xlab="Year", ylab="Mean Temperature")

rm(list=ls(all=TRUE))


### Bigfoot Map  Note: Some of these packages no longer work

bigfoot <- read.csv("http://www.stat.ufl.edu/~winner/data/bigfoot_state.csv",
                    header=TRUE)
attach(bigfoot); names(bigfoot)

# install.packages("usmap")
library(ggplot2)
library(usmap)

bigfoot$fips <- fips(bigfoot$State)

## Figure 2.8
plot_usmap(data = bigfoot, values = "Bigfoot", color = "red") +
  scale_fill_continuous(low="white", high="black", name = "Bigfoot Sightings",
```

```
      label = scales::comma) +
  theme(legend.position = "right")

rm(list=ls(all=TRUE))


### Example 2.4

## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

(table.RsnStop <- table(RsnStop))
round(table.RsnStop / sum(table.RsnStop), 5)

rm(list=ls(all=TRUE))

### Example 2.5 (NHL BMI portion)

### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

### obtain the population size from number of rows of data frame
N <- NROW(nhl)
### obtain the total of the BMI values
sum.bmi <- sum(bmi.nhl)
### mean = sum / N
mean.bmi <- sum.bmi/N
### Obtain sorted bmi's
bmi.nhl.sort <- sort(bmi.nhl)
### Print first and last few cases to confirm ordered
cbind(head(bmi.nhl.sort), tail(bmi.nhl.sort))
### If N is even, average middle 2 cases, otherwise take middle case
median.bmi <- ifelse(N%%2==0,(bmi.nhl.sort[N/2]+bmi.nhl.sort[N/2+1])/2,
      bmi.nhl.sort[(N+1)/2])
bmi.cent.out <- cbind(N, sum.bmi, mean.bmi, median.bmi)
colnames(bmi.cent.out) <- c("N", "sum", "mean", "median")
round(bmi.cent.out, 4)

### Use built-in mean and median functions
mean(bmi.nhl)
median(bmi.nhl)

rm(list=ls(all=TRUE))

### Examples 2.6-2.8 (NHL BMI portion)
### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)

bmi.max <- max(bmi.nhl)     # Highest BMI
bmi.min <- min(bmi.nhl)     # Lowest BMI
range <- bmi.max - bmi.min  # Compute Range
bmi.75 <- quantile(bmi.nhl,.75)    # BMI 75%-ile
bmi.25 <- quantile(bmi.nhl,.25)    # BMI 25%-ile
IQR <-  bmi.75 - bmi.25    # Compute IQR
N <- length(bmi.nhl)       # Use "length" function to get N
```

```
mu <- mean(bmi.nhl)        # Use "mean" function to get mu
sum.dev2 <- sum((bmi.nhl - mu)^2)  # Numerator of Variance
sigma2 <- sum.dev2/N       # Population Variance
s2 <- sum.dev2/(N-1)       # Sample Variance
sigma <- sqrt(sigma2)      # Population Standard Deviation
s <- sqrt(s2)              # Sample Standard Deviation
var(bmi.nhl)               # Sample Variance with "var" function
(N-1)*var(bmi.nhl)/N       # Pop variance with "var" function
sd(bmi.nhl)                # Sample Std Dev with "sd" function
sqrt((N-1)/N)*sd(bmi.nhl)  # Population Std Dev with "sd" function
## Proportion of Individual w/in 1 and 2 SDs of mean
mu.pm.1sd <- sum(bmi.nhl >= mu-sigma & bmi.nhl <= mu+sigma) / N
mu.pm.2sd <- sum(bmi.nhl >= mu-2*sigma & bmi.nhl <= mu+2*sigma) / N

bmi.var.out1 <- cbind(bmi.min, bmi.max, range, bmi.25, bmi.75, IQR)
bmi.var.out2 <- cbind(mu, sum.dev2, sigma2, s2, sigma, s,
       mu.pm.1sd, mu.pm.2sd)
colnames(bmi.var.out1) <- c("min", "max", "range", "LQ", "UQ", "IQR")
colnames(bmi.var.out2) <- c("mean", "sum(dev^2)", "sigma^2", "s^2",
    "sigma", "s", "P(mu+/-1sigma)", "P(mu+/-2sigma)")
round(bmi.var.out1, 3)
round(bmi.var.out2, 3)

mu <- mean(bmi.nhl)        # Use "mean" function to get mu
sum.dev2 <- sum((bmi.nhl - mu)^2)  # Numerator of Variance
sigma2 <- sum.dev2/N       # Population Variance
sigma <- sqrt(sigma2)      # Population Standard Deviation
bmi.median <- median(bmi.nhl)
mad <- median(abs(bmi.nhl - bmi.median))   # Median absolute deviation
mad_6745 <- mad/0.6745                     # Approximating sigma
cv <- sigma/mu                             # Coefficient of Variation

mad.cv.out <- cbind(bmi.median, mad, mad_6745, sigma, mu, cv)
colnames(mad.cv.out) <- c("median", "MAD", "MAD/0.6745", "sigma", "mu", "CV")
round(mad.cv.out, 3)

mu3 <-  (sum((bmi.nhl-mu)^3)/N)
skew <- mu3/(sigma^3)
SE.skew <- sqrt(6*N*(N-1)/((N-2)*(N-1)*(N+3)))

mu4 <-  (sum((bmi.nhl-mu)^4)/N)
kurt <- mu4/(sigma^4)-3
SE.kurt <- sqrt(24*N*(N-1)^2/((N-3)*(N-2)*(N-1)*(N+3)*(N+5)))

skew.kurt.out <- cbind(mu3, skew, SE.skew,
      mu4, kurt, SE.kurt)
colnames(skew.kurt.out) <- c("mu3", "Skewness", "SE{Skew}",
          "mu4", "Kurtosis", "SE{Kurt}")
round(skew.kurt.out, 4)

rm(list=ls(all=TRUE))

### Examples 2.6-2.8 (Rock and Roll Marathon portion)
## Read data from website and attach data frame and obtain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

f.mph <- mph[Gender=="F"]
N.f <- length(f.mph)
mean.f <- mean(f.mph)
var.f <- var(f.mph)
sd.f <- sd(f.mph)
mu.pm.1sd.f <- sum(f.mph >= mean.f - sd.f & f.mph <= mean.f + sd.f) / N.f
```

```
mu.pm.2sd.f <- sum(f.mph >= mean.f - 2*sd.f & f.mph <= mean.f + 2*sd.f) / N.f

m.mph <- mph[Gender=="M"]
N.m <- length(m.mph)
mean.m <- mean(m.mph)
var.m <- var(m.mph)
sd.m <- sd(m.mph)
mu.pm.1sd.m <- sum(m.mph >= mean.m - sd.m & m.mph <= mean.m + sd.m) / N.m
mu.pm.2sd.m <- sum(m.mph >= mean.m - 2*sd.m & m.mph <= mean.m + 2*sd.m) / N.m

rr.var.out.f <- cbind(N.f, mean.f, var.f, sd.f, mu.pm.1sd.f, mu.pm.2sd.f)
rr.var.out.m <- cbind(N.m, mean.m, var.m, sd.m, mu.pm.1sd.m, mu.pm.2sd.m)
rr.var.out <- rbind(rr.var.out.f, rr.var.out.m)
rownames(rr.var.out) <- c("Females", "Males")
colnames(rr.var.out) <- c("N", "mean", "sigma^2", "sigma",
                "P(mu+/-1sigma)", "P(mu+/-2sigma)")

round(rr.var.out, 3)

f.mph <- mph[Gender=="F"]
N.f <- length(f.mph)
mu.f <- mean(f.mph)
median.f <- median(f.mph)
sigma.f <- sd(f.mph) * sqrt((N.f-1)/N.f)
mad.f <- median(abs(f.mph-median.f))
mad.f_6745 <- mad.f / 0.6745
cv.f <- sigma.f/mean.f

m.mph <- mph[Gender=="M"]
N.m <- length(m.mph)
mu.m <- mean(m.mph)
median.m <- median(m.mph)
sigma.m <- sd(m.mph) * sqrt((N.m-1)/N.m)
mad.m <- median(abs(m.mph-median.m))
mad.m_6745 <- mad.m / 0.6745
cv.m <- sigma.m/mean.m

mad.cv.out.f <- cbind(median.f, mad.f, mad.f_6745, sigma.f, mu.f, cv.f)
mad.cv.out.m <- cbind(median.m, mad.m, mad.m_6745, sigma.m, mu.m, cv.m)
mad.cv.out <- rbind(mad.cv.out.f, mad.cv.out.m)
rownames(mad.cv.out) <- c("Females", "Males")
colnames(mad.cv.out) <- c("median", "MAD", "MAD/0.6745", "sigma", "mu", "CV")
round(mad.cv.out, 3)


m3.f <- sum((f.mph-mean.f)^3)/N.f
skew.f <- m3.f / sd.f^3
m4.f <- sum((f.mph-mean.f)^4)/N.f
kurt.f <- (m4.f/sd.f^4)-3
SE.skew.f <- sqrt(6*N.f*(N.f-1)/((N.f-2)*(N.f-1)*(N.f+3)))
SE.kurt.f <- sqrt(24*N.f*(N.f-1)^2/((N.f-3)*(N.f-2)*(N.f-1)*(N.f+3)*(N.f+5)))

skew.kurt.out.f <- cbind(m3.f, skew.f, SE.skew.f,
      m4.f, kurt.f, SE.kurt.f)

m3.m <- sum((m.mph-mean.m)^3)/N.m
skew.m <- m3.m / sd.m^3
m4.m <- sum((m.mph-mean.m)^4)/N.m
kurt.m <- (m4.m/sd.m^4)-3
SE.skew.m <- sqrt(6*N.m*(N.m-1)/((N.m-2)*(N.m-1)*(N.m+3)))
SE.kurt.m <- sqrt(24*N.m*(N.m-1)^2/((N.m-3)*(N.m-2)*(N.m-1)*(N.m+3)*(N.m+5)))

skew.kurt.out.f <- cbind(m3.f, skew.f, SE.skew.f,
      m4.f, kurt.f, SE.kurt.f)
```

```
skew.kurt.out.m <- cbind(m3.m, skew.m, SE.skew.m,
      m4.m, kurt.m, SE.kurt.m)
skew.kurt.out <- rbind(skew.kurt.out.f, skew.kurt.out.m)
rownames(skew.kurt.out) <- c("Females", "Males")
colnames(skew.kurt.out) <- c("mu3", "Skewness", "SE{Skew}",
          "mu4", "Kurtosis", "SE{Kurt}")
round(skew.kurt.out, 4)


m.mph <- mph[Gender=="M"]
(N.m <- length(m.mph))
(mean.m <- mean(m.mph))
(var.m <- var(m.mph))
(sd.m <- sd(m.mph))
sum(m.mph >= mean.m - sd.m & m.mph <= mean.m + sd.m) / N.m
sum(m.mph >= mean.m - 2*sd.m & m.mph <= mean.m + 2*sd.m) / N.m
(cv.m <- sd.m/mean.m)
(mad.m <- median(abs(m.mph-mean.m)))
(m3.m <- sum((m.mph-mean.m)^3)/N.m)
(skew.m <- m3.m / sd.m^3)
(m4.m <- sum((m.mph-mean.m)^4)/N.m)
(kurt.m <- (m4.m/sd.m^4)-3)

rm(list=ls(all=TRUE))

### Example 2.9
## Read data off web page, attach file as data frame, and list variable names
bh <- read.csv("http://www.stat.ufl.edu/~winner/data/blues_hand.csv")
attach(bh); names(bh)

region <- factor(region)
levels(region) <- c("East", "Delta", "Texas")
thumbSty <- factor(thumbSty)
levels(thumbSty) <- c("Alternating", "Utility", "Dead")

## Obtain Table of Counts (Row=Region, Column=Thumb Style)
(reg_ts <- table(region, thumbSty))
## Obtain Row (1) and Column (2) Marginal Totals
margin.table(reg_ts,1)
margin.table(reg_ts,2)
## Obtain Proportions across all Cells
reg_ts/sum(reg_ts)
## Obtain Row Proportions (Thumb Style w/in Region)
prop.table(reg_ts,1)
## Obtain Column Proportions (Region w/in Thumb Style)
prop.table(reg_ts,2)

## Obtain Cluster (Grouped) and Stacked Bar Plots
## t(prop.table(reg_ts,1)) takes transpose so that group var is Region
## Figure 2.9
par(mfrow=c(1,2))
barplot(t(prop.table(reg_ts,1)),beside=T,legend=colnames(reg_ts),
ylim=c(0,1), main="Grouped Bar Plot - Blues Guitarists")
barplot(t(prop.table(reg_ts,1)),beside=F,legend=colnames(reg_ts),
   main="Stacked Bar Plot - Blues Guitarists", ylim=c(0,1.40))

# install.packages("vcd")
library(vcd)
table(post1906,region,thumbSty)
## Figure 2.10
mosaic(~post1906+region+thumbSty, data=bh, shade=TRUE, legend=TRUE)

rm(list=ls(all=TRUE))
```

```
### Example 2.10

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

## Obtain the densities (for plotting) of mph by gender
d.F <- density(mph[Gender=="F"])
d.M <- density(mph[Gender=="M"])

## Figure 2.11 - Density Plots for Female and Male mph
# win.graph(height=5.5, width=7.0)
plot(d.F,xlim=c(4,11),xlab="Speed(mph)",ylab="Density",
  main="Kernel Density Plot of Marathon Speeds")
lines(d.M,lty=2)
legend(9,0.45,c("Females","Males"),lty=c(1,2))

## Figure 2.12 - Combined histogram
library(ggplot2)
## win.graph(height=5.5, width=7.0)
ggplot(rr.mar, aes(x=mph,fill=Gender)) +
   geom_histogram(binwidth=0.1)

rm(list=ls(all=TRUE))

### Example 2.11

sw1 <- read.csv("http://www.stat.ufl.edu/~winner/data/software1.csv")
attach(sw1); names(sw1)

cor(sizeProj,effortProj)
cor(sizeProj[1:16],effortProj[1:16])

## Figure 2.13
par(mfrow=c(2,2))
plot(sizeProj,effortProj,xlab="size",ylab="effort",
   main="Original Scale, All Projects")
abline(lm(effortProj~sizeProj))
plot(sizeProj[1:16],effortProj[1:16],xlab="size",ylab="effort",
   main="Original Scale, Project 17 Removed")
abline(lm(effortProj[1:16]~sizeProj[1:16]))

cor(log(sizeProj),log(effortProj))
cor(log(sizeProj[1:16]),log(effortProj[1:16]))

plot(log(sizeProj),log(effortProj),xlab="ln(size)",ylab="ln(effort)",
   main="Log Scale, All Projects")
abline(lm(log(effortProj) ~ log(sizeProj)))
plot(log(sizeProj[1:16]),log(effortProj[1:16]),xlab="ln(size)",ylab="ln(effort)",
   main="Log Scale, Project 17 Removed")
abline(lm(log(effortProj[1:16]) ~ log(sizeProj[1:16])))

rm(list=ls(all=TRUE))

### Example 2.12

rs1 <- read.csv("http://www.stat.ufl.edu/~winner/data/rockstrength.csv")
attach(rs1); names(rs1)

## Figure 2.14 - Scatterplot matrix of UCS, hb, gs, ga (columns 2,6,7,8 of rs1)
plot(rs1[,c(2,6,7,8)])

## Obtain correlation matrix of UCS, hb, gs, ga  (columns 2,6,7,8 of rs1)
```

```
cor(rs1[,c(2,6,7,8)])

rm(list=ls(all=TRUE))


### Example 2.13
nasRace <- read.fwf("http://www.stat.ufl.edu/~winner/data/nascarr.dat",
  widths=c(3,6,4,4,9,7,9,9,7,5,3,4,4,9,7,8,5,38),col.names=c("seriesRace",
  "year","yearRace","numCar","payout","cpiU","spearman","kendall","trkLength",
  "lapsComp","roadRace","cautionFlag","leadChange","winTime","trkLat","trkLong",
  "trkCode","trkName"))
attach(nasRace)

length(spearman)
summary(spearman)

## Figure 2.15
hist(spearman,breaks=seq(-.5,1,.02),prob=T,
main="NASCAR Start/Finish - Spearman's rho")
lines(density(spearman))

rm(list=ls(all=TRUE)
```

Chapter 3

Probability

In this chapter, we describe the concepts of probability, random variables, probability distributions, and sampling distributions. There are three commonly used interpretations of probability: classical, relative frequency, and subjective. Probability is the basis of all methods of statistical inference covered in this course and its sequel.

3.1 Terminology and Basic Probability Rules

The **classical** interpretation of probability involves listing (or using counting rules to quantify) all possible outcomes of a random process, often referred to as an "experiment." It is often (but not necessarily) assumed that each outcome is equally likely. If a coin is tossed once, it can land either "heads" or "tails," and unless there is reason to believe otherwise, we would assume the probability of each possible outcome is 1/2. If a dice is rolled, the possible numbers on the "up face" are {1,2,3,4,5,6}. Again, unless some external evidence leads us to believe otherwise, we would assume each side has a probability of landing as the "up face" is 1/6. When dealing a 5 card hand from a well shuffled 52 card deck, there are $\frac{52!}{5!(52-5)!} = 2,598,960$ possible hands. Clearly that would be impossible to enumerate, but with counting rules it is still fairly easy to assign probabilities to different types of hands.

An **event** is a pre-specified outcome of an experiment/random process. It can be made up of a single element or a group of elements of the sample space. If the sample space is made up of N elements and the event of interest constitutes N_E elements of the sample space, the probability of the event is $p_E = N_E/N$, when all elements are equally likely. If elements are not equally likely, p_E is the sum of the probabilities of the elements constituting the event (where the sum of all the N probabilities is 1).

The **relative frequency** interpretation of probability corresponds to how often an event of interest would occur if an experiment were conducted repeatedly. If an unbalanced dice were tossed a very large number of times, we could observe the fractions of times each number was the "up face." With modern computing power, simulations can be run to approximate probabilities of complex events, which could never be able to be obtained via a model of a sample space.

In cases where a sample space can not be enumerated or an experiment can not be repeated, individuals often resort to assessing **subjective** probabilities. For instance, in considering whether the price of a stock will increase over a specific time horizon, individuals may speculate on the probability based on any market information available at the time of the assessment. Different individuals may have different probabilities for the same event. Many studies have been conducted to assess people's abilities and heuristics used to assign probabilities to events (see e.g. Kahneman, Slovic, and Tversky (1982) [30]), for a large collection of research on the topic.

Three useful counting tools are the **multiplication rule**, **permutations** and **combinations**. The multiplication rule is useful when the experiment is made up of k stages, where stage i can end in one of m_i outcomes. Permutations are used when sampling k items from n items without replacement, and order matters. Combinations are similar to permutations with the exception that order does not matter. The total possible outcomes for each of these rules is given below.

$$\text{Multiplication Rule: } m_1 \times m_2 \times \cdots \times m_k = \prod_{i=1}^{k} m_i$$

$$\text{Permutations: } P_k^n = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!} \qquad 0! \equiv 1$$

$$\text{Combinations: } C_k^n = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \cdots \times 1} = \frac{n!}{k!(n-k)!}$$

Note that there are $k!$ possible orderings of the k items selected from n items, which is why there are fewer combinations than permutations.

Example 3.1: Lotteries and Competitions

The Florida lottery has many "products" for consumers (flalottery.com). The Pick 4 game is conducted twice per day and pays out up to $5000 per drawing. Participants choose 4 digits from 0-9 (digits can be repeated). Thus at each of $k = 4$ stages, there are $m = 10$ potential digits. Thus there are $10(10)(10)(10) = 10{,}000$ possible sequences (order matters in payouts).

In a race among 10 "identical" mice of a given strain, there are $P_3^{10} = 10(9)(8) = 720$ possible orderings of 1st, 2nd, and 3rd place. In the 2017 Kentucky Derby, there were 22 horses in the race. Starting positions are taken by "pulling names out of a hat." Thus, there are $22! = 1.124 \times 10^{21}$ possible orderings of the horses to the starting positions. This is 10.4 billion times as many people who had ever lived on the earth as of 2011 according to the Population Reference Bureau (www.prb.com).

The Florida Lotto game, held every Wednesday and Saturday night, involves selecting 6 numbers without replacement from the integers 1,...,53; where order does not matter. There are $C_6^{53} = \frac{53!}{6!47!} = 22,957,480$ possible drawings.

3.1.1 Basic Probability

Let A and B be events of interest with corresponding probabilities $P(A)$ and $P(B)$, respectively. The **Union** of events A and B is the event that either A and/or B occurs and is denoted $A \cup B$. Events A and B are

	B	$\overline{B}$	Total
A	909	67	976
$\overline{A}$	2528	142	2670
Total	3437	209	3646

Table 3.1: Counts of UFO's by Shape Type and nation of sighting

mutually exclusive if they can not both occur as an experimental outcome. That is, if A occurs, B cannot occur, and vice versa. The **Complement** of event A, is the event that A does not occur and is denoted by $\overline{A}$ or sometimes A'. The **Intersection** of events A and B is the event that both A and B occur, and is denoted as $A \cap B$ or simply AB. In terms of probabilities, we have the following rules.

Union: $P(A \cup B) = P(A)+P(B)-P(AB)$ Mutually Exclusive: $P(AB) = 0$ Complement: $P\left(\overline{A}\right) = 1-P(A)$

The probability of an event A or B, without any other information, is referred to as its **unconditional** or **marginal** probability. When information is known whether or not another event has (or has not) occurred is referred to as its **conditional** probability. If the unconditional probability of A and its conditional probability given B has occurred are equal, then the events A and B are said to be **independent**. The rules for obtaining conditional probabilities (assuming $P(A) > 0$ and $P(B) > 0$) are given below, as well as probabilities under independence.

$$\text{Prob. of A Given B: } P(A|B) = \frac{P(AB)}{P(B)} \qquad \text{Prob. of B Given A: } P(B|A) = \frac{P(AB)}{P(A)}$$

$$P(AB) = P(A)P(B|A) = P(B)P(A|B)$$

A and B independent: $P(A) = P(A|B) = P\left(A|\overline{B}\right)$ $P(B) = P(B|A) = P\left(B|\overline{A}\right)$ $P(AB) = P(A)P(B)$

Example 3.2: UFO Sightings

Based on 3646 UFO sightings on the UFO Research Database (www.uforesearchdb.com), we define A to be the event that a UFO is classified as being shaped as an orb/sphere or circular or a disk and event B that the sighting is in the USA. Table 3.1 gives a cross-tabulation of the counts for this "population."

$$P(A) = \frac{976}{3646} = .2677 \qquad P(B) = \frac{3437}{3646} = .9427 \qquad P(AB) = \frac{909}{3646} = .2493 \qquad P(A \cup B) = .2677+.9427-.2493 = .9611$$

$$P(A|B) = \frac{.2493}{.9427} = \frac{909}{2528} = .2645 \qquad P\left(A|\overline{B}\right) = \frac{67}{209} = .3206 \qquad P(B|A) = \frac{.2493}{.2677} = \frac{909}{976} = .9314$$

Note that the event that a UFO is classified as orb/sphere or circular or a disk is not independent of whether it was sighted in the USA. There is a higher probability for these types of shapes to be sighted outside the USA (.3206) than in the USA (.2645).

∇

Example 3.3: Women's and Men's Marathon Speeds

For the Rock and Roll marathon speeds considered previously, we classify events as follow. Event F is that the runner is Female, event S_5 is the event that a runner's speed is less than or equal to 5 miles per hour, and S_7 is the event that the runner's speed is greater than or equal to 7 miles per hour. Counts of runners by gender and speed are given in Table 3.2. Note that the middle row represents the intersection of the compliments of events S_5 and S_7 and represents the runners with speeds between 5 and 7 miles per hour. We compute various probabilities below.

$$P(F) = \frac{1045}{2499} = .4182 \qquad P\left(\overline{F}\right) = 1-.4182 = \frac{1454}{2499} = .5818 \qquad P(S_5) = \frac{326}{2499} = .1305 \qquad P(S_7) = \frac{464}{2499} = .1857$$

$$P\left(\overline{S_5} \cap \overline{S_7}\right) = 1-.1305-.1857 = \frac{1709}{2499} = .6839 \qquad P(F \cap S_5) = \frac{172}{2499} = .0688 \qquad P\left(\overline{F} \cap S_5\right) = \frac{154}{2499} = .0616$$

$$P(F \cap S_7) = \frac{106}{2499} = .0424 \qquad P\left(\overline{F} \cap S_7\right) = \frac{358}{2499} = .1433 \qquad P\left(F \cap \overline{S_5} \cap \overline{S_7}\right) = \frac{767}{2499} = .3069$$

$$P\left(\overline{F} \cap \overline{S_5} \cap \overline{S_7}\right) = \frac{942}{2499} = .3770 \qquad P(S_5|F) = \frac{.0688}{.4182} = \frac{172}{1045} = .1646 \qquad P(S_7|F) = \frac{.0424}{.4182} = \frac{106}{1045} = .1014$$

$$\left(\overline{S_5} \cap \overline{S_7}|F\right) = \frac{.3069}{.4182} = \frac{767}{1045} = .7340$$

∇

3.1.2 Bayes' Rule

Bayes' rule is used in a wide range of areas to update probabilities (and probability distributions) in light of new information (data). In the case of updating probabilities of particular events, we start with a set

	F	$\overline{F}$	Total
S_5	172	154	326
$\overline{S_5} \cap \overline{S_7}$	767	942	1709
S_7	106	358	464
Total	1045	1454	2499

Table 3.2: Counts of Speeds (mph) by Gender - 2015 Rock and Roll Marathon

of events $A_1, \ldots, A_k$ that represent a **partition** of the sample space. That means that each element in the sample space must fall in exactly one A_i. In probability terms this means the following statements hold.

$$i \neq j: \quad P(A_i \cap A_j) = 0 \qquad\qquad P(A_1) + \cdots P(A_k) = 1$$

The probability $P(A_i)$ is referred to as the **prior probability** of the i^{th} portion of the partition, and in some contexts are referred to as **base rates**. Let C be an event, such that $0 < P(C) < 1$, with known conditional probabilities $P(C|A_i)$. This leads to being able to "update" the probability that A_i occurred, given knowledge that C has occurred, the **posterior probability** of the i^{th} portion of the partition. This is simply (in this context) an application of conditional probability making use of formulas given above and the fact that there is a partition of the sample space.

$$P(A_i \cap C) = P(A_i)P(C|A_i) \qquad\qquad P(C) = \sum_{i=1}^{k} P(A_i \cap C) = \sum_{i=1}^{k} P(A_i)P(C|A_i)$$

$$\Rightarrow \qquad P(A_i|C) = \frac{P(A_i \cap C)}{P(C)} = \frac{P(A_i)P(C|A_i)}{\sum_{i=1}^{k} P(A_i)P(C|A_i)} \quad i = 1, ..., k$$

Example 3.4: Women's and Men's Marathon Speeds

Treating the three speed ranges ($A_1 \equiv\leq 5, \quad A_2 \equiv 5-7, \quad A_3 \equiv\geq 7$) as a partition of the sample space, we can update the probabilities of the runner's speed range, given knowledge of gender. The prior probabilities are $P(A_1) = 326/2499 = .1305$, $P(A_2) = 1709/2499 = .6839$, and $P(A_3) = 464/2499 = .1857$. The relevant probabilities are given below to obtain the posterior probabilities of the speed ranges, given the runner's gender.

$$P(A_1) = \frac{326}{2499} = .1305 \qquad P(F|A_1) = \frac{172}{326} = .5276 \qquad P(A_1 \cap F) = P(A_1)P(F|A_1) = \left(\frac{326}{2499}\right)\left(\frac{172}{326}\right) = .0688$$

$$P(A_2) = \frac{1709}{2499} = .6839 \qquad P(F|A_2) = \frac{767}{1709} = .4488 \qquad P(A_2 \cap F) = P(A_2)P(F|A_2) = \left(\frac{1709}{2499}\right)\left(\frac{767}{1709}\right) = .3069$$

$$P(A_3) = \frac{464}{2499} = .1857 \qquad P(F|A_3) = \frac{106}{464} = .2284 \qquad P(A_3 \cap F) = P(A_3)P(F|A_3) = \left(\frac{464}{2499}\right)\left(\frac{106}{464}\right) = .0424$$

$$P(F) = \sum_{i=1}^{3} P(A_i \cap F) = .0688 + .3069 + .0424 = .4182 \qquad P(A_1|F) = \frac{.0688}{.4182} = .1646$$

$$P(A_2|F) = \frac{.3069}{.4182} = .7340 \qquad P(A_3|F) = \frac{.0424}{.4182} = .1014$$

Note that these can be computed very easily from the counts in Table 3.2 by taking the cell counts over the column totals, as can be seen for the males.

$$P(M) = \frac{1454}{2499} = .5818 \qquad P(A_1|M) = \frac{154}{1454} = .1059 \qquad P(A_2|M) = \frac{942}{1454} = .6479 \qquad P(A_3|M) = \frac{358}{1454} = .2462$$

∇

Example 3.5: Drug Testing Accuracy

As a second example based on assessed probabilities, Barnum and Gleason (1964), [6], considered drug tests among workers. They had four sources of prevalence of recreational drug users based on published data sources (2.4% (.024), 3.1% (.031), 8.2% (.082), and 20.2% (.202)). Further, based on studies of test accuracy at the time, they had the probability that a drug user (correctly) tests positive is 0.80, and the probability a non-drug user (incorrectly) tests positive is 0.02. Let D be the event that a worker is a drug user, and T^+ be the event that a worker tests positive for drug use.

Consider the case where $P(D) = .024$. We are interested in the probability a worker who tests positive is a drug user. Note that we do not have this probability stated above. The relevant probabilities and calculations are given below.

$$P(D) = .024 \qquad P\left(\overline{D}\right) = 1 - .024 = .976 \qquad P\left(T^+|D\right) = .80 \qquad P\left(T^+|\overline{D}\right) = .02$$

$$P\left(D \cap T^+\right) = .024(.80) = .01920 \qquad P\left(\overline{D} \cap T^+\right) = .976(.02) = .01952 \qquad P\left(T^+\right) = .01920 + .01952 = .03872$$

$$P\left(D|T^+\right) = \frac{.01920}{.03872} = .4959 \qquad P\left(\overline{D}|T^+\right) = \frac{.01952}{.03872} = .5041$$

Thus a positive result on the test implies slightly less than a 50:50 chance the worker uses drugs. As the prevalence increases, this probability increases, see Table 3.3.

∇

$P(D)$	$P\left(D \cap T^+\right)$	$P\left(\overline{D} \cap T^+\right)$	$P\left(T^+\right)$	$P\left(D\|T^+\right)$
.024	.01920	.01952	.03872	.4959
.031	.02480	.01938	.04418	.5613
.082	.06560	.01836	.08396	.7813
.202	.16160	.01596	.17756	.9101

Table 3.3: Probability a Positive Drug test corresponds to a drug user as a function of Prevalence of Drug Use

3.2 Random Variables and Probability Distributions

When an experiment is conducted, or an observation is made, the outcome will not be known in advance, and is considered to be a **random variable**. Random variables can be qualitative or quantitative. Qualitative variables are generally modeled as a list of outcomes and their corresponding counts, as in contingency tables and cross-tabulations. Quantitative random variables are numeric outcomes and are classified as being either discrete or continuous, as described previously in describing data.

A **probability distribution** gives the values a random variable can take on and their corresponding probabilities (discrete case) or density (continuous case). Probability distributions can be given in tabular, graphic, or formulaic form. Some commonly used families of distributions are described below.

3.3 Discrete Random Variables

Discrete random variables can take on a finite, or countably infinite, set of outcomes. We label the random variable as Y, and its specific outcomes as $y_1, y_2, \ldots, y_k$. Note that in some cases there is no upper limit for k. We denote the probabilities of the outcomes as $P\left(Y = y_i\right) = p\left(y_i\right)$, with the following restrictions.

$$0 \leq p\left(y_i\right) \leq 1 \qquad \sum_{i=1}^{k} p\left(y_i\right) = 1 \qquad F\left(y_t\right) = P\left(Y \leq y_t\right) = \sum_{i=1}^{t} p\left(y_i\right) \quad t = 1, \ldots, k$$

Here $F(y)$ is called the **cumulative distribution function (cdf)**. This is a monotonic "step" function for discrete random variables, and ranges from 0 to 1.

Example 3.6: NASCAR Race Finish Positions - 1975-2003

For the NASCAR race data in Winner (2006) [52], each driver was classified by their starting position and their finishing position in the 898 races (34884 driver/races). For each race, we identify the number of racers who start in the top 10, that finish in the top 3. This random variable (Y) can take on the values $y =$ 0, 1, 2, or 3. That is, none of the people who start toward the front (top 10) finish in the top 3, or one, or two, or three. Table 3.4 gives the counts, probabilities, cumulative probabilities, and calculations used later to numerically describe the empirical population distribution. The probability of either 2 or 3 drivers who started in the top 10 finish in the top 3, is over 3/4 (.3987+.3708=.7695).

y	# races	$p(y)$	$F(y)$	$yp(y)$	$y^2p(y)$
0	37	.0412	.0412	0.0000	0.0000
1	170	.1893	.2305	0.1893	0.1893
2	358	.3987	.6292	0.7974	1.5948
3	333	.3708	1.0000	1.1124	3.3372
Total	898	1		2.0991	5.1213

Table 3.4: Probability Distribution for Number of Top 10 Starters finishing in Top 3 positions, NASCAR races 1975-2003

R Output

```
## Output
> (t.strt10Fin3 <- table(strt10Fin3))  ### Count 0,1,2,3 Top 3 finishers
strt10Fin3
  0   1   2   3
 37 170 358 333
> t.strt10Fin3 / sum(t.strt10Fin3)     ### Turn counts to proportions
strt10Fin3
         0          1          2          3
0.04120267 0.18930958 0.39866370 0.37082405
```

∇

Population Numerical Descriptive Measures

Three widely used numerical descriptive measures corresponding to a population are the **population mean**, μ, the **population variance**, σ^2, and the **population standard deviation**, σ. While we have previously covered these based on a population of measurements, we now base them on a probability distribution. Their formulas are given below.

$$\text{Mean: } E\{Y\} = \mu_Y = y_1p(y_1) + \cdots + y_kp(y_k) = \sum_y yp(y)$$

$$\text{Variance: } V\{Y\} = E\{(Y-\mu_Y)^2\} = \sigma_Y^2 = (y_1-\mu_Y)^2p(y_1) + \cdots + (y_k-\mu_Y)^2p(y_k) = \sum_y (y-\mu_Y)^2p(y) =$$

$$= \sum_y y^2p(y) - \mu_Y^2 \qquad \text{Standard Deviation: } \sigma_Y = +\sqrt{\sigma_Y^2}$$

Example 3.7: NASCAR Race Finish Positions - 1975-2003

If we repeatedly sampled a race from this population, observed and saved the number of the top 10 starters who finished in the top 3, the long run mean would be μ_Y, and a "typical" distance from the mean would be σ_Y. From Table 3.4, the necessary calculations to compute μ_Y, σ_Y^2, and σ_Y are given.

$$\mu_Y = \sum_y yp(y) = 2.0991 \qquad \sigma_Y^2 = \sum_y (y-\mu_Y)^2 p(y) = \sum_y y^2 p(y) - \mu_Y^2 = 5.1213 - 2.0991^2 = 0.7151$$

$$\sigma_Y = +\sqrt{0.7151} = 0.8456$$

A sample of 10000 races is taken from this population (equivalently done by taking 10000 integers between 1 and 898 WITH replacement), observing the number of top 3 finishers for each race. Then the mean and standard deviation of those numbers are computed.

R Output

```
## Output
> mean(strt10Fin3[sample.race])
[1] 2.0816
> sd(strt10Fin3[sample.race])
[1] 0.8542918
```

Note that the mean of the 10000 sampled races is close to the population mean (2.0816 vs 2.0991) and sample standard deviation is close to the corresponding population value (0.8543 vs 0.8456). If a different (or no) seed had been used, the samples, and thus their means and standard deviations would change as well.

∇

Some useful rules among **linear** functions of random variables are given here. Suppose Y is a random variable with mean and variance μ_Y and σ_Y^2, respectively. Further, suppose that a and b are constants (not random). Then we have the following results.

$$E\{a+bY\} = \sum_y (a+by)p(y) = a\sum_y p(y) + b\sum_y yp(y) = a(1) + b\mu_Y = a + b\mu_Y$$

$$V\{a+bY\} = \sum_y ((a+by) - (a+b\mu_Y))^2 p(y) = b^2 \sum_y (y-\mu_Y)^2 p(y) = b^2\sigma_Y^2 \qquad \sigma_{a+bY} = |b|\sigma_Y$$

Examples where these can be applied involve transforming from inches to centimeters (1 inch = 2.54 cm, 1 cm = 1/2.54=0.3937 inch), from pounds to kilograms (1 kilogram = 2.204623 pounds) and from degrees Fahrenheit to Celsius ($\deg F = 32 + 1.8 \deg C$). These rules do not work for values raised to powers, exponentials, or logarithms, although some approximations exist.

Example 3.8: NHL Hockey Player BMI and Marathon Speeds

Previously, we obtained the population mean and variance for NHL player body mass indices. Now we obtain the mean, variance, and standard deviation of their weights (pounds) and heights (inches), and

convert them to kilograms and centimeters, respectively. The mean weight is 202.42 pounds, and the variance is 228.60 pounds^2. To convert from pounds to kilos, we have to divide pounds by 2.2, that is $K = (1/2.204623)P = 0.453592P$. Thus, we obtain the following quantities.

$$\mu_K = 0.453592\mu_P = 0.453592(202.42) = 91.92 \qquad \sigma^2_K = (0.453592)^2\sigma^2_P = (0.453592)^2(228.60) = 47.03$$
$$\sigma_K = \sqrt{47.03} = 6.86$$

The population mean and variance of heights are 73.26 inches and 4.26 inches^2, respectively. To convert inches to centimeters, we have to multiply by 2.54, that is $C = 2.54I$. Thus, we obtain the following quantities.

$$\mu_C = 2.54\mu_I = 2.54(73.26) = 186.08 \qquad \sigma^2_C = (2.54)^2\sigma^2_I = (2.54)^2(4.26) = 27.48 \qquad \sigma_C = \sqrt{27.48} = 5.24$$

Note that in the metric system, the weights in kilograms are less variable than weights in pounds, while the heights in centimeters are more variable than than heights in inches.

For the female marathon runners, the mean and variance of their speeds were 5.84 mph and 0.69 mph^2, respectively. One mile represents 1.60394 kilometers, so that so that a person who runs M miles in 1 hour, runs $K = 1.60394M$ kilometers in one hour. This leads to the following quantities.

$$\mu_K = 1.60394(5.84) = 9.37 \qquad \sigma^2_K = (1.60394)^2(0.69) = 1.78 \qquad \sigma_K = \sqrt{1.78} = 1.33$$

∇

In many settings, we are interested in linear functions of a sequence of random variables: $Y_1, \ldots, Y_n$. Typically, we have fixed coefficients $a_1, \ldots, a_n$, and $E\{Y_i\} = \mu_i$, $V\{Y_i\} = \sigma^2_i$, and $\text{COV}\{Y_i, Y_j\} = \sigma_{ij}$.

$$W = \sum_{i=1}^n a_i Y_i \qquad E\{W\} = \mu_W = \sum_{i=1}^n a_i\mu_i \qquad V\{W\} = \sum_{i=1}^n a_i^2\sigma_i^2 + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^n a_i a_j \sigma_{ij}$$

If, as in many, but by no means all, cases, the Y_i values are independent ($\sigma_{ij} = 0$), the variance simplifies to $V\{W\} = \sum_{i=1}^n a_i^2\sigma_i^2$. A special case is when we have two random variables: X and Y, and a linear function $W = aX + bY$ for fixed constants. We have means μ_X, μ_Y, standard deviations σ_X, σ_Y, covariance σ_{XY}, and correlation ρ_{XY}.

$$W = aX + bY \qquad E\{W\} = a\mu_X + b\mu_Y \qquad V\{W\} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma_{XY} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\rho_{XY}\sigma_X\sigma_Y$$

Some special cases include where we have: $a = 1, b = 1$ (sums), and $a = 1, b = -1$ (differences). This leads to the following results.

$$E\{X+Y\} = \mu_X + \mu_Y \qquad V\{X+Y\} = \sigma_X^2 + \sigma_Y^2 + 2\rho_{XY}\sigma_X\sigma_Y$$
$$E\{X-Y\} = \mu_X - \mu_Y \qquad V\{X-Y\} = \sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y$$

Example 3.9: Movie "Close Up" Scenes

Barry Salt has classified film shots along an ordinal scale for a "population" of 398 movies. The levels are (BCU=Big Close Up, CU=Close Up, MCU=Medium Close Up, MLS=Medium Long Shot, LS=Long Shot, and VLS=Very Long Shot). We consider X to be the number of Big Close Up's and Y to be the number of Close Up's in a film. For this population, $\mu_X = 28.84$, $\mu_Y = 79.23$, $\sigma_X = 31.48$, $\sigma_Y = 61.37$, and $\rho_{XY} = 0.51$. We obtain the population mean, variance, and standard deviations of the sum of Big Close Up's and Close Up's $(X+Y)$ and the difference between Big Close Up's and Close Up's $(X-Y)$.

$E\{X+Y\} = 28.84+79.23 = 108.07 \quad V\{X+Y\} = 31.48^2+61.37^2+2(0.51)(31.48)(61.37) = 6727.83 \quad \sigma_{X+Y} = 82.02$

$E\{X-Y\} = 28.84-79.23 = -50.39 \quad V\{X-Y\} = 31.48^2+61.37^2-2(0.51)(31.48)(61.37) = 2786.70 \quad \sigma_{X-Y} = 52.79$

`Source: http://www.cinemetrics.lv/salt.php`

∇

3.3.1 Common Families of Discrete Probability Distributions

Here we consider some commonly used families of discrete probability distributions, namely the Binomial, Poisson, and Negative Binomial families. These are used in many situations where data are counts of numbers of events occurring in an experiment.

Binomial Distribution

A binomial "experiment" is based on a series of Bernoulli trials with the following characteristics.

- The experiment consists of n trials or observations.
- Trial outcomes are independent of one another.
- Each trial can end in one of two possible outcomes, often labeled **S**uccess or **F**ailure.
- The probability of Success, π is constant across all trials.
- The random variable, Y, is the number of Successes in the n trials

Note that many experiments are well approximated by this model, and thus it has wide applicability. One problem that has been considered in great detail is the assumption of independence from trial to trial. A classic paper that looked at the "hot hand" in basketball shooting has led to many studies in sports involving the topic is Gilovich, Vallone, and Tversy, 1985, [23].

The probability of any sequence of y Successes and $n-y$ Failures is $\pi^y(1-\pi)^{n-y}$ for $y = 0, 1, \ldots, n$. The number of ways to observe y successes in n trials makes use of combinations described previously. The number of ways of choosing y positions from $1, 2, \ldots, n$ is $C_y^n = \frac{n!}{y!(n-y)!} = \binom{n}{y}$. For instance, there is only one way observing either 0 or n Successes, there are n ways of observing 1 or $n-1$ Successes, and so on. This leads to the following probability distribution for $Y \sim Bin(n, \pi)$.

$$P(Y = y) = p(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y} \quad y = 0, 1, \ldots, n \qquad \sum_{y=0}^{n} p(y) = (\pi + (1-\pi))^n = 1^n = 1$$

Statistical packages and spreadsheets have functions for computing probabilities for the Binomial (and all distributions covered in these notes). In R, the function **dbinom(y,n,π)** returns $P(Y = y) = p(y)$ (the probability "density") when $Y \sim Bin(n, \pi)$.

To obtain the mean and variance of the Binomial distribution, consider the n independent trials individually (these are referred to as **Bernoulli** trials). Let $S_i = 1$ if trial i is a success, and $S_i = 0$ if it is a failure. Then Y, the number of Successes is the sum of the independent S_i values, leading to the following results.

$$E\{S_i\} = 1\pi + 0(1-\pi) = \pi \qquad E\{S_i^2\} = 1^2\pi + 0^2(1-\pi) = \pi \qquad V\{S_i\} = E\{S_i^2\} - (E\{S_i\})^2 = \pi - \pi^2 = \pi(1-\pi)$$

$$Y = \sum_{i=1}^{n} S_i \quad \Rightarrow \quad E\{Y\} = \mu_Y = \sum_{i=1}^{n} E\{S_i\} = n\pi \qquad V\{Y\} = \sigma_Y^2 = \sum_{i=1}^{n} V\{S_i\} = n\pi(1-\pi) \qquad \sigma_Y = \sqrt{n\pi(1-\pi)}$$

Example 3.10: Experiments of Mobile Phone Telepathy

A set of experiments was conducted to determine whether people displayed evidence of telepathy in receiving mobile phone calls (Sheldrake, Smart, and Avraamides, 2015, [47]). Each subject received 6 calls from one of two potential callers. Each subject predicted which caller was calling. Assuming random guessing, the number of successful predictions should be Binomial, with $n = 6$ trials, and probability of Success $\pi = 0.5$, since there were two potential callers. The probabilities of 0,1,2,...,6 successes for a subject in the experiment are given below. A plot of the probability distribution is given in Figure 3.1.

$$\frac{6!}{0!(6-0)!} = \frac{6!}{6!(6-6)!} = 1 \quad \frac{6!}{1!(6-1)!} = \frac{6!}{5!(6-5)!} = 6 \quad \frac{6!}{2!(6-2)!} = \frac{6!}{4!(6-4)!} = 15 \quad \frac{6!}{3!(6-3)!} = 20$$

$$.5^y(1-.5)^{6-y} = .5^6 = .015625$$

y	$\pi = 0.50 : p(y)$	$\pi = 0.56 : p(y)$	$\pi = 0.50$: Expected #	$\pi = 0.56$: Expected #	Observed #
0	.015625	.007256	1.72	0.80	1
1	.093750	.055412	10.31	6.10	5
2	.234375	.176310	25.78	19.39	18
3	.312500	.299193	34.38	32.91	37
4	.234375	.285594	25.78	31.42	31
5	.093750	.145393	10.31	15.99	15
6	.015625	.030841	1.72	3.39	3
Total	1	1	110	110	110

Table 3.5: Probability Distribution for Number of successful prediction for mobile telephone telepathy study

$$p(0) = p(6) = .015625 \quad p(1) = p(5) = .09375 \quad p(2) = p(4) = .234375 \quad p(3) = .3125$$

R Output

```
### Output
> (p_y <- dbinom(y, 6, 0.5))  ## Obtain p(y) for y=0,1,...,6
[1] 0.015625 0.093750 0.234375 0.312500 0.234375 0.093750 0.015625
```

The mean, variance, and standard deviation of the number of Successful predictions in the $n = 6$ trials under this model are as follow.

$$\mu_Y = n\pi = 6(0.5) = 3 \qquad \sigma_Y^2 = n\pi(1-\pi) = 6(0.5)(1-0.5) = 1.5 \qquad \sigma_Y = \sqrt{1.5} = 1.2247$$

For the Sheldrake, et al study, [47], 110 subjects completed 6 trials each (660 total trials). There were a total of 369 hits (there appears to be a typo saying 370 in their Table 3). This corresponds to a proportion of 369/660=.559, in other words, these subjects in aggregate showed better than expected success in predicting callers. Table 3.5 gives the probability distributions for $\pi = 0.50$ and $\pi = 0.56$, along with expected counts under the two models and the observed counts ($N = 110$ subjects).

∇

Poisson Distribution

In many applications, researchers observe the counts of a random process in some fixed amount of time or space. The random variable Y is a count that can take on any non-negative integer. One important aspect of the Poisson family is that the mean and variance are the same. This is one aspect that does not work for all applications. We use the notation: $Y \sim Poi(\lambda)$. The probability distribution, mean and variance of Y are:

$$p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, \ldots; \quad \lambda > 0 \qquad E\{Y\} = \mu_Y = \lambda \qquad V\{Y\} = \sigma_Y^2 = \lambda$$

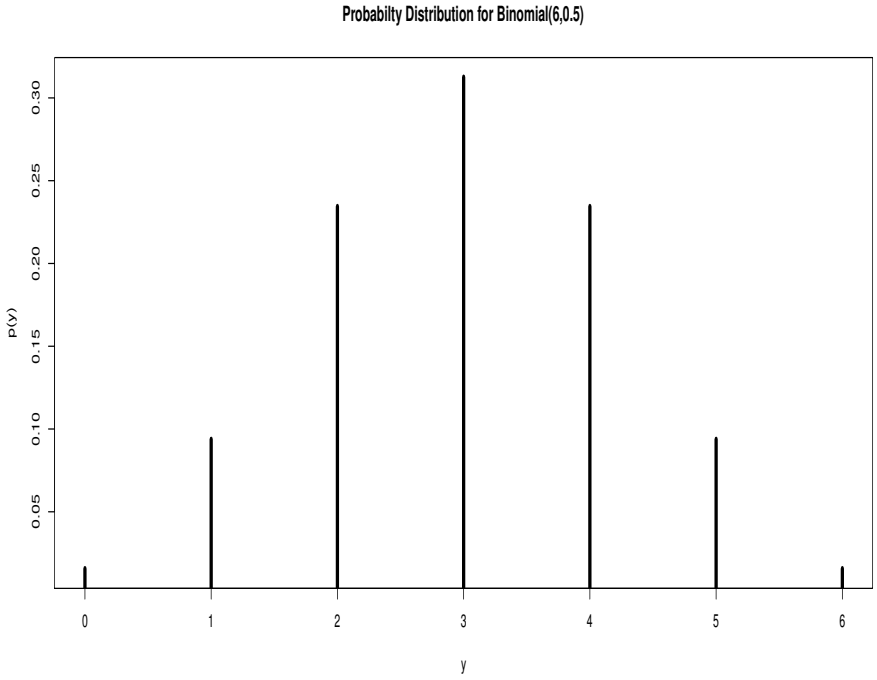

Figure 3.1: Probability Distribution for Mobile Telephone Telepathy experiment assuming random guessing, $Y \sim$Bin(6,0.5)

Note that $\lambda > 0$. The Poisson arises by dividing the time/space into n "infinitely" small areas, each having either 0 or 1 Success, with Success probability $\pi = \lambda/n$. Then Y is the number of areas having a success.

$$p(y) = \frac{n!}{y!(n-y)!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} = \frac{n(n-1)\cdots(n-y+1)}{y!}\left(\frac{\lambda}{n}\right)^y\left(1-\frac{\lambda}{n}\right)^{n-y} =$$

$$= \frac{1}{y!}\left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\cdots\left(\frac{n-y+1}{n}\right)\lambda^y\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-y}$$

The limit as n goes to ∞ is:

$$\lim_{n\to\infty} p(y) = \frac{1}{y!}(1)(1)\cdots(1)\lambda^y e^{-\lambda}(1) = p(y) = \frac{e^{-\lambda}\lambda^y}{y!} \quad y = 0, 1, 2...$$

The mean and variance for the Poisson distribution are both λ. This restriction can be problematic in many applications, and the Negative Binomial distribution (described below) is often used when the variance exceeds the mean.

Example 3.11: E Coli Bacterial Cell Counts

A study considered the distribution of bacterial cell counts for various bacteria strains in single-cell studies (Koyama, et al, 2016 [33]). There were 8 strains, and the authors observed counts for 96 cells under target means of $\lambda = 1$ and $\lambda = 2$ in an experimental study. They found that the observed counts were highly consistent with the Poisson models. The theoretical probability distributions are given as follow.

$$\lambda = 1: \quad p(y) = \frac{e^{-1}1^y}{y!} = \frac{e^{-1}}{y!} \quad y = 0, 1, 2, \ldots \qquad \lambda = 2: \quad p(y) = \frac{e^{-2}2^y}{y!} \quad y = 0, 1, 2, \ldots$$

∇

Example 3.12: London Bomb Hits in World War II

A widely reported application of the Poisson Distribution involves the counts of the number of bombs hitting among 576 areas of $0.5km^2$ in south London during WWII (Clarke (1946), [15], also reported in Feller (1950), [22]). There were a total of 537 bombs hit with a mean of $537/576 = .9323$. Table 3.6 gives the counts, and their expected counts ($576p(y)$) for the occurrences of 0 bombs, 1 bomb, ..., ≥ 5 bombs (the last cell involves 1 area which was hit 7 times).

Negative Binomial Distribution

The negative binomial distribution is used in two quite different contexts. The first is where a binomial type experiment is being conducted, except instead of having a fixed number of trials, the experiment is completed when the r^{th} success occurs. The random variable Y is the number of trials needed until the r^{th} success, and can take on any integer value greater than or equal to r. The probability distribution, its mean and variance are given below.

$$p(y) = \binom{y-1}{r-1}\pi^r(1-\pi)^{y-r} \qquad E\{Y\} = \mu_Y = \frac{r}{\pi} \qquad V\{Y\} = \sigma_Y^2 = \frac{r(1-\pi)}{\pi^2}.$$

y	$p(y)$	Expected #	Observed #
0	.3936	226.71	229
1	.3670	211.39	211
2	.1711	98.55	93
3	.0532	30.64	35
4	.0124	7.14	7
≥ 5	.0027	1.56	1
Total	1	576	576

Table 3.6: Probability Distribution for Number of bombs hitting within 576 areas on a grid in the south of London during World War II

A second use of the negative binomial distribution is as a model for count data. It arises from a mixture of Poisson models. In this setting it has 2 parameters and is more flexible than the Poisson (which has the variance equal to the mean), and can take on any non-negative integer value. In this form, the negative binomial distribution and its mean and variance can be written as follow (see e.g. Agresti (2002) [1] and Cameron and Trivedi (2005) [11]).

$$f(y;\mu,\alpha) = \frac{\Gamma\left(\alpha^{-1}+y\right)}{\Gamma\left(\alpha^{-1}\right)\Gamma\left(y+1\right)}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu}{\alpha^{-1}+\mu}\right)^{y} \qquad \Gamma(w) = \int_0^\infty x^{w-1}e^{-x}dx = (w-1)\,\Gamma\left(w-1\right).$$

$$E\left\{Y\right\} = \mu \qquad V\left\{Y\right\} = \mu\left(1+\alpha\mu\right).$$

Example 3.13: Number of Comets Observed per Year - 1789-1888

The number of comets observed per year for the century 1789-1888 inclusive were reported by Chambers, 1889, [12] and included in a large number of datasets by Thorndike, 1926, [50]. The annual number of comets ranged from 0 (19 years) to 9 (1 year), with frequency counts and computations for the mean and variance given in Table 3.7, treating this as a population of years. The mean and variance are given below, along with "method of moments" estimates for μ and α for the Negative Binomial distribution.

$$\mu_Y = \sum_y yp(y) = 2.58 \qquad \sigma_Y^2 = \sum_y y^2 p(y) - \mu_Y^2 = 11.36 - 2.58^2 = 4.70$$

$$\sigma^2 = \mu(1+\alpha\mu) \quad \Rightarrow \quad \alpha = \frac{\sigma^2/\mu - 1}{\mu} = \frac{4.70/2.58 - 1}{2.58} = 0.32$$

The Negative Binomial appears to fit better than a Poisson distribution with mean 2.58, based on observed and expected counts, this will be quantified later.

∇

y	# comets	$p(y)$	$yp(y)$	$y^2p(y)$	Exp(Poi)	Exp(NegBin)
0	19	.19	0.00	0.00	7.58	15.22
1	19	.19	0.19	0.19	19.55	21.54
2	17	.17	0.34	0.68	25.22	20.11
3	14	.14	0.42	1.26	21.69	15.54
4	13	.13	0.52	2.04	13.99	10.76
5	8	.08	0.40	2.00	7.22	6.93
6	4	.04	0.24	1.44	3.10	4.24
7	2	.02	0.14	0.98	1.14	2.50
8	3	.03	0.24	1.92	0.37	1.43
≥ 9	1	.01	0.09	0.81	0.14	1.73
Total	100	1	2.58	11.36	100	100

Table 3.7: Probability Distribution for Number of Comets Observed for years 1789-1888

3.4 Continuous Random Variables

Continuous random variables can take on any values along a continuum. Their distributions are described as densities, with probabilities being assigned as areas under the curve. Unlike discrete random variables, individual points have no probability assigned to them. While discrete probabilities and means and variances make use of summation, continuous probabilities and means and variances are obtained by integration. The following rules and results are used for continuous random variables and probability distributions. We use $f(y)$ to denote a probability density function and $F(y)$ to dentote the cumulative distribution function.

$$f(y) \geq 0 \qquad \int_{-\infty}^{\infty} f(y)dy = 1 \qquad P(a \leq Y \leq b) = \int_a^b f(y)dy \qquad F(y) = \int_{-\infty}^{y} f(t)dt$$

$$E\{Y\} = \mu_Y = \int_{-\infty}^{\infty} yf(y)dy \qquad V\{Y\} = \sigma_Y^2 = \int_{-\infty}^{\infty} (y-\mu_Y)^2 f(y)dy = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu_Y^2 \qquad \sigma_Y = +\sqrt{\sigma_Y^2}$$

3.4.1 Common Families of Continuous Probability Distributions

Three commonly applied families of distributions for describing populations of continuous measurements are the **normal**, **gamma**, and **beta** families, although there are many other families also used in practice.

The normal distribution is symmetric and mound-shaped. It has two parameters: a mean and variance (the standard deviation is often used in software packages). Many variables have distributions that are modeled well by the normal distribution, and many estimators have **sampling distributions** that are approximately normal. The gamma distribution has a density over positive values that is skewed to the right. There are many applications where data are skewed with a few extreme observations, such as the marathon running times observed previously. The gamma distribution also has two parameters associated with it. The beta distribution is often used to model data that are proportions (or can be extended to any finite length interval). The beta distribution also has two parameters. All of these families can take on a wide range of shapes by changing parameter values.

Probabilities, quantiles, densities, and random number generators for specific distributions and parameter values can be obtained from many statistical software packages and spreadsheets such as EXCEL. We will use R throughout these notes.

Normal Distribution

The normal distributions, also known as the Gaussian distributions, are a family of symmetric mound-shaped distributions. The distribution has 2 parameters: the mean μ and the variance σ^2, although often it is indexed by its standard deviation σ. We use the notation $Y \sim N(\mu, \sigma)$. The probability density function, the mean and variance are:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \quad -\infty < y < \infty, -\infty < \mu < \infty, \sigma > 0 \quad E\{Y\} = \mu_Y = \mu \quad V\{Y\} = \sigma_Y^2 = \sigma^2$$

The mean μ defines the center (median and mode) of the distribution, and the standard deviation σ is a measure of the spread ($\mu - \sigma$ and $\mu + \sigma$ are the inflection points). Despite the differences in location and spread of the different distributions in the normal family, probabilities with respect to standard deviations from the mean are the same for all normal distributions. For $-\infty < z_1 < z_2 < \infty$, we have:

$$P(\mu + z_1\sigma \le Y \le \mu + z_2\sigma) = \int_{\mu+z_1\sigma}^{\mu+z_2\sigma} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) dy = \int_{z1}^{z_2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \Phi(z_2) - \Phi(z_1).$$

Here Z is **standard normal**, a normal distribution with mean 0, and variance (standard deviation) 1. $\Phi(z^*)$ is the cumulative distribution function of the standard normal distribution, up to the point z^*:

$$\Phi(z^*) = \int_{-\infty}^{z^*} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

These probabilities and critical values can be obtained directly or indirectly from standard tables, statistical software, or spreadsheets. Note that:

$$Y \sim N(\mu, \sigma) \qquad \Rightarrow \qquad Z = \frac{Y-\mu}{\sigma} \sim N(0,1).$$

This makes it possible to use the standard normal table to obtain probabilities and quantiles for any normal distribution. Plots of three normal distributions are given in Figure 3.2.

Approximately 68% (.6826) of the probability lies within 1 standard deviation from the mean, 95% (.9544) lies within 2 standard deviations, and virtually all (.9970) lies within 3 standard deviations.

Example 3.14: NHL Player Body Mass Indices

Previously, we saw that the Body Mass Indices (BMI) of National Hockey League players for the 2013-2014 season were mound shaped with a mean of 26.50 and standard deviation 1.45. Figure 3.3 gives a histogram along with the corresponding normal density. There is a tendency to observe more actual BMI's in the center than the normal distribution would imply, but the normal model seems to be reasonable.

Consider the following quantiles (.10, .25, .50, .75, .90) for the NHL data and the corresponding N(26.50, 1.45) distribution. Also consider the probabilities of the following ranges ($< 26.50 - 2(1.45) = 23.60, >$

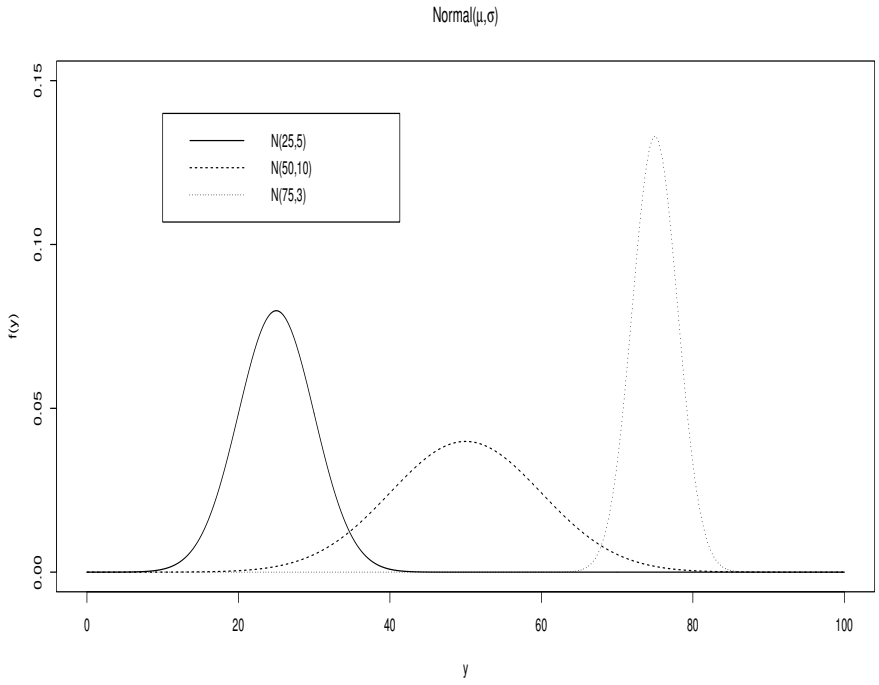

Figure 3.2: Three Normal Densities

$26.50 + 2(1.45) = 29.40$, and $(25.05 = 26.50 - 1.45, 26.50 + 1.45 = 27.95))$ for the NHL data and the normal distribution.

R Output

```
### Output
> round(q.out, 3)
              10%   25%     50    75%    90%
Theoretical 24.637 25.52 26.500 27.481 28.363
Empirical   24.702 25.62 26.516 27.439 28.342
>
> round(p.out, 4)
            <mu-2sigma (mu-sigma,mu+sigma) >mu+2sigma
Theoretical     0.0228              0.6827     0.0228
Empirical       0.0265              0.7057     0.0279
```

The quantiles and probabilities are very similar, showing the normal model is a reasonable approximation to the distribution of NHL BMI values.

∇

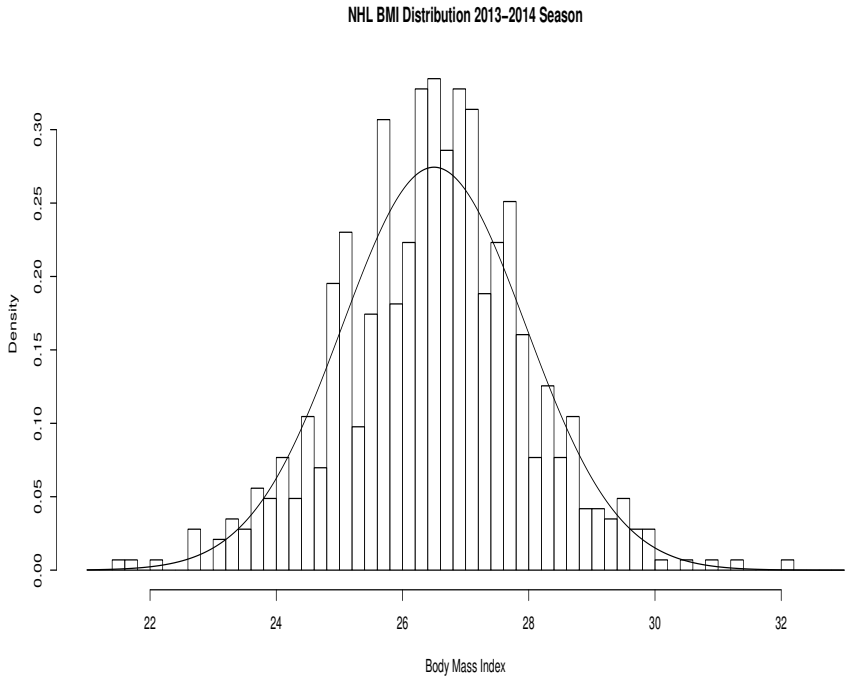

Figure 3.3: NHL Body Mass Indices and Normal Distribution

Gamma Distribution

The gamma family of distributions are used to model non-negative random variables that are often right-skewed. There are two widely used parameterizations. The first given here is in terms of *shape* and *scale* parameters.

$$f(y) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} y^{\alpha-1} e^{-y/\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \alpha\beta \qquad V\{Y\} = \sigma_Y^2 = \alpha\beta^2$$

Here, $\Gamma(\alpha)$ is the gamma function $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$ and is built-in to virtually all statistical packages and spreadsheets. It also has two simple properties.

$$\alpha > 1: \quad \Gamma(\alpha) = (\alpha - 1)\,\Gamma(\alpha - 1) \qquad\qquad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

Thus, if α is an integer, $\Gamma(\alpha) = (\alpha - 1)!$. The second parameterization given here is in terms of *shape* and *rate* parameters.

$$f(y) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-y\beta} \quad y \geq 0, \alpha > 0, \beta > 0 \qquad E\{Y\} = \mu_Y = \frac{\alpha}{\beta} \qquad V\{Y\} = \sigma_Y^2 = \frac{\alpha}{\beta^2}$$

Note that different software packages use the different parameterizations in generating samples and giving tail-areas and critical values. For instance, EXCEL uses the first parameterization and R uses the second. Figure 3.4 displays three gamma densities of various shapes.

Example 3.15: Rock and Roll Marathon Speeds

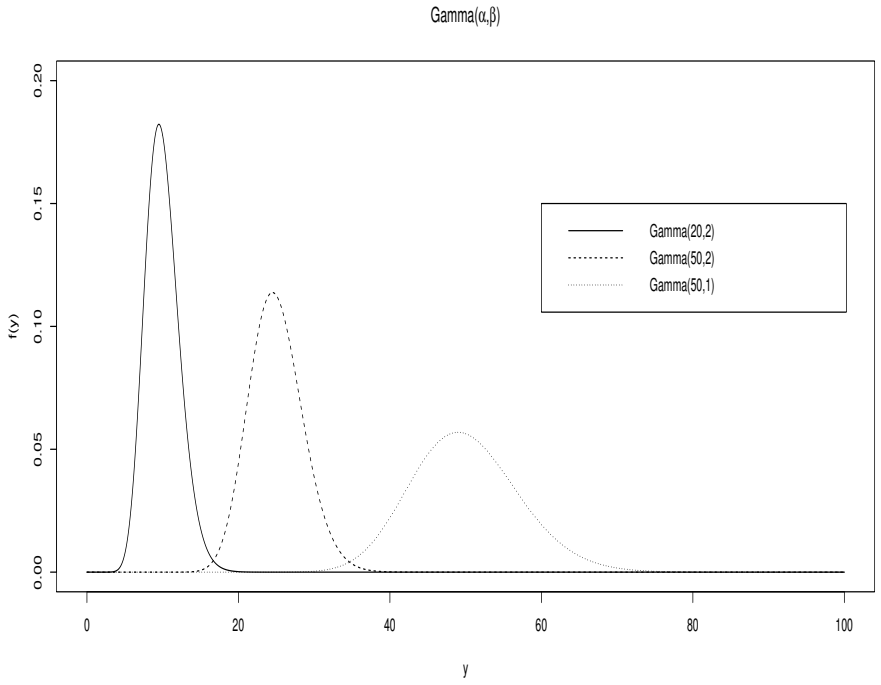

Figure 3.4: Three Gamma Densities

As seen previously, when considering females and males separately, the distributions of running speeds are all positive, and skewed to the right. The means for females and males were 5.8398 and 6.3370, respectively; and the variances were 0.6906 and 1.1187, respectively. Using the second formulation of the gamma distribution, with $\mu = \alpha/\beta$ and $\sigma^2 = \alpha/\beta^2$, we obtain the following parameter values for the two distributions based on the method of moments.

$$\frac{\mu^2}{\sigma^2} = \frac{(\alpha/\beta)^2}{\alpha/\beta^2} = \alpha \qquad \frac{\mu}{\sigma^2} = \frac{\alpha/\beta}{\alpha/\beta^2} = \beta$$

$$\text{Females: } \alpha_F = \frac{5.8398^2}{0.6906} = 49.38 \qquad \beta_F = \frac{5.8398}{0.6906} = 8.46$$

$$\text{Males: } \alpha_M = \frac{6.3370^2}{1.1187} = 35.90 \qquad \beta_M = \frac{6.3370}{1.1187} = 5.66$$

Histograms of the actual speeds and the corresponding Gamma densities are given in Figure 3.5. Similar to what was done for the NHL BMI measurements, we compare the theoretical quantiles for the female and male speeds with the actual quantiles, and compare theoretical probabilities for females and males with observed probabilities. There is very good agreement between the quantiles. The extreme probabilities do

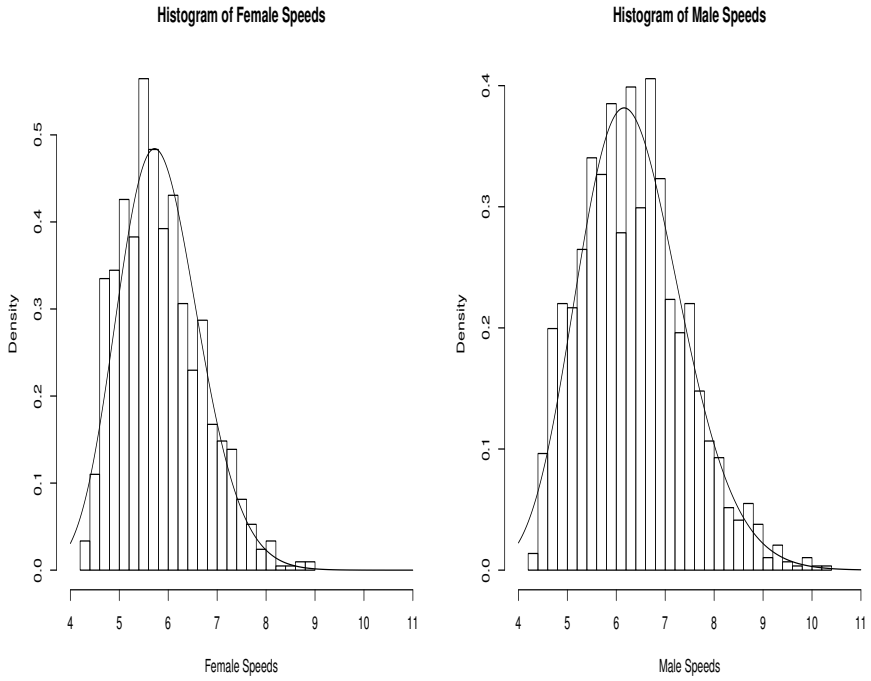

Figure 3.5: Rock and Roll Marathon speeds and Gamma Distributions for Females and Males

not match up as well, but still show fairly good agreement, with exception of no actual cases falling more than 2 standard deviations below the means.

R Output

```
## Output
> round(q.out, 3)
                     10%   25%    50   75%   90%
Theoretical/Female 4.803 5.260 5.800 6.377 6.927
Empirical/Female   4.811 5.203 5.711 6.357 7.015
Theoretical/Male   5.025 5.595 6.278 7.015 7.725
Empirical/Male     4.970 5.561 6.277 6.986 7.718

> round(p.out, 4)
                 <mu-2sigma (mu-sigma,mu+sigma) >mu+2sigma
Theoretical/Female     0.0146              0.6843     0.0298
Empirical/Female       0.0000              0.6622     0.0364
Theoretical/Male       0.0131              0.6850     0.0309
Empirical/Male         0.0000              0.6651     0.0365
```

∇

Two special cases are the exponential family, where $\alpha = 1$ and the Chi-square family, with $\alpha = \nu/2$ and $\beta = 2$ for integer valued ν. For the exponential family, based on the second parameterization, the symbol β

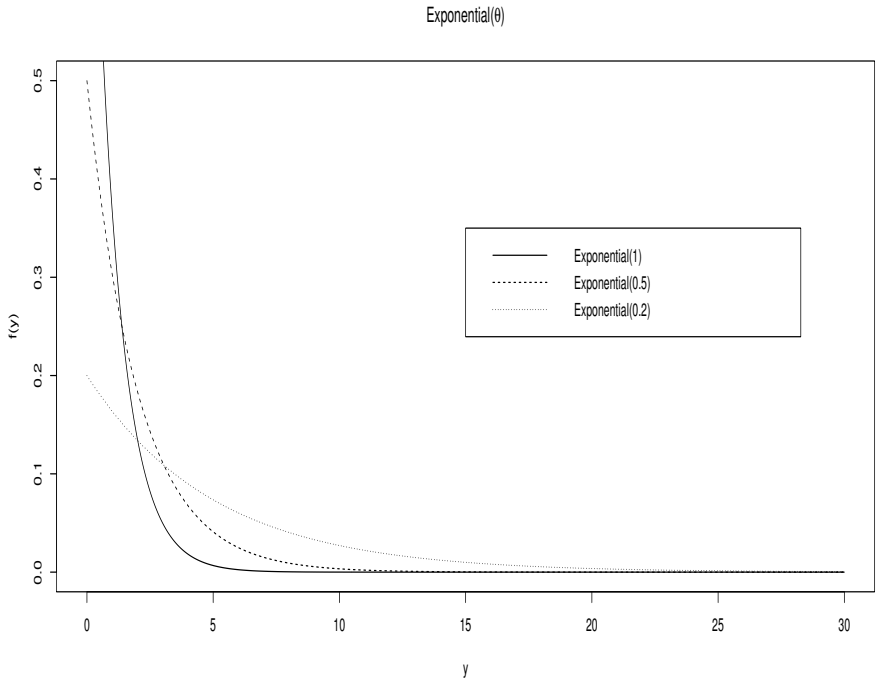

Figure 3.6: Three Exponential Densities

is often replaced by θ.

$$f(y) = \theta e^{-y\theta} \qquad E\{Y\} = \mu_Y = \frac{1}{\theta} \qquad V\{Y\} = \sigma_Y^2 = \frac{1}{\theta^2}.$$

Probabilities for the exponential distribution are trivial to obtain as $F(y^*) = 1 - e^{-y^*\theta}$. Figure 3.6 gives three exponential distributions.

For the chi-square family, based on the first parameterization, we have the following.

$$f(y) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\nu/2}} y^{\frac{\nu}{2}-1} e^{-y/2} \qquad E\{Y\} = \mu_Y = \nu \qquad V\{Y\} = \sigma_Y^2 = 2\nu$$

Here, ν is the **degrees of freedom** and we denote the distribution as: $Y \sim \chi^2_\nu$. Upper and lower critical values of the chi-square distribution are available in tabular form, and in statistical packages and spreadsheets. Probabilities, quantiles, densities, and random samples can be obtained with statistical packages and spreadsheets. The chi-square distribution is widely used in statistical testing as will be seen later. Figure 3.7 gives three Chi-square distributions.

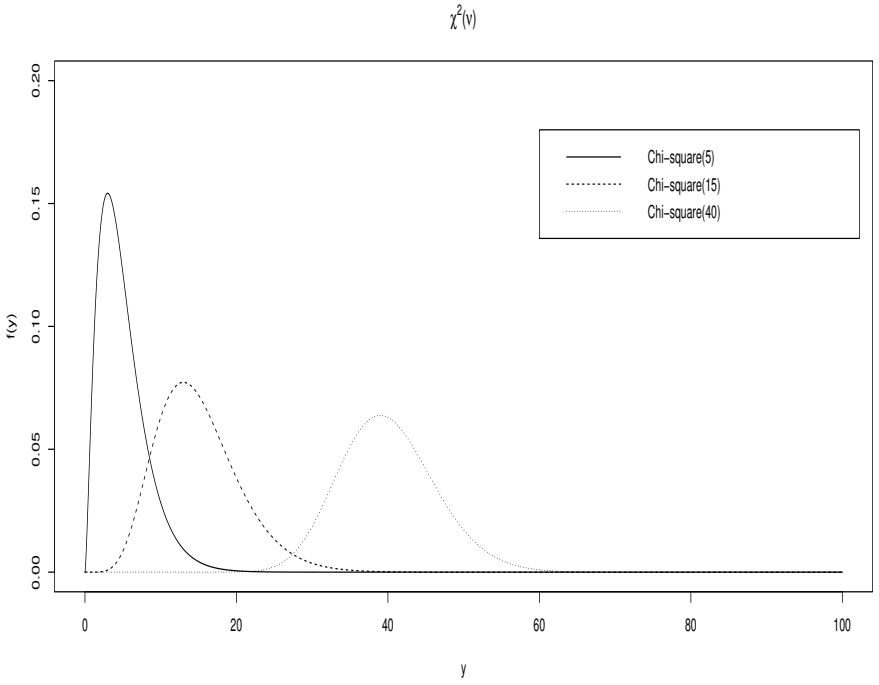

Figure 3.7: Three Chi-Square Densities

Beta Distribution

The Beta distribution can be used to model data that are proportions (or percentages divided by 100). The traditional model for the Beta distribution is given below.

$$f(y;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1} \qquad 0<y<1; \quad \alpha>0,\beta>0 \qquad \int_0^1 w^a(1-w)^b\,dw=\frac{\Gamma(a+1)\Gamma(b+1)}{\Gamma(a+b+2)}$$

Note that the Uniform distribution is a special case, with $\alpha=\beta=1$. The mean and variance of the Beta distribution are given here.

$$E\{Y\}=\frac{\alpha}{\alpha+\beta} \qquad V\{Y\}=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$$

An alternative formulation of the distribution involves a re-parameterization as follows.

$$\mu=\frac{\alpha}{\alpha+\beta} \qquad \phi=\alpha+\beta \quad \Rightarrow \quad \alpha=\mu\phi \qquad \beta=(1-\mu)\phi$$

$$V\{Y\}=\sigma^2=\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}=\frac{\mu(1-\mu)\phi^2}{\phi^2(\phi+1)}=\frac{\mu(1-\mu)}{\phi+1} \quad \Rightarrow \quad \phi=\frac{\mu(1-\mu)}{\sigma^2}-1$$

Figure 3.8 gives three Beta distributions.

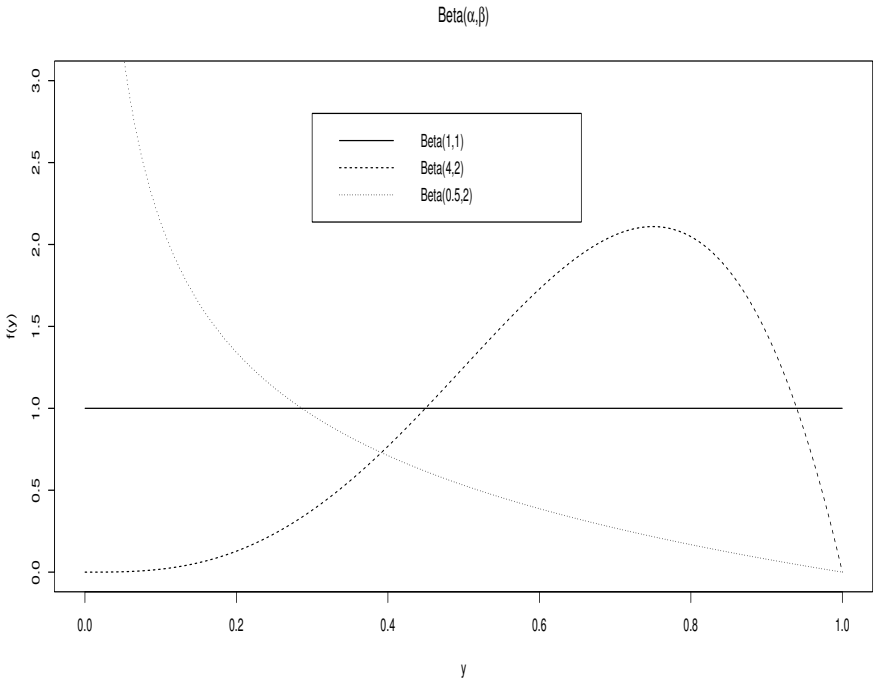

Figure 3.8: Three Beta Densities

Example 3.16: NBA 3-Point Field Goal Proportion by Team/Game - 2016/2017 Regular Season

During the NBA 2016/2017 regular season, each of the 30 teams played 82 games, for a total of 2460 team/games. For each team/game, the 3-Point field goal proportion is obtained by dividing the the number made by the number attempted. The number attempted per team/game ranged from 7 to 61, with mean and median of 27, and standard deviation of 6.7. Among the proportions made, the mean and standard deviation are 0.3566 and 0.0947, respectively. These lead to the following parameters based on the method of moments.

$$\phi = \frac{0.3566(1-0.3566)}{0.0947^2} - 1 = 24.60 \qquad \alpha = 24.60(.3566) = 8.77 \qquad \beta = 24.60(1-.3566) = 15.83$$

A histogram of the data and the corresponding Beta density are given in Figure 3.9. As with the previous examples, we compare the theoretical quantiles and probabilities for the beta denstities with the actual values for this population. They show considerable agreement.

R Output

```
## Output
> round(q.out, 3)
             10%   25%    50   75%   90%
Theoretical 0.237 0.289 0.353 0.420 0.482
```

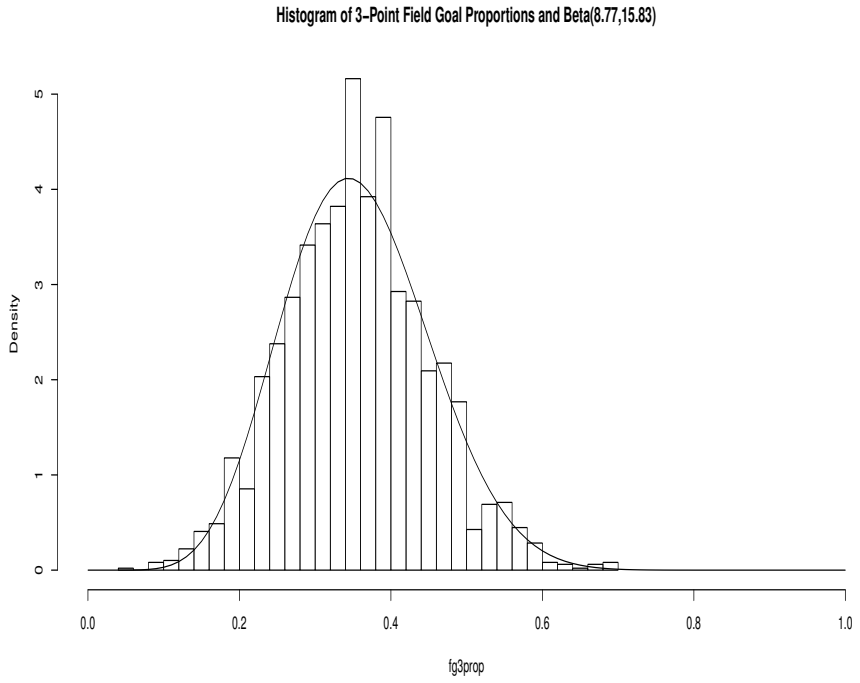

Figure 3.9: Three Point Field Goal proportions by team/game - NBA 2016/2017 regular season

```
Empirical   0.238 0.294 0.355 0.415 0.478

> round(p.out, 4)
           <mu-2sigma (mu-sigma,mu+sigma) >mu+2sigma
Theoretical     0.0139              0.6742     0.0282
Empirical       0.0236              0.6829     0.0297
```

∇

3.4.2 Functions of Normal Random Variables

First, note that if $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$. Many software packages present Z-tests as (Wald) χ^2-tests.

Suppose $Y_1, ..., Y_n$ are independent with $Y_i \sim N(\mu, \sigma)$ for $i = 1, \ldots, n$. Then the sample mean and sample variance are computed as follow.

$$\overline{Y} = \frac{\sum_{i=1}^n Y_i}{n} \qquad S^2 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{n-1}$$

In this case, we obtain the following sampling distributions for the mean and a function of the variance.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \qquad \overline{Y}, \quad \frac{(n-1)S^2}{\sigma^2} \text{ are independent.}$$

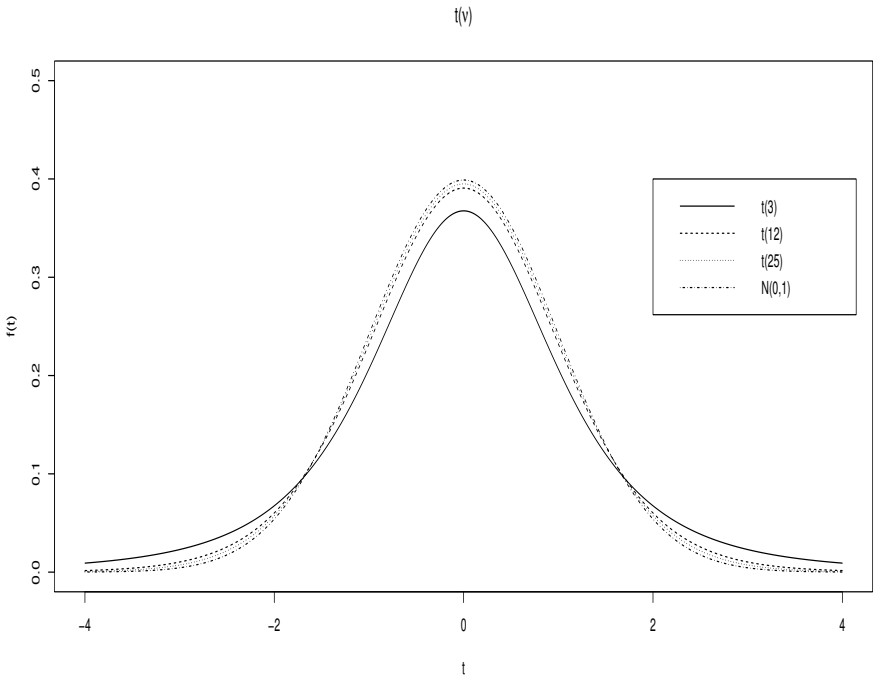

Figure 3.10: Three t-densities and z

Note that in general, if $Y_1, ..., Y_n$ are normally distributed (and not necessarily with the same mean and/or variance), any linear function of them will be normally distributed, with mean and variance given previously in the section with linear functions of random variables.

Two distributions associated with the normal and chi-square distributions are **Student's** t and F. Student's t-distribution is similar to the standard normal ($N(0,1)$), except that it is indexed by its degrees of freedom and that it has heavier tails than the standard normal. As its degrees of freedom approach infinity, its distribution converges to the standard normal. Let $Z \sim N(0,1)$ and $W \sim \chi^2_\nu$, where Z and W are independent. Then, we have the following result.

$$Y \sim N(\mu, \sigma) \quad \Rightarrow \quad Z = \frac{Y-\mu}{\sigma} \sim N(0,1) \qquad\qquad T = \frac{Z}{\sqrt{W/\nu}} \sim t_\nu$$

where the probability density, mean, and variance for Student's t-distribution are:

$$f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\sqrt{\nu\pi}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}} \qquad E\{T\} = \mu_T = 0 \qquad V\{T\} = \frac{\nu}{\nu-2} \quad \nu > 2$$

and we use the notation $T \sim t_\nu$. Three t-distributions, along with the standard normal (z) distribution are shown in Figure 3.10.

Now consider the sample mean and variance, and the fact they are independent.

$$\overline{Y} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad Z = \frac{\overline{Y}-\mu}{\sqrt{\frac{\sigma^2}{n}}} = \sqrt{n}\frac{\overline{Y}-\mu}{\sigma} \sim N(0,1)$$

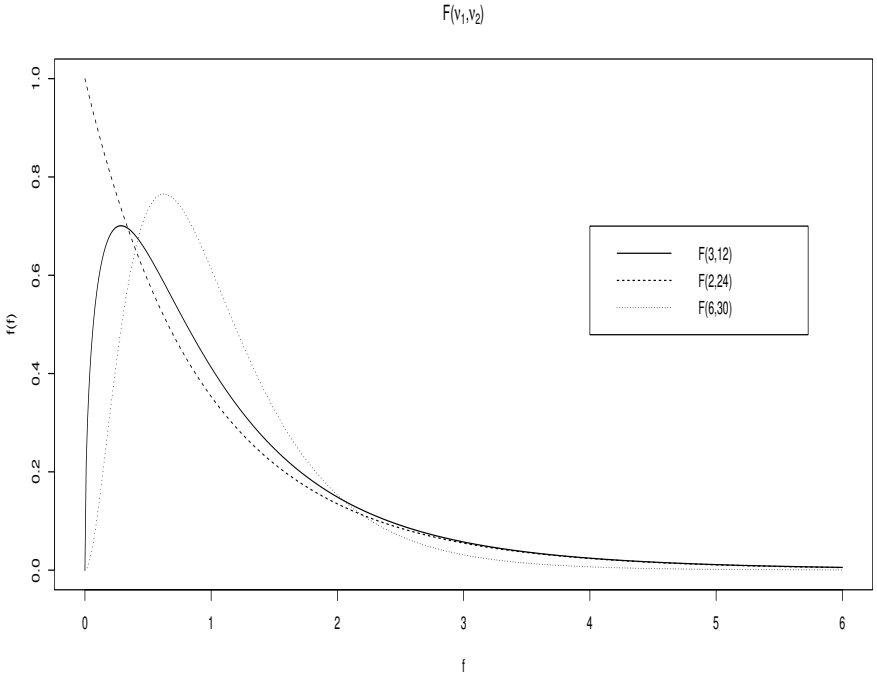

Figure 3.11: Three F-densities

$$W = \frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow \quad \sqrt{\frac{W}{\nu}} = \sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}} = \frac{S}{\sigma}$$

$$\Rightarrow \quad T = \frac{Z}{\sqrt{W/\nu}} = \frac{\sqrt{n}\frac{\overline{Y}-\mu}{\sigma}}{\frac{S}{\sigma}} = \sqrt{n}\frac{\overline{Y}-\mu}{S} \sim t_{n-1}$$

The F-distribution arises often in Regression and Analysis of Variance applications. If $W_1 \sim \chi^2_{\nu_1}$, $W_2 \sim \chi^2_{\nu_2}$, and W_1, W_2 are independent, then:

$$F = \frac{\left[\frac{W_1}{\nu_1}\right]}{\left[\frac{W_2}{\nu_2}\right]} \sim F_{\nu_1,\nu_2}.$$

where the probability density, mean, and variance for the F-distribution are given below as a function of the specific point $F = f$.

$$f(f) = \left[\frac{\Gamma\left(\left(\frac{\nu_1+\nu_2}{2}\right)\right)\nu_1^{\nu_1/2}\nu_2^{\nu_2/2}}{\Gamma(\nu_1/2)\,\Gamma(\nu_2/2)}\right]\left[\frac{f^{\nu_1/2-1}}{(\nu_1 f+\nu_2)^{(\nu_1+\nu_2)/2}}\right]$$

$$E\{F\} = \mu_F = \frac{\nu_1}{\nu_2-2} \quad \nu_2 > 2 \qquad V\{F\} = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)(\nu_2-4)} \quad \nu_2 > 4$$

Three F-distributions are given in Figure 3.11.

Critical values for the t, χ^2, and F-distributions are given in statistical textbooks and webpages. Probabilities, quantiles, densities, and random samples can be obtained from many statistical packages and

spreadsheets. Technically, the t, χ^2, and F distributions described here are **central** t, **central** χ^2, and **central** F distributions. These will be made use of repeatedly when making inferences regarding population parameters.

3.5 Sampling Distributions and the Central Limit Theorem

Sampling distributions are the probability distributions of sample statistics across different random samples from a population. That is, if we take many random samples, compute the statistic for each sample, then save that value, what would be the distribution of those saved statistics? In particular, if we are interested in the sample mean $\overline{Y}$, or the sample proportion with a characteristic $\hat{\pi}$, we know the following results, based on independence of elements within a random sample.

$$\text{Sample Mean: } E\{Y_i\} = \mu \quad V\{Y_i\} = \sigma^2 \quad E\{\overline{Y}\} = E\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = n\left(\frac{1}{n}\right)\mu = \mu$$

$$V\{\overline{Y}\} = V\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 V\{Y_i\} = n\left(\frac{1}{n}\right)^2\sigma^2 = \frac{\sigma^2}{n}$$

$$SE\{\overline{Y}\} = \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$

$$\text{Sample Proportion: } E\{Y_i\} = \pi \quad V\{Y_i\} = \pi(1-\pi) \quad E\{\hat{\pi}\} = E\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = n\left(\frac{1}{n}\right)\pi = \pi$$

$$V\{\hat{\pi}\} = V\left\{\sum_{i=1}^{n}\left(\frac{1}{n}\right)Y_i\right\} = \sum_{i=1}^{n}\left(\frac{1}{n}\right)^2 V\{Y_i\} = n\left(\frac{1}{n}\right)^2\pi(1-\pi) = \frac{\pi(1-\pi)}{n}$$

$$SE\{\hat{\pi}\} = \sigma_{\hat{\pi}} = \sqrt{\frac{\pi(1-\pi)}{n}}$$

The standard deviation of the sampling distribution of a sample statistic (aka estimator) is referred to as its **standard error**. Thus $SE\{\overline{Y}\} = \sigma_{\overline{Y}}$ is the standard error of the sample mean, and $SE\{\hat{\pi}\} = \sigma_{\hat{\pi}}$ is the standard error of the sample proportion.

When the data are normally distributed, the sampling distribution of the sample mean is also normal. When the data are not normally distributed, as the sample size increases, the sampling distribution of the sample mean or proportion tends to normality. The "rate" of convergence to normality depends on how "non-normal" the underlying distribution is. The mathematical arguments for these results are **Central Limit Theorems**.

$$\text{Sample Mean: } \overline{Y} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad \text{Sample Proportion: } \hat{\pi} \dot{\sim} N\left(\pi, \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

Example 3.17: Sampling Distributions - NHL BMI, Female Marathon Speeds, Charlotte Traffic Stops

We consider the sampling distributions of sample means for the NHL player Body Mass Indices, Female Rock and Roll Marathon Speeds, and Charlotte, N.C. traffic stops (proportion of stops due to speed violations, category 7). For the NHL BMI data, the population mean is $\mu = 26.500$ and standard deviation is $\sigma = 1.454$. As the underlying distribution is approximately normal, the sampling distribution of the mean is approximately normal, regardless of the sample size. We take 10000 random samples of size $n = 9$, computing and saving the sample mean for each sample. The theoretical and empirical (based on the 10000 random samples) mean and standard error of the sample means are given below and a histogram with the normal density are shown in Figure 3.12.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 26.500 \quad \sigma_{\overline{Y}} = \frac{1.454}{\sqrt{9}} = 0.485 \qquad \text{Empirical: } \overline{\overline{y}} = 26.504 \quad s_{\overline{y}} = 0.485$$

The mean and standard deviation are very close to the corresponding theoretical values (they won't always be this close, as sampling error exists).

For the female marathon speeds, we saw that the distribution was skewed to the right, and well modeled by a gamma distribution with mean $\mu = 5.84$ and standard deviation $\sigma = 0.83$. We take 10000 random samples of $n = 16$ from this population, computing and saving the sample mean from each sample. The theoretical and empirical (based on the 10000 random samples) mean and standard error of the sample means are given below and a histogram with the normal density are shown in Figure 3.13.

$$\text{Theory: } \mu_{\overline{Y}} = \mu = 5.840 \quad SE\{\overline{Y}\} = \frac{0.831}{\sqrt{16}} = 0.208 \qquad \text{Empirical: } \overline{\overline{y}} = 5.839 \quad SE\{\overline{y}\} = 0.206$$

Again, we see very strong agreement between the empirical and theoretical values (as we should). Also, note that the sampling distribution is very well approximated by the N(5.840,0.208) in the graph.

Finally, we consider the proportions of traffic stops due to speeding for the Charlotte, NC traffic stops, based on 10000 random samples of $n = 50$. For the population, $\pi = 22222/79884 = .2782$. The theoretical and empirical results are given below, and the histogram is given in Figure 3.14.

$$\text{Theory: } \mu_{\hat{\pi}} = \pi = .2782 \quad \sigma_{\hat{\pi}} = \sqrt{\frac{.2782(1-.2782)}{50}} = .0634 \qquad \text{Empirical: } \overline{\hat{\pi}} = .2775 \quad SE\{\hat{\pi}\} = .0633$$

The empirical mean and standard error, again, are in strong agreement with their theoretical values. Note that the sample proportion is discrete as it can only take on values .00, .02, ..., .98, 1.00, as $n = 50$. The histogram is clearly bell-shaped like a normal distribution. As n gets larger $\hat{\pi}$ becomes more "continuous."

∇

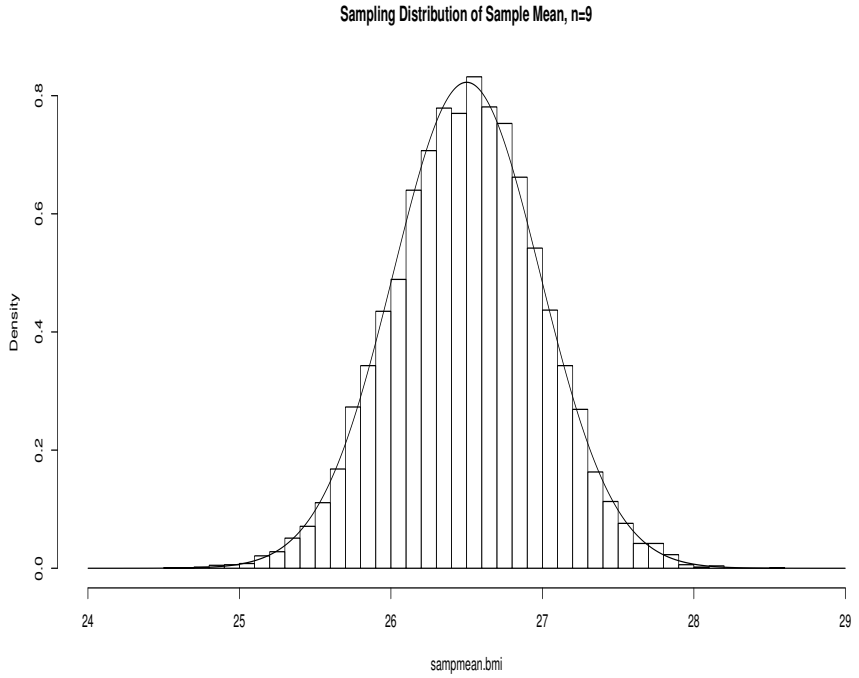

Figure 3.12: Sampling distribution for sample means (n=9) for NHL Body Mass Index

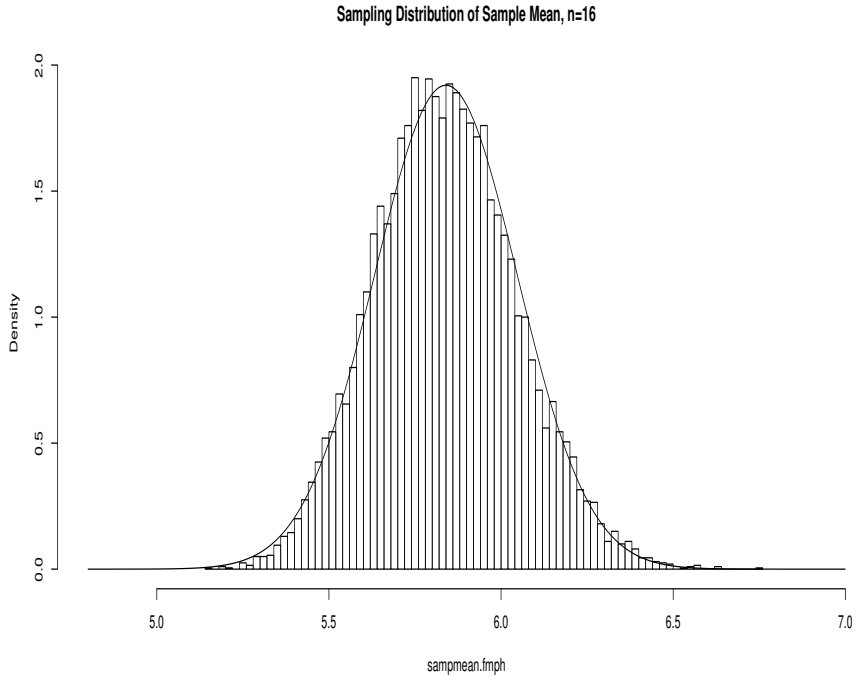

Figure 3.13: Sampling Distribution for sample means (n=16) for Female Rock and Roll Marathon speeds

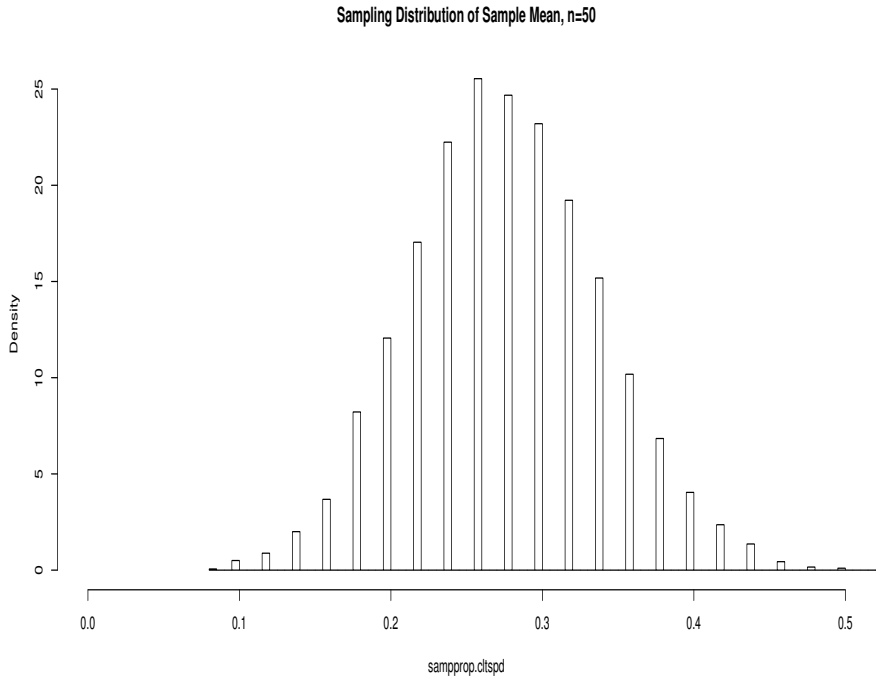

Figure 3.14: Sampling Distribution for sample proportions (n=50) for Charlotte traffic stops for speeding

3.6 R Code for Chapter 3

```
### Chapter 3

### Examples 3.6-3.7

## Read Driver Level Data into "nascard"
nascard <- read.fwf("http://www.stat.ufl.edu/~winner/data/nascard.dat",
   width=c(3,6,4,4,4,5,9,4,11,32), col.names=c("serRace", "year",
   "yrRace", "finPos", "strtPos", "lapsComp", "winnings", "numCars",
   "carMake", "driver"))

names(nascard)
nascard[1:100, c(1,4,5)]

## Subset rows of nascard w/ driver starting in 1st 10 positions in "start10"
## Save only columns: series Race, start and finish positions
start10 <- nascard[nascard$strtPos <= 10,c("serRace","strtPos","finPos")]
nrow(start10)
start10[1:30,]

(nraces <- length(unique(start10$serRace)))   ### number races=898

strt10Fin3 <- rep(0, nraces)    ### Initialize Top 3 Fins per race
strt10Len <- rep(0, nraces)
### Count # of top 10 starters finish in top 3
for (i in 1:nraces) {
  strt10Fin3[i] <- sum(start10[start10$serRace==i,]$finPos <=3)
  strt10Len[i] <- length(start10[start10$serRace==i,]$finPos)
```

```
}

strt10Len
strt10Fin3

(t.strt10Fin3 <- table(strt10Fin3))  ### Count 0,1,2,3 Top 3 finishers
t.strt10Fin3 / sum(t.strt10Fin3)     ### Turn counts to proportions

set.seed(12345)
sample.race <- sample(x=1:nraces, size=10000, replace=TRUE)
mean(strt10Fin3[sample.race])
sd(strt10Fin3[sample.race])

rm(list=ls(all=TRUE))

### Example 3.10

y <- 0:6 ## Values that Y can take on: y=0,1,...,6
(p_y <- dbinom(y, 6, 0.5)) ## Obtain p(y) for y=0,1,...,6

### Plot probabilities (type="h" is histogram) - Figure 3.1
plot(y, p_y, type="h", lwd=5, ylab="p(y)",
main="Probabilty Distribution for Binomial(6,0.5)")

rm(list=ls(all=TRUE))

### Figure 3.2 - Normal plots (3 on same graph)

y.seq <- seq(0,100,0.01)
mu <- c(25,50,75)
sigma <- c(5,10,3)

par(mfrow=c(1,1))
plot(y.seq, dnorm(y.seq,mu[1],sigma[1]), type="l", ylim=c(0,0.15),
 xlab="y", ylab="f(y)",
 main=expression(paste("Normal(", mu,",", sigma,")")))
lines(y.seq, dnorm(y.seq,mu[2],sigma[2]), lty=2)
lines(y.seq, dnorm(y.seq,mu[3],sigma[3]), lty=3)
legend(10,0.14, c("N(25,5)", "N(50,10)","N(75,3)"), lty=1:3)

rm(list=ls(all=TRUE))

### Example 3.14
### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)
N <- length(bmi.nhl)
(mean.bmi.nhl <- mean(bmi.nhl))
(sd.bmi.nhl <- sd(bmi.nhl)*sqrt((N-1)/N))
bmi <- seq(21,33,.01)

### Obtain histogram - Figure 3.3
hist(bmi.nhl, breaks=seq(21,33,0.2), xlab="Body Mass Index", freq=FALSE,
main="NHL BMI Distribution 2013-2014 Season")
lines(bmi, dnorm(bmi, mean.bmi.nhl, sd.bmi.nhl))

## Quantiles: Theoretical Normal, Empirical Distribution
q.the <- qnorm(c(.10,.25,.50,.75,.90), mean.bmi.nhl, sd.bmi.nhl)
q.emp <- quantile(bmi.nhl, c(.10,.25,.50,.75,.90))
q.out <- rbind(q.the, q.emp)
rownames(q.out) <- c("Theoretical", "Empirical")
```

```
colnames(q.out) <- c("10%", "25%", "50", "75%", "90%")
round(q.out, 3)

## Probabilities: Theoretical Normal, Actual Distribution
# Theoretical
p.the1 <- pnorm(mean.bmi.nhl-2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
p.the3 <- 1-pnorm(mean.bmi.nhl+2*sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
p.the2 <- pnorm(mean.bmi.nhl+sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl) -
          pnorm(mean.bmi.nhl-sd.bmi.nhl, mean.bmi.nhl, sd.bmi.nhl)
# Empirical
p.emp1 <- sum(bmi.nhl <= mean.bmi.nhl-2*sd.bmi.nhl)/N
p.emp3 <- sum(bmi.nhl >= mean.bmi.nhl+2*sd.bmi.nhl)/N
p.emp2 <- sum(bmi.nhl >= mean.bmi.nhl-sd.bmi.nhl &
                  bmi.nhl <= mean.bmi.nhl+sd.bmi.nhl)/N
p.out <- rbind(cbind(p.the1, p.the2, p.the3), cbind(p.emp1, p.emp2, p.emp3))
rownames(p.out) <- c("Theoretical", "Empirical")
colnames(p.out) <- c("<mu-2sigma","(mu-sigma,mu+sigma)",">mu+2sigma")
round(p.out, 4)


rm(list=ls(all=TRUE))

### Figure 3.4 - Gamma plots (3 on same graph)

y.seq <- seq(0,100,0.01)
alpha <- c(20,50,50)
beta <- c(2,2,1)

# win.graph(height=5.5, width=7.0)
par(mfrow=c(1,1))
plot(y.seq, dgamma(y.seq,alpha[1],beta[1]), type="l", ylim=c(0,0.20),
 xlab="y", ylab="f(y)",
 main=expression(paste("Gamma(", alpha,",", beta,")")))
lines(y.seq, dgamma(y.seq,alpha[2],beta[2]), lty=2)
lines(y.seq, dgamma(y.seq,alpha[3],beta[3]), lty=3)
legend(60,0.15, c("Gamma(20,2)", "Gamma(50,2)","Gamma(50,1)"), lty=1:3)

rm(list=ls(all=TRUE))

### Example 3.15

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)
## Obtain mean and standard deviation by gender
tapply(mph,Gender,mean)
tapply(mph,Gender,median)
tapply(mph,Gender,var)
tapply(mph,Gender,sd)
## Obtain the Gamma parameters (for plotting) of mph by gender
(alpha.f <- mean(mph[Gender=="F"])^2 / var(mph[Gender=="F"]))
(alpha.m <- mean(mph[Gender=="M"])^2 / var(mph[Gender=="M"]))
(beta.f <- mean(mph[Gender=="F"]) / var(mph[Gender=="F"]))
(beta.m <- mean(mph[Gender=="M"]) / var(mph[Gender=="M"]))

## Figure 3.5
## Set up a 1x2 grid for plots
par(mfrow=c(1,2))
## Histograms for Female and Male mph
hist(mph[Gender=="F"],breaks=25,main="Histogram of Female Speeds",
xlab="Female Speeds", xlim=c(4,11), freq=FALSE)
x.seq <- seq(4,11,.01)
lines(x.seq, dgamma(x.seq, alpha.f, beta.f))
```

```
hist(mph[Gender=="M"],breaks=25,main="Histogram of Male Speeds",
xlab="Male Speeds", xlim=c(4,11), freq=FALSE)
lines(x.seq, dgamma(x.seq, alpha.m, beta.m))
## Quantiles: Theoretical Gamma, Empirical Distribution
q.the.f <- qgamma(c(.10,.25,.50,.75,.90), alpha.f, beta.f)
q.emp.f <- quantile(mph[Gender=="F"], c(.10,.25,.50,.75,.90))
## Quantiles: Theoretical Gamma, Empirical Distribution
q.the.m <- qgamma(c(.10,.25,.50,.75,.90), alpha.m, beta.m)
q.emp.m <- quantile(mph[Gender=="M"], c(.10,.25,.50,.75,.90))

q.out <- rbind(q.the.f, q.emp.f, q.the.m, q.emp.m)
rownames(q.out) <- c("Theoretical/Female", "Empirical/Female",
                     "Theoretical/Male", "Empirical/Male")
colnames(q.out) <- c("10%", "25%", "50", "75%", "90%")
round(q.out, 3)

## Probabilities: Theoretical Normal, Actual Distribution
# Theoretical Female
(mean.mph.female <- alpha.f / beta.f)
(sd.mph.female <- sqrt(alpha.f) / beta.f)
(N.female <- length(mph[Gender=="F"]))
p.the1.f <- pgamma(mean.mph.female-2*sd.mph.female, alpha.f, beta.f)
p.the3.f <- 1-pgamma(mean.mph.female+2*sd.mph.female, alpha.f, beta.f)
p.the2.f <- pgamma(mean.mph.female+sd.mph.female, alpha.f, beta.f) -
          pgamma(mean.mph.female-sd.mph.female, alpha.f, beta.f)
# Actual Female
p.emp1.f <- sum(mph[Gender=="F"] <= mean.mph.female-2*sd.mph.female)/N.female
p.emp3.f <- sum(mph[Gender=="F"] >= mean.mph.female+2*sd.mph.female)/N.female
p.emp2.f <- sum(mph[Gender=="F"] >= mean.mph.female-sd.mph.female &
                mph[Gender=="F"] <= mean.mph.female+sd.mph.female)/N.female
# Theoretical Male
(mean.mph.male <- alpha.m / beta.m)
(sd.mph.male <- sqrt(alpha.m) / beta.m)
(N.male <- length(mph[Gender=="M"]))
p.the1.m <- pgamma(mean.mph.male-2*sd.mph.male, alpha.m, beta.m)
p.the3.m <- 1-pgamma(mean.mph.male+2*sd.mph.male, alpha.m, beta.m)
p.the2.m <- pgamma(mean.mph.male+sd.mph.male, alpha.m, beta.m) -
            pgamma(mean.mph.male-sd.mph.male, alpha.m, beta.m)
# Actual Male
p.emp1.m <- sum(mph[Gender=="M"] <= mean.mph.male-2*sd.mph.male)/N.male
p.emp3.m <- sum(mph[Gender=="M"] >= mean.mph.male+2*sd.mph.male)/N.male
p.emp2.m <- sum(mph[Gender=="M"] >= mean.mph.male-sd.mph.male &
                mph[Gender=="M"] <= mean.mph.male+sd.mph.male)/N.male

p.out <- rbind(cbind(p.the1.f, p.the2.f, p.the3.f),
               cbind(p.emp1.f, p.emp2.f, p.emp3.f),
               cbind(p.the1.m, p.the2.m, p.the3.m),
               cbind(p.emp1.m, p.emp2.m, p.emp3.m))
rownames(p.out) <- c("Theoretical/Female", "Empirical/Female",
                     "Theoretical/Male", "Empirical/Male")
colnames(p.out) <- c("<mu-2sigma","(mu-sigma,mu+sigma)",">mu+2sigma")
round(p.out, 4)

rm(list=ls(all=TRUE))

###### Exponential, Chi-Square, Beta Distribution Plots (3 per graph)

## Figure 3.6
y.seq <- seq(0,30,0.01)
alpha <- c(1,1,1)
beta <- c(1,0.5,0.2)

par(mfrow=c(1,1))
plot(y.seq, dexp(y.seq,beta[1]), type="l", ylim=c(0,0.50),
```

```
 xlab="y", ylab="f(y)",
 main=expression(paste("Exponential(", theta,")")))
lines(y.seq, dexp(y.seq,beta[2]), lty=2)
lines(y.seq, dexp(y.seq,beta[3]), lty=3)
legend(15,0.35, c("Exponential(1)", "Exponential(0.5)","Exponential(0.2)"),
               lty=1:3)

rm(list=ls(all=TRUE))

## Figure 3.7
y.seq <- seq(0,100,0.01)
nu <- c(5,15,40)

par(mfrow=c(1,1))
plot(y.seq, dchisq(y.seq,nu[1]), type="l", ylim=c(0,0.20),
 xlab="y", ylab="f(y)",
 main=expression(paste(chi^2,"(", nu,")")))
lines(y.seq, dchisq(y.seq,nu[2]), lty=2)
lines(y.seq, dgamma(y.seq,nu[3]), lty=3)
legend(60,0.18, c("Chi-square(5)", "Chi-square(15)",
                  "Chi-square(40)"), lty=1:3)

rm(list=ls(all=TRUE))


## Figure 3.8
y.seq <- seq(0,1,0.001)
alpha <- c(1,4,0.5)
beta <- c(1,2,2)

par(mfrow=c(1,1))
plot(y.seq, dbeta(y.seq,alpha[1],beta[1]), type="l", ylim=c(0,3.0),
 xlab="y", ylab="f(y)",
 main=expression(paste("Beta(", alpha,",", beta,")")))
lines(y.seq, dbeta(y.seq,alpha[2],beta[2]), lty=2)
lines(y.seq, dbeta(y.seq,alpha[3],beta[3]), lty=3)
legend(0.3,2.8, c("Beta(1,1)", "Beta(4,2)","Beta(0.5,2)"), lty=1:3)

rm(list=ls(all=TRUE))

### Example 3.16

nba2017 <- read.csv("http://www.stat.ufl.edu/~winner/data/nba_teamgame_20167.csv")
attach(nba2017); names(nba2017)
# Regular Season Games Only (GameType=1)
fg3prop <- fg3m[GameType==1]/fg3a[GameType==1]
summary(fg3a[GameType==1])
sd(fg3a[GameType==1])
# Function to compute phi, alpha, beta from mean, sd
betaShRtMeanSD <- function(mean, sd) {
if (mean <= 0 | sd <= 0) return("FAIL")
phi <- (mean*(1-mean) / sd^2) - 1
alpha <- mean*phi
beta <- (1-mean) * phi
return(list(phi=phi, alpha=alpha, beta=beta))
}
(fg3ab <- betaShRtMeanSD(mean(fg3prop),
sd(fg3prop)))
(mean.fg3 <- mean(fg3prop))
(sd.fg3 <- sd(fg3prop))
(N.fg3 <- length(fg3prop))

## Figure 3.9
hist(fg3prop, xlim=c(0,1), freq=FALSE, breaks=35,
```

```
main="Histogram of 3-Point Field Goal Proportions and Beta(8.77,15.83)")
x <- seq(0,1.0,0.01)
lines(x,dbeta(x, fg3ab$alpha, fg3ab$beta))

## Quantiles: Theoretical Beta, Actual Distribution
q.the <- qbeta(c(.10,.25,.50,.75,.90), fg3ab$alpha, fg3ab$beta)
q.emp <-quantile(fg3prop, c(.10,.25,.50,.75,.90))

q.out <- rbind(q.the, q.emp)
rownames(q.out) <- c("Theoretical", "Empirical")
colnames(q.out) <- c("10%", "25%", "50", "75%", "90%")
round(q.out, 3)


## Probabilities: Theoretical Beta, Empirical Distribution
# Theoretical
p.the1 <- pbeta(mean.fg3-2*sd.fg3, fg3ab$alpha, fg3ab$beta)
p.the3 <- 1-pbeta(mean.fg3+2*sd.fg3, fg3ab$alpha, fg3ab$beta)
p.the2 <- pbeta(mean.fg3+sd.fg3, fg3ab$alpha, fg3ab$beta) -
          pbeta(mean.fg3-sd.fg3, fg3ab$alpha, fg3ab$beta)
# Empirical
p.emp1 <- sum(fg3prop <= mean.fg3-2*sd.fg3)/N.fg3
p.emp3 <- sum(fg3prop >= mean.fg3+2*sd.fg3)/N.fg3
p.emp2 <- sum(fg3prop >= mean.fg3-sd.fg3 &
              fg3prop <= mean.fg3+sd.fg3)/N.fg3

p.out <- rbind(cbind(p.the1, p.the2, p.the3), cbind(p.emp1, p.emp2, p.emp3))
rownames(p.out) <- c("Theoretical", "Empirical")
colnames(p.out) <- c("<mu-2sigma","(mu-sigma,mu+sigma)",">mu+2sigma")
round(p.out, 4)

rm(list=ls(all=TRUE))

### Plots of t- and F- densties (3 per graph)

y.seq <- seq(-4,4,0.01)
nu <- c(3,12,25)

## Figure 3.10
par(mfrow=c(1,1))
plot(y.seq, dt(y.seq,nu[1]), type="l", ylim=c(0,0.50),
 xlab="t", ylab="f(t)",
 main=expression(paste("t(", nu,")")))
lines(y.seq, dt(y.seq,nu[2]), lty=2)
lines(y.seq, dt(y.seq,nu[3]), lty=3)
lines(y.seq, dnorm(y.seq,0,1), lty=4)
legend(2,0.4, c("t(3)", "t(12)","t(25)","N(0,1)"), lty=1:4)

rm(list=ls(all=TRUE))

y.seq <- seq(0,6,0.01)
nu1 <- c(3,2,6)
nu2 <- c(12,24,30)

## Figure 3.11
par(mfrow=c(1,1))
plot(y.seq, df(y.seq,nu1[1],nu2[1]), type="l", ylim=c(0,1.0),
 xlab="f", ylab="f(f)",
 main=expression(paste("F(", nu[1],",", nu[2],")")))
lines(y.seq, df(y.seq,nu1[2],nu2[2]), lty=2)
lines(y.seq, df(y.seq,nu1[3],nu2[3]), lty=3)
legend(4,0.7, c("F(3,12)", "F(2,24)","F(6,30)"), lty=1:3)

rm(list=ls(all=TRUE))
```

```
### Example 3.17

### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
bmi.nhl <- 703 * Weight / (Height^2)
N <- length(bmi.nhl)
mean(bmi.nhl)
sd(bmi.nhl) * sqrt((N-1)/N)
set.seed(34567)
num.sim <- 10000
num.sample <- 9
sampmean.bmi <- rep(0,num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.sample, replace=F)
sampmean.bmi[i] <- mean(bmi.nhl[sample])
}
mean(sampmean.bmi)
sd(sampmean.bmi)

## Figure 3.12
hist(sampmean.bmi, breaks=50, xlim=c(24,29), freq=F,
main="Sampling Distribution of Sample Mean, n=9")
bmi.seq <- seq(24,29,0.01)
lines(bmi.seq,dnorm(bmi.seq,mean(bmi.nhl),sd(bmi.nhl)/sqrt(num.sample)))

detach(nhl)
rm(list=ls(all=TRUE))


## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)
f.mph <- mph[Gender == "F"]
mean(f.mph)
sd(f.mph)
N <- length(f.mph)
num.sim <- 10000
num.sample <- 16
sampmean.fmph <- rep(0,num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.sample, replace=F)
sampmean.fmph[i] <- mean(f.mph[sample])
}
mean(sampmean.fmph)
sd(sampmean.fmph)

## Figure  3.13
hist(sampmean.fmph, breaks=100, xlim=c(4.80,7.00), freq=F,
main="Sampling Distribution of Sample Mean, n=16")
fmph.seq <- seq(4.80,7.00,0.01)
lines(fmph.seq,dnorm(fmph.seq,mean(f.mph),sd(f.mph)/sqrt(num.sample)))

detach(rr.mar)
rm(list=ls(all=TRUE))


## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)
```

```
table(RsnStop)
N <- length(RsnStop)
table(RsnStop)/N
num.sim <- 10000
num.sample <- 50
sampprop.cltspd <- rep(0,num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.sample, replace=F)
sampprop.cltspd[i] <- sum(RsnStop[sample] == 7) / num.sample
}
mean(sampprop.cltspd)
sd(sampprop.cltspd)

## Figure 3.14
hist(sampprop.cltspd, breaks=100, xlim=c(0,0.50), freq=F,
main="Sampling Distribution of Sample Mean, n=50")

rm(list=ls(all=TRUE))
```

Chapter 4

Inferences for Population Means and Medians

Researchers often are interested in making statements regarding unknown population means and medians based on sample data. There are two common methods for making inferences: **Estimation** and **Hypothesis Testing**. The two methods are related and make use of the sampling distribution of the sample mean when making statements regarding the population mean.

Estimation can provide a single "best" prediction of the population mean, a **point estimate**, or it can provide a range of values that hopefully encompass the true population mean, an **interval estimate**. Hypothesis testing involves setting an a priori (null) value for the unknown population mean, and measuring the extent to which the sample data contradict that value. Note that a confidence interval provides a credible set of values for the unknown population mean, and can be used to test whether or not the population mean is the null value. Both methods involve uncertainty as we are making statements regarding a population based on sample data.

4.1 Estimation

For large samples, the sample mean has an approximately normal sampling distribution centered at the population mean, μ, and a standard error $\sigma/\sqrt{n}$. When the data are normally distributed, the sampling distribution is normal for all sample sizes. For normal distributions, 95% of its density lies in the range (mean +/- 1.96 SD). Thus, when we take a random sample, we obtain the following probability statement regarding the sample mean.

$$\overline{Y} \dot{\sim} N\left(\mu, SE\{\overline{Y}\} = \frac{\sigma}{\sqrt{n}}\right) \quad \Rightarrow \quad P\left(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \overline{Y} \leq \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \approx 1-\alpha \qquad P\left(Z \geq z_a\right) = a$$

$$\Rightarrow \quad 1-\alpha \approx P\left(-z_{\alpha/2} \leq \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

Some commonly used coverage probabilities $(1-\alpha)$ are given here, along with the corresponding z values.

$$1-\alpha = .90 \Rightarrow \alpha = .10 \Rightarrow \frac{\alpha}{2} = .05 \Rightarrow z_{.05} = 1.645 \quad 1-\alpha = .95 \Rightarrow z_{.025} = 1.96 \quad 1-\alpha = .99 \Rightarrow z_{.005} = 2.576$$

Note that in the probability statements above, μ is a fixed, unknown constant in practice, and $\overline{Y}$ is a random variable that varies from sample to sample. The probability refers to the fraction of the samples that will provide sample means such that the lower and upper bounds "cover" μ. Also, in practice, σ will be unknown and need to be replaced by the sample standard deviation.

A Large-Sample $(1-\alpha)100\%$ Confidence Interval for a Population Mean μ is given below, where $\overline{y}$ and s are the observed mean and standard deviation from a random sample of size n and $\hat{SE}\{\overline{Y}\}$ represents the **estimated standard error** .

$$\overline{y} \pm z_{\alpha/2}\hat{SE}\{\overline{Y}\} \qquad \overline{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

When the data are normally distributed, for small samples (although this has shown to work well for other distributions), replace $z_{\alpha/2}$ with $t_{\alpha/2,n-1}$.

$$\overline{y} \pm t_{\alpha/2,n-1}\hat{SE}\{\overline{Y}\} \qquad \overline{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

Any software package or spreadsheet that is used to obtain a confidence interval for a mean (or difference between two means) will always use the version based on the t-distribution. There will be settings, when making confidence intervals for parameters, that there is no justification for using the t-distribution, and we will make use the z-distribution, as does statistical software packages.

Example 4.1: NHL Players' BMI

The Body Mass Indices for the NHL players are approximately normally distributed with mean $\mu = 26.500$ and standard deviation $\sigma = 1.454$. We take 10000 random samples of size $n = 12$, implying a standard error of $\sigma_{\overline{Y}} = 1.454/\sqrt{12} = 0.420$. We count the number of the 10000 sample means that lie in the ranges $\mu \pm z_{\alpha/2}\sigma_{\overline{Y}}$ for the three values of $1-\alpha$ given above.

Of the 10000 sample means, 8975 (89.75%) lied within $\mu \pm 1.645(.420)$, 9512 (95.12%) within $\mu \pm 1.96(.420)$, and 9902 (99.02%) within $\mu \pm 2.576(.420)$. Had we constructed intervals of the form $\overline{y} \pm z_{\alpha/2}(.420)$ for each sample mean, the coverage rates for μ would have been the same values (89.75%, 95.12%, 99.02%).

When the population standard error $SE\{\overline{Y}\} = \sigma/\sqrt{n}$ is replaced by the estimated standard error $\hat{SE}\{\overline{Y}\} = s/\sqrt{n}$, which varies from sample to sample, we find the coverage rates of the intervals decrease.

When constructing intervals of the form $\overline{y} \pm z_{\alpha/2} s/\sqrt{n}$, the coverage rates fall to 86.78%, 92.29%, and 97.58%, respectively. This is a by-product of the fact that the sampling distribution of the standard deviation is skewed right, and its median is below its mean. Whenever the sample standard deviation is small, the width of the constructed interval is shortened. When using the estimated standard error, replace $z_{\alpha/2}$ with the corresponding critical value for the t-distribution, with $n-1$ degrees of freedom: $t_{\alpha/2,n-1}$. For this case, with $n = 12$, we obtain $t_{.05,11} = 1.796$, $t_{.025,11} = 2.201$, and $t_{.005,11} = 3.106$. When z is replaced by the corresponding t values, the coverage rates for the constructed intervals with the estimated standard errors reach their nominal rates: 89.79%, 95.22%, and 99.15%, respectively.

For the first random sample of the 10000 generated, we observe $\overline{y} = 25.838$ and $s = 1.717$. The 95% Confidence Interval for μ based on the first sample is obtained as follows.

$$\overline{y} \pm t_{.025,n-1}\frac{s}{\sqrt{n}} \quad \equiv \quad 25.838 \pm 2.201\left(\frac{1.717}{\sqrt{12}}\right) \quad \equiv \quad 25.838 \pm 1.091 \quad \equiv \quad (24.747, 26.929)$$

Thus, this interval does contain $\mu = 26.500$.

R Output

```
### Output
> round(cover.out,4)
                 90% Confidence 95% Confidence 99% Confidence
Z - True SE              0.8975         0.9512         0.9902
Z - Estimated SE         0.8678         0.9229         0.9758
t - Estimated SE         0.8979         0.9522         0.9915
```

∇

Often, researchers choose the sample size so that the **margin of error** will not exceed some fixed level E with high confidence. That is, we want the difference between the sample and population means to be within E with confidence level $1-\alpha$. This means the width of a $(1-\alpha)100\%$ Confidence Interval will be $2E$. This can be done in one calculation based on using the z distribution, or more conservatively, by trivial iteration based on the t-distribution. Either way, we must have an approximation of σ based on previous research or a pilot study.

$$z: \quad E_z = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad n = \left(\frac{z_{\alpha/2}\sigma}{E_z}\right)^2 \qquad\qquad t: \text{Smallest } n \text{ such that } E_t \leq t_{\alpha/2,n-1}\frac{\sigma}{\sqrt{n}}$$

Example 4.2: Estimating Population Mean Male Marathon Speed

Suppose we want to estimate the population mean of the male Rock and Roll marathon running speeds within $E = 0.20$ miles per hour with 95% confidence. We treat the standard deviation as known, $\sigma = 1.058$. The calculation for the sample size based on the z-distribution is given below, followed by R commands that iteratively solve for n based on the t-distribution.

$$z: \quad z_{.025} = 1.96 \quad n = \left(\frac{1.96(1.058)}{0.20}\right)^2 = 107.5 \approx 108$$

R Output

```
## Output

> cbind(n, E.t)
       n       E.t
[1,] 110 0.1999336
```

Since n was needed to be so large, $z_{.025}$ and $t_{.025,n-1}$ are very close, and both methods give virtually the same n (108 and 110).

4.2 Hypothesis Testing

In hypothesis testing, a sample of data is used to determine whether a population mean is equal to some pre-specified level μ_0. It is rare, except in some situations to test whether the mean is some specific value based on historical level, or government or corporate specified level to have a null value to test. These tests are more common when comparing two or more populations or treatments and determining whether their means are equal. The elements of a hypothesis test are given below.

Null Hypothesis (H_0) Statement regarding a parameter that is to be tested. It always includes an equality, and the test is conducted assuming its truth.

Alternative (Research) Hypothesis (H_A) Statement that contradicts the null hypothesis. Includes "greater than" ($>$), "less than" ($<$),or "not equal too" ($\neq$)]

Test Statistic (T.S.) A statistic measuring the discrepancy between the sample statistic and the parameter value under the null hypothesis (where the equality holds).

Rejection Region (R.R.) Values of the Test Statistic for which the Null Hypothesis is rejected. Depends on the significance level of the test.

***P*-value** Probability under the null hypothesis (at the equality) of observing a Test Statistic as extreme or more extreme than the observed Test Statistic. Also known as the observed significance level.

Type I Error Rejecting the Null Hypothesis when in fact it is true. The Rejection Region is chosen so that this has a particular small probability ($\alpha = P$(Type I Error) is the **significance level** and is often set at 0.05).

Type II Error Failing to reject the Null Hypothesis when it is false. Depends on the true value of the parameter. Sample size is often selected so that it has a particular small probability for an important difference. $\beta = P$(Type II Error).

Power The probability the Null Hypothesis is rejected. When H_0 is true the power is $\pi = \alpha$, when H_A is true, it is $\pi = 1 - \beta$.

The testing procedure for a mean is based on the sampling distribution of $\overline{Y}$ being approximately normal with mean μ_0 under the null hypothesis. Also, when the data are normal the difference between the sample mean and μ_0 divided by its estimated standard error is distributed as t with $n-1$ degrees of freedom under the null hypothesis.

$$\overline{Y} \dot{\sim} N\left(\mu_0, SE\{\overline{Y}\} = \frac{\sigma}{\sqrt{n}}\right) \qquad \frac{\overline{Y}-\mu_0}{\hat{SE}\{\overline{Y}\}} = \frac{\overline{Y}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

When the absolute value of the t-statistic is large, there is evidence against the null hypothesis. Once a sample is taken (observed), and the sample mean $\overline{y}$ and sample standard deviation s are observed, the test is conducted as follows for 2-tailed, upper tailed, and lower tailed alternatives.

$$\text{2-tailed: } H_0: \mu = \mu_0 \quad H_A: \mu \neq \mu_0 \quad \text{T.S.: } t_{obs} = \frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } |t_{obs}| \geq t_{\alpha/2,n-1} \quad P = 2P\left(t_{n-1} \geq |t_{obs}|\right)$$

$$\text{Upper tailed: } H_0: \mu \leq \mu_0 \quad H_A: \mu > \mu_0 \quad \text{T.S.: } t_{obs} = \frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } t_{obs} \geq t_{\alpha,n-1} \quad P = P\left(t_{n-1} \geq t_{obs}\right)$$

$$\text{Lower tailed: } H_0: \mu \geq \mu_0 \quad H_A: \mu < \mu_0 \quad \text{T.S.: } t_{obs} = \frac{\overline{y}-\mu_0}{s/\sqrt{n}} \quad \text{R.R.: } t_{obs} \leq -t_{\alpha,n-1} \quad P = P\left(t_{n-1} \leq t_{obs}\right)$$

The form of the rejection regions are given for 2-tailed, Upper and Lower tailed tests in Figure 4.1. These are based on $\alpha = 0.05$, and $n = 16$. The vertical lines lie at $t_{.975,15} = -t_{.025,15} = -2.131$ and $t_{.025,15} = 2.131$ for the 2-tailed test, $t_{.05,15} = 1.753$ for the Upper tailed test, and $t_{.95,15} = -t_{.05,15} = -1.753$ for the Lower tailed test.

When the Null Hypothesis is false, the test statistic is distributed as non-central t with non-centrality parameter given below.

$$H_0: \mu = \mu_0 \qquad \text{In reality: } \mu = \mu_A \neq \mu_0 \qquad \Delta = \frac{\mu_A - \mu_0}{\sigma/\sqrt{n}} \qquad t = \frac{\overline{Y}-\mu_0}{S/\sqrt{n}} \dot{\sim} t_{n-1,\Delta}$$

Power probabilities, which depend on whether the test is 2-tailed or 1-tailed can be obtained from statistical software packages, such as R, but not directly in EXCEL.

$$\text{2-tailed tests: } \pi = P\left(t_{n-1,\Delta} \leq -t_{\alpha/2,n-1}\right) + P\left(t_{n-1,\Delta} \geq t_{\alpha/2,n-1}\right)$$

$$\text{Lower tailed tests: } \pi = P\left(t_{n-1,\Delta} \leq -t_{\alpha,n-1}\right) \qquad \text{Upper tailed tests: } \pi = P\left(t_{n-1,\Delta} \geq t_{\alpha,n-1}\right)$$

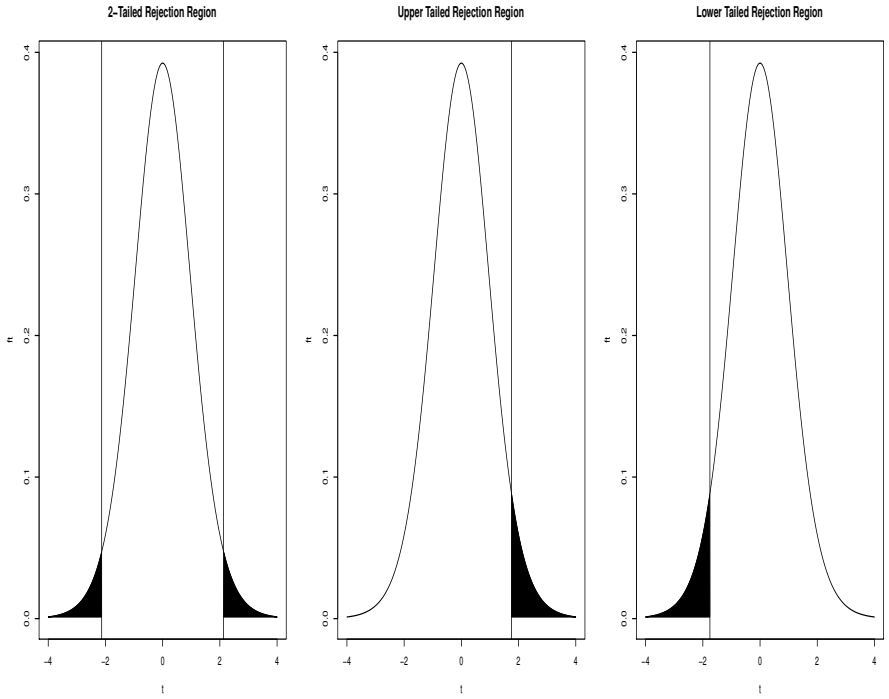

Figure 4.1: Rejection Regions for 2-tailed, Upper and Lower tailed tests, with $\alpha = 0.05$ and $n = 16$

While it is rare to use hypothesis testing regarding a single mean (except in the case where data are paired differences within individual units), the procedure is demonstrated based on male Rock and Roll marathon speeds with several vlaues of μ_0.

Example 4.3: Male Rock and Roll Marathon Speeds

For the males participating in the Rock and Roll marathon, the population mean speed was $\mu = 6.337$ miles per hour with standard deviation of $\sigma = 1.058$. We will demonstrate hypothesis testing regarding a single mean by first testing $H_0 : \mu = 6.337$ versus $H_A : \mu \neq 6.337$, based on random samples of $n = 40$. Since the null hypothesis is true, if the test is conducted with a Type I Error rate of $\alpha = 0.05$, the test should reject the null in approximately 5% of samples. The distribution of the test statistic is t with $n - 1 = 39$ degrees of freedom. Further, the P-values should approximate a Uniform distribution between 0 and 1. Note that 482 (4.82%) of the 10000 samples reject the null hypothesis, in agreement with what is to be expected. A histogram of the observed test statistics, along with the t-density, and the P-values and the the Uniform density are given in Figure 4.2. The two vertical bars on the t-statistic plot are at $\pm t_{.025,39} = \pm 2.023$.

Next consider cases where the null hypothesis is not true. Consider $H_{01} : \mu = 6$ versus $H_{A1} : \mu \neq 6$ and $H_{02} : \mu = 6.5$ versus $H_{A2} : \mu \neq 6.5$. Since the null value for H_{02} is closer to the true value $\mu_A = 6.337$ than the null value for H_{01}, we expect that we will reject H_{02} less often for tests based on the same sample size. That is, the power is higher for H_{01} than H_{02}. The non-centrality parameters and the corresponding power values are given below, based on samples of $n = 40$.

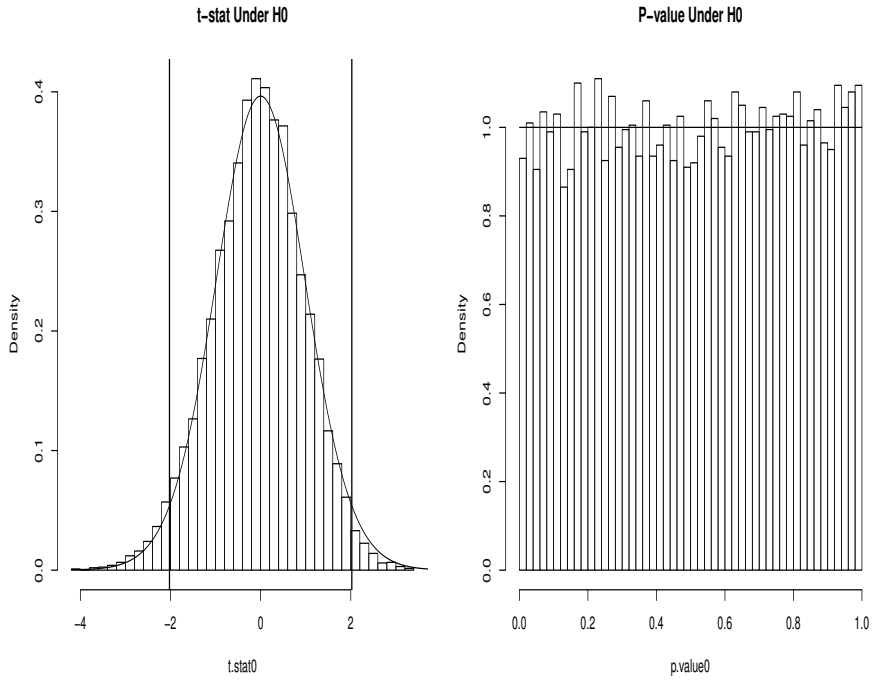

Figure 4.2: t-statistics and P-values for testing $H_0 : \mu = 6.337$

$$\Delta_1 = \frac{6.337 - 6.0}{1.058/\sqrt{40}} = 2.015 \qquad \pi_1 = .5022 \qquad\qquad \Delta_2 = \frac{6.337 - 6.5}{1.058/\sqrt{40}} = -0.974 \qquad \pi_2 = .1583$$

Based on 10000 random samples from the male marathon speeds, 49.93% rejected $H_0 : \mu = 6$, and for another set of 10000 random samples, 17.05% rejected $H_0 : \mu = 6.5$. The histogram of the test statistics and the non-central t-distribution are given in Figure 4.3 for testing $H_0 : \mu = 6$.

R Output

```
## Output

> round(power.out, 4)
          Delta Theoretical Power Empirical Power
mu0=6.33  0.0000           0.0500          0.0482
mu0=6.00  2.0150           0.5022          0.4993
mu0=6.50 -0.9748           0.1583          0.1705
```

∇

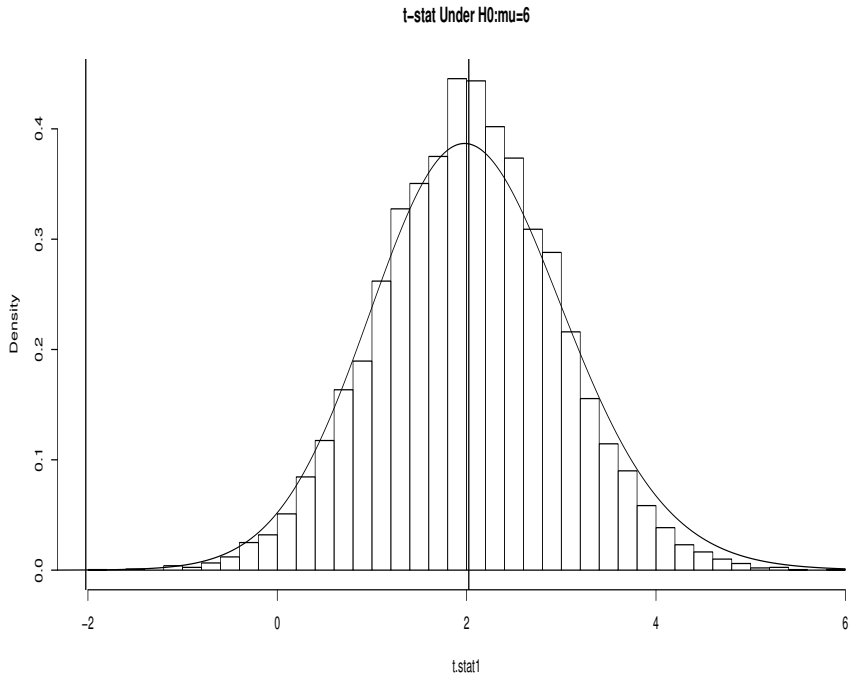

Figure 4.3: t-statistics and non-central t-distribution for testing $H_0 : \mu = 6.0$

4.2.1 Choosing Sample Size for Fixed Power for an Alternative

Once an important difference $\mu_A - \mu_0$ is determined, and an estimate of σ is obtained, the functions involving the non-central t-distribution can be used iteratively to find the n that makes the power large enough. The algorithm goes as follows for 2-tailed tests.

1. Choose an important difference $\mu_A - \mu_0$ and σ. Or alternatively make the difference in units of σ: $(\mu_A - \mu_0)/\sigma$.
2. Start with a small value for n, and compute the critical values for the t-test: $CV_{LO} = -t_{\alpha/2,n-1}$, $CV_{HI} = t_{\alpha/2,n-1}$.
3. Compute $\Delta = (\mu_0 - \mu_A)/\left(\sigma/\sqrt{n}\right)$.
4. Obtain the probability the test statistic falls in the Rejection Region, based on the non-central t-distribution, with $n-1$ degrees of freedom, and non-centrality parameter Δ: Power = pt(CV_{LO},$n-1$,Δ) + (1-pt(CV_{HI},$n-1$,Δ))
5. Continue increasing n until Power exceeds some specified value (typically 0.80 or higher).

Example 4.4: Male Rock and Roll Marathon Speeds

Suppose we would like to be able to detect a difference between μ_A and μ_0 of 0.25 with power of $\pi = 0.8$ when the test is conducted at $\alpha = 0.05$. In this case, recall $\sigma = 1.058$. Start with $n = 3$.

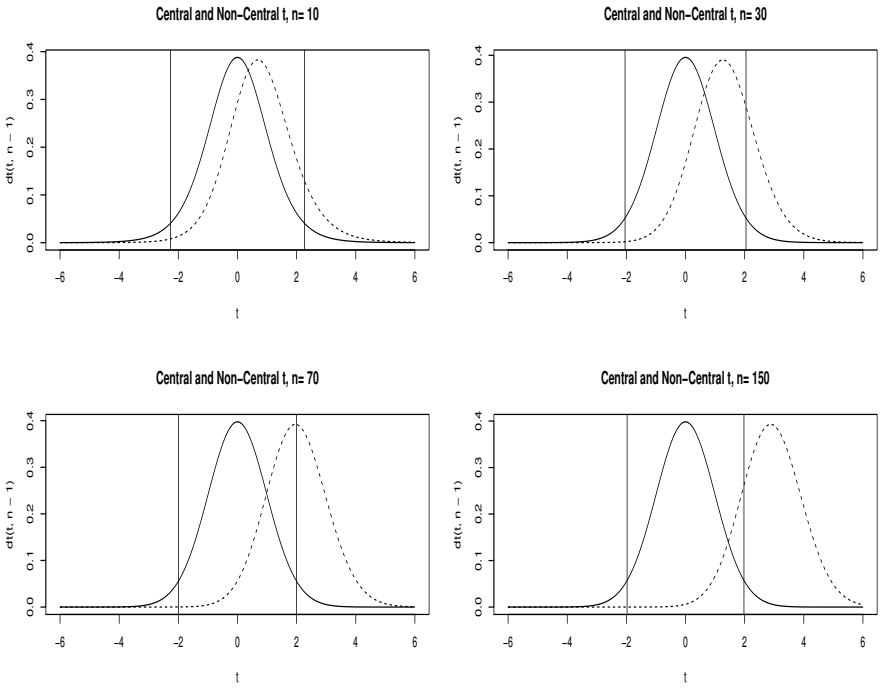

Figure 4.4: Central and non-Central t-distributions for n=10, 30, 70, 150, $\mu_0 - \mu_A = 0.25$, and $\sigma = 1.058$

$$t_{.025,3-1} = 4.303 \qquad \Delta = \frac{0.25}{1.058/\sqrt{3}} = \sqrt{3}\frac{0.25}{1.058} = 0.409$$

$$\pi = P\left(t_{3-1,0.409} \leq -4.303\right) + P\left(t_{3-1,0.409} \geq 4.303\right) = .0577$$

Keep increasing n, which affects the critical t-values (making them smaller in absolute value) and increasing Δ, thus increasing the power of the test, until $\pi \geq 0.80$. It ends up that we would need a sample of $n = 143$ to meet the power requirement. The target difference is very small (0.25) relative to the standard deviation (1.058) which is why such a large sample would be needed. A plot of the central and non-central t-distributions for n=10, 30, 70, and 150 is given in Figure 4.4. The vertical bars give the critical values for the $\alpha = 0.05$ level test.

R Output

```
## Output

> (power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta)))
[1] 0.05772603
> cbind(n, power)
       n     power
[1,] 143 0.8013787
```

∇

4.3 Inferences Concerning the Population Median

The population median represents the 50^{th} percentile of the distribution. For each sampled observation, there is a 0.5 probability that it is larger (or smaller) than the median. The number of observations of a random sample of size n that are above (or below) the median is binomial with n trials, and probability of success $\pi = 0.5$. Let $B_{\alpha/2,n}$ be the smallest number such that $P\left(Y \leq B_{\alpha/2,n} | Y \sim Bin(n, 0.5)\right) \leq \alpha/2$. Then the probability that the number of sample observations falling above or below the median will lie in the range $\left(L_{\alpha/2} = B_{\alpha/2,n} + 1, U_{\alpha/2} = n - B_{\alpha/2,n}\right)$ will be greater than or equal to $1 - \alpha$. This leads to a $(1 - \alpha)100\%$ Confidence Interval for the population median to be the range encompassed by the $(L_{\alpha/2})^{th}$ ordered observation to the $(U_{\alpha/2})^{th}$ ordered observation.

A large-sample approximation based on the normal distribution involves taking the range encompassed by the observations within ranks $(n/2) \pm n^{1/2}$. This is a result of the standard error of Y being $\sqrt{n(0.5)(1 - 0.5)}$, and using mean plus/minus 2 standard errors for approximate 95% confidence.

Example 4.5: Movie Average Shot Lengths

Barry Sands has compiled a population of 11001 films and their average shot length (ASL, in seconds). The distribution of ASL is highly skewed to the right, with a population median of 6.4 (the mean is 7.74). A histogram of the ASL's is given in Figure 4.5, it has been truncated at 100 (due to distortion if the full distribution is given), with 8 cases falling between 100 and 1000. The thick vertical line is the population median.

Consider samples of $n = 20$. For the $Bin(20, 0.5)$ distribution, the following cumulative probabilities are obtained.

$$P(Y \leq 4) = .0059 \qquad P(Y \leq 5) = .0207 \qquad P(Y \leq 6) = .0577$$

$$\Rightarrow \qquad B_{\alpha/2,n} = 5 \quad L_{\alpha/2} = 5 + 1 = 6 \quad U_{\alpha/2} = 20 - 5 = 15$$

Thus, once we order the the 20 sampled films, we would take the range encompassed by the 6^{th} through the 15^{th} films.

The following random sample (ordered) was obtained in R.

```
> (ASL.sample.order <- sort(ASL.sample)) ## Sample values sorted
 [1]  3.80  3.80  4.42  4.42  4.67  5.13  5.56  5.80  5.81  6.36  6.56  6.80
[13]  7.00  7.10  7.80  9.47  9.50  9.60  9.80 13.17
> cbind(ASL.sample.order[6],ASL.sample.order[15]) ## 6th and 15th selected
     [,1] [,2]
[1,] 5.13  7.8
```

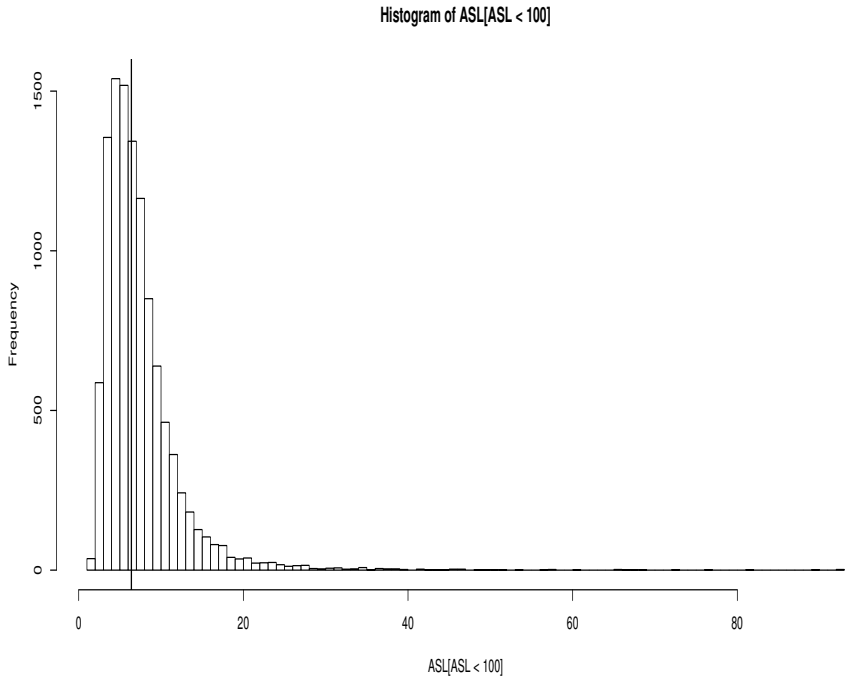

Figure 4.5: Average Shot Length (ASL) for a population of 11001 films.

For this sample, we obtain the 95% Confidence Interval: (5.13, 7.80), which does contain the population median (6.4). We now obtain 10000 random samples of size $n = 20$, and count the number that contain 6.4. Note that due to the "discreteness" of the distribution, $\alpha = 2(.0207) = .0414$, so we expect slightly more than 95% of the intervals to contain 6.4. Based on the 10000 random samples, 9629 (96.29%) contain the population mean.

Had we used the large-sample approximation here, which is questionable, with $n = 20$, we would have $n/2 = 10$, and $\sqrt{n} = 4.47$, and $L_{.025} \approx 10 - 4.47 = 5.53 = 5$ and $U_{.025} \approx 10 + 4.47 = 15$. We would still be selecting the 6^{th} and 15^{th} ordered values.

R Output

```
## Output

> pbinom(0:20,20,0.5)
 [1] 9.536743e-07 2.002716e-05 2.012253e-04 1.288414e-03 5.908966e-03
 [6] 2.069473e-02 5.765915e-02 1.315880e-01 2.517223e-01 4.119015e-01
[11] 5.880985e-01 7.482777e-01 8.684120e-01 9.423409e-01 9.793053e-01
[16] 9.940910e-01 9.987116e-01 9.997988e-01 9.999800e-01 9.999990e-01
[21] 1.000000e+00
> (ASL.sample.order <- sort(ASL.sample)) ## Sample values sorted
 [1]  3.80  3.80  4.42  4.42  4.67  5.13  5.56  5.80  5.81  6.36  6.56  6.80
[13]  7.00  7.10  7.80  9.47  9.50  9.60  9.80 13.17
> cbind(ASL.sample.order[6],ASL.sample.order[15]) ## 6th and 15th selected
     [,1] [,2]
[1,] 5.13  7.8
```

```
> sum(med.ci[,1] <= med.pop & med.ci[,2] >= med.pop) / num.sim
[1] 0.9629
```

∇

For a hypothesis test of whether the population median is some particular value (as with the mean, this is rare except in paired data experiments), we can use the **sign test**. The test makes use of the count of the number of observations exceeding the null value of the median being a binomial random variable with n trials, and probability of success $\pi = 0.5$ under the null hypothesis $H_0 : M = M_0$. There can be 2-tailed or Upper/Lower tailed alternatives. In each case, let B_{obs} be the count of the number of observations above M_0.

2-tailed tests: $H_0 : M = M_0 \quad H_A : M \neq M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \leq B_{\alpha/2,n}$ or $B_{obs} \geq n - B_{\alpha/2,n}$

Upper tailed tests: $H_0 : M \leq M_0 \quad H_A : M > M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \geq n - B_{\alpha,n}$

Lower tailed tests: $H_0 : M \geq M_0 \quad H_A : M < M_0 \quad T.S. : B_{obs} \quad R.R. : B_{obs} \leq B_{\alpha,n}$

For large-samples, the approximate normality of the Binomial can be used, and under the null hypothesis, the number of observations exceeding M_0 is approximately normal with mean $n/2$ and standard deviation $\sqrt{n(0.5)(1-0.5)} = 0.5\sqrt{n}$. Then we can obtain a z-statistic for the tests.

$$T.S. : z_{obs} = \frac{B_{obs} - (n/2)}{0.5\sqrt{n}} \quad R.R.(2) : |z_{obs}| \geq z_{\alpha/2} \quad R.R.(U) : z_{obs} \geq z_{\alpha} \quad R.R.(L) : z_{obs} \leq z_{1-\alpha} = -z_{\alpha}$$

Example 4.6: Movie Average Shot Lengths

Suppose we wanted to test whether the population median average shot length (ASL),M, differs from $M_0 = 5$ seconds (for some reason). Based on the sample of $n = 20$ films obtained previously, we have the following ASL values.

```
> (ASL.sample.order <- sort(ASL.sample))
 [1]  3.80  3.80  4.42  4.42  4.67  5.13  5.56  5.80  5.81  6.36  6.56  6.80
[13]  7.00  7.10  7.80  9.47  9.50  9.60  9.80 13.17
```

The test statistic is $B_{obs} = 15$. Depending on whether the goal is a 2-tailed or 1-tailed test we have that $P(Y \leq 5) = .0207$ and $P(Y \leq 6) = .0577$ for $Y \sim Bin(n = 20, \pi = .5)$. Thus, for a 2-tailed test, reject the null $H_0 : M = M_0$ for a 2-tailed test (with $\alpha = 0.05$), if $B_{obs} \leq 5$ or if $B_{obs} \geq 20 - 5 = 15$. Because the

probability that $Y \leq 6$ exceeds $\alpha = 0.05$, the Upper tail rejection region would be $R.R. : B_{obs} \geq 15$ and the Lower tail rejection region would be $R.R. : B_{obs} \leq 5$. Note that the 2-sided P-value is $P = 2P(Y \geq 15|Y \sim Bin(20, 0.5)) = 2P(Y \leq 5) = 2(.0207) = .0414$.

The large-sample z-statistic would be computed as follows.

$$z_{obs} = \frac{15 - (20/2)}{0.5\sqrt{20}} = \frac{5}{2.236} = 2.236 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 2.236) = 2(.0127) = .0254$$

The reason for the discrepancy between the P-values is the discreteness of the binomial and the continuity of the normal approximation. Some authors suggest the following continuity correction. The subtracting of the 0.5 is to get all the area over 15 for binomial, since 15 is above its expected value. This results in virtually the exact same P-value. As n gets large, the correction makes little difference.

$$z_{obs} = \frac{15 - (20/2) - 0.5}{0.5\sqrt{20}} = \frac{4.5}{2.236} = 2.013 \qquad \text{2-tailed } P\text{-value: } P = 2P(Z \geq 2.013) = 2(.0221) = .0442$$

4.4 The Bootstrap

In many applications, individual measurements are not normally distributed and the sample size is not large enough to justify the use of the Central Limit Theorem. Further, in many practical settings, the sampling distribution of an estimator is unknown (such as the coefficient of variation). The bootstrap makes use of the sample that is obtained (and is assumed to be representative of the population of measurements) to approximate the sampling distribution of the estimator of interest. The classic reference is Efron and Tibshirani (1993) [20], and for an introduction to Mathematical Statistics based on resampling methods, see Chihara and Hesterberg (2011) [14].

The process involves resampling from the sample data, with replacement, many times and computing the estimate for each resample, and saving the values. The samples are each of size n. Note that when estimating the sampling distribution of the sample mean, the mean of the resampled means will be very close to the sample mean of the original sample. That implies that the bootstrap will not directly estimate the mean of the sampling distribution (which is the population mean). The spread, bias, and skewness of the bootstrap distribution do reflect those of the target sampling distribution, where bias refers to the difference between the mean of the bootstrap distribution and the population mean.

4.4.1 Bootstrap Inferences Concerning the Population Mean

When trying to estimate a population mean (particularly with nonnormal data with a small sample size), a bootstrap prediction interval for the population mean μ can be obtained from the central $(1-\alpha)100\%$ values of the bootstrap sample estimates. This is a very simple approach, as all that is needed to be computed and saved are the sample means from each of the resamples (see e.g. Chihara and Hesterberg (2011), [14] Section 5.3). Once the means are obtained, the $\alpha/2$ and $1-\alpha/2$ quantiles are identified. Note that this interval will not typically be symmetric around the sample mean, unless the sample data are highly symmetric.

Example 4.7: Movie Average Shot Lengths

Suppose we wish to estimate the population mean of movie average shot lengths (ASL). The distribution is highly skewed, refer back to Figure 4.5. We first take a sample of $n = 25$ films, then draw $B = 10000$ random resamples with replacement from the 25 sampled films, and compute the sample mean for each resample, labeled $\overline{y}_i^*$ for the i^{th} resample. Finally we obtain the 2.5%-ile and 97.5%-ile from the resample means, for an interval that we can be approximately 95% confident will contain μ.

R Output

```
## Output

> ASL.sample1
 [1]  4.70 31.91  4.68  4.10 13.00  3.29  3.30  4.58 19.70  5.77 14.50  5.97
[13] 14.00  7.30  5.90  3.67  5.50 13.20  4.08  5.70  4.04  4.46  5.00  4.33
[25]  2.30

> round(boot1.out, 4)
         N     mu   sigma  n   ybar      s
[1,] 11001 7.7394 12.7654 25 7.7992 6.6888

> round(boot1a.out, 4)
     Samples   Mean 2.5%   97.5%     SD t-Lower t-Upper P(<t-L) P(>t-U)
2.5%   10000 7.7998 5.58 10.5992 1.2919  5.1334 10.4662  0.0049  0.0308
```

The sample mean for the original sample is $\overline{y} = 7.7992$ which exceeds the population mean $\mu = 7.7394$, although different samples could be below, close to, or above μ due to sampling error. The mean of the $B = 10000$ resample means is $\overline{\overline{y}^*} = 7.7998$, which is very close to $\overline{y} = 7.7992$, but not as close to μ, as would be expected due to the sampling process of the bootstrap. The approximate 95% prediction interval for μ is (5.5800, 10.5992) which does include $\mu = 7.7394$. The standard deviation of the resample means (1.2919) is referred to as the bootstrap standard error. Note that the prediction interval is not of the form $\overline{\overline{y}^*} \pm t_{.025,25-1} s_{\overline{y}*}$, which is of the following form, where $t_{.025,24} = 2.064$.

$$7.7998 \pm 2.064(1.2919) \quad \equiv \quad 7.7998 \pm 2.6665 \quad \equiv \quad (5.1333, 10.4663)$$

Of the 10000 sample means, 0.49% of the sample means fall below the lower bound 5.1333, and 3.08% fall above the upper bound 10.4663. The "t-type" interval goes outside both of the lower and upper bounds of the bootstrap interval. The lower bound is 2.218 bootstrap standard errors below the mean of the resample means, and the upper bound is 1.910 standard errors above it. In some cases the asymmetry will be larger.

∇

This approach of obtaining an approximate Confidence Interval for a parameter works well for many types of estimators/parameters. It is particularly useful when the bootstrap sample estimators have an approximately continuous distribution. When the distribution of bootstrap sample estimators have a discrete sampling distribution, the method does not work well. Consider estimating the population median in the average shot length example. Once we have our sample of $n = 25$ films, the median is the "middle" ASL of the 25 (13th) ordered films. When we take bootstrap samples, the median will always be one of the 25 ASL's in the original sample. Thus, there are only 25 possible values the sample median that each resample can take on.

A second approach that is specific to estimating a population mean makes use of a t-type statistic computed for each resample. This is referred to as **Bootstrap t Confidence Intervals**, (see e.g. Chihara and Hesterberg (2011), [14] Section 7.5). In this method, once the original sample of size n is taken, obtain the sample mean $\overline{y}$ and standard deviation s. Then for each of B resamples, compute the mean $\overline{y}_i^*$ and standard deviation s_i^*, where i represents the i^{th} resample. Then compute a t-type statistic for each resample, making use of the original sample mean as follows.

$$t_i^* = \frac{\overline{y}_i^* - \overline{y}}{s_i^*/\sqrt{n}} = \sqrt{n}\left(\frac{\overline{y}_i^* - \overline{y}}{s_i^*}\right) \quad i = 1, \ldots, B$$

Once the B values of t_i^* are computed, obtain the $\alpha/2$ quantile and the $(1-\alpha/2)$ quantiles, say (Q_L^*, Q_U^*). Note that Q_L^* will be negative and Q_U^* will be positive, and not necessarily of the same magnitude. The $(1-\alpha)100\%$ Confidence Interval for μ will be of the following form.

$$\text{Lower Bound: } \overline{y} - Q_U^* \frac{s}{\sqrt{n}} \qquad \text{Upper Bound: } \overline{y} - Q_L^* \frac{s}{\sqrt{n}}$$

Example 4.8: Movie Average Shot Lengths

We apply this method to the same sample and resamples used previously.

R Output

```
## Output

      mu   sigma   ybar      s    Q_L*   Q_U*   mu_L    mu_U
  7.7394 12.7654 7.7992 6.6888 -3.4285 1.5824 5.6824 12.3857
```

A histogram of the t^* values and the t_{24} density are given in Figure 4.6. The distribution of the t^* values is skewed left, with Q_L^* and Q_U^* being -3.4285 and 1.5824, respectively for $\alpha = 0.05$. The original sample mean and standard deviation are 7.7992 and 6.6888 respectively leading to the following 95% Confidence Interval for μ.

$$\left(7.7992 - 1.5824\frac{6.6888}{\sqrt{25}}, 7.7992 - (-3.4285)\frac{6.6888}{\sqrt{25}}\right) \equiv (7.7992 - 2.1169, 7.7992 + 4.5865) \equiv (5.6823, 12.3857)$$

Note that the interval is not symmetric about $\overline{y}$, it adds a larger term for the upper end than the term it subtracts for the lower end. This reflects the fact that the data are right-skewed. The bootstrap estimate of the **bias** is the difference from the average of the resample means and the overall sample mean: $\overline{\overline{y}^*} - \overline{y} = 0.0006$. This bias is very small relative to the standard error of the bootstrap estimator: .0006/1.2919=.00046. The 95% Confidence Interval using just the original sample mean, standard deviation, and the t-distribution is given for comparison.

$$7.7992 \pm 2.064\frac{6.6888}{\sqrt{25}} \quad \equiv \quad 7.7992 \pm 2.7611 \quad \equiv \quad (5.0381, 10.5603)$$

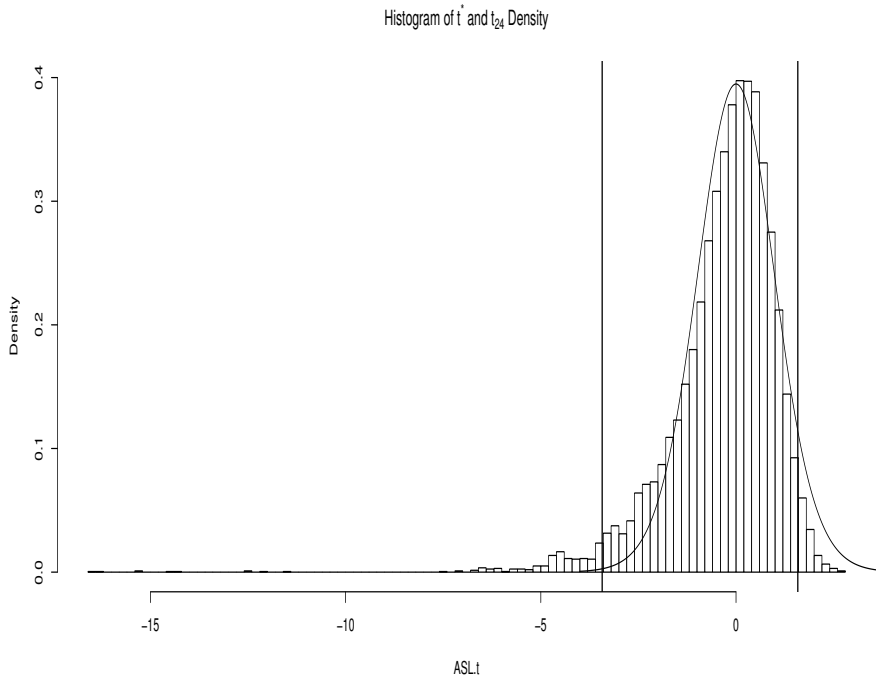

Figure 4.6: Histogram of t^* values and t_{24} Density - ASL Data

∇

Example 4.9: Average Shot Lengths - Comparing the Three Methods

Finally, 1000 random samples were taken from the ASL data. The two Bootstrap methods were performed on 1000 resamples from each (original) random sample and their 95% Confidence Intervals were obtained, as were the 1000 t-based intervals from the (original) samples. The first bootstrap method (middle 95% of the resample means) contained $\mu = 7.7394$ in 851 of the 1000 original samples (85.1% coverage). The second bootstrap method (based on constructed t-statistics around the sample mean) contained μ in 906 of the 1000 original samples (90.6% coverage). The normal based t-interval contained μ in 864 of the 1000 original samples (86.4%). All three performed below the nominal 95% level. This is due to the very large amount of skew in the data (largest ASL is 1000, while the population mean is less than 8), as well as the relatively small sample size ($n = 25$).

R Output

```
### Output

> round(boot3.out, 4)
     Boot Method 1 Boot Method2 Normal t
[1,]         0.851        0.906    0.864
```

∇

4.5 R Code for Chapter 4

```
### Chapter 4

## Figure 4.1 - Plot Rejection Regions for t-test

t <- seq(-4,4,.01)
ft <- dt(t,15)
LB2 <- qt(.025,15)
UB2 <- qt(.975,15)
UB1 <- qt(.95,15)
LB1 <- qt(.05,15)

par(mfrow=c(1,3))

plot(t,ft,type="l",main="2-Tailed Rejection Region")
abline(v=c(LB2,UB2))
polygon(c(t[t <= LB2],LB2),c(ft[t <= LB2],ft[t == -4]),col="black")
polygon(c(t[t >= UB2],UB2),c(ft[t >= UB2],ft[t == 4]),col="black")

plot(t,ft,type="l",main="Upper Tailed Rejection Region")
abline(v=UB1)
polygon(c(t[t >= UB1],UB1),c(ft[t >= UB1],ft[t == 4]),col="black")

plot(t,ft,type="l",main="Lower Tailed Rejection Region")
abline(v=LB1)
polygon(c(t[t <= LB1],LB1),c(ft[t <= LB1],ft[t == -4]),col="black")

rm(list=ls(all=TRUE))



### Example 4.1

### Read data and set up data frame
nhl <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv")
attach(nhl); names(nhl)

### Compute BMI
N <- NROW(nhl)
bmi.nhl <- 703 * Weight / (Height^2)
set.seed(98765)
num.sim <- 10000
n.sample <- 12
samp.mean <- rep(0, num.sim)
samp.sd <- rep(0, num.sim)
mu.bmi <- mean(bmi.nhl)
sd.bmi <- sd(bmi.nhl) * sqrt((N-1)/N)
std.err.bmi <- sd.bmi/sqrt(n.sample)
for (i in 1:num.sim) {
sample <- sample(1:N, n.sample, replace=FALSE)
samp.mean[i] <- mean(bmi.nhl[sample])
samp.sd[i] <- sd(bmi.nhl[sample])
}
cbind(samp.mean[1], samp.sd[1])
```

```
z_050 <- qnorm(.95); z_025 <- qnorm(.975); z_005 <- qnorm(.995)
cover10a <- sum(samp.mean >= mu.bmi - z_050*std.err.bmi &
  samp.mean <= mu.bmi + z_050*std.err.bmi) / num.sim
cover05a <- sum(samp.mean >= mu.bmi - z_025*std.err.bmi &
  samp.mean <= mu.bmi + z_025*std.err.bmi) / num.sim
cover01a <- sum(samp.mean >= mu.bmi - z_005*std.err.bmi &
  samp.mean <= mu.bmi + z_005*std.err.bmi) / num.sim
samp.se <- samp.sd / sqrt(n.sample)
cover10b <- sum(samp.mean >= mu.bmi - z_050*samp.se &
  samp.mean <= mu.bmi + z_050*samp.se) / num.sim
cover05b <- sum(samp.mean >= mu.bmi - z_025*samp.se &
  samp.mean <= mu.bmi + z_025*samp.se) / num.sim
cover01b <- sum(samp.mean >= mu.bmi - z_005*samp.se &
  samp.mean <= mu.bmi + z_005*samp.se) / num.sim

t_050 <- qt(.95,11); t_025 <- qt(.975,11); t_005 <- qt(.995,11)
cover10c <- sum(samp.mean >= mu.bmi - t_050*samp.se &
  samp.mean <= mu.bmi + t_050*samp.se) / num.sim
cover05c <- sum(samp.mean >= mu.bmi - t_025*samp.se &
  samp.mean <= mu.bmi + t_025*samp.se) / num.sim
cover01c <- sum(samp.mean >= mu.bmi - t_005*samp.se &
  samp.mean <= mu.bmi + t_005*samp.se) / num.sim

cover.out <- rbind(cbind(cover10a, cover05a, cover01a),
                   cbind(cover10b, cover05b, cover01b),
                   cbind(cover10c, cover05c, cover01c))
rownames(cover.out) <- c("Z - True SE",
               "Z - Estimated SE", "t - Estimated SE")
colnames(cover.out) <- c("90% Confidence", "95% Confidence", "99% Confidence")
round(cover.out,4)

rm(list=ls(all=TRUE))

### Example 4.2

E <- 0.20
sigma <- 1.058
alpha <- 0.05
n <- 1
E.t <- E+1
# Keep increasing $n$ until E.t < E
while (E.t >= E) {
n <- n+1
E.t <- qt(1-alpha/2,n-1)*sigma/sqrt(n)
}
cbind(n, E.t)

rm(list=ls(all=TRUE))

### Example 4.3

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)
male.mph <- mph[Gender == "M"]
N <- length(male.mph)

## Figure 4.2
mu0 <- mean(male.mph)
sigma <- sd(male.mph)
set.seed(13579)
num.sim <- 10000
num.samp <- 40
```

```
cv.lo <- qt(.025,num.samp-1)
cv.hi <- qt(.975,num.samp-1)
t.stat0 <- rep(0, num.sim)
p.value0 <- rep(0, num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.samp, replace=FALSE)
ybar <- mean(male.mph[sample])
s <- sd(male.mph[sample])
t.stat0[i] <- (ybar - mu0) / (s / sqrt(num.samp))
p.value0[i] <- 2*(1-pt(abs(t.stat0[i]),num.samp-1))
}
rejrate0.emp <- sum(p.value0 <= 0.05) / num.sim

par(mfrow=c(1,2))

hist(t.stat0, breaks=50, freq=FALSE, main="t-stat Under H0")
xt <- seq(-4,4,.01)
lines(xt, dt(xt,num.samp-1))
abline(v=cv.lo,lwd=2)
abline(v=cv.hi,lwd=2)

hist(p.value0, breaks=50, freq=FALSE, main="P-value Under H0")
xp <- seq(0,1,0.01)
lines(xp,dbeta(xp,1,1))

## End of Figure 4.2

mu01 <- 6.0
Delta1 <- (mu0 - mu01) / (sigma/sqrt(num.samp)) ## Old mu_0 is new mu_A
power1 <- pt(cv.lo, num.samp-1, Delta1) + (1-pt(cv.hi, num.samp-1, Delta1))
mu02 <- 6.5
Delta2 <- (mu0 - mu02) / (sigma/sqrt(num.samp)) ## Old mu_0 is new mu_A
power2 <- pt(cv.lo, num.samp-1, Delta2) + (1-pt(cv.hi, num.samp-1, Delta2))
set.seed(1234)
t.stat1 <- rep(0, num.sim)
p.value1 <- rep(0, num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.samp, replace=FALSE)
ybar <- mean(male.mph[sample])
s <- sd(male.mph[sample])
t.stat1[i] <- (ybar - mu01) / (s / sqrt(num.samp))
p.value1[i] <- 2*(1-pt(abs(t.stat1[i]),num.samp-1))
}
rejrate1.emp <- sum(t.stat1 <= cv.lo | t.stat1 >= cv.hi) / num.sim

## Figure 4.3
par(mfrow=c(1,1))
hist(t.stat1, breaks=50, freq=FALSE, main="t-stat Under H0:mu=6")
xt <- seq(-4,6,.01)
lines(xt, dt(xt,num.samp-1,Delta1))
abline(v=cv.lo,lwd=2)
abline(v=cv.hi,lwd=2)

## End of Figure 4.3

set.seed(5678)
t.stat2 <- rep(0, num.sim)
p.value2 <- rep(0, num.sim)
for (i in 1:num.sim) {
sample <- sample(1:N, num.samp, replace=FALSE)
ybar <- mean(male.mph[sample])
s <- sd(male.mph[sample])
t.stat2[i] <- (ybar - mu02) / (s / sqrt(num.samp))
p.value2[i] <- 2*(1-pt(abs(t.stat2[i]),num.samp-1))
```

```
}
rejrate2.emp <- sum(t.stat2 <= cv.lo | t.stat2 >= cv.hi) / num.sim

power.out <- rbind(cbind(0, 0.05, rejrate0.emp),
                   cbind(Delta1, power1, rejrate1.emp),
                   cbind(Delta2, power2, rejrate2.emp))
rownames(power.out) <- c("mu0=6.33", "mu0=6.00", "mu0=6.50")
colnames(power.out) <- c("Delta", "Theoretical Power", "Empirical Power")
round(power.out, 4)


rm(list=ls(all=TRUE))


### Example 4.4

n <- 3
alpha <- 0.05
mu_diff <- 0.25
sigma <- 1.058
Delta <- mu_diff/(sigma/sqrt(n))
CV_LO <- qt(alpha/2,n-1)
CV_HI <- qt(1-alpha/2,n-1)
(power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta)))
while (power <= 0.80) {
n <- n+1
Delta <- mu_diff/(sigma/sqrt(n))
CV_LO <- qt(alpha/2,n-1)
CV_HI <- qt(1-alpha/2,n-1)
power <- pt(CV_LO,n-1,Delta) + (1-pt(CV_HI,n-1,Delta))
}
cbind(n, power)

## Figure 4.4
par(mfrow=c(2,2))

t <- seq(-6,6,0.01)
mu_diff <- 0.25
sigma <- 1.058
for (n in c(10, 30, 70, 150)) {
Delta <- mu_diff/(sigma/sqrt(n))
plot(t,dt(t,n-1),type="l",main=paste("Central and Non-Central t, n=",n))
lines(t,dt(t,n-1,Delta),lty=2)
abline(v=qt(.025,n-1))
abline(v=qt(.975,n-1))
}

rm(list=ls(all=TRUE))

### Example 4.5

avshotlen <- read.csv(
"http://www.stat.ufl.edu/~winner/data/movie_avshotlength.csv")
attach(avshotlen); names(avshotlen)
mean(ASL)
median(ASL)
sum(ASL > 100)

## Figure 4.5

par(mfrow=c(1,1))
hist(ASL[ASL < 100], breaks=100)
abline(v=median(ASL),lwd=2)
```

```
pbinom(0:20,20,0.5)
set.seed(4321)
N <- length(ASL)
sample1 <- sample(1:N, 20, replace=FALSE)
ASL.sample <- ASL[sample1]
(ASL.sample.order <- sort(ASL.sample)) ## Sample values sorted
cbind(ASL.sample.order[6],ASL.sample.order[15]) ## 6th and 15th selected
set.seed(7654)
num.sim <- 10000
num.samp <- 100
med.ci <- matrix(rep(0, 2*num.sim),ncol=2)
for (i in 1:num.sim) {
sample <- sample(1:N, 20, replace=FALSE)
med.ci[i,1] <- sort(ASL[sample])[6]
med.ci[i,2] <- sort(ASL[sample])[15]
}
med.pop <- median(ASL)
sum(med.ci[,1] <= med.pop & med.ci[,2] >= med.pop) / num.sim

rm(list=ls(all=TRUE))

### Example 4.7

avshotlen <- read.csv(
"http://www.stat.ufl.edu/~winner/data/movie_avshotlength.csv")
attach(avshotlen); names(avshotlen)
N <- length(ASL)
mu <- mean(ASL)
sigma <- sd(ASL)
## Obtain the original random sample of n=25
set.seed(34567)
samp.size <- 25
sample1 <- sample(1:N,samp.size,replace=F)
ASL.sample1 <- ASL[sample1]
ASL.sample1

boot1.out <- cbind(N, mu, sigma, samp.size,
             mean(ASL.sample1), sd(ASL.sample1))
colnames(boot1.out) <- c("N", "mu", "sigma", "n", "ybar", "s")
round(boot1.out, 4)


## Figure 4.6
par(mfrow=c(1,1))
qqnorm(ASL.sample1); qqline(ASL.sample1)
shapiro.test(ASL.sample1)

### Method 1 - Chihara/Hesterberg Section 5.3, pp. 113-114
set.seed(24680)
num.boot.inner <- 10000
ASL.mean <- rep(0,num.boot.inner)
for (i2 in 1:num.boot.inner) {
x <- sample(ASL.sample1, samp.size, replace=T)
ASL.mean[i2] <- mean(x)
}

q.mean.025 <- quantile(ASL.mean,.025)
q.mean.975 <- quantile(ASL.mean,.975)
mean.mean <- mean(ASL.mean)
sd.mean <- sd(ASL.mean)
LBt.mean <- mean(ASL.mean) - qt(.975,samp.size-1)*sd(ASL.mean)
UBt.mean <- mean(ASL.mean) + qt(.975,samp.size-1)*sd(ASL.mean)
range.lo.mean <-sum(ASL.mean < mean(ASL.mean) -
```

```
    qt(.975,samp.size-1)*sd(ASL.mean)) / num.boot.inner
range.hi.mean <- sum(ASL.mean > mean(ASL.mean) +
    qt(.975,samp.size-1)*sd(ASL.mean)) / num.boot.inner

boot1a.out <- cbind(num.boot.inner, mean.mean, q.mean.025, q.mean.975,
      sd.mean, LBt.mean, UBt.mean, range.lo.mean, range.hi.mean)
colnames(boot1a.out) <- c("Samples", "Mean", "2.5%", "97.5%",
     "SD", "t-Lower", "t-Upper", "P(<t-L)", "P(>t-U)")
round(boot1a.out, 4)

rm(list=ls(all=TRUE))


### Example 4.8

### Part 1
avshotlen <- read.csv(
"http://www.stat.ufl.edu/~winner/data/movie_avshotlength.csv")
attach(avshotlen); names(avshotlen)
N <- length(ASL)
mu <- mean(ASL)
sigma <- sd(ASL)

## Obtain the original random sample of n=25
set.seed(34567)
samp.size <- 25
sample1 <- sample(1:N,samp.size,replace=F)
ASL.sample1 <- ASL[sample1]

### Method 2 - Chihara/Hesterberg Section 7.5, pp. 195-198
# ASL.t computes and saves t*
set.seed(24680)
ybar.sample1 <- mean(ASL.sample1)
s.sample1 <- sd(ASL.sample1)
num.boot.inner <- 10000
ASL.t <- rep(0,num.boot.inner)
ASL.mean <- rep(0,num.boot.inner)

for (i2 in 1:num.boot.inner) {
x <- sample(ASL.sample1, samp.size, replace=T)
ASL.t[i2] <- (mean(x) - ybar.sample1) / (sd(x) / sqrt(samp.size))
ASL.mean[i2] <- mean(x)
}

Q_L <- quantile(ASL.t,0.025)
Q_U <- quantile(ASL.t,0.975)
mu_L <- ybar.sample1 - Q_U*s.sample1/sqrt(samp.size)
mu_U <- ybar.sample1 - Q_L*s.sample1/sqrt(samp.size)

boot2.out <- cbind(mu, sigma, ybar.sample1, s.sample1, Q_L, Q_U, mu_L, mu_U)
colnames(boot2.out) <- c("mu", "sigma", "ybar", "s", "Q_L*", "Q_U*",
                "mu_L", "mu_U")
round(boot2.out, 4)


## Figure 4.7

par(mfrow=c(1,1))
hist(ASL.t,breaks=80,freq=F,
main=expression(paste("Histogram of ",t^"*"," and ",t[24]," Density")))
t.seq <- seq(-4,4,.01)
lines(t.seq,dt(t.seq,samp.size-1))
abline(v=c(Q_L,Q_U),lwd=2)
```

```
rm(list=ls(all=TRUE))

### Part 2
avshotlen <- read.csv(
"http://www.stat.ufl.edu/~winner/data/movie_avshotlength.csv")
attach(avshotlen); names(avshotlen)
N <- length(ASL)
mu <- mean(ASL)
sigma <- sd(ASL)
## Initialize mean/sd and CI holders - 1000 outer (original) samples
set.seed(13579)
num.boot.outer <- 1000
ASL.mean.sd <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.boot1 <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.boot2 <- matrix(rep(0,2*num.boot.outer),ncol=2)
ASL.tnorm <- matrix(rep(0,2*num.boot.outer),ncol=2)
samp.size <- 25
sqrt.n <- sqrt(samp.size)
t.24 <- qt(c(.025,.975),samp.size-1)
### Begin outer loop
for (i1 in 1:num.boot.outer) {
sample1 <- sample(1:N,samp.size,replace=F)
ASL.sample1 <- ASL[sample1] ### Original Samples
ASL.mean.sd[i1,1] <- mean(ASL.sample1) ### Save mean in column 1
ASL.mean.sd[i1,2] <- sd(ASL.sample1) ### Save sd in column 2
### Begin inner (bootstrap) loop
num.boot.inner <- 1000
ASL.mean <- rep(0,num.boot.inner)
ASL.t <- rep(0,num.boot.inner)
for (i2 in 1:num.boot.inner) {
x <- sample(ASL.sample1, samp.size, replace=T)
ASL.mean[i2] <- mean(x)
ASL.t[i2] <- (mean(x) - ASL.mean.sd[i1,1]) /
(sd(x) /sqrt.n)
}    ### Close inner loop
ASL.boot1[i1,] <- quantile(ASL.mean,c(.025,.975))
ASL.boot2[i1,] <- ASL.mean.sd[i1,1] -
quantile(ASL.t,c(.975,.025)) * ASL.mean.sd[i1,2]/sqrt.n
ASL.tnorm[i1,] <- ASL.mean.sd[i1,1] + t.24 *
ASL.mean.sd[i1,2]/sqrt.n
}   ### Close outer loop

## Obtain coverage probabilities
cov.bm1 <- sum(ASL.boot1[,1] <= mu & ASL.boot1[,2] >= mu)/num.boot.outer
cov.bm2 <- sum(ASL.boot2[,1] <= mu & ASL.boot2[,2] >= mu)/num.boot.outer
cov.tnorm <- sum(ASL.tnorm[,1] <= mu & ASL.tnorm[,2] >= mu)/num.boot.outer

boot3.out <- cbind(cov.bm1, cov.bm2, cov.tnorm)
colnames(boot3.out) <- c("Boot Method 1", "Boot Method2", "Normal t")
round(boot3.out, 4)

rm(list=ls(all=TRUE))
```

Chapter 5

Comparing Two Populations' Means and Medians

While estimating the mean or median of a population is important, many more applications involve comparing two or more treatments or populations. There are two commonly used designs: **independent samples** and **paired samples**. Independent samples are used in controlled experiments when a sample of experimental units is obtained, and randomly assigned to one of two treatments or conditions. That is, each unit receives only one of the two treatments. These are often referred to as **Completely Randomized** or **Parallel Groups** or **Between Subjects** designs in various fields of study. Paired samples can involve the same experimental unit receiving each treatment, or units being matched based on external criteria, then being randomly assigned to the two treatments within pairs. These are often referred to as **Randomized Block** or **Crossover** or **Within Subjects** designs.

In observational studies, independent samples can be taken from two existing populations, or elements within two populations can be matched based on external criteria and observed. In each case, the goal is to make inferences concerning the difference between the two means or medians based on sample data.

5.1 Independent Samples

In the case of independent samples, assume we sample n_1 units or subjects in treatment 1 which has a population mean response μ_1 and population standard deviation σ_1. Further, a sample of n_2 elements from treatment 2 is obtained where the population mean is μ_2 and standard deviation is σ_2. Measurements within and between samples are independent. Regardless of the distributions of the individual measurements, we have the following results based on linear functions of random variables, in terms of the means of the two random samples. The notation used is Y_{1j} is the j^{th} unit (replicate) from sample 1, and Y_{2j} is the j^{th} unit (replicate) from sample 2. In the case of independent samples, these two random variables are independent.

$$\overline{Y}_1 = \frac{\sum_{j=1}^{n_1} Y_{1j}}{n_1} = \sum_{j=1}^{n_1} \left(\frac{1}{n_1}\right) Y_{1j} \quad \Rightarrow \quad E\{\overline{Y}_1\} = \mu_1 \quad V\{\overline{Y}_1\} = \frac{\sigma_1^2}{n_1} \quad E\{\overline{Y}_2\} = \mu_2 \quad V\{\overline{Y}_2\} = \frac{\sigma_2^2}{n_2}$$

$$E\{\overline{Y}_1 - \overline{Y}_2\} = E\{\overline{Y}_1\} - E\{\overline{Y}_2\} = \mu_1 - \mu_2$$

$$V\{\overline{Y}_1 - \overline{Y}_2\} = \sigma^2_{\overline{Y}_1 - \overline{Y}_2} = V\{\overline{Y}_1\} + V\{\overline{Y}_2\} - 2\text{COV}\{\overline{Y}_1, \overline{Y}_2\} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} + 0 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$SE\{\overline{Y}_1 - \overline{Y}_2\} = \sigma_{\overline{Y}_1 - \overline{Y}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

If the data are normally distributed, $\overline{Y}_1 - \overline{Y}_2$ is also normally distributed. If the data are not normally distributed, $\overline{Y}_1 - \overline{Y}_2$ will be approximately normally distributed in large samples. As in the case of a single mean, how large of samples are needed depends on the shape of the underlying distributions.

The problem arises again that the variances will be unknown and must be estimated. For large sample sizes n_1 and n_2, we have the following approximation for the sampling distribution of the following quantity, where the sample variances replace the true population variances.

$$\frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} N(0,1)$$

$$\Rightarrow \quad P\left(\left(\overline{Y}_1 - \overline{Y}_2\right) + z_{1-\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq \left(\overline{Y}_1 - \overline{Y}_2\right) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right) \approx 1 - \alpha$$

Example 5.1: NHL and EPL Players' BMI

Body Mass Indices for all National Hockey League (NHL) and English Premier League (EPL) football players for the 2013/4 season were obtained. Identifying the NHL as league 1 and EPL as league 2 we have the following population parameters.

$$N_1 = 717 \quad \mu_1 = 26.500 \quad \sigma_1 = 1.454 \qquad N_2 = 526 \quad \mu_2 = 23.019 \quad \sigma_2 = 1.711$$

A plot of the two population histograms, along with normal densities is given in Figure 5.1. Both distributions are well approximated by the normal distribution, with the NHL having a substantially higher mean and EPL having a slightly higher standard deviation.

We take 100000 independent random samples of sizes $n_1 = n_2 = 20$ from the two populations, each time computing and saving $\overline{y}_1, s_1, \overline{y}_2, s_2$. A histogram of the 100000 sample mean differences and the superimposed Normal density with mean $\mu_1 - \mu_2 = 3.481$ and standard error 0.502 (calculation given below) is shown in Figure 5.2. The mean of the 100000 mean differences $\overline{y}_1 - \overline{y}_2$ is 3.482 with standard deviation (standard error) 0.493. Both are very close to their theoretical values (as they should be). Then we compute the following quantity (and interval), counting the number of samples for which it contains $\mu_1 - \mu_2$, and its average estimated variance (squared standard error).

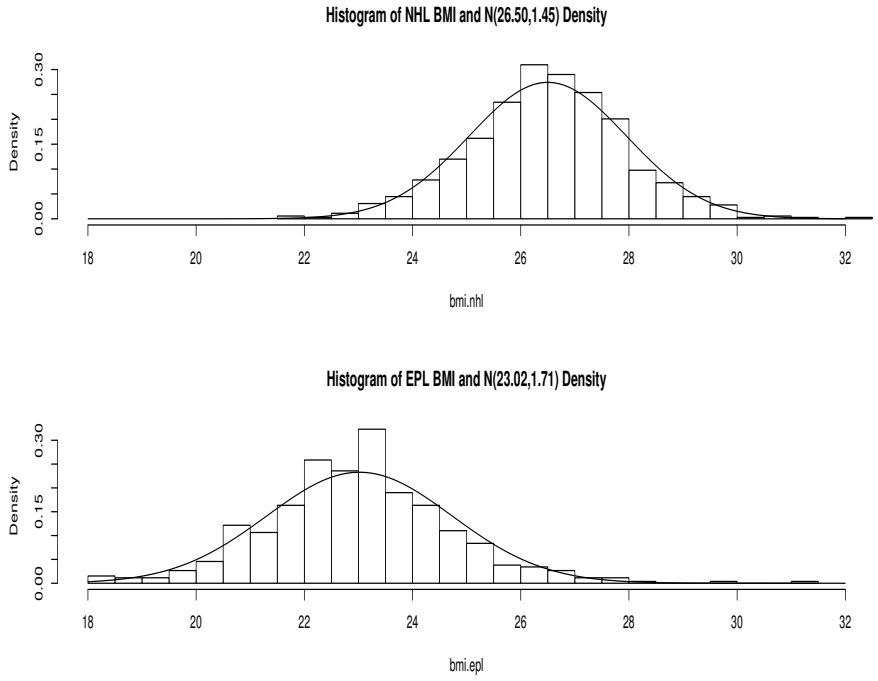

Figure 5.1: Distributions of NHL and EPL players Body Mass Index

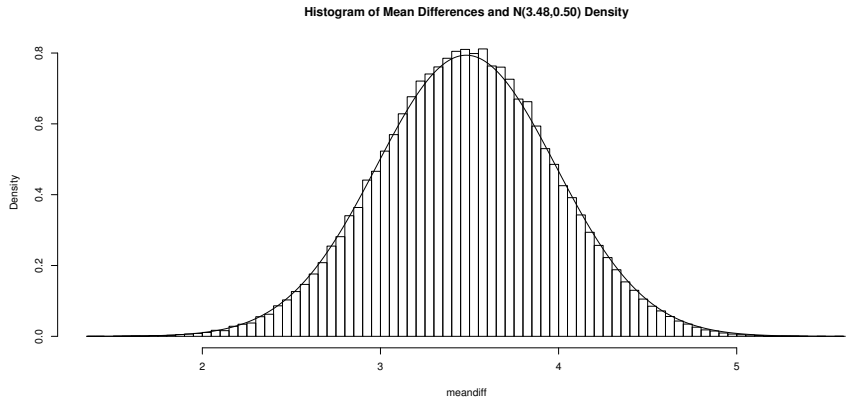

Figure 5.2: 100000 sample mean differences ($n_1 = n_2 = 20$) for NHL and EPL BMI values and Normal Density

$$(\overline{y}_1 - \overline{y}_2) \pm 1.96\sqrt{\frac{s_1^2}{20} + \frac{s_2^2}{20}} \qquad \mu_1 - \mu_2 = 26.500 - 23.019 = 3.481$$

$$SE\left\{\overline{Y}_1 - \overline{Y}_2\right\} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1.454^2}{20} + \frac{1.711^2}{20}} = 0.502$$

The mean of the 10000 sample mean differences is 3.479 compared to the theoretical mean difference of 3.481. The standard deviation of the sample mean differences is 0.493, compared to the theoretical standard error of 0.502.

Of the intervals constructed from each sample mean difference and its estimated standard error (using s_1, s_2 in place of σ_1, σ_2), the interval contains the true mean difference (3.481) for 94.698% of the samples, very close to the nominal 95% coverage rate. If we replace $z_{.025} = 1.96$ with the more appropriate $t_{.025,n_1+n_2-2} = t_{.025,38} = 2.0244$, the coverage rate increases to 95.395%. Note that virtually all software packages will automatically use t in place of z, however, there are various statistical methods that always use the z case.

The average of the estimated variance of $\overline{y}_1 - \overline{y}_2$: $s_1^2/n_1 + s_2^2/n_2$ is 0.2527, while its theoretical value is $\sigma_1^2/n_1 + \sigma_2^2/n_2 = 0.2521$. Note that the variance of the estimated difference is unbiased, not so for the standard error.

R Output

```
### Output

> round(md.out, 3)
      mu1    mu2 sigma1 sigma2  n mu1-mu2 SE{Yb1-Yb2} Mean(yb1-yb2) SD(yb1-yb2) cover(z) cover(t)
[1,] 26.5 23.019  1.454  1.711 20   3.481       0.502         3.479       0.493    0.947    0.954
```

∇

This logic leads to a large-sample test and Confidence Interval regarding $\mu_1 - \mu_2$ once estimates $\overline{y}_1, s_1, \overline{y}_2, s_2$ have been observed in an experiment or observational study. The Confidence Interval and test are given below. Typically, $z_{\alpha/2}$ is replaced with $t_{\alpha/2,\nu}$, where ν is the degrees of freedom, which depends on assumptions involving the variances (see below).

$$\text{Large Sample } (1-\alpha)100\% \text{ CI for } \mu_1 - \mu_2\text{: } (\overline{y}_1 - \overline{y}_2) \pm z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{2-tail: } H_0 : \mu_1 - \mu_2 = \Delta_0 \quad H_A : \mu_1 - \mu_2 \neq \Delta_0 \quad TS : z_{obs} = \frac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : |z_{obs}| \geq z_{\alpha/2} \quad P = 2P(Z \geq |z_{obs}|)$$

Upper tail: $H_0 : \mu_1 - \mu_2 \leq \Delta_0 \quad H_A : \mu_1 - \mu_2 > \Delta_0 \quad TS : z_{obs} = \dfrac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : z_{obs} \geq z_\alpha \quad P = P(Z \geq z_{obs})$

Lower tail: $H_0 : \mu_1 - \mu_2 \geq \Delta_0 \quad H_A : \mu_1 - \mu_2 < \Delta_0 \quad TS : z_{obs} = \dfrac{(\overline{y}_1 - \overline{y}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad RR : z_{obs} \leq z_\alpha \quad P = P(Z \leq z_{obs})$

Example 5.2: Gender Classification from Physical Measurements

A study in forensics used measurements of the length and breadth of the scapula from samples of 95 male and 96 female Thai adults (Peckmann, Scott, Meek, Mahakkanukrauh (2017), [42]). The measurements were length and breadth of glenoid cavity (LGC and BGC, in mm), respectively. Summary data for the two samples for BGC are given below.

$$n_m = 95 \quad \overline{y}_m = 27.87 \quad s_m = 2.04 \qquad n_f = 96 \quad \overline{y}_f = 23.77 \quad s_f = 1.85$$

$$\overline{y}_m - \overline{y}_f = 27.87 - 23.77 = 4.10 \qquad \hat{SE}\{\overline{Y}_m - \overline{Y}_f\} = \sqrt{\frac{2.04^2}{95} + \frac{1.85^2}{96}} = 0.282$$

A 95% Confidence Interval for the population mean difference, $\mu_m - \mu_f$ is given below.

$$(\overline{y}_m - \overline{y}_f) \pm z_{.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad \equiv \quad 4.10 \pm 1.960(0.282) \quad \equiv \quad 4.10 \pm 0.553 \quad \equiv \quad (3.55, 4.65)$$

The interval is very far away from 0, making us very confident that the population mean is higher for males than females. To test whether the population means differ (which they clearly do from the Confidence Interval), we conduct the following 2-tailed test with $\alpha = 0.05$.

$$H_0 : \mu_m - \mu_f = 0 \quad H_A : \mu_m - \mu_f \neq 0 \quad T.S. : z_{obs} = \frac{4.10 - 0}{0.282} = 14.54 \quad R.R. : |z_{obs}| \geq 1.960 \quad P = 2P(Z \geq 14.54) \approx 0$$

∇

5.2 Small–Sample Tests

In this section we cover small–sample tests without going through the detail given for the large–sample tests. In each case, we will be testing whether or not the means (or medians) of two distributions are equal.

There are two considerations when choosing the appropriate test: (1) Are the population distributions of measurements approximately normal? and (2) Was the study conducted as an independent samples (parallel groups) or paired samples (crossover) design? The appropriate test for each situation is given in Table 5.1. We will describe each test with the general procedure and an example.

The two tests based on non–normal data are called **nonparametric tests** and are based on ranks, as opposed to the actual measurements. When distributions are skewed, samples can contain measurements that are extreme (usually large). These extreme measurements can cause problems for methods based on means and standard deviations, but will have less effect on procedures based on ranks.

	Design Type	
	Parallel Groups	Crossover
Normally Distributed Data	2–Sample t–test	Paired t–test
Non–Normally Distributed Data	Wilcoxon Rank Sum test (Mann–Whitney U–Test)	Wilcoxon Signed–Rank Test

Table 5.1: Statistical Tests for small–sample 2 group situations

5.2.1 Independent Samples (Completely Randomized Designs)

Completely Randomized Designs are designs where the samples from the two populations are independent. That is, subjects are either assigned at random to one of two treatment groups (possibly active drug or placebo), or possibly selected at random from one of two populations (as in Example 5.1, where we had NHL and EPL players and in Example 5.2 where they measured males and females). In the case where the two populations of measurements are normally distributed, the 2–sample t–test is used. Note that it also works well for reasonably large sample sizes when the measurements are not normally distributed. This procedure is very similar to the large–sample test from the previous section, where only the critical values for the rejection region changes. In the case where the populations of measurements are not approximately normal, the Wilcoxon Rank–Sum test (or, equivalently the Mann–Whitney U–test) is commonly used. These tests are based on comparing the average ranks across the two groups when the measurements are ranked from smallest to largest, across groups.

2–Sample Student's t–test for Normally Distributed Data

This procedure is similar to the large–sample test, except the critical values for the rejection regions and Confidence Intervals are based on the t–distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom and the variances are "pooled" (see below). We will assume the two population variances are equal in the 2–sample t–test. If they are not, simple adjustments can be made to obtain an appropriate test, which will be given below. We then 'pool' the 2 sample variances to get an estimate of the common variance $\sigma^2 = \sigma_1^2 = \sigma_2^2$. This estimate, that we will call s_p^2 is calculated as follows:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

The test of hypothesis concerning $\mu_1 - \mu_2$ is conducted as follows:

1. $H_0 : \mu_1 - \mu_2 = 0$

2. $H_A : \mu_1 - \mu_2 \neq 0$ or $H_A : \mu_1 - \mu_2 > 0$ or $H_A : \mu_1 - \mu_2 < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$
4. R.R.: $|t_{obs}| > t_{\alpha/2,n_1+n_2-2}$ or $t_{obs} > t_{\alpha,n_1+n_2-2}$ or $t_{obs} < -t_{\alpha,n_1+n_2-2}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P\left(t_{n_1+n_2-2} > |t_{obs}|\right)$ or $P\left(t_{n_1+n_2-2} > t_{obs}\right)$ or $P\left(t_{n_1+n_2-2} < t_{obs}\right)$ (again, depending on which alternative you are using).

Example 5.3: Comparison of Two Instructional Methods

A study was conducted (Rusanganwa (2013) [44]) to compare two instructional methods: multimedia (treatment 1) and traditional (treatment 2) for teaching physics to undergraduate students in Rwanda. Subjects were assigned at random to the two treatments. Each subject received only one of the two methods. The numbers of subjects who completed the courses and took two exams were $n_1 = 13$ for the multimedia course and $n_2 = 19$ for the traditional course. The primary response was the post-course score on an examination. We will conduct the test $H_0 : \mu_1 - \mu_2 = 0$ vs $H_A : \mu_1 - \mu_2 \neq 0$, where the null hypothesis is no difference in the effects of the two methods. The summary statistics are given below.

$$n_1 = 13 \quad \overline{y}_1 = 11.10 \quad s_1 = 3.47 \qquad n_2 = 19 \quad \overline{y}_2 = 8.35 \quad s_2 = 2.45$$

First, compute s_p^2, the pooled variance:

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = \frac{(13-1)(3.47)^2 + (19-1)(2.45)^2}{13+19-2} = \frac{252.54}{30} = 8.42 \quad (s_p = 2.90)$$

Now conduct the (2-sided) test as described above with $\alpha = 0.05$ significance level:

- $H_0 : \mu_1 - \mu_2 = 0$
- $H_A : \mu_1 - \mu_2 \neq 0$
- T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{(11.10-8.35)}{\sqrt{8.42\left(\frac{1}{13}+\frac{1}{19}\right)}} = \frac{2.75}{1.04} = 2.633$
- R.R.: $|t_{obs}| \geq t_{\alpha/2,n_1+n_2-2} = t_{.05/2,13+19-2} = t_{.025,30} = 2.042$
- P-value: $2P\left(t_{30} \geq |t_{obs}|\right) = 2P\left(t_{30} \geq 2.633\right) = 0.0132$

Based on this test, reject H_0 (for any $\alpha \geq .0132$), and conclude that the population mean post course scores differ under these two conditions. The 95% Confidence Interval for $\mu_1 - \mu_2$ is $2.75 \pm 2.042(1.04) \equiv (0.62, 4.88)$ which does not contain 0.

Below we use generated samples that have the same means and standard deviation and use **t.test** function in R to conduct the 2-sample t-test.

R Commands and Output

```
## Commands

rp <- read.csv("http://www.stat.ufl.edu/~winner/data/rwanda_physics.csv")
attach(rp); names(rp)
t.test(score ~ trt.y, var.equal=T) # t-test with single y-var and trt id

## Output

> t.test(score ~ trt.y, var.equal=T)

        Two Sample t-test

data:  score by trt.y
t = 2.6323, df = 30, p-value = 0.01327
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.6163295 4.8826179
sample estimates:
mean in group 1 mean in group 2
      11.100000        8.350526
```

∇

When the population variances are not equal, there is no justification for pooling the sample variances to better estimate the common variance σ^2. In this case the estimated standard error of $\overline{Y}_1 - \overline{Y}_2$ is $\sqrt{s_1^2/n_1 + s_2^2/n_2}$. An adjustment is made to the degrees of freedom for an approximation to a t-distribution of the t-statistic.

$$\frac{\left(\overline{Y}_1 - \overline{Y}_2\right) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \dot{\sim} t_\nu \qquad \nu = \frac{\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^2}{\left[\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}\right]}$$

The test is referred to as **Welch's Test**, and the degrees of freedom **Satterthwaite's Approximation**. Statistical software packages automatically compute the approximate degrees of freedom. The approximation extends to more complex models as well. Once the samples are obtained, and the sample means and standard deviations are computed, the $(1-\alpha)100\%$ Confidence Interval for $\mu_1 - \mu_2$ is computed as follows.

$$(\overline{y}_1 - \overline{y}_2) \pm t_{\alpha/2,\nu}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \nu = \frac{\left[\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right]^2}{\left[\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}\right]}$$

The test of hypothesis concerning $\mu_1 - \mu_2$ is conducted as follows:

1. $H_0 : \mu_1 - \mu_2 = 0$
2. $H_A : \mu_1 - \mu_2 \neq 0$ or $H_A : \mu_1 - \mu_2 > 0$ or $H_A : \mu_1 - \mu_2 < 0$ (which alternative is appropriate should be clear from the setting).

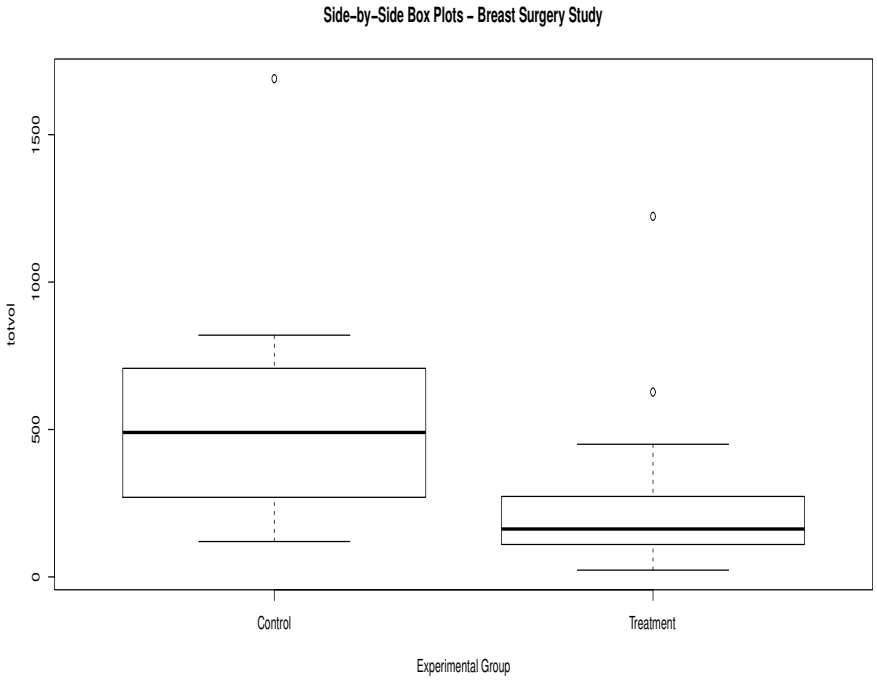

Figure 5.3: Abdominal drainage in breast reconstruction surgery, DIEP procedure with and without abdominal suture quilting.

3. T.S.: $t_{obs} = \frac{(\overline{y}_1 - \overline{y}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

4. R.R.: $|t_{obs}| \geq t_{\alpha/2,\nu}$ or $t_{obs} \geq t_{\alpha,\nu}$ or $t_{obs} \leq -t_{\alpha,\nu}$ (which R.R. depends on which alternative hypothesis you are using).

5. p-value: $2P\left(t_\nu \geq |t_{obs}|\right)$ or $P\left(t_\nu \geq t_{obs}\right)$ or $P\left(t_\nu \leq t_{obs}\right)$ (again, depending on which alternative you are using).

Example 5.4: Abdominal Quilting to Reduce Drainage in Breast Reconstruction Surgery

A study considered the effect of abdominal suture quilting on abdominal drainage during breast reconstruction surgery (Liang, et al, (2016), [34]). A group of $n_1 = 27$ subjects (controls) received the standard DIEP procedure, while a group of $n_2 = 26$ subjects (treatment) received the DIEP procedure along with the suture quilting. The response measured was the amount of abdominal drainage during the surgery (in ml). The summary data are given below, note that the sample standard deviations are substantially different, and these are relatively large sample sizes. Side-by-side box plots are given in Figure 5.3.

$$n_1 = 27 \quad \overline{y}_1 = 527.78 \quad s_1 = 322.07 \qquad n_2 = 26 \quad \overline{y}_2 = 238.31 \quad s_2 = 242.66$$

The estimated mean difference, standard error, and degrees of freedom are computed below.

$$\overline{y}_1 - \overline{y}_2 = 527.78 - 238.31 = 289.47 \qquad \hat{SE}\{\overline{Y}_1 - \overline{Y}_2\} = \sqrt{\frac{322.07^2}{27} + \frac{242.66^2}{26}} = 78.14$$

$$\nu = \frac{\left[\frac{322.07^2}{27} + \frac{242.66^2}{26}\right]^2}{\left[\frac{(322.07^2/27)^2}{27-1} + \frac{(242.66^2/26)^2}{26-1}\right]} = 48.25 \qquad t_{.025,48.25} = 2.010$$

The 95% Confidence Interval for $\mu_1 - \mu_2$ and test statistic and P-value for testing $H_0 : \mu_1 - \mu_2 = 0$ versus $H_A : \mu_1 - \mu_2 \neq 0$ are given below. There is strong evidence that the suture quilting reduces blood loss during surgery.

$$\text{95\% CI for } \mu_1 - \mu_2\text{: } 289.47 \pm 2.010(78.14) \quad \equiv \quad 289.47 \pm 157.06 \quad \equiv \quad (132.41, 446.53)$$

$$\text{T.S.: } t_{obs} = \frac{289.47}{78.14} = 3.705 \qquad P\left(t_{48.25} \geq 3.705\right) = .0005$$

R Commands and Output

```
## Commands

quilt <- read.csv("http://www.stat.ufl.edu/~winner/data/breast_diep.csv")
attach(quilt); names(quilt)

trt.f <- factor(trt)
levels(trt.f) <- c("Control", "Treatment")
t.test(totvol ~ trt.f, var.equal=F)

## Output

> t.test(totvol ~ trt, var.equal=F)

        Welch Two Sample t-test

data:  totvol by trt
t = 3.7043, df = 48.25, p-value = 0.0005452
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 132.3707 446.5695
sample estimates:
mean in group 1 mean in group 2
       527.7778        238.3077
```

∇

Wilcoxon Rank-Sum Test for Non-Normally Distributed Data

The idea behind this test is as follows. Take samples of n_1 measurements from population 1 and n_2 measurements from population 2. Rank the n_1+n_2 measurements from 1 (smallest) to n_1+n_2 (largest), adjusting for ties by averaging the ranks the measurements would have received if they were different. Then compute T_1, the rank sum for measurements from population 1, and T_2, the rank sum for measurements from population 2. This test is mathematically equivalent to the Mann–Whitney U–test. To test for differences between the two population distributions, we use the following procedure, where to be able to use the commonly used table on the class webpage, $n_1 \geq n_2$. Before describing the procedure, define the following quantities.

$$T_{\text{Total}} = T_1 + T_2 = 1 + 2 + \cdots + (n_1 + n_2) = \frac{(n_1 + n_2)(n_1 + n_2 + 1)}{2}$$

$$T_1^{\min} = 1+2+\cdots+n_1 = \frac{n_1(n_1+1)}{2} \qquad T_2^{\min} = 1+2+\cdots+n_2 = \frac{n_2(n_2+1)}{2} \qquad T_2^{\max} = T_{\text{Total}} - T_1^{\min}$$

1. H_0 : The two population medians are equal ($M_1 = M_2$)
2. H_A : The medians are not equal ($M_1 \neq M_2$)
3. T.S.: $T = T_2$ (The rank sum for the group with smaller sample size)
4. R.R.: $T \leq T_0$ or $T \geq T_2^{\max} - \left(T_0 - T_2^{\min}\right)$, where values of T_0 given in tables in many statistics texts and on the web for various levels of α and sample sizes.

For one-sided tests to show that the distribution of population 1 is shifted to the right or left of population 2, use the following procedures (again, using with $n_1 \geq n_2$).

1. H_0 : The median for population 1 is less than or equal the median for population 2 ($M_1 \leq M_2$)
2. H_A : The median for population 1 is larger than the median for population 2 ($M_1 > M_2$)
3. T.S.: $T = T_2$
4. R.R.: $T \leq T_0$, where values of T_0 are given in tables in many statistics texts and on the web for various levels of α and various sample sizes.

1. H_0 : The median for population 1 is greater than or equal the median for population 2 ($M_1 \geq M_2$)
2. H_A : The median for population 1 is smaller than the median for population 2 ($M_1 < M_2$)
3. T.S.: $T = T_2$
4. R.R.: $T \geq T_2^{\max} - \left(T_0 - T_2^{\min}\right)$, where values of T_0 are given in tables in many statistics texts and on the web for various levels of α and various sample sizes.

Example 5.5: Apple Procyanidin B-2 for Hair Growth

A study was conducted to determine whether procyanidin B-2 from apples is effective in hair growth (Kamimura, Takahishi, and Watanabe (2000), [32]). Based on a small trial, with $n_1 = 19$ treatment subjects

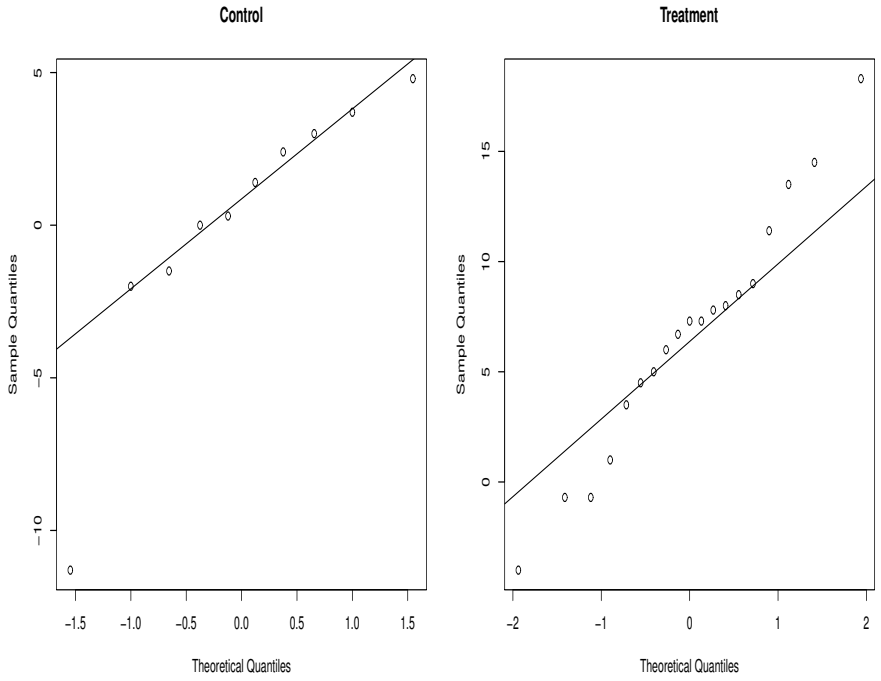

Figure 5.4: Change in total hairs - procyanidin B-2 from apples study

and $n_2 = 10$ control subjects, Table 5.2 gives the 6 month change in total hairs, along with their ranks from smallest (most negative) to largest. Note that $n_1 + n_2 = 19 + 10 = 29$. Normal probability plots are given in Figure 5.4, there is evidence of outlying cases in each group.

$$T_{\text{Total}} = T_1 + T_2 = 1 + 2 + \cdots + 29 = \frac{29(30)}{2} = 435 \qquad T_1^{\min} = 1 + 2 + \cdots + 19 = \frac{19(20)}{2} = 190$$

$$T_2^{\min} = 1 + 2 + \cdots + 10 = \frac{10(11)}{2} = 55 \qquad T_2^{\max} = 435 - 190 = 245$$

For a 2-tailed test with $\alpha = 0.05$, based on sample sizes of $n_1 = 19$ and $n_2 = 10$, the lower critical value is $T_0 = 107$ (see class webpage). The upper critical value is $245 - (107 - 55) = 245 - 52 = 193$. Thus, reject the null hypothesis of equal medians (no differences in effects) if the rank sum for treatment 2 (control, with the smaller sample size) is below 107 or above 193. Since 86 is (well) below 107, conclude the medians differ (and that $M_c < M_t$). If this had been conducted as a 1-tailed test (alternative being higher median for treatment group), the critical value would have been $T_0 = 113$.

R Commands and Output

```
## Commands

apphair <- read.table("http://www.stat.ufl.edu/~winner/data/apple_hair.dat",
  header=F, col.names=c("hair.trt","total0","total6","totaldiff",
```

Trt	TotalDif	Rank	Trt	TotalDif	Rank
1	0.3	8	2	3.5	13
1	1.4	10	2	5	17
1	3	12	2	7.3	20.5
1	3.7	14	2	18.3	29
1	-1.5	4	2	14.5	28
1	-2	3	2	6.7	19
1	0	7	2	9	25
1	4.8	16	2	-0.7	5.5
1	2.4	11	2	7.8	22
1	-11.3	1	2	-4	2
			2	6	18
			2	4.5	15
			2	8	23
			2	11.4	26
			2	1	9
			2	7.3	20.5
			2	8.5	24
			2	-0.7	5.5
			2	13.5	27
Total		$T_c = 86$			$T_t = 349$
Average		$T_c/n_c = 8.60$			$T_t/n_t = 18.37$

Table 5.2: Total Growth measurements (and ranks) for Procyanidin B-2 from Apple Hair Growth Experiment

```
  "term0", "term6", "termdiff"))
attach(apphair)

wilcox.test(totaldiff ~ hair.trt)

## Output

> wilcox.test(totaldiff ~ hair.trt)

        Wilcoxon rank sum test with continuity correction

data:  totaldiff by hair.trt
W = 31, p-value = 0.003565
alternative hypothesis: true location shift is not equal to 0

Warning message:
In wilcox.test.default(x = c(0.3, 1.4, 3, 3.7, -1.5, -2, 0, 4.8,  :
  cannot compute exact p-value with ties
```

Note that W represents the difference between the Rank Sum for each group and its minimum (low average rank group) or maximum (high average rank group) possible value. Making use of the notation above, W is defined below.

$$W = min\left(T_2 - T_2^{\min}, T_2^{\max} - T_2\right) = min(86 - 55 = 31, 245 - 86 = 159) = 31$$

∇

For large samples, it's difficult to find tables that contain the critical values (this example pushed the limits, in fact). The rank sums are approximately normal in large samples, so a normal approximation can be used. Let T be the rank sum for group 1 (the test is symmetric, so the statistic will have the same absolute value, no matter which group gets labeled as 1). The expected value and standard deviation of T under the null hypothesis $M_1 = M_2$ and the test statistic are given here.

$$n_\cdot = n_1 + n_2 \qquad T = T_1 \qquad \mu_T = \frac{n_1(n_\cdot + 1)}{2} \qquad \sigma_T = \sqrt{\frac{n_1 n_2 (n_\cdot + 1)}{12}} \qquad z_{obs} = \frac{T - \mu_T}{\sigma_T}$$

The critical values for the Rejection Region are based on whether the test is 2-tailed or upper tailed and α, as in other large-sample z-tests.

$$H_A : M_1 \neq M_2 \quad R.R. |z_{obs}| \geq z_{\alpha/2} \quad P = 2P(Z \geq |z_{obs}|) \qquad H_A : M_1 > M_2 \quad R.R. z_{obs} \geq z_\alpha \quad P = P(Z \geq z_{obs})$$

Example 5.5: Apple Procyanidin B-2 for Hair Growth

To use the large-sample approximation, let the treatment group be treatment 1 (again, the conclusions do not depend on this for a 2-tailed test).

$$n_1 = 19 \quad n_2 = 10 \quad n_\cdot = 29 \quad T = 349 \quad \mu_T = \frac{19(30)}{2} = 285 \quad \sigma_T = \sqrt{\frac{19(10)(30)}{12}} = 21.79$$

$$z_{obs} = \frac{349 - 285}{21.79} = 2.937 \quad P = 2P(Z \geq 2.937) = .0033$$

The rank sum for the treatment group is much larger than we would have expected under the null hypothesis of no treatment effect.

∇

5.2.2 Paired Sample Designs

In paired samples (aka crossover or within subjects) designs, subjects receive each treatment, thus acting as their own control. They may also have been matched based on some characteristics. Procedures based on these designs take this into account, and are based in determining differences between treatments after "removing" variability in the subjects (or pairs). When it is possible to conduct them, paired sample designs are more powerful than independent sample designs in terms of being able to detect a difference (reject H_0) when differences truly exist (H_A is true), for a fixed sample size and when measurements within subjects or pairs are positively correlated.

Paired t–test for Normally Distributed Data

In paired sample designs, each subject (or pair) receives each treatment. In the case of two treatments being compared, we compute the difference in the two measurements within each subject (or pair), and test whether or not the population mean difference is 0. When the differences are normally distributed, we use the paired t–test to determine if differences exist in the mean response for the two treatments. Then this is simply a 1-sample problem on the differences.

Let Y_1 be the score in condition 1 for a randomly selected subject, and Y_2 be the score in condition 2 for the subject. Let $D = Y_1 - Y_2$ be the difference. Further, suppose the following assumptions and their corresponding results. Note that the differences across subjects (or pairs) are considered to be independent.

$$E\{Y_1\} = \mu_1 \qquad V\{Y_1\} = \sigma_1^2 \qquad E\{Y_2\} = \mu_2 \qquad V\{Y_2\} = \sigma_2^2 \qquad \text{COV}\{Y_1, Y_2\} = \sigma_{12}$$

$$\Rightarrow \quad E\{D\} = \mu_1 - \mu_2 = \mu_D \qquad\qquad V\{D\} = \sigma_D^2 = \sigma_1^2 + \sigma_2^2 - 2\sigma_{12}$$

$$\overline{D} = \frac{\sum_{i=1}^n D_i}{n} \qquad E\{\overline{D}\} = \mu_D \qquad V\{\overline{D}\} = \sigma^2_{\overline{D}} = \frac{\sigma_D^2}{n} \qquad SE\{\overline{D}\} = \sigma_{\overline{D}} = \frac{\sigma_D}{\sqrt{n}}$$

$$\text{For large } n\text{: } \overline{D} \dot{\sim} N\left(\mu_D, SE\{\overline{D}\} = \frac{\sigma_D}{\sqrt{n}}\right)$$

Normality holds for any sample size if the individual measurements (or the differences) are normally distributed.

It should be noted that in the paired case $n_1 = n_2$ by definition. That is, there will always be equal sized samples when the experiment is conducted properly. There will be $n = n_1 = n_2$ differences, even though there were $2n = n_1 + n_2$ measurements made. From the n differences obtained in a sample, the mean and standard deviation are obtained, and will labeled as $\overline{d}$ and s_d.

$$\overline{d} = \frac{\sum_{i=1}^n d_i}{n} \qquad s_d^2 = \frac{\sum_{i=1}^n (d_i - \overline{d})^2}{n-1} \qquad s_d = \sqrt{s_d^2} \qquad \hat{SE}\{\overline{D}\} = s_{\overline{D}} = \frac{s_d}{\sqrt{n}}$$

A $(1-\alpha)100\%$ Confidence Interval for the population mean difference μ_D is given below.

$$\overline{d} \pm t_{\alpha/2,n-1}\hat{SE}\{\overline{D}\} \qquad \equiv \qquad \overline{d} \pm t_{\alpha/2,n-1}\frac{s_d}{\sqrt{n}}$$

The test is conducted as follows.

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$

2. $H_A : \mu_D \neq 0$ or $H_A : \mu_D > 0$ or $H_A : \mu_D < 0$ (which alternative is appropriate should be clear from the setting).
3. T.S.: $t_{obs} = \frac{\overline{d}}{\widehat{SE}\{\overline{D}\}} = \frac{\overline{d}}{\left(\frac{s_d}{\sqrt{n}}\right)}$
4. R.R.: $|t_{obs}| \geq t_{\alpha/2,n-1}$ or $t_{obs} \geq t_{\alpha,n-1}$ or $t_{obs} \leq -t_{\alpha,n-1}$ (which R.R. depends on which alternative hypothesis you are using).
5. p-value: $2P(t_{n-1} \geq |t_{obs}|)$ or $P(t_{n-1} \geq t_{obs})$ or $P(t_{n-1} \leq t_{obs})$ (again, depending on which alternative you are using).

Example 5.6: Comparison of Two Analytic Methods for Determining Wine Isotope

A study was conducted to compare two analytic methods for determining $^{87}Sr/^{86}Sr$ isotope ratios in wine samples (Durante, et al (2015), [19]). These are used in geographic tracing of wine. The two methods are microwave (method 1) and low temperature (method 2). The data, and the differences (microwave - lowtemp) are given in Table 5.3.

sample id	microwave	lowtemp	diff(m-l)
1	0.70866	0.70861	0.000050000
2	0.708762	0.708792	-0.00003000
3	0.708725	0.708734	-0.00000900
4	0.708668	0.708662	0.000006000
5	0.708675	0.70867	0.000005000
6	0.708702	0.708713	-0.00001100
7	0.708647	0.708661	-0.00001400
8	0.708677	0.708667	0.000010000
9	0.709145	0.709176	-0.00003100
10	0.709017	0.709024	-0.00000700
11	0.70882	0.708814	0.000006000
12	0.709402	0.709364	0.000038000
13	0.709374	0.709378	-0.00000400
14	0.709508	0.709517	-0.00000900
15	0.70907	0.709063	0.000007000
16	0.709061	0.709079	-0.00001800
17	0.709096	0.709039	0.000057000
18	0.70872	0.7087	0.000020000
Mean	0.708929	0.708926	0.000003667
SD	0.000287	0.000288	0.000024646

Table 5.3: $^{87}SR/^{86}SR$ Isotope ratios for 18 wine samples by Microwave and Low Temperature Methods

As there are $n = 18$ differences, the degrees of freedom are $n - 1 = 17$.The 95% Confidence Interval for μ_D is computed below, where $t_{.025,17} = 2.110$. First, the mean and standard deviation of the differences are multiplied by 100000 (remove first 5 0s after decimal) to reduce the risk of calculation error. This is legitimate as the mean and standard deviation are of the same units. This leads to $\overline{d}^* = 0.3667$ and $s_d^* = 2.46466$.

$$0.3667 \pm 2.110\frac{2.4646}{\sqrt{18}} \quad \equiv \quad 0.3667 \pm 2.110(0.5809) \quad \equiv \quad 0.3667 \pm 1.2257 \quad \equiv \quad (-0.8590, 1.5924)$$

In the original units the interval is of the form of (-.00000859,.000015924). Since the interval contains 0, there is no evidence that one method tends to score higher (or lower) than the other on average.

The test of whether there is a difference in the true mean determinations between the two methods (with $\alpha = 0.05$) is conducted by completing the steps outlined below.

1. $H_0 : \mu_1 - \mu_2 = \mu_D = 0$
2. $H_A : \mu_D \neq 0$
3. T.S.: $t_{obs} = \frac{0.3667}{\left(\frac{2.4646}{\sqrt{18}}\right)} = \frac{0.3667}{0.5809} = 0.631$
4. R.R.: $t_{obs} > t_{\alpha/2,n-1} = t_{.025,17} = 2.110$
5. P-value: $2P\left(t_{17} \geq 0.631\right) = .5364$

There is definitely no evidence that the two methods differ in terms of determinations of wine isotope ratios.

R Commands and Output

```
## Commands

wine1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wine_isotope.csv")
attach(wine1); names(wine1)

## t.test Function
t.test(microwave, lowtemp, paired=TRUE)

## Output

> round(wine.out, 6)
        ybar1       s1    ybar2       s2 cor(y1,y2)
[1,] 0.708929 0.000287 0.708926 0.000288   0.996329

> round(diff.out,9)
          mean         SD   Std Err         t   P(>|t|)         LB         UB
[1,] 3.667e-06 2.4646e-05 5.809e-06 0.6311987 0.5363058 -8.589e-06 1.5923e-05

> t.test(microwave, lowtemp, paired=TRUE)

        Paired t-test

data:  microwave and lowtemp
t = 0.6312, df = 17, p-value = 0.5363
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -8.589364e-06  1.592270e-05
sample estimates:
mean of the differences
           3.666667e-06
```

∇

Wilcoxon Signed–Rank Test for Paired Data

A nonparametric test that is often conducted in paired sample designs is the Wilcoxon Signed-Rank test. Like the paired t-test, the signed-rank test takes into account that the two treatments are being assigned to the same subject (or pair). The test is based on the difference in the measurements within each subject (pair). Any subjects (pairs) with differences of 0 (measurements are equal under both treatments) are removed and the sample size is reduced. The test statistic is computed as follows.

1. For each pair, subtract measurement 2 from measurement 1.
2. Take the absolute value of each of the differences, and rank from 1 (smallest) to n (largest), adjusting for ties by averaging the ranks they would have had if not tied.
3. Compute T^+, the rank sum for the positive differences from step 1, and T^-, the rank sum for the negative differences.

To test whether or not the population distributions are identical, use the following procedure:

1. H_0 : The two population distributions have equal Medians ($M_1 = M_2$)
2. H_A : The Medians Differ ($M_1 \neq M_2$)
3. T.S.: $T = \min(T^+, T^-)$
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts and on the web.

For a one-sided test, if you wish to show that the distribution of population 1 is shifted to the right of population 2 ($M_1 > M_2$), the procedure is as follows:

1. H_0 : The two population distributions have equal Medians ($M_1 = M_2$)
2. H_A : Distribution 1 is shifted to the right of distribution 2 ($M_1 > M_2$)
3. T.S.: $T = T^-$
4. R.R.: $T \leq T_0$, where T_0 is a function of n and α and given in tables in many statistics texts and on the web.

Note that if the goal is to test with the alternative $M_1 < M_2$, use the above procedure with T^+ replacing T^-. The idea behind this test is to determine whether the differences tend to be positive ($M_1 > M_2$) or negative ($M_1 < M_2$), where differences are 'weighted' by their magnitude.

Example 5.7: Water Consumption by Cats under Still and Flowing Sources

A small pilot study was conducted to compare the daily amount of water consumed (mL) by cats when presented with still or flowing water (Pachel and Neilson (2010) [41]). Each of $n = 9$ cats was observed 2 days each under each condition, and the mean for each condition was computed for each cat. Data are given

Cat (i)	still	flowing	d_i=still-flowing	$\lvert d_i \rvert$	rank($\lvert d_i \rvert$)
1	157.5	164.5	-7	7	2
2	84.5	51.5	33	33	6
3	134.0	250.0	-116	116	9
4	74.0	139.0	-65	65	7
5	108.0	113.0	-5	5	1
6	107.5	124.5	-17	17	4
7	106.0	95.5	10.5	10.5	3
8	163.0	70.5	92.5	92.5	8
9	54.0	30.5	23.5	23.5	5

Table 5.4: Average daily water consumed by cats in still and flowing conditions

in Table 5.4, along with ranks. We will test whether there is evidence that the true medians differ (even though this is clearly a very small sample).

Based on Table 5.4, T^+ (the sum of the ranks for positive differences) and T^- (the sum of the ranks of the negative differences), as well as the test statistic T, are computed as follows.

$$T^+ = 6+3+8+5 = 22 \qquad T^- = 2+9+7+1+4 = 23 \qquad T = \min(T^+, T^-) = \min(22, 23) = 22$$

Note that short of there having been a tie, this is the closest T^+ and T^- could be. Using the previously given steps, the test for differences in the medians of the true distributions for the 2 water conditions is given below.

1. H_0 : The two population medians ($M_1 = M_2$)
2. H_A : One distribution is shifted to the right of the other ($M_1 \neq M_2$)
3. T.S.: $T = \min(T^+, T^-) = 22$
4. R.R.: $T \leq T_0$, where $T_0 = 5$ is based on 2-sided alternative , $\alpha = 0.05$, and $n = 9$.

Since $T = 22$ does not fall in the rejection region, fail to reject H_0, and fail to conclude that the medians differ. Note that the P-value is thus larger than 0.05, since we fail to reject H_0 (in fact it is 1).

R Commands and Output

```
## Commands

still <- c(157.5, 84.5, 134, 74, 108, 107.5, 106, 163, 54)
flowing <- c(164.5, 51.5, 250, 139, 113, 124.5, 95.5, 70.5, 30.5)
wilcox.test(still, flowing, paired=TRUE)

## Output

> wilcox.test(still, flowing, paired=TRUE)

        Wilcoxon signed rank test

data:  still and flowing
V = 22, p-value = 1
alternative hypothesis: true location shift is not equal to 0
```

∇

In large-samples, the rank-sums T^+ and T^- have approximately normal sampling distributions. By definition, $T^+ + T^- = 1 + \cdots + n = \frac{n(n+1)}{2}$. Under the null hypothesis $H_0 : M_1 = M_2$, the mean and variance for T^+ and T^- are given below.

$$\mu_T = \frac{n(n+1)}{4} \qquad \sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} \qquad z_{obs} = \frac{T - \mu_T}{\sigma_T}$$

The usual rules for rejection regions and P-values apply. If the alternative is $H_A : M_1 > M_2$ use $T = T^+$ and reject H_0 if $z_{obs} \geq z_\alpha$. If the alternative is $H_A : M_1 < M_2$ use $T = T^-$ and reject H_0 if $z_{obs} \geq z_\alpha$. For $H_A : M_1 \neq M_2$, use either T^+ or T^-, and reject if $|z_{obs}| \geq z_{\alpha/2}$.

Example 5.8: Efficiency Comparison of Recreational and Professional Bettors

An economic study was conducted, comparing recreational and professional bettors' efficiencies (Bruce, Johnson, and Peirson (2012), [10]). They considered race attendees as Recreational bettors and remote (online) bettors as Professional bettors. The authors had aggregate returns (amount won divided by amount bet) data for both groups on $n = 2057$ races. The difference (remote - attendee) was obtained for each race. There were 963 negative differences (attendees outperformed remote bettors) and 1094 positive differences. The rank sum information is given below.

$$T^+ = 1167023.5 \quad T^- = 949629.5 \quad T^+ + T^- = 2116653 = 1 + \cdots + 2057 \quad \mu_T = \frac{2057(2058)}{4} = 1058326.5$$

$$\sigma_T = \sqrt{\frac{2057(2057+1)(2(2057)+1)}{24}} = 26941.34 \qquad z_{obs} = \frac{1167023.5 - 1058326.5}{26941.34} = 4.03$$

There is strong evidence of a difference in the two groups. Note the authors also present the mean and the standard deviation of the differences. The 95% Confidence Interval for μ_D is (.0287,.0671), an advantage in aggregate return of about 2.9% to 6.7%.

$$\overline{y}_r = 0.8659 \quad \overline{y}_a = 0.8180 \quad \overline{d} = 0.0479 \quad s_d = 0.4448 \quad \frac{s_d}{\sqrt{n}} = \frac{0.4448}{\sqrt{2057}} = 0.0098$$

$$0.0479 \pm 1.96(0.0098) \equiv (0.0287, 0.0671)$$

∇

5.3 Power and Sample Size Considerations

In this section, issues of power and sample size are considered in the 2-Sample Location problem. Power refers to the probability of rejecting the null hypothesis. When H_0 is true, it should be α, and when the

alternative is true, it will depend on the magnitude of the difference, the variability and the sample sizes. Once power has been considered empirically, sample size computations will be made based on distributional results.

5.3.1 Empirical Study of Power

To compare the power of the independent sample t-test and the Wilcoxon Rank-Sum test, consider the populations of NHL/EPL players' BMI and the Female and Male marathon runner's speeds. The BMI distributions were approximately normal, while the marathon speeds were right skewed.

Example 5.9: Small-Sample Inference Comparing BMI for NHL and EPL Players

The means and standard deviations of the BMI levels for NHL and EPL players are given below, along with the mean and variance of the sampling distribution of $\overline{Y}_n - \overline{Y}_e$. Note that as each distribution is approximately normal, its sampling distribution will be very close to a normal distribution, even with relatively small samples. Further, the variances are not equal, although they are not too far apart. Refer back to Figure 5.2 for a histogram of 100000 random samples' mean differences of $n_1 = n_2 = 20$.

$$\text{BMI: } \mu_n = 26.50 \quad \sigma_n = 1.45 \quad \mu_e = 23.02 \quad \sigma_e = 1.71 \quad E\{\overline{Y}_n - \overline{Y}_e\} = 26.50 - 23.02 = 3.48$$

$$V\{\overline{Y}_n - \overline{Y}_e\} = \frac{1.45^2}{n_n} + \frac{1.71^2}{n_e}$$

We compare the coverage rates of small sample Confidence Intervals based on equal variance and unequal variance assumptions, as well as their widths for samples of $n_n = n_e = 10$. The unequal variance case will always be wider, as the sample mean difference and estimated standard error will be the same as the equal variance case, but will have fewer degrees of freedom. Due to the equivalence of the 2-tailed test and Confidence Interval for testing $H_0 : \mu_n - \mu_e = 0$, the empirical power of the two methods are observed as well. The process is conducted as follows.

1. Sample 10 players from NHL and 10 players from EPL
2. Compute $\overline{y}_n$, s_n, $\overline{y}_e$, s_e
3. Compute the sample mean difference $\overline{y}_n - \overline{y}_e$ and its estimated standard error $\hat{SE}\{\overline{Y}_n - \overline{Y}_e\} = \sqrt{\frac{s_n^2}{10} + \frac{s_e^2}{10}}$
4. Compute the approximate degrees of freedom for the unequal variance case (Satterthwaite's approximation)
5. Obtain the 95% Confidence Intervals for $\mu_n - \mu_e$
6. Determine whether the Confidence Intervals contain 3.48 (true value) and whether they contain 0 (Testing $\mu_n - \mu_e = 0$)
7. Obtain the width of the intervals

The equal variance Confidence Intervals contained $\mu_n - \mu_e = 3.48$ in 95.18% of the samples, the unequal variance CI's covered in 95.37% of the samples. Based on equal sample sizes, (and will typically always be the case) the unequal case will always have wider intervals and thus higher coverage rates at the cost of being wider. The average width of the equal variance CI's was 2.9235 versus 2.9539 for the unequal case. The unequal case was only about 1% wider on average due to how similar the population standard deviations are. The equal variance case rejected $H_0 : \mu_n - \mu_e = 0$ in favor of $H_A : \mu_n - \mu_e \neq 0$ in 99.01% of the samples, while the unequal variance case did so in 98.89%. Neither ever rejected with a negative t-statistic. The mean difference was very large relative to the standard deviations for the two leagues, so it's not surprising to have such high power.

R Output

```
## Output

> round(bmisim.out1, 2)
     mu_nhl mu_epl mu_nhl-mu_epl sigma_nhl sigma_epl n_nhl n_epl SE{Ybar_n-Ybar_e)
[1,]   26.5  23.02          3.48      1.45      1.71    10    10              0.71

> round(bmisim.out2, 4)
     EV Cover UV Cover EV Width UV Width EV Diff > 0 UV Diff > 0
[1,]   0.9518   0.9537   2.9235   2.9539      0.9901      0.9889
```

∇

Example 5.10: Small-Sample Inference for Female and Male Marathon Speeds

Comparisons among Female and Male marathon speeds are now made. Unlike the NHL/EPL Body Mass Indices, these speeds are not approximately normally distributed, but are rather skewed to the right, refer to Figure 3.5. The population means and standard deviations are given below, along with the mean and standard error of the sampling distribution of the sample mean $\overline{Y}_f - \overline{Y}_m$.

$$\mu_f = 5.840 \quad \sigma_f = 0.831 \qquad \mu_m = 6.337 \quad \sigma_m = 1.058$$

$$E\{\overline{Y}_f - \overline{Y}_m\} = -0.497 \qquad SE\{\overline{Y}_f - \overline{Y}_m\} = \sqrt{\frac{0.831^2}{n_f} + \frac{1.058^2}{n_m}}$$

We will consider fairly small samples, $n_f = n_m = 6$, and first repeat the comparisons made in BMI example, and further compare the t-tests with the Wilcoxon Rank-Sum test in terms of power for testing $H_0 : \mu_f - \mu_m \geq 0$ vs $H_A : \mu_f - \mu_m < 0$. The equal variance Confidence Interval covered $\mu_f - \mu_m = -0.497$ in 94.85% of samples, the unequal case covered in 95.36%, so even with these small samples, and the skewed distributions, the t-based Confidence Intervals performed well. In terms of concluding $H_A : \mu_f - \mu_m < 0$, the equal variance t-test correctly rejected H_0 in 20.73% of samples, the unequal variance t-test in 19.67%, and the Wilcoxon Rank-Sum test in 18.91%.

R Output

```
## Output

> round(rrsim.out1, 2)
     mu_f mu_m mu_f-mu_m sigma_f sigma_m n_f n_m SE{Ybar_f-Ybar_m)
[1,] 5.84 6.34      -0.5    0.83    1.06   6   6              0.55

> round(rrsim.out2, 4)
     EV Cover UV Cover EV Reject UV Reject Rank-Sum Reject
[1,]   0.9485   0.9536    0.2072    0.1967          0.1891
```

∇

5.3.2 Power Computations

To obtain the sample sizes needed to detect an important difference in means, the non-central t-distribution can be used in a similar manner to what was done for the one-sample problem. The only difference is that instead of looking for an important difference from some pre-specified null mean, we are interested in the difference between two population means. First, consider the case of independent samples. This is generally done under the assumption of equal variances.

$$H_0 : \mu_1 - \mu_2 = 0 \quad H_A : \mu_1 - \mu_2 = (\mu_1 - \mu_2)_A \neq 0 \quad \Delta = \frac{(\mu_1 - \mu_2)_A}{\sqrt{\sigma^2\left(\frac{2}{n}\right)}} \qquad t = \frac{\overline{Y}_1 - \overline{Y}_2}{\sqrt{S_p^2\left(\frac{2}{n}\right)}} \sim t_{2(n-1),\Delta}$$

If σ is known (or well approximated), researchers can choose an important difference $(\mu_1 - \mu_2)_A$, and determine the sample size that gives a reasonable power π to detect it based on a test with significance level α. In other situations, an important **effect size** $\delta = (\mu_1 - \mu_2)_A / \sigma$ can be obtained, which measures the difference in means in standard deviation units. Once the important effect size is chosen, beginning with small n, the power π is determined and the process continues until the desired power is obtained. The process works as follows for a 2-tailed test.

1. Determine important effect size $\delta = (\mu_1 - \mu_2)_A / \sigma$ and set the significance level α and desired power π.
2. Starting with (say) $n_1 = n_2 = n = 2$, obtain the degrees of freedom $2(n-1)$ and critical value $t_{\alpha/2,2(n-1)}$.
3. Compute the non-centrality parameter $\Delta = \frac{\delta}{\sqrt{2/n}}$.
4. Obtain π_n: the probability the non-central t is greater than $t_{\alpha/2,2(n-1)}$ or less than $-t_{\alpha/2,2(n-1)}$.
5. If π_n exceeds the desired π, stop. Otherwise, increment n by 1 and repeat the process.

In the case of 1-tailed tests, the Rejection Region is in only 1-tail, with area α and only one of the tail area probabilities is computed.

Example 5.11: Power Calculation for Comparison of Female and Male Marathon Speeds

Using numbers similar to those observed in the populations of marathon runners, suppose we want to be able to detect a difference $(\mu_f - \mu_m)_A = -0.5$ and that $\sigma_f = \sigma_m = \sigma = 0.94$ (we are just averaging the true standard deviations for computational purposes). We then obtain the following results. Start with $n_f = n_m = n = 6$, since the power was so low (approximately 0.20) for the lower-tailed t-test in Example 5.10.

$$\delta = \frac{-0.50}{0.94} = -0.532 \quad \Delta_6 = \frac{-0.532}{\sqrt{2/6}} = -0.921 \quad df = 2(6-1) = 10 \quad -t_{.05,10} = -1.812$$

For the lower-tailed test $H_A : \mu_f - \mu_m < 0$, for these sample sizes, reject the null of no difference if $t_{obs} \leq -1.812$. Now find the probability under the non-central t-density with $2(6-1) = 10$ degrees of freedom and non-centrality parameter -0.921 that is below -1.812. The power turns out to be 0.216 (see R output below). Using the R functions **qt** for quantiles and **pt** for lower tail probabilities (cumulative distribution function), the relevant probabilities (powers) can be obtained. Samples of size $n_f = n_m = 45$ would be needed for the power to reach 0.8.

R Output

```
## Output

> round(power.out1,3)
     alpha pi* n df (mu1-mu2)_A sigma  delta  Delta -t(.05,df) power
[1,]  0.05 0.8 6 10        -0.5  0.94 -0.532 -0.921     -1.812 0.216

> cbind(n.out, power.out)
      n.out power.out
 [1,]     7 0.2402697
 [2,]     8 0.2636261
        ...
[38,]    44 0.7968277
[39,]    45 0.8047651
```

Had this been a 2-tailed test with $H_A : \mu_f - \mu_m \neq 0$, the Rejection Region would be $|t_{obs}| \geq t_{\alpha/2,2(n-1)}$. Below are the R Commands and Output that computes the power for the 2-tailed test (it only contains the initial calculation, the loop part is similar to the lower-tail test). Samples of $n = 57$ females and males would be needed for the power to reach 0.80.

R Output

```
## Output

> round(power.out2,3)
     alpha pi* n df (mu1-mu2)_A sigma  delta  Delta t(.025,df) power
[1,]  0.05 0.8 6 10        -0.5  0.94 -0.532 -0.921      2.228 0.133

> cbind(n.out, power.out)
      n.out power.out
 [1,]     7 0.1505426
      ...
```

```
[50,]    56 0.7967349
[51,]    57 0.8037961
```

∇

In terms of the paired t-test, when testing $H_0 : \mu_D = 0$ vs $H_A : \mu_D \neq 0$, there may be a specific difference μ_{DA} that would like to be detected with a specified power π. This is very similar to the 1-sample problem in the previous chapter. Define the following terms, where μ_{DA} is the mean difference under H_A and σ_D is the standard deviation of the differences.

$$t_{obs} = \frac{\overline{d}}{s_d/\sqrt{n}} = \sqrt{n}\frac{\overline{d}}{s_d} \qquad \delta = \frac{\mu_{DA}}{\sigma_D} \qquad \Delta = \sqrt{n}\delta$$

Again δ is the effect size and Δ is the non-centrality parameter. The degrees of freedom for the paired t-test is $n - 1$. The process generalizes directly from the independent samples method described above.

Example 5.12: Water Consumption by Cats under Still and Flowing Sources

In the pilot study of cats drinking flowing versus still water, the standard deviation of the differences was approximately 60 ml. Suppose the researchers would like to detect a true mean difference of $\mu_{DA} = 30$ mL with power $\pi = 0.75$. In this setting $\delta = 30/60 = 0.5$ and $\Delta = \sqrt{n}(0.5)$. Beginning with the authors' original sample of $n = 9$, we obtain the power then iterate until $\pi \geq 0.75$. The R program and output are given below, for $n = 9$, $\pi = 0.263$. A sample of $n = 30$ would be needed to reach $\pi = 0.75$.

R Output

```
## Output

> round(power.out3,3)
     alpha  pi* n df (mu1-mu2)_A sigma delta Delta t(.025,df) power
[1,]  0.05 0.75 9  8        -0.5  0.94   0.5   1.5      2.306 0.263

> cbind(n.out, power.out)
      n.out power.out
 [1,]    10 0.2931756
      ...
[20,]    29 0.7386963
[21,]    30 0.7539647
```

∇

5.4 Methods Based on Resampling

In this section, two methods for comparing two means are considered. These are the **Bootstrap** and **Randomization/Permutation Tests**.

5.4.1 The Bootstrap

The bootstrap method is the same principle as in the one-sample case. In terms of independent samples, take resamples within each group with replacement, then take the difference between the two group means in each subsample. This will be illustrated below. In terms of paired samples, the one-sample methods are used on the observed paired differences from the original sample.

For the Bootstrap t Intervals, for each resample, compute $\overline{y}^*_{1i}, s^*_{1i}, \overline{y}^*_{2i}, s^*_{2i}$ for the i^{th} resample, and compute t^*_i as below, where $n_1, \overline{y}_1, s_1, n_2, \overline{y}_2, s_2$ are the sizes, means, and standard deviations of the original samples.

$$t^*_i = \frac{(\overline{y}^*_{1i} - \overline{y}^*_{2i}) - (\overline{y}_1 - \overline{y}_2)}{\sqrt{\frac{s^{*2}_{1i}}{n_1} + \frac{s^{*2}_{2i}}{n_2}}} \quad i = 1, \ldots, B$$

Once the B t^*_i statistics are obtained the $\alpha/2$ and $1-\alpha/2$ quantiles are obtained and labeled Q^*_L and Q^*_U, respectively. The $(1-\alpha)100\%$ Bootstrap t CI for $\mu_1 - \mu_2$ is of the following form.

$$(\overline{y}_1 - \overline{y}_2) - Q^*_U\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}} \quad , \quad (\overline{y}_1 - \overline{y}_2) - Q^*_L\sqrt{\frac{s^2_1}{n_1} + \frac{s^2_2}{n_2}}$$

Example 5.13: Anthropometric Measurements of Lahoul and Kulu Kanets in Punjab

A study sampled 30 Lahoul Kanet adults and 60 Kulu Kanet adults, making various physical measurements (Holland (1902) [26]). The author reported on 7 characteristics among each subject. Consider the variable cubit (cm), given in Table 5.5. The summary statistics from the samples are given below.

$$n_L = 30 \quad \overline{y}_L = 44.657 \quad s_L = 2.056 \qquad n_K = 60 \quad \overline{y}_K = 45.298 \quad s_K = 1.692 \qquad \overline{y}_L - \overline{y}_K = -0.641$$

We take 10000 resamples of 30 Lahoul and 60 Kulu Kanets, obtaining the means for each group and the difference. Then, obtaining the bootstrap mean and standard error for the differences, along with the bootstrap percentile intervals from the 2.5 and 97.5 percentiles of the resampled mean differences. The mean of the 10000 mean differences is -0.645, the bootstrap standard error is 0.430, and the 95% bootstrap percentile Confidence Interval is (-1.483, 0.192). A histogram of the resample mean differences and a normal probability plot are given in Figure 5.5.

R Output

```
## Output

> round(boot.out1, 3)
     ybar_L ybar_K yb_L-yb_K   s_L   s_K Mean(MeanDiff) SD(MD) Q.025(MD) Q.975(MD)
     44.657 45.298    -0.642 2.056 1.692        -0.645   0.43    -1.483     0.192
```

Lahoul	Lahoul	Kulu	Kulu	Kulu	Kulu
45.2	44.3	44.8	46.6	44.9	43.2
46.9	46.6	45.7	43.3	46.1	45.7
44.7	42.4	44.4	44.9	47.5	46.4
46.3	42.7	45.8	44.6	44.9	49.3
43.4	44.9	44.6	45.3	49.2	46.1
43.3	42.3	44.3	44.6	43.7	44.7
39.6	43.5	45.4	47.8	46.0	45.1
45.6	42.9	44.3	44.0	43.7	43.4
43.6	46.8	44.8	47.8	45.4	45.6
44.2	46.2	43.2	47.8	45.0	47.7
47.4	43.9	46.5	44.9	42.8	50.3
48.2	46.8	45.0	45.1	47.1	42.1
45.0	43.3	46.8	44.3	45.7	46.2
45.4	42.5	41.9	45.5	45.2	44.1
42.9	48.9	44.9	43.8	42.5	45.6

Table 5.5: Cubit lengths (cm) for samples of 30 Lahoul Kanets and 60 Kulu Kanets

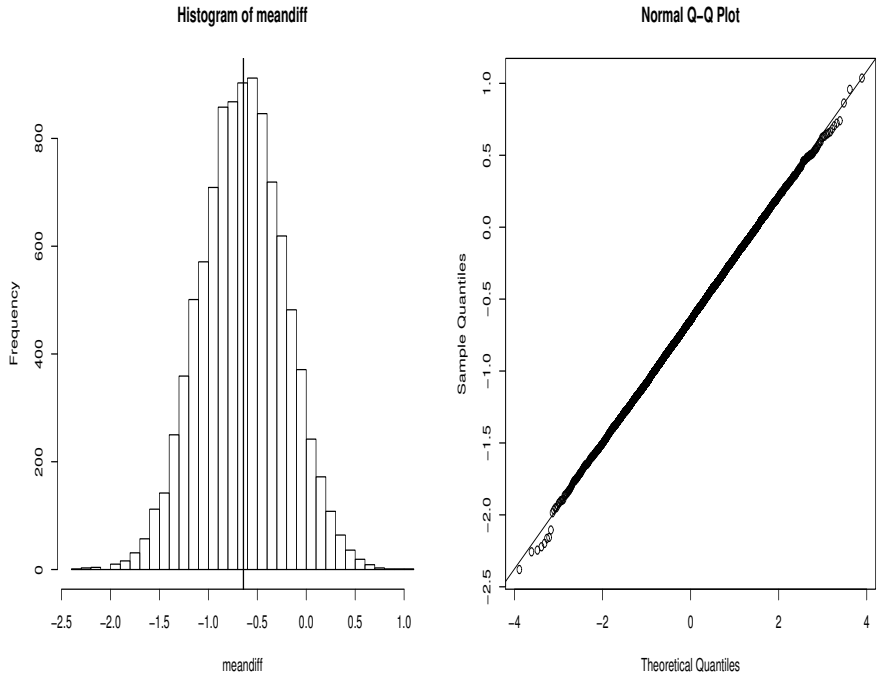

Figure 5.5: Histogram and Normal Probability Plot for Bootstrap Resample Mean Differences (Lahout - Kulu)

For the 95% Bootstrap t Confidence Interval, the .025 quantile of t^* is $Q_L^* = -1.762$ and the .975 quantile is $Q_U^* = 1.794$ and the resulting 95% Confidence Interval is (-1.421, 0.123).

R Output

```
## Output

> round(boot.out2, 3)
     ybar_L ybar_K yb_L-yb_K   s_L   s_K n_L n_K SE{diff}    Q_L   Q_U     LB    UB
     44.657 45.298    -0.642 2.056 1.692  30  60    0.434 -1.762 1.794 -1.421 0.123
```

∇

5.4.2 Randomization/Permutation Tests

Randomization/Permutation tests consider the observed responses as being made up of a treatment/population mean and a random error term. That is, $Y_{ij} = \mu_i + \epsilon_{ij}, i = 1, 2; j = 1, \ldots, n_{ij}$. The random error term is unique to the experimental unit that it corresponds to, and could be due to any number of factors. If there are no differences in the treatment/population means ($\mu_1 = \mu_2$), then all of the observed values could have come from either treatment/population on any number of randomizations by the experimenter or nature. The process of randomization and permutation tests is as follows for the independent sample t-test.

1. Compute a statistic from the original data that measures a discrepancy between the sample data and the null hypothesis, such as $\overline{y}_1 - \overline{y}_2$.
2. Generate many permutations (N) of the original samples to the two groups and compute and save the statistic for each permutation.
3. Count the number of permutations for which the statistic is as or more extreme than the original sample's value.
4. The P-value is (Count+1)/(N+1) the proportion of the statistics as or more extreme than the original (including the original).

Example 5.14: Cubit Lengths of Lahout and Kulu Kanets

To illustrate the test, consider the lengths of the cubits of the Lahout and Kulu Kanets. In Example 5.14, the mean difference from the original samples was $\overline{y}_L - \overline{y}_K = -0.641$. Suppose there is no difference in the two cultures' tendencies to generate different cubit lengths and they are due to randomness among individuals who "nature" randomized to the cultures. Then consider 9999 permutations of these 90 cubit lengths to the n_L Lahouts and n_K Kulus. Of $N = 9999$ permutation samples, 1207 were as large as the observed difference in absolute value, for a P-value of (1207+1)/(9999+1) = .1208. Thus, there is no evidence to reject the null hypothesis that $\mu_L = \mu_K$. A histogram of the permutation mean differences with a vertical line at the observed mean difference is given in Figure 5.6.

R Output

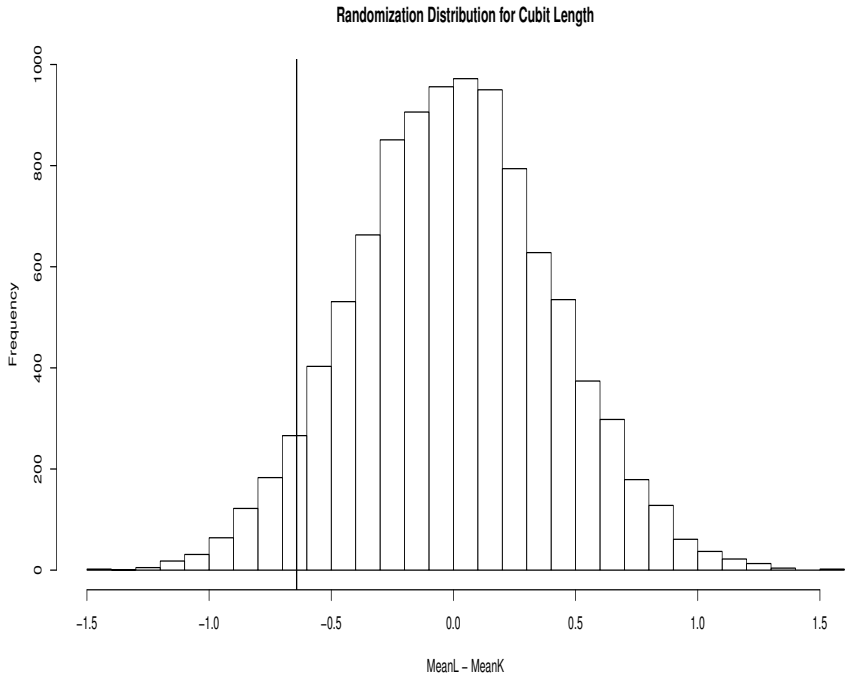

Figure 5.6: Randomization Distribution for Lahout and Kulu Kanet cubit measurements

```
## Output

> round(perm.out1, 4)
      ybar_L  ybar_K Test Stat Extreme Perms P-value
[1,] 44.6567 45.2983   -0.6417          1182  0.1183
```

∇

For paired samples, if there is no difference in the means of the two treatments, then the 2 observed measurements on each unit or pair could have just as easily appeared under either of the two treatments. The process for the Randomization/Permutation test goes as follows.

1. Compute a statistic from the original data that measures a discrepancy between the sample data and the null hypothesis, such as $\overline{d}$.
2. Generate many permutations (N) of the signs of the observed differences, where for each unit, its sign is changed with probability 0.5 (in effect switching the observed scores for the two treatments). Compute and save the mean difference $\overline{d}^*$.
3. Count the number of permutations for which the statistic is as or more extreme than the original sample's value.
4. The P-value is (Count+1)/(N+1) the proportion of the statistics as or more extreme than the original (including the original).

Example 5.15: Home Field Advantage in English Premier League Football (2012)

The English Premier League has 20 football clubs. Each club plays the remaining 19 clubs twice each season (once at home, once away). If clubs are labeled in alphabetical order from 1:20, then let $y_{1jk} = H_j - A_k$ $j < k$ be the score differential (Home-Away) when club j played at home versus club k. Further, let $y_{2jk} = A_j - H_k$ $j < k$ be the score differential (Away-Home) when club j played away versus club k. Then:

$$d_{jk} = y_{1jk} - y_{2jk} = (H_j - A_k) - (A_j - H_k) = (H_j + H_k) - (A_j + A_k)$$

That is, d_{jk} represents the total home versus away differential for the two matches played between clubs j and k. There are $\begin{pmatrix} 20 \\ 2 \end{pmatrix} = 190$ pairs of clubs. If there is no home field differential, then $\mu_D = 0$. Here we conduct a 2-tailed permutation test for a home field differential. There is overwhelming evidence of a home field advantage. None of the permutation means is close to the observed mean $\overline{d} = 0.6368$. A histogram of the randomization distribution and observed mean differential (vertical line) is given in Figure 5.7.

R Output

```
## Output

> round(perm.out2, 4)
       n Observed TS # Exceed 1-tail # Exceed 2-tail 1-tailed P-value 2-tailed P-value
[1,] 190      0.6368               0               0            1e-04            1e-04
```

∇

5.5 R Code for Chapter 5

```
### Chapter 5

### Example 5.1

bmi.sim <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_nba_ebl_bmi.csv")
attach(bmi.sim); names(bmi.sim)
N.nhl <- 717 # # of NHL players
N.epl <- 526 # # of EPL players
bmi.nhl <- NHL_BMI[1:N.nhl]
bmi.epl <- EPL_BMI[1:N.epl]
(mu.nhl <- mean(bmi.nhl)); (sigma.nhl <- sd(bmi.nhl)*sqrt((N.nhl-1)/N.nhl))
(mu.epl <- mean(bmi.epl)); (sigma.epl <- sd(bmi.epl)*sqrt((N.epl-1)/N.epl))

## Figure 5.1

par(mfrow=c(2,1))
hist(bmi.nhl, breaks=30, xlim=c(18,32), freq=F,
main="Histogram of NHL BMI and N(26.50,1.45) Density")
bmi.x <- seq(18,32,.01)
lines(bmi.x, dnorm(bmi.x, mu.nhl, sigma.nhl))
```

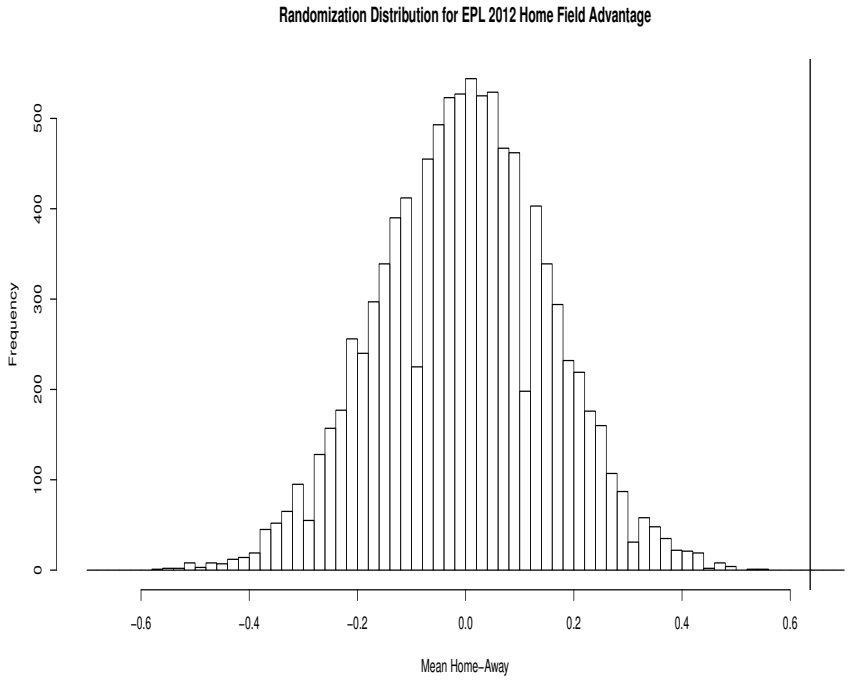

Figure 5.7: Randomization Distribution for Home-Away Mean Differential - EPL 2012

```
hist(bmi.epl, breaks=30, xlim=c(18,32), freq=F,
main="Histogram of EPL BMI and N(23.02,1.71) Density")
lines(bmi.x, dnorm(bmi.x, mu.epl, sigma.epl))

### Take 100000 Independent samples of n1=n2=20 and obtain ybar1-ybar2
num.sim <- 100000
n.nhl <- 20
n.epl <- 20
(mu.meandiff <- mu.nhl - mu.epl)
(sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl))
set.seed(6677)
ybar.s.nhl <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.epl <- matrix(rep(0,2*num.sim),ncol=2)
for (i in 1:num.sim) {
y1 <- sample(bmi.nhl,n.nhl,replace=F)
y2 <- sample(bmi.epl,n.nhl,replace=F)
ybar.s.nhl[i,1] <- mean(y1)
ybar.s.nhl[i,2] <- sd(y1)
ybar.s.epl[i,1] <- mean(y2)
ybar.s.epl[i,2] <- sd(y2)
}
meandiff <- ybar.s.nhl[,1] - ybar.s.epl[,1]
mean.md <- mean(meandiff)
sd.md <- sd(meandiff)

se.meandiff <- sqrt(ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)
mean.var.md <- mean(se.meandiff^2)
sigma.meandiff^2
diff.lo.z <- meandiff + qnorm(.025,0,1) * se.meandiff
diff.hi.z <- meandiff + qnorm(.975,0,1) * se.meandiff
cov.z <- sum(diff.lo.z <= mu.meandiff & diff.hi.z >= mu.meandiff) / num.sim
```

```
diff.lo.t <- meandiff + qt(.025,n.nhl+n.epl-2) * se.meandiff
diff.hi.t <- meandiff + qt(.975,n.nhl+n.epl-2) * se.meandiff
cov.t <- sum(diff.lo.t <= mu.meandiff & diff.hi.t >= mu.meandiff) / num.sim

md.out <- cbind(mu.nhl, mu.epl, sigma.nhl, sigma.epl, n.nhl, mu.nhl-mu.epl,
      sigma.meandiff, mean.md, sd.md, cov.z, cov.t)
colnames(md.out) <- c("mu1","mu2","sigma1","sigma2","n","mu1-mu2",
         "SE{Yb1-Yb2}", "Mean(yb1-yb2)", "SD(yb1-yb2)",
         "cover(z)","cover(t)")
round(md.out, 3)

## Figure 5.2

par(mfrow=c(1,1))
hist(meandiff, breaks=100, xlim=c(min(meandiff)-0.01, max(meandiff)+0.01),
freq=F, main="Histogram of Mean Differences and N(3.48,0.50) Density")
diff.x <- seq(min(meandiff)-0.01, max(meandiff)+0.01,length.out=1000)
lines(diff.x, dnorm(diff.x, mu.meandiff, sigma.meandiff))
## End of Figure 5.2

rm(list=ls(all=TRUE))

### Example 5.3

rp <- read.csv("http://www.stat.ufl.edu/~winner/data/rwanda_physics.csv")
attach(rp); names(rp)
t.test(score ~ trt.y, var.equal=T) # t-test with single y-var and trt id

rm(list=ls(all=TRUE))

### Example 5.4

quilt <- read.csv("http://www.stat.ufl.edu/~winner/data/breast_diep.csv")
attach(quilt); names(quilt)

trt.f <- factor(trt)
levels(trt.f) <- c("Control", "Treatment")

## Figure 5.3

par(mfrow=c(1,1))
plot(totvol ~ trt.f, xlab="Experimental Group",
   main="Side-by-Side Box Plots - Breast Surgery Study")
## End Figure 5.3

t.test(totvol ~ trt.f, var.equal=F)

rm(list=ls(all=TRUE))

### Example 5.5

apphair <- read.table("http://www.stat.ufl.edu/~winner/data/apple_hair.dat",
header=F, col.names=c("hair.trt","total0","total6","totaldiff",
"term0", "term6", "termdiff"))
attach(apphair)
hair.trt.f <- factor(hair.trt, levels=1:2, labels=c("placebo", "PC2"))

## Figure 5.4

plot(totaldiff ~ hair.trt.f)
par(mfrow=c(1,2))
qqnorm(totaldiff[hair.trt==1],main="Control")
qqline(totaldiff[hair.trt==1])
qqnorm(totaldiff[hair.trt==2],main="Treatment")
```

```
qqline(totaldiff[hair.trt==2])

wilcox.test(totaldiff ~ hair.trt)

rm(list=ls(all=TRUE))

### Example 5.6

wine1 <- read.csv("http://www.stat.ufl.edu/~winner/data/wine_isotope.csv")
attach(wine1); names(wine1)
wine.out <- cbind(mean(microwave), sd(microwave), mean(lowtemp), sd(lowtemp),
   cor(microwave,lowtemp))
colnames(wine.out) <- c("ybar1", "s1", "ybar2", "s2", "cor(y1,y2)")
round(wine.out, 6)

## Brute Force Computations
diff <- microwave - lowtemp ## Obtain differences
n.diff <- length(diff) ## Obtain n of diffs
mean.diff <- mean(diff) ## Obtain mean of diffs
sd.diff <- sd(diff) ## Obtain SD of diffs
se.diff <- sd.diff/sqrt(length(diff)) ## Obtain Std Error of mean
t.diff <- mean.diff/se.diff ## t-statistic
pt.diff <- 2*(1-pt(abs(t.diff),n.diff-1))## P-value
t.025 <- qt(.975,n.diff-1) ## Critical t-value
muD.LO <- mean.diff-t.025*se.diff ## Lower Bound CI
muD.HI <- mean.diff+t.025*se.diff ## Upper Bound CI
diff.out <- cbind(mean.diff, sd.diff, se.diff, t.diff, pt.diff, muD.LO,
muD.HI)
colnames(diff.out) <- c("mean","SD","Std Err", "t", "P(>|t|)","LB","UB")
round(diff.out,9)
## t.test Function
t.test(microwave, lowtemp, paired=TRUE)

rm(list=ls(all=TRUE))

### Example 5.7

still <- c(157.5, 84.5, 134, 74, 108, 107.5, 106, 163, 54)
flowing <- c(164.5, 51.5, 250, 139, 113, 124.5, 95.5, 70.5, 30.5)
wilcox.test(still, flowing, paired=TRUE)

rm(list=ls(all=TRUE))

### Example 5.9

bmi.sim <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_nba_ebl_bmi.csv")
attach(bmi.sim); names(bmi.sim)
## Obtain populations and mu and sigma for each
N.nhl <- 717 # # of NHL players
N.epl <- 526 # # of EPL players
bmi.nhl <- NHL_BMI[1:N.nhl]
bmi.epl <- EPL_BMI[1:N.epl]
mu.nhl <- mean(bmi.nhl); sigma.nhl <- sd(bmi.nhl)
mu.epl <- mean(bmi.epl); sigma.epl <- sd(bmi.epl)
## Set up and run samples and ybar and s arrays
num.sim <- 100000
n.nhl <- 10
n.epl <- 10
mu.meandiff <- mu.nhl - mu.epl
sigma.meandiff <- sqrt(sigma.nhl^2/n.nhl + sigma.epl^2/n.epl)

bmisim.out1 <- cbind(mu.nhl, mu.epl, mu.meandiff, sigma.nhl, sigma.epl,
   n.nhl, n.epl, sigma.meandiff)
colnames(bmisim.out1) <- c("mu_nhl", "mu_epl", "mu_nhl-mu_epl",
```

```
   "sigma_nhl", "sigma_epl", "n_nhl", "n_epl", "SE{Ybar_n-Ybar_e)")
round(bmisim.out1, 2)

set.seed(1122)
ybar.s.nhl <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.epl <- matrix(rep(0,2*num.sim),ncol=2)
for (i in 1:num.sim) {
y1 <- sample(bmi.nhl,n.nhl,replace=F)
y2 <- sample(bmi.epl,n.nhl,replace=F)
ybar.s.nhl[i,1] <- mean(y1)
ybar.s.nhl[i,2] <- sd(y1)
ybar.s.epl[i,1] <- mean(y2)
ybar.s.epl[i,2] <- sd(y2)
}
## End of sampling
## Generate sample mean differences SEs and CIs
## ev=equal variances, uv=unequal variances
meandiff <- ybar.s.nhl[,1] - ybar.s.epl[,1]
se.meandiff <- sqrt(ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)
df.uv1 <- (ybar.s.nhl[,2]^2/n.nhl + ybar.s.epl[,2]^2/n.epl)^2
df.uv2 <- ((ybar.s.nhl[,2]^2/n.nhl)^2/(n.nhl-1)) +
((ybar.s.epl[,2]^2/n.epl)^2/(n.epl-1))
df.uv <- df.uv1 / df.uv2
df.ev <- n.nhl + n.epl - 2
meandiff.LB.ev <- meandiff + qt(.025,df.ev) * se.meandiff
meandiff.UB.ev <- meandiff + qt(.975,df.ev) * se.meandiff
meandiff.LB.uv <- meandiff + qt(.025,df.uv) * se.meandiff
meandiff.UB.uv <- meandiff + qt(.975,df.uv) * se.meandiff
## Obtain Coverage rates, widths, power (H0:mu1-mu2=0)
cov.ev <- sum(meandiff.LB.ev <= mu.meandiff &
              meandiff.UB.ev >= mu.meandiff) / num.sim
cov.uv <- sum(meandiff.LB.uv <= mu.meandiff &
              meandiff.UB.uv >= mu.meandiff) / num.sim
width.ev <- mean(meandiff.UB.ev-meandiff.LB.ev)
width.uv <- mean(meandiff.UB.uv-meandiff.LB.uv)
ci.ev.gt0 <- sum(meandiff.LB.ev >= 0) / num.sim
ci.uv.gt0 <- sum(meandiff.LB.uv >= 0) / num.sim
ci.ev.lt0 <- sum(meandiff.UB.ev <= 0) / num.sim
ci.uv.lt0 <- sum(meandiff.UB.uv <= 0) / num.sim

bmisim.out2 <- cbind(cov.ev, cov.uv, width.ev, width.uv, ci.ev.gt0, ci.uv.gt0)
colnames(bmisim.out2) <- c("EV Cover", "UV Cover", "EV Width",
    "UV Width", "EV Diff > 0", "UV Diff > 0")
round(bmisim.out2, 4)

rm(list=ls(all=TRUE))

### Example 5.10

## Read data from website and attach data frame and obtain variable names
rr.mar <- read.csv(
"http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)
f.mph <- mph[Gender=="F"]
m.mph <- mph[Gender=="M"]
mu.f <- mean(f.mph); sigma.f <- sd(f.mph)
mu.m <- mean(m.mph); sigma.m <- sd(m.mph)
num.sim <- 100000
n.f <- 6; n.m <- 6
mu.meandiff <- mu.f - mu.m
sigma.meandiff <- sqrt(sigma.f^2/n.f + sigma.m^2/n.m)

rrsim.out1 <- cbind(mu.f, mu.m, mu.meandiff, sigma.f, sigma.m,
   n.f, n.m, sigma.meandiff)
```

```
colnames(rrsim.out1) <- c("mu_f", "mu_m", "mu_f-mu_m",
   "sigma_f", "sigma_m", "n_f", "n_m", "SE{Ybar_f-Ybar_m)")
round(rrsim.out1, 2)

set.seed(3344)
ybar.s.f <- matrix(rep(0,2*num.sim),ncol=2)
ybar.s.m <- matrix(rep(0,2*num.sim),ncol=2)
ranksum.fm <- matrix(rep(0,2*num.sim),ncol=2)
for (i in 1:num.sim) {
y1 <- sample(f.mph,n.f,replace=F)
y2 <- sample(m.mph,n.m,replace=F)
ybar.s.f[i,1] <- mean(y1)
ybar.s.f[i,2] <- sd(y1)
ybar.s.m[i,1] <- mean(y2)
ybar.s.m[i,2] <- sd(y2)
ranksum.fm [i,1] <- sum(rank(c(y1,y2))[1:n.f])
ranksum.fm [i,2] <- sum(rank(c(y1,y2))[(n.f+1):(n.f+n.m)])
}

meandiff <- ybar.s.f[,1] - ybar.s.m[,1]
se.meandiff <- sqrt(ybar.s.f[,2]^2/n.f + ybar.s.m[,2]^2/n.m)
df.uv1 <- (ybar.s.f[,2]^2/n.f + ybar.s.m[,2]^2/n.m)^2
df.uv2 <- ((ybar.s.f[,2]^2/n.f)^2/(n.f-1)) +
((ybar.s.m[,2]^2/n.m)^2/(n.m-1))
df.uv <- df.uv1 / df.uv2
df.ev <- n.f + n.m - 2
meandiff.LB.ev <- meandiff + qt(.025,df.ev) * se.meandiff
meandiff.UB.ev <- meandiff + qt(.975,df.ev) * se.meandiff
meandiff.LB.uv <- meandiff + qt(.025,df.uv) * se.meandiff
meandiff.UB.uv <- meandiff + qt(.975,df.uv) * se.meandiff
## Obtain Coverage rates, widths, power (H0:mu1-mu2=0 HA:mu1-mu2<0)
cov.ev <- sum(meandiff.LB.ev <= mu.meandiff &
              meandiff.UB.ev >= mu.meandiff) / num.sim
cov.uv <- sum(meandiff.LB.uv <= mu.meandiff &
              meandiff.UB.uv >= mu.meandiff) / num.sim
width.ev <- mean(meandiff.UB.ev-meandiff.LB.ev)
width.uv <- mean(meandiff.UB.uv-meandiff.LB.uv)
t.uv.ev <- meandiff / se.meandiff
rr.t.uv <- qt(.05,df.uv)
rr.t.ev <- qt(.05,df.ev)
rr.t1.w <- 28 ## From Wilcoxon Rank-sum w/ n1=n2=6
reject.ev <- sum(t.uv.ev <= rr.t.ev) / num.sim
reject.uv <- sum(t.uv.ev <= rr.t.uv) / num.sim
reject.wrs <- sum(ranksum.fm[,1] <= rr.t1.w) / num.sim

rrsim.out2 <- cbind(cov.ev, cov.uv, reject.ev, reject.uv, reject.wrs)
colnames(rrsim.out2) <- c("EV Cover", "UV Cover", "EV Reject",
               "UV Reject", "Rank-Sum Reject")
round(rrsim.out2, 4)

rm(list=ls(all=TRUE))

### Example 5.11

## Set parameters, alpha, chosen power, for starting sample size (n0)
m1_m2_A <- -0.50
sigma <- 0.94
n0 <- 6
df <- 2 * (n0-1)
alpha <- 0.05
power.star <- 0.80
delta <- m1_m2_A / sigma
Delta <- delta / sqrt(2/n0)
crit_val <- qt(.05, df)
```

```
power.lt <- pt(crit_val, df, Delta)

power.out1 <- cbind(alpha, power.star, n0, df, m1_m2_A, sigma, delta, Delta,
    crit_val, power.lt)
colnames(power.out1) <- c("alpha", "pi*", "n", "df", "(mu1-mu2)_A", "sigma",
    "delta", "Delta", "-t(.05,df)", "power")
round(power.out1,3)

## Set up holders for power and sample size and row and sample size start values
power.out <- numeric()
n.out <- numeric()
i <- 0
n <- n0
## Loop until power exceeds chosen power
while (power.lt < power.star) {
i <- i+1
n <- n+1
crit_val <- qt(alpha,2*(n-1))
power.lt <- pt(crit_val,2*(n-1),delta/sqrt(2/n))
power.out[i] <- power.lt
n.out[i] <- n
}
## Print Sample sizes and corresponding powers
cbind(n.out, power.out)

####### 2-Tailed Test
## Set parameters, alpha, chosen power, for starting sample size (n0)
m1_m2_A <- -0.50
sigma <- 0.94
n0 <- 6
df <- 2 * (n0-1)
alpha <- 0.05
power.star <- 0.80
delta <- m1_m2_A / sigma
Delta <- delta / sqrt(2/n0)
crit_val_lo <- qt(.05/2, df)
crit_val_hi <- qt(1-.05/2, df)
power.2t <- pt(crit_val_lo, df, Delta) +
  1-pt(crit_val_hi, df, Delta)

power.out2 <- cbind(alpha, power.star, n0, df, m1_m2_A, sigma, delta, Delta,
    crit_val_hi, power.2t)
colnames(power.out2) <- c("alpha", "pi*", "n", "df", "(mu1-mu2)_A", "sigma",
    "delta", "Delta", "t(.025,df)", "power")
round(power.out2,3)

rm(list=ls(all=TRUE))

### Example 5.12

mu_DA <- 30
sigma_D <- 60
n0 <- 9
df <- n0 - 1
alpha <- .05
power.star <- 0.75
delta <- mu_DA / sigma_D
Delta <- sqrt(n0) * delta
crit_val_lo <- qt(alpha/2, df)
crit_val_hi <- qt(1-alpha/2, df)
power.2t <- pt(crit_val_lo, df, Delta) +
   (1-pt(crit_val_hi, df, Delta))

power.out3 <- cbind(alpha, power.star, n0, df, m1_m2_A, sigma, delta, Delta,
```

```
    crit_val_hi, power.2t)
colnames(power.out3) <- c("alpha", "pi*", "n", "df", "(mu1-mu2)_A", "sigma",
    "delta", "Delta", "t(.025,df)", "power")
round(power.out3,3)

power.out <- numeric()
n.out <- numeric()
i <- 0
n <- n0
## Loop until power exceeds chosen power
while (power.2t < power.star) {
i <- i+1
n <- n+1
crit_val_lo <- qt(.05/2,n-1)
crit_val_hi <- qt(1-.05/2,n-1)
power.2t <- pt(crit_val_lo,n-1,sqrt(n)*delta) +
(1-pt(crit_val_hi,n-1,sqrt(n)*delta))
power.out[i] <- power.2t
n.out[i] <- n
}
## Print Sample sizes and corresponding powers
cbind(n.out, power.out)

rm(list=ls(all=TRUE))

### Example 5.13

## Part 1
kanet <- read.fwf("http://www.stat.ufl.edu/~winner/data/kanet.dat",
width=c(18,2,rep(8,7)), col.names=c("name","kgroup","age","stature",
"armspan","sitheight","knlheight","cubit","leftfoot"))
attach(kanet)

cubit
cubit.mean <- as.vector(tapply(cubit,kgroup,mean))
cubit.sd <- as.vector(tapply(cubit,kgroup,sd))
L.cubit <- cubit[kgroup==1]
K.cubit <- cubit[kgroup==2]
set.seed(97531)
num.boot <- 10000
boot.ybar1 <- rep(0,num.boot)
boot.ybar2 <- rep(0,num.boot)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
for (i in 1:num.boot) {
y1 <- sample(L.cubit, n.L, replace=T)
y2 <- sample(K.cubit, n.K, replace=T)
boot.ybar1[i] <- mean(y1); boot.ybar2[i] <- mean(y2)
}
meandiff <- boot.ybar1-boot.ybar2

boot.out1 <- cbind(cubit.mean[1], cubit.mean[2], cubit.mean[1]-cubit.mean[2],
   cubit.sd[1], cubit.sd[2], mean(meandiff), sd(meandiff), quantile(meandiff,.025),
   quantile(meandiff,.975))
colnames(boot.out1) <- c("ybar_L", "ybar_K", "yb_L-yb_K", "s_L", "s_K",
     "Mean(MeanDiff)", "SD(MD)", "Q.025(MD)", "Q.975(MD)")
round(boot.out1, 3)

## Figure 5.5

par(mfrow=c(1,2))
hist(meandiff,breaks=30)
abline(v=(mean(L.cubit)-mean(K.cubit)),lwd=2)
qqnorm(meandiff); qqline(meandiff)
```

```
## Part 2
## Bootstrap t CIs - Chihara and Hesterberg, Sec.7.5, p.198-200
set.seed(97531)
num.boot <- 10000
boot.ybar.s.L <- matrix(rep(0,2*num.boot),ncol=2)
boot.ybar.s.K <- matrix(rep(0,2*num.boot),ncol=2)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
mean.L <- mean(L.cubit)
mean.K <- mean(K.cubit)
sd.L <- sd(L.cubit)
sd.K <- sd(K.cubit)
SE.diff <- sqrt((sd.L^2/n.L) + (sd.K^2/n.K))

for (i in 1:num.boot) {
y1 <- sample(L.cubit, n.L, replace=T)
y2 <- sample(K.cubit, n.K, replace=T)
boot.ybar.s.L[i,1] <- mean(y1); boot.ybar.s.K[i,1] <- mean(y2)
boot.ybar.s.L[i,2] <- sd(y1); boot.ybar.s.K[i,2] <- sd(y2)
}
t.star <- ((boot.ybar.s.L[,1]-boot.ybar.s.K[,1])-(mean.L-mean.K)) /
   sqrt((boot.ybar.s.L[,2]^2/n.L)+(boot.ybar.s.K[,2]^2/n.L))
Q_L <- quantile(t.star, 0.025)
Q_U <- quantile(t.star, 0.975)
boot.LB <- (mean.L - mean.K) - Q_U * SE.diff
boot.UB <- (mean.L - mean.K) - Q_L * SE.diff

boot.out2 <- cbind(mean.L, mean.K, mean.L - mean.K, sd.L, sd.K, n.L, n.K, SE.diff,
   Q_L, Q_U, boot.LB, boot.UB)
colnames(boot.out2) <- c("ybar_L", "ybar_K", "yb_L-yb_K", "s_L", "s_K", "n_L", "n_K",
   "SE{diff}", "Q_L", "Q_U", "LB", "UB")
round(boot.out2, 3)

rm(list=ls(all=TRUE))

### Example 5.14

kanet <- read.fwf("http://www.stat.ufl.edu/~winner/data/kanet.dat",
width=c(18,2,rep(8,7)), col.names=c("name","kgroup","age","stature",
"armspan","sitheight","knlheight","cubit","leftfoot"))
attach(kanet)
L.cubit <- cubit[kgroup==1]
K.cubit <- cubit[kgroup==2]
TS.obs <- mean(L.cubit) - mean(K.cubit)

## Set up and obtain Permutation Samples
set.seed(24680)
num.perm <- 9999
TS <- rep(0,num.perm)
n.L <- length(L.cubit)
n.K <- length(K.cubit)
n.LK <- n.L + n.K
for (i in 1:num.perm) {
perm <- sample(1:n.LK,n.LK,replace=F) # Permutation of 1:90
ybar1 <- mean(cubit[perm[1:n.L]]) # First 30 assigned L
ybar2 <- mean(cubit[perm[(n.L+1):n.LK]]) # Last 60 assigned K
TS[i] <- ybar1 - ybar2
}
## Count # permutations where |TS| >= |TS.obs| and obtain 2-tail P-value
num.exceed <- sum(abs(TS) >= abs(TS.obs))
p.val.2tail <- (num.exceed+1) / (num.perm+1)

perm.out1 <- cbind(mean(L.cubit), mean(K.cubit), TS.obs, num.exceed, p.val.2tail)
```

```
colnames(perm.out1) <- c("ybar_L", "ybar_K", "Test Stat", "Extreme Perms", "P-value")
round(perm.out1, 4)

## Figure 5.6

par(mfrow=c(1,1))
hist(TS,breaks=30, xlab="MeanL - MeanK",
main="Randomization Distribution for Cubit Length")
abline(v=TS.obs,lwd=2)

rm(list=ls(all=TRUE))

### Example 5.15

epl2012 <- read.csv("http://www.stat.ufl.edu/~winner/data/epl_2012_home_perm.csv",
header=T)
attach(epl2012); names(epl2012)

### Obtain Sample Size and Test Statistic (Average of d.jk)
n <- length(d.jk)
TS.obs <- mean(d.jk)

### Choose the number of samples and initialize TS, and set seed
N <- 9999; TS <- rep(0,N); set.seed(86420)

### Loop through samples and compute each TS
for (i in 1:N) {
ds.jk <- d.jk # Initialize d*.jk = d.jk
u <- runif(n)-0.5 # Generate n U(-0.5,0.5)s
u.s <- sign(u) # -1 if u.s < 0, +1 if u.s > 0
ds.jk <- u.s * ds.jk
TS[i] <- mean(ds.jk) # Compute Test Statistic for this sample
}
summary(TS)
num.exceed1 <- sum(TS >= TS.obs) # Count for 1-sided (Upper Tail) P-value
num.exceed2 <- sum(abs(TS) >= abs(TS.obs)) # Count for 2-sided P-value
p.val.1sided <- (num.exceed1 + 1)/(N+1) # 1-sided p-value
p.val.2sided <- (num.exceed2 + 1)/(N+1) # 2-sided p-value

perm.out2 <- cbind(n, TS.obs, num.exceed1, num.exceed2, p.val.1sided, p.val.2sided)
colnames(perm.out2) <- c("n", "Observed TS", "# Exceed 1-tail", "# Exceed 2-tail",
    "1-tailed P-value", "2-tailed P-value")
round(perm.out2, 4)


### Draw histogram of distribution of TS, with vertical line at TS.obs
##Figure 5.7

hist(TS,breaks=seq(-.7,.7,.02), xlab="Mean Home-Away",
main="Randomization Distribution for EPL 2012 Home Field Advantage")
abline(v=TS.obs,lwd=2)

rm(list=ls(all=TRUE))
```

Chapter 6

Estimating and Testing Variances

When making inferences regarding means, even if the data themselves are not normal, the sampling distribution of $\overline{Y}$ is approximately normal for reasonably large samples. When making inferences regarding variances, the normality assumption is more stringent, and if data are not normally distributed, robust methods are used (see Levene's test below). First, we consider estimation and testing a single variance, then comparing two variances, and finally comparing $k \geq 2$ variances.

6.1 Estimation and Testing for a Single Variance

When the data are normal (and independent), then a multiple of the sample variance follows a Chi-square distribution with $n-1$ degrees of freedom. That is, we have the following results.

$$Y_1, ..., Y_n \sim N(\mu, \sigma) \quad \Rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \quad \Rightarrow \quad P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = 1-\alpha$$

This leads to a rule for a $(1-\alpha)100\%$ Confidence Interval for σ^2 (and thus for σ), as well as a test of whether σ^2 is equal to some null value σ^2_0. Consider a Confidence Interval, then a test, where S^2 is a random variable (sample variance), and s^2 is a particular value from an observed sample.

$$1-\alpha = P\left(\chi^2_{1-\alpha/2,n-1} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{\alpha/2,n-1}\right) = P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} \geq \sigma^2 \geq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}\right)$$

$$\Rightarrow \quad (1-\alpha)100\% \text{ Confidence Interval for } \sigma^2\text{: } \left[\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}\right]$$

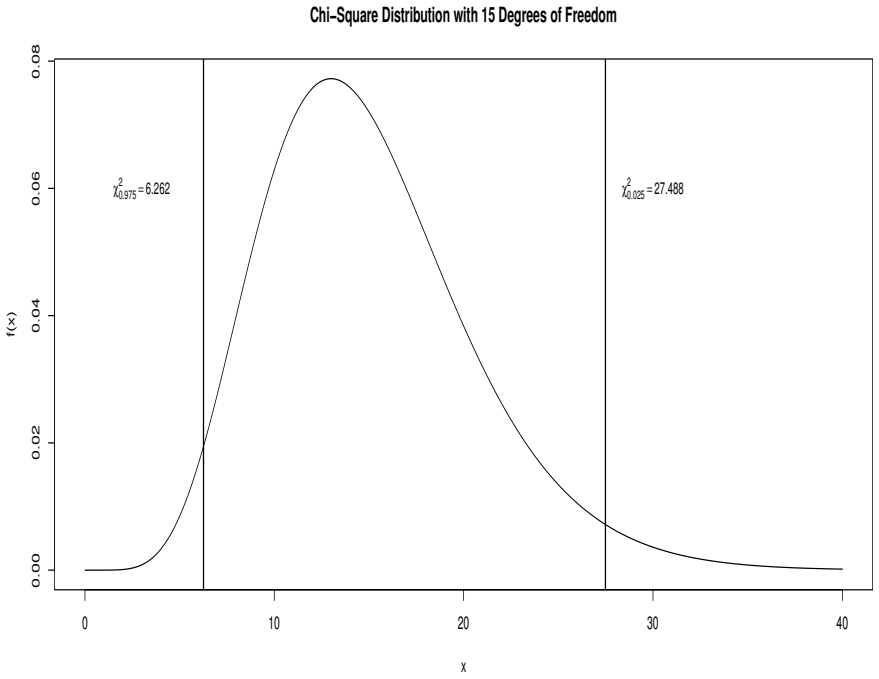

Figure 6.1: Chi-Square Distribution with 15 degrees of freedom

$$\text{2-Tailed test: } H_0 : \sigma^2 = \sigma_0^2 \qquad H_A : \sigma^2 \neq \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : \left\{X^2_{obs} \leq \chi^2_{1-\alpha/2,n-1}\right\} \cup \left\{X^2_{obs} \geq \chi^2_{\alpha/2,n-1}\right\} \quad P = 2\min\left[P\left(\chi^2_{n-1} \leq X^2_{obs}\right), P\left(\chi^2_{n-1} \geq X^2_{obs}\right)\right]$$

$$\text{Upper-Tail test: } H_0 : \sigma^2 \leq \sigma_0^2 \qquad H_A : \sigma^2 > \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : X^2_{obs} \geq \chi^2_{\alpha,n-1} \qquad P = P\left(\chi^2_{n-1} \geq X^2_{obs}\right)$$

$$\text{Lower-Tail test: } H_0 : \sigma^2 \geq \sigma_0^2 \qquad H_A : \sigma^2 < \sigma_0^2 \qquad TS : X^2_{obs} = \frac{(n-1)s^2}{\sigma_0^2}$$

$$RR : X^2_{obs} \leq \chi^2_{1-\alpha,n-1} \qquad P = P\left(\chi^2_{n-1} \leq X^2_{obs}\right)$$

Clearly, as in the case of a single mean, most practical situations will involve estimation rather than testing unless there is some focal null value σ_0^2 of interest. A plot of the Chi-Square distribution with 15 degrees of freedom along with $\chi^2_{.975,15} = 6.262$ and $\chi^2_{.025,15} = 27.488$ is given in Figure 6.1.

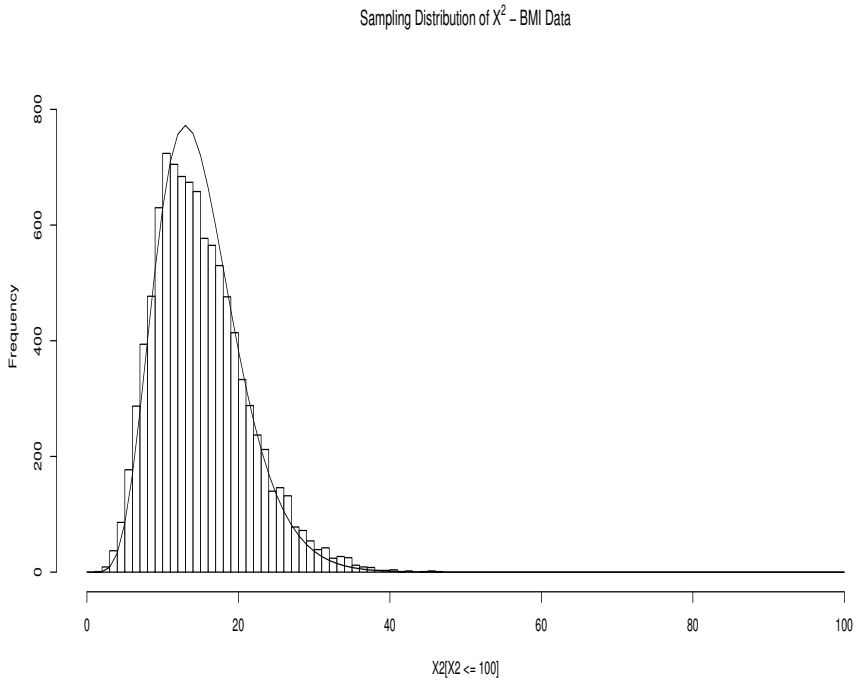

Figure 6.2: Histogram of Scaled Variance and Chi-Square(15) Density - NHL BMI data

Example 6.1: NHL Body Mass Indices

For the NHL body mass index measurements, the population mean and variance are 26.500 and $1.454^2 = 2.114$, respectively. Then 10000 random samples of size $n = 16$ are obtained and s^2 is computed for each sample. The 95% Confidence Interval for σ^2 is calculated for each sample. A histogram of the quantity $(n-1)s^2/\sigma^2$ is given in Figure 6.2. There are fewer values under the peak and more in the tails than would be expected if BMI's were exactly normally distributed. The first sample of the 10000 samples yielded a sample variance of $s^2 = 1.916$. This leads to the 95% Confidence Interval computed below.

$$n = 16 \quad s^2 = 1.916 \quad \chi^2_{.975,15} = 6.262 \quad \chi^2_{.025,15} = 27.488 \quad \left[\frac{15(1.916)}{27.488}, \frac{15(1.916)}{6.262}\right] \equiv [1.046, 4.590]$$

Despite the not so wonderful Chi-Square approximation in Figure 6.2, the coverage rate of the 10000 Confidence Intervals is 92.6%, not so far from the nominal 95%.

R Output

```
## Output

> cbind(stddev[1], stddev[1]^2)
         [,1]     [,2]
[1,] 1.384297 1.916278
```

```
> round(var.out1, 3)
     sigma^2  n df Mean(X2) Var(X2) chisq(.025,15) chisq(.975,15) cover prob
[1,]   2.113 16 15   15.026  36.737          6.262         27.488      0.926

> round(q.out, 3)
             10%    25%     50    75%    90%
Theoretical 6.262 11.037 14.339 18.245 27.488
Empirical   5.681 10.555 14.174 18.553 29.159
```

∇

6.2 Comparing Two Variances

In this section, we consider two tests. The first, based on data being normally distributed is the F**-test**. The second, which does not assume normality is a **Jacknife test**. There are other tests that extend to more than two groups, considered in the next section, that can also be used to compare two groups.

6.2.1 F-Test

When there are two populations, and there are independent samples, inference is made regarding the ratio σ_1^2/σ_2^2. When the populations of measurements are normally distributed, the following results are obtained.

$$Y_{i1},\ldots,Y_{in_i} \sim N(\mu_i,\sigma_i) \quad i=1,2 \qquad W_i = \frac{(n_i-1)S_i^2}{\sigma_i^2} \sim \chi^2_{n_i-1} \qquad \frac{W_1/(n_1-1)}{W_2/(n_2-1)} = \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$$

This leads to the following probability statements and a Confidence Interval and test for σ_1^2/σ_2^2.

$$1-\alpha = P\left(F_{1-\alpha/2,n_1-1,n_2-1} \leq \frac{S_1^2/S_2^2}{\sigma_1^2/\sigma_2^2} \leq F_{\alpha/2,n_1-1,n_2-1}\right) = P\left(\frac{S_1^2/S_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}} \geq \frac{\sigma_1^2}{\sigma_2^2} \geq \frac{S_1^2/S_2^2}{F_{\alpha/2,n_1-1,n_2-1}}\right)$$

Once samples are taken and s_1 and s_2 are calculated, a Confidence Interval and a test of whether $\sigma_1^2=\sigma_2^2$ can be obtained.

$$(1-\alpha)100\%\text{ CI for }\frac{\sigma_1^2}{\sigma_2^2}: \left[\frac{s_1^2/s_2^2}{F_{\alpha/2,n_1-1,n_2-1}}, \frac{s_1^2/s_2^2}{F_{1-\alpha/2,n_1-1,n_2-1}}\right] \qquad F_{1-\alpha/2,n_1-1,n_2-1} = \frac{1}{F_{\alpha/2,n_2-1,n_1-1}}$$

$$\text{2-Tailed test: } H_0:\sigma_1^2=\sigma_2^2 \qquad H_A:\sigma_1^2\neq\sigma_2^2 \qquad TS: F_{obs}=\frac{s_1^2}{s_2^2}$$

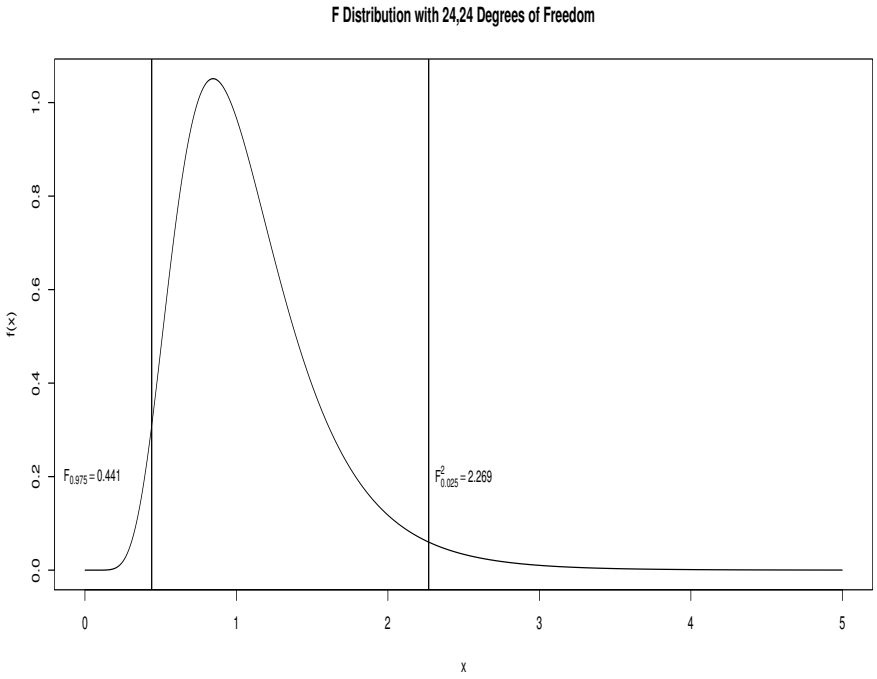

Figure 6.3: F-distribution with $\nu_1 = \nu_2 = 24$

$$RR: \left\{F_{obs} \leq F_{1-\alpha/2,n_1-1,n_2-1}\right\} \cup \left\{F_{obs} \geq F_{\alpha/2,n_1-1,n_2-1}\right\} \quad P = 2\min\left[P\left(F_{n_1-1,n_2-1} \leq F_{obs}\right), P\left(F_{n_1-1,n_2-1} \geq F_{obs}\right)\right]$$

$$\text{Upper Tailed test: } H_0: \sigma_1^2 \leq \sigma_2^2 \qquad H_A: \sigma_1^2 > \sigma_2^2 \qquad TS: F_{obs} = \frac{s_1^2}{s_2^2}$$

$$RR: \{F_{obs} \geq F_{\alpha,n_1-1,n_2-1}\} \qquad P = P\left(F_{n_1-1,n_2-1} \geq F_{obs}\right)$$

$$\text{Lower Tailed test: } H_0: \sigma_1^2 \geq \sigma_2^2 \qquad H_A: \sigma_1^2 < \sigma_2^2 \qquad TS: F_{obs} = \frac{s_1^2}{s_2^2}$$

$$RR: \{F_{obs} \leq F_{1-\alpha,n_1-1,n_2-1}\} \qquad P = P\left(F_{n_1-1,n_2-1} \leq F_{obs}\right)$$

A plot of the F-distribution with 24 numerator and 24 denominator degrees of freedom, along with $F_{.975,24,24} = 0.441$ and $F_{.025,24,24} = 2.269$ is given in Figure 6.3.

Example 6.2: Female and Male Rock and Roll Marathon Speeds

Although the distributions are right-skewed, we will construct 10000 95% Confidence Intervals for σ_f^2/σ_m^2, and tests of $H_0: \sigma_f^2 = \sigma_m^2$. As was seen previously, $\sigma_f = 0.831$, $\sigma_m = 1.058$, and thus $\sigma_f^2/\sigma_m^2 = 0.617$. Despite the non-normality of the distributions, the 95% Confidence Intervals covered 0.617 in 94.87% of the

samples. The F-test rejected the null hypothesis in 21.96% of the samples (less than $F_{.975}$ in 21.85%, and greater than $F_{.025}$ in in 0.11%). In the first sample, we have the following results.

$$n_f = n_m = 25 \quad F_{.975,24,24} = 0.441 \quad F_{.025,24,24} = 2.269 \quad s_f = 0.748 \quad s_m = 1.277 \quad \frac{s_f^2}{s_m^2} = 0.343$$

$$95\% \text{ Confidence Interval: } \left[\frac{0.343}{2.269}, \frac{0.343}{0.441}\right] \equiv [0.151, 0.778] \qquad TS : F_{obs} = 0.343$$

For the first of the 10000 samples, the Confidence Interval contains the true variance ratio, and the test rejects the null hypothesis that the variances are equal ($F_{obs} = 0.343 < F_{.975,25,24} = 0.441$). The quantiles given below are the empirical quantiles of $\left(s_f^2/s_m^2\right)/0.617$, which match up very well with the theoretical quantiles of the $F_{24,24}$ distribution.

R Output

```
## Output

> round(var.out2, 4)
     sigma_F^2 sigma_M^2 Var Ratio  Cover Rej Lo Rej Hi Rej Tot
[1,]    0.6906    1.1187    0.6173 0.9487 0.2185 0.0011  0.2196

> round(q.out, 3)
             10%   25%    50   75%   90%
Theoretical 0.441 0.757 1.000 1.321 2.269
Empirical   0.432 0.746 0.997 1.340 2.245
```

∇

Example 6.3: Physical Properties of Rocks from 3 Locations in Iran

A study compared Anhydrite rock properties at 3 locations in Iran (Mehrgini, et al (2016), [38]). There were 8 samples at each site. Data have been generated to preserve the means, standard deviations, minimums, and maximums for the locations and are given in Table 6.1.

Obtain a 95% Confidence Interval for the variance ratio (Ghotvand/Chamsir) and test whether their population variances are equal ($s_g = 45.15, s_c = 30.25, n_g = n_c = 8$).

$$\frac{s_g^2}{s_c^2} = \frac{45.15^2}{30.25^2} = 2.228 \quad F_{.975,7,7} = 0.200 \quad F_{.025,7,7} = 4.995$$

$$95\% \text{ CI for } \frac{\sigma_g^2}{\sigma_c^2}\text{: } \left[\frac{2.228}{4.995}, \frac{2.228}{0.200}\right] \equiv [0.446, 11.140] \qquad TS : F_{obs} = 2.228 \quad RR : \{F_{obs} \leq 0.200\} \cup \{F_{obs} \geq 4.995\}$$

Ghotvand	Chamshir	Khersan
2800.0	2851.0	2790.0
2803.6	2879.8	2796.4
2811.0	2854.2	2817.0
2825.6	2862.1	2853.6
2847.7	2861.1	2804.4
2857.4	2918.4	2832.0
2906.7	2873.4	2854.5
2916.0	2932.0	2860.0

Table 6.1: Density of Rock Samples (kg/m^3) from 3 Locations in Iran

There is insufficient evidence to conclude that the population variances differ for these two locations (these are very small samples). The R commands and output are given below.

R Commands and Output

```
## Commands

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)

## Using var.test directly on y.g and y.c
var.test(y.g, y.c)

## Output

> round(F.out,3)
     F-stat F(.975) F(.025) P-value Lower  Upper
[1,]  2.231     0.2   4.995   0.312 0.447 11.142
>
> var.test(y.g, y.c)

        F test to compare two variances

data:  y.g and y.c
F = 2.2306, num df = 7, denom df = 7, p-value = 0.3118
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.4465729 11.1416011
sample estimates:
ratio of variances
          2.230591
```

∇

6.2.2 Jacknife Test

Various methods exist for testing equal variances when data are nonnormal. In particular **Levene's Test** is widely used and built in to many (if not all) statistical packages. Levene's test will be described in the

next section on comparing 2 or more groups. A **Jacknife** based test is a nonparametric test for dispersion that does not assume equal medians (see e.g. Hollander and Wolfe (1999) [27], Section 5.2). The algorithm works as follows.

1. Compute s_1^2 and s_2^2 based on the full samples.
2. Drop observations 1-at-a-time and compute $\overline{y}_{1(j)}, s^2_{1(j)}, \overline{y}_{2(j)}, s^2_{2(j)}$ for $i = 1, 2$ $j = 1, \ldots, n_i$ where the (j) represents that the j^{th} observation was removed.
3. Compute $\ln\left(s_i^2\right)$ and $\ln\left(s^2_{i(j)}\right)$ for each observation j and sample i.
4. Compute $D_{i(j)} = n_i \ln\left(s_i^2\right) - (n_i - 1)\ln\left(s^2_{i(j)}\right) \quad i = 1, 2; j = 1, \ldots, n_i$
5. Compute $\overline{D}_i$ and $S^2_{Di} = \sum_{j=1}^{n_i}\left(D_{i(j)} - \overline{D}_i\right)^2/(n_i(n_i-1))$
6. Compute $z_D = \left(\overline{D}_1 - \overline{D}_2\right)/\sqrt{S^2_{D1} + S^2_{D2}}$
7. Compare z_D with the critical values of the standard normal distribution (or when n_1 and n_2 small, the $t_{n_1+n_2-2}$ distribution).

Example 6.4: Physical Properties of Rocks from 3 Locations in Iran

We apply the jacknife method to the rock density data from Example 6.3. The z_D statistic is 1.021 with P-values based on Z of .3074 and based on t_{14} of .3247. There is no evidence of a difference in population variances for the two locations. The P-values are very similar to that from the F-test.

R Output

```
## Output

> round(cbind(var1.jack, var2.jack),2)
     var1.jack var2.jack
[1,]   2145.38    949.25
[2,]   2353.96    770.72
[3,]   1676.90   1060.43
[4,]   2378.16   1005.37
[5,]   2299.44   1012.00
[6,]   2036.28   1066.28
[7,]   1445.38    531.36
[8,]   1975.66    917.07

> round(jk.out, 4)
     Dbar_1 Dbar_2  S2_D1 S2_D2      z   P(z)   P(t)
[1,] 7.7107 6.9687 0.1904 0.338 1.0207 0.3074 0.3247
```

∇

6.3 Comparing $k \geq 2$ Variances

Methods for comparing 2 or more variances include **Bartlett's Test**, **Hartley's** F_{max} **Test**, **Levene's Test**, and an extension of the **Jacknife Test**. A comparison of various testing procedures is given in Lim and Loh (1996) [35]. The first two tests are theoretically based on normal distributions, while Levene's and the Jacknife tests are robust to nonnormal distributions. Hartley's test requires equal sample sizes. Bartlett's test is very general in terms of applicability to different modeling situations.

6.3.1 Bartlett's Test

There are k estimated variances $s_1^2, \ldots, s_k^2$, and associated with the i^{th} variance estimate is a degrees of freedom ν_i. Obtain a pooled estimate of the common variance under $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ and conduct the test as follows for this particular case of comparing k population variances.

$$\nu_i = n_i - 1 \quad \nu_\cdot = \sum_{i=1}^{k} \nu_i \qquad s^2 = \frac{\sum_{i=1}^{k} \nu_i s_i^2}{\nu_\cdot} \qquad C = 1 + \frac{1}{3(k-1)} \left[\left(\sum_{i=1}^{k} \frac{1}{\nu_i} \right) - \left(\frac{1}{\nu_\cdot} \right) \right]$$

$$TS : X_B^2 = \frac{1}{C} \left[\nu_\cdot \ln\left(s^2\right) - \sum_{i=1}^{k} \nu_i \ln\left(s_i^2\right) \right] \qquad RR : X_B^2 \geq \chi_{\alpha,k-1}^2 \qquad P = P\left(\chi_{k-1}^2 \geq X_B^2\right)$$

Bartlett's test can be used in many different applications for comparing variances, for instance it can be used to test whether variances differ when linear regression models are being fit separately for different groups. The theoretical foundation of the method is that the distributions are normal.

Example 6.5: Physical Properties of Rocks from 3 Locations in Iran

We extend the comparison of two regions to comparing all three regions. We find that there is no evidence to reject $H_0 : \sigma_g^2 = \sigma_c^2 = \sigma_k^2$, with a P-value of .398.

$$n_g = n_c = n_k = 8 \quad \nu_g = \nu_c = \nu_k = 7 \quad \nu_\cdot = 21 \qquad s_g^2 = 2038.894 \quad s_c^2 = 914.060 \quad s_k^2 = 783.676$$

$$s^2 = \frac{7(2038.894) + 7(914.060) + 7(783.676)}{21} = 1245.543 \qquad C = 1 + \frac{1}{3(3-1)} \left[\left(\frac{1}{7} + \frac{1}{7} + \frac{1}{7} \right) - \frac{1}{21} \right] = 1.0653$$

$$21 \ln(1245.543) - (7 \ln(2038.894) + 7 \ln(914.060) + 7 \ln(783.676)) = 1.9595 \qquad TS : X_B^2 = \left(\frac{1}{1.0653} \right) (1.9595) = 1.843$$

$$RR : X_B^2 \geq \chi_{.05,3-1}^2 = 5.991 \qquad P = P\left(\chi_2^2 \geq 1.843\right) = .398$$

R Commands and Output

```
## Commands

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)
n.g <- length(y.g); n.c <- length(y.c); n.k <- length(y.k)
## Using bartlett.test directly on y.g, y.c, and y.k
## Combine y.g, y.c, y.k into a single variable
y <- c(y.g, y.c, y.k)
## Create a variable that contains the locations of elements
loc.y <- c(rep(1,n.g),rep(2,n.c),rep(3,n.k))
bartlett.test(y ~ loc.y)

## Output

> bartlett.test(y ~ loc.y)

        Bartlett test of homogeneity of variances

data:  y by loc.y
Bartlett's K-squared = 1.8425, df = 2, p-value = 0.398
```

∇

6.3.2 Hartley's F_{max} Test

This test is very easy to implement, but is based on normally distributed data, equal sample sizes, and requires a special table. The table is available on the class website. Critical values are obtained for $\alpha = 0.05$ or 0.01, the sample variance degrees of freedom within groups $(n-1)$ and the number of groups k. The test statistic is simply the ratio of the largest to smallest sample variance among the groups. When $k = 2$, it is equivalent to the F-test covered previously.

Example 6.6: Physical Properties of Rocks from 3 Locations in Iran

Based on $\alpha = 0.05$, $k = 3$, and $n - 1 = 8 - 1 = 7$, we find the critical value is 6.94, so we reject $H_0 : \sigma_g^2 = \sigma_c^2 = \sigma_k^2$ if the ratio of the largest to smallest sample variance exceeds 6.94. For this example, $F_{max} = 2038.894/783.676 = 2.60$. Again, there is no evidence that the population variances of rock densities differ among the 3 locations.

∇

6.3.3 Levene's Test

Levene's test is not based on data being normally distributed and is robust to outliers. The test makes use of the Analysis of Variance (see Chapter 7) on absolute deviations from the group median, and is described below where $n_\cdot = n_1 + \cdots + n_k$.

$$z_{ij} = |y_{ij} - \tilde{y}_i| \quad i = 1, \ldots, k; j = 1, \ldots, n_i \quad \tilde{y}_i = \text{median}\,(y_{i1}, \ldots, y_{in_i}) \quad \overline{z}_{i.} = \frac{\sum_{j=1}^{n_i} z_{ij}}{n_i} \quad \overline{z}_{..} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} z_{ij}}{n_.}$$

$$TS: F_L = \frac{\sum_{i=1}^{k} n_i \left(\overline{z}_{i.} - \overline{z}_{..}\right)^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \overline{z}_{i.}\right)^2/(n_. - k)} \quad RR: F_L \geq F_{\alpha,k-1,n_.-k} \quad P = P\left(F_{k-1,n_.-k}\right) \geq F_L$$

Note that while variances are based on squared deviations from the mean, Levene's test is based on absolute deviations from the median. Some versions do use the mean in place of the median.

Example 6.7: Physical Properties of Rocks from 3 Locations in Iran

The computations needed for Levene's test are given in Table 6.2 for the Iran rock density data. There are $k = 3$ groups (locations) and $n_. = 8 + 8 + 8 = 24$ total observations. Again, we find no evidence of population variances being different.

$$\overline{z}_{..} = \frac{35.95 + 21.9 + 24.0375}{3} = 27.296$$

$$\sum_{i=1}^{k} n_i \left(\overline{z}_{i.} - \overline{z}_{..}\right)^2 = 8\left[(35.95 - 27.296)^2 + (21.9 - 27.296)^2 + (24.0375 - 27.296)^2\right] = 917.01$$

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(z_{ij} - \overline{z}_{i.}\right)^2 = 4632.42 + 3574.04 + 881.02 = 9087.48$$

$$TS: F_L = \frac{917.01/(3-1)}{9087.48/(24-3)} = \frac{458.505}{432.737} = 1.060 \quad RR: F_L \geq F_{.05,2,21} = 3.4668 \quad P = P(F_{2,21} \geq 1.090) = .3644$$

R Commands and Output

```
## Commands

## Using levene.test directly on y.g, y.c, and y.k
## Combine y.g, y.c, y.k into a single variable
y <- c(y.g, y.c, y.k)
## Create a variable that contains the locations of elements
loc.y <- c(rep(1,n.g),rep(2,n.c),rep(3,n.k))
loc.y <- factor(loc.y)
install.packages("car")
```

id(j)	Ghotvand($i=1$)	Chamshir($i=2$)	Khersan($i=3$)	z_{1j}	z_{2j}	z_{3j}
1	2800	2851	2790	36.65	16.75	34.5
2	2803.6	2879.8	2796.4	33.05	12.05	28.1
3	2811	2854.2	2817	25.65	13.55	7.5
4	2825.6	2862.1	2853.6	11.05	5.65	29.1
5	2847.7	2861.1	2804.4	11.05	6.65	20.1
6	2857.4	2918.4	2832	20.75	50.65	7.5
7	2906.7	2873.4	2854.5	70.05	5.65	30
8	2916	2932	2860	79.35	64.25	35.5
Median	2836.65	2867.75	2824.5	29.35	12.8	28.6
Mean	2846	2879	2825.988	35.95	21.9	24.0375
SumSq	14272.26	6398.42	5485.729	4632.42	3574.04	881.0187
Variance	2038.894	914.06	783.6755	661.7743	510.5771	125.8598

Table 6.2: Density of Rock Samples (kg/m^3) from 3 Locations in Iran - Calculations for Levene's test

```
library(car)
leveneTest(y, loc.y, "median")

## Output

> round(levene.out,4)
     F-stat DF1 DF2 F(.05) P-value
[1,] 1.0595   2  21 3.4668  0.3644

> leveneTest(y, loc.y, "median")
Levene's Test for Homogeneity of Variance (center = "median")
      Df F value Pr(>F)
group  2  1.0595 0.3644
      21
```

6.3.4 Jacknife Test

This extends the 2-sample Jacknife test to $k \geq 2$ groups. Similar notation is used, where for each group ($j = 1, \ldots, k$), the sample mean and variance is obtained with each observation deleted one-at-a-time and used to compute an F-statistic in a manner similar to Levene's test. Let $s_1^2, \ldots, s_k^2$ be the sample variances for the full samples for the k groups. The test is conducted as follows.

$$\overline{y}_{i(j)} = \frac{\sum_{j' \neq j} y_{ij'}}{n_i - 1} \quad s^2_{i(j)} = \frac{\sum_{j' \neq j} \left(y_{ij'} - \overline{y}_{i(j)}\right)^2}{n_i - 2} \quad D_{ij} = n_i \ln\left(s_i^2\right) - (n_i - 1) \ln\left(s^2_{i(j)}\right) \quad i = 1, \ldots, k; j = 1, \ldots, n_i$$

$$\overline{D}_{i\cdot} = \frac{\sum_{j=1}^{n_i} D_{ij}}{n_i} \quad i = 1, \ldots, k \qquad \overline{D}_{\cdot\cdot} = \frac{\sum_{i=1}^{k} \sum_{j=1}^{n_i} D_{ij}}{n_\cdot}$$

$$TS: F_J = \frac{\sum_{i=1}^{k} n_i \left(\overline{D}_{i\cdot} - \overline{D}_{\cdot\cdot}\right)^2 / (k-1)}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(D_{ij} - \overline{D}_{i\cdot}\right)^2 / (n_\cdot - k)} \quad RR: F_J \geq F_{\alpha, k-1, n_\cdot - k} \quad P = P\left(F_{k-1, n_\cdot - k}\right) \geq F_J$$

The test statistic is $F_J = 1.3558$ with a P-value of .2794. This result is consistent with the other methods.

R Output

```
## Output

> round(F.J.out, 4)
     F-stat df1 df2 F(.05) P-value
[1,] 1.3558   2  21 3.4668  0.2794
```

∇

6.4 R Code for Chapter 6

```
### Chapter 6

### Figure 6.1

## Chi-square Distribution

y <- seq(0,40,.01)
df <- 15
fy <- dchisq(y, df)
X2.LO <- qchisq(.025, df)
X2.HI <- qchisq(.975, df)

plot(y,fy,type="l", xlim=c(0,40),
xlab="x", ylab="f(x)",
## main=expression(paste(chi^2,"(",nu,"=15) Distribution")))
main=("Chi-Square Distribution with 15 Degrees of Freedom"))
abline(v=X2.LO, lwd=2)
abline(v=X2.HI, lwd=2)
text(3, 0.06, substitute(chi[.975]^2==X2.LO,
   list(X2.LO=round(X2.LO,3))), cex=0.9)
text(30, 0.06, substitute(chi[.025]^2==X2.HI,
   list(X2.HI=round(X2.HI,3))), cex=0.9)

rm(list=ls(all=TRUE))

### Example 6.1

nhl_ht_wt <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_ht_wt.csv",
     header=T)
attach(nhl_ht_wt); names(nhl_ht_wt)
set.seed(54321)
bmi <- 703*Weight/(Height^2) ### Create bmi from Height and Weight
N.bmi <- length(bmi) ### Population size
mu.bmi <- mean(bmi) ### Population mean
sigma.bmi <- sd(bmi)*sqrt((N.bmi-1)/N.bmi) ### Population SD (Uses N as denominator, not N-1)
ybar <- numeric(10000); stddev <- numeric(10000) # Create vectors to save sample means, SDs
for (i in 1:10000) {
y <- sample(bmi, 16) ### Sample n=16 bmi w/out replacement
ybar[i] <- mean(y)
```

```
stddev[i] <- sd(y)
}
cbind(stddev[1], stddev[1]^2)
X2 <- (16-1)*(stddev/sigma.bmi)^2
# mean(ybar); sd(ybar)

q.emp <- quantile(X2,c(.025,.25,.5,.75,.975))
q.the <- qchisq(c(.025,.25,.5,.75,.975),16-1)
cover <- sum(X2>=qchisq(.025,15) & X2<=qchisq(.975,15)) / length(X2)

var.out1 <- cbind(sigma.bmi^2, 16, 16-1, mean(X2), var(X2), qchisq(.025,15), qchisq(.975,15), cover)
colnames(var.out1) <- cbind("sigma^2", "n", "df", "Mean(X2)", "Var(X2)", "chisq(.025,15)",
     "chisq(.975,15)", "cover prob")
round(var.out1, 3)

q.out <- rbind(q.the, q.emp)
rownames(q.out) <- c("Theoretical", "Empirical")
colnames(q.out) <- c("10%", "25%", "50", "75%", "90%")
round(q.out, 3)

## Figure 6.2

yx2lim <- 1.1*10000*max(dchisq(0:100,16-1))
hist(X2[X2 <= 100],breaks=0:100, ylim=c(0,yx2lim),
main=expression(paste("Sampling Distribution of ",X^2," - BMI Data")))
lines(0:100,1*10000*dchisq(0:100,16-1))

rm(list=ls(all=TRUE))

### Figure 6.3
###   F-distribution

y <- seq(0,5,.01)
df1 <- 24;  df2 <- 24
fy <- df(y, df1, df2)
F.LO <- qf(.025, df1, df2)
F.HI <- qf(.975, df1, df2)

plot(y,fy,type="l", xlim=c(0,5),
xlab="x", ylab="f(x)",
## main=expression(paste(chi^2,"(",nu,"=15) Distribution")))
main=("F Distribution with 24,24 Degrees of Freedom"))
abline(v=F.LO, lwd=2)
abline(v=F.HI, lwd=2)
text(0.05, 0.2, substitute(F[.975]==F.LO,
   list(F.LO=round(F.LO,3))), cex=0.9)
text(2.5, 0.2, substitute(F[.025]^2==F.HI,
   list(F.HI=round(F.HI,3))), cex=0.9)

rm(list=ls(all=TRUE))


### Example 6.2

rr.mar <- read.csv("http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv",
header=T)
attach(rr.mar); names(rr.mar)
f.mph <- mph[Gender=="F"] ### Subsets Females from population
m.mph <- mph[Gender=="M"] ### Subsets Males from population
f.mu <- mean(f.mph)
f.sigma <- sd(f.mph)
m.mu <- mean(m.mph)
m.sigma <- sd(m.mph)
var.ratio <- (f.sigma/m.sigma)^2
```

```
## Begin sampling
set.seed(45678)
ybar.f <- numeric(10000); sd.f <- numeric(10000)
ybar.m <- numeric(10000); sd.m <- numeric(10000)
for (i in 1:10000) {
y.f <- sample(f.mph, 25)
y.m <- sample(m.mph, 25)
ybar.f[i] <- mean(y.f); sd.f[i] <- sd(y.f)
ybar.m[i] <- mean(y.m); sd.m[i] <- sd(y.m)
}
cbind(sd.f[1], sd.m[1])
F.fm <- ((sd.f**2)/(f.sigma**2))/((sd.m**2)/(m.sigma**2))

q.emp <- quantile(F.fm,c(.025,.25,.5,.75,.975))
q.the <- qf(c(.025,.25,.5,.75,.975),25-1,25-1)
CI.LO <- (sd.f/sd.m)^2/qf(.975,24,24)
CI.HI <- (sd.f/sd.m)^2/qf(.025,24,24)
cover <- sum(CI.LO <= var.ratio & CI.HI >= var.ratio) / 10000
F.stat <- (sd.f/sd.m)^2
rej.LO <- sum(F.stat <= qf(.025,24,24)) / 10000
rej.HI <- sum(F.stat >= qf(.975,24,24)) / 10000

var.out2 <- cbind(f.sigma^2, m.sigma^2, var.ratio, cover, rej.LO, rej.HI, rej.LO+rej.HI)
colnames(var.out2) <- c("sigma_F^2", "sigma_M^2", "Var Ratio", "Cover", "Rej Lo",
    "Rej Hi", "Rej Tot")
round(var.out2, 4)

q.out <- rbind(q.the, q.emp)
rownames(q.out) <- c("Theoretical", "Empirical")
colnames(q.out) <- c("10%", "25%", "50", "75%", "90%")
round(q.out, 3)

rm(list=ls(all=TRUE))

### Example 6.3


y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)
## Brute Force
n.g <- length(y.g); n.c <- length(y.c)
s2.g <- var(y.g); s2.c <- var(y.c)
F.025 <- qf(.975,n.g-1,n.c-1)
F.975 <- qf(.025,n.g-1,n.c-1)
F.obs <- s2.g / s2.c
F.LB <- F.obs / F.025
F.UB <- F.obs / F.975
F.P <- 2\min(pf(F.obs,n.g-1,n.c-1),1-pf(F.obs,n.g-1,n.c-1))
F.out <- cbind(F.obs, F.975, F.025, F.P, F.LB, F.UB)
colnames(F.out) <- c("F-stat", "F(.975)", "F(.025)","P-value", "Lower", "Upper")
round(F.out,3)
## Using var.test directly on y.g and y.c
var.test(y.g, y.c)

rm(list=ls(all=TRUE))

### Example 6.4

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)
n1 <- length(y.g) ## Sample size for group 1
n2 <- length(y.c) ## Sample size for group 2
```

```
var1 <- var(y.g) ## Sample variance for group 1
var2 <- var(y.c) ## Sample variance for group 2
var1.jack <- rep(0,n1) ## Holder for jacknifed variances for grp 1
var2.jack <- rep(0,n2) ## Holder for jacknifed variances for grp 2
for (i1 in 1:n1) var1.jack[i1] <- var(y.g[-i1]) ## Jacknifed var1
for (i2 in 1:n2) var2.jack[i2] <- var(y.c[-i2]) ## Jacknifed var2
round(cbind(var1.jack, var2.jack),2)
D1 <- n1*log(var1) - (n1-1)*log(var1.jack) ## D1 stat for each rock
D2 <- n2*log(var2) - (n2-1)*log(var2.jack) ## D2 stat for each rock
D1.mean <- mean(D1)
D2.mean <- mean(D2)
D1.var_n <- var(D1)/n1
D2.var_n <- var(D2)/n2
z.D <- (D1.mean-D2.mean) / sqrt(D1.var_n+D2.var_n)
p.z <- 2*(1-pnorm(abs(z.D),0,1))
p.t <- 2*(1-pt(abs(z.D),n1+n2-2))

jk.out <- cbind(D1.mean, D2.mean, D1.var_n, D2.var_n, z.D, p.z, p.t)
colnames(jk.out) <- c("Dbar_1", "Dbar_2", "S2_D1", "S2_D2", "z", "P(z)", "P(t)")
round(jk.out, 4)

rm(list=ls(all=TRUE))

### Example 6.5

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)

## Brute Force
num.groups <- 3
n.g <- length(y.g); n.c <- length(y.c); n.k <- length(y.k)
(s2.g <- var(y.g)); (s2.c <- var(y.c)); (s2.k <- var(y.k))
df.g <- n.g-1; df.c <- n.c-1; df.k <- n.k-1
(s2 <- (df.g*s2.g + df.c*s2.c + df.k*s2.k) / (df.g + df.c + df.k))
df <- df.g + df.c + df.k
(C.bart <- 1 + (1/(3*(num.groups-1))) * ((1/df.g + 1/df.c + 1/df.k) - 1/df))
X2.bart <- (1/C.bart)*(df*log(s2) -
(df.g*log(s2.g) + df.c*log(s2.c) + df.k*log(s2.k)))
X2.05 <- qchisq(.95,num.groups-1)
X2.p <- 1 - pchisq(X2.bart, num.groups-1)
bart.out <- cbind(X2.bart, num.groups-1, X2.05, X2.p)
colnames(bart.out) <- c("X2-stat", "DF", "X2(.05)", "P-value")
round(bart.out,3)

## Using bartlett.test directly on y.g, y.c, and y.k
## Combine y.g, y.c, y.k into a single variable
y <- c(y.g, y.c, y.k)
## Create a variable that contains the locations of elements
loc.y <- c(rep(1,n.g),rep(2,n.c),rep(3,n.k))
bartlett.test(y ~ loc.y)

rm(list=ls(all=TRUE))

### Example 6.7

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)

## Brute Force
num.groups <- 3
n.g <- length(y.g); n.c <- length(y.c); n.k <- length(y.k)
med.g <- median(y.g); med.c <- median(y.c); med.k <- median(y.k)
```

```
z.g <- abs(y.g-med.g); z.c <- abs(y.c-med.c); z.k <- abs(y.k-med.k)
mean.z.g <- mean(z.g); mean.z.c <- mean(z.c); mean.z.k <- mean(z.k)
mean.z <- (n.g*mean.z.g + n.c*mean.z.c + n.k*mean.z.k) / (n.g+n.c+n.k)
ssz.between <- n.g*(mean.z.g-mean.z)^2 + n.c*(mean.z.c-mean.z)^2 +
n.k*(mean.z.k-mean.z)^2
ssz.within <- sum((z.g-mean.z.g)^2) + sum((z.c-mean.z.c)^2) +
sum((z.k-mean.z.k)^2)
df.between <- num.groups-1
df.within <- (n.g+n.c+n.k) - num.groups
F.L <- (ssz.between/df.between) / (ssz.within/df.within)
F.05 <- qf(.95,df.between,df.within)
F.p <- 1-pf(F.L,df.between,df.within)
levene.out <- cbind(F.L,df.between,df.within,F.05,F.p)
colnames(levene.out) <- c("F-stat", "DF1","DF2", "F(.05)", "P-value")
round(levene.out,4)

## Using levene.test directly on y.g, y.c, and y.k
## Combine y.g, y.c, y.k into a single variable
y <- c(y.g, y.c, y.k)
## Create a variable that contains the locations of elements
loc.y <- c(rep(1,n.g),rep(2,n.c),rep(3,n.k))
loc.y <- factor(loc.y)
install.packages("car")
library(car)
leveneTest(y, loc.y, "median")

rm(list=ls(all=TRUE))

### Example 6.8

y.g <- c(2811.0, 2857.4, 2906.7, 2847.7, 2825.6, 2803.6, 2916.0, 2800.0)
y.c <- c(2854.2, 2918.4, 2873.4, 2861.1, 2862.1, 2879.8, 2932.0, 2851.0)
y.k <- c(2817.0, 2832.0, 2854.5, 2804.4, 2853.6, 2796.4, 2860.0, 2790.0)
n1 <- length(y.g) ## Sample size for group 1
n2 <- length(y.c) ## Sample size for group 2
n3 <- length(y.k) ## Sample size for group 3
n.all <- n1 + n2 + n3
num.grp <- 3
var1 <- var(y.g) ## Sample variance for group 1
var2 <- var(y.c) ## Sample variance for group 2
var3 <- var(y.k) ## Sample variance for group 3
var1.jack <- rep(0,n1) ## Holder for jacknifed variances for grp 1
var2.jack <- rep(0,n2) ## Holder for jacknifed variances for grp 2
var3.jack <- rep(0,n3) ## Holder for jacknifed variances for grp 3

for (i1 in 1:n1) var1.jack[i1] <- var(y.g[-i1]) ## Jacknifed var1
for (i2 in 1:n2) var2.jack[i2] <- var(y.c[-i2]) ## Jacknifed var2
for (i3 in 1:n3) var3.jack[i3] <- var(y.k[-i3]) ## Jacknifed var3
round(cbind(var1.jack, var2.jack, var3.jack),2)
D1 <- n1*log(var1) - (n1-1)*log(var1.jack) ## D1 stat for each rock
D2 <- n2*log(var2) - (n2-1)*log(var2.jack) ## D2 stat for each rock
D3 <- n3*log(var3) - (n3-1)*log(var3.jack) ## D3 stat for each rock
D1.mean <- mean(D1)
D2.mean <- mean(D2)
D3.mean <- mean(D3)
D.mean <- (n1*D1.mean + n2*D2.mean + n3*D3.mean) / (n.all)
df1 <- num.grp-1; df2 <- n.all-num.grp
SS1 <- n1*(D1.mean-D.mean)^2 + n2*(D2.mean-D.mean)^2 + n3*(D3.mean-D.mean)^2
SS2 <- sum((D1-D1.mean)^2) + sum((D2-D2.mean)^2) + sum((D3-D3.mean)^2)
F.J <- (SS1/df1) / (SS2/df2)
F.J.RR <- qf(.95,df1,df2)
F.J.p <- 1-pf(F.J,df1,df2)
F.J.out <- cbind(F.J,df1,df2,F.J.RR,F.J.p)
colnames(F.J.out) <- c("F-stat","df1","df2","F(.05)","P-value")
```

```
round(F.J.out, 4)

rm(list=ls(all=TRUE))
```

Chapter 7

Experimental Design and the Analysis of Variance

Chapter 5 covered methods to make comparisons between the means of a numeric response variable for two treatments or groups. The cases were considered where the experiment was conducted as an independent samples (aka parallel groups, between subjects) design, as well as a paired (aka crossover, within subjects) design. Procedures were covered that assume normally distributed data, as well as nonparametric methods that can be used when data are not normally distributed.

This chapter will introduce methods that can be used to compare more than two groups (that is, when the explanatory variable has more than two levels). In this chapter, we will refer to explanatory variable as a **factor**, and their levels as **treatments**. The following situations will be covered.

- 1–Factor, Independent Samples Designs (Completely Randomized Design)
- 1– Treatment Factor, Paired Designs (Randomized Block Design, Latin Square)

In all situations, there will be a numeric response variable, and at least one categorical (or possibly numeric, with several levels) independent variable. The goal will always be to compare mean (or median) responses among several populations. When all factor levels for a factor are included in the experiment, the factor is said to be **fixed**. When a sample of a larger population of factor levels are included, the factor is said to be **random**. Only fixed effects designs are considered here.

7.1 Completely Randomized Design (CRD) For Independent Samples

In the Completely Randomized Design, there is one factor that is controlled. This factor has k levels (which are often treatment groups), and n_i units are measured for the i^{th} level of the factor. Observed responses

are defined as y_{ij}, representing the measurement on the j^{th} experimental unit (subject), receiving the i^{th} treatment. We will write this in model form based on random responses as follows where the factor is **fixed** (all levels of interest are included in the experiment).

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} = \mu_i + \epsilon_{ij} \quad i = 1, \ldots, k; \quad j = 1, \ldots, n_i$$

Here, μ is the overall mean measurement across all treatments, α_i is the effect of the i^{th} treatment ($\mu_i = \mu + \alpha_i$), and ϵ_{ij} is a random error component that has mean 0 and variance σ^2. This ϵ_{ij} allows for the fact that there will be variation among the measurements of different subjects (units) receiving the same treatment. A common parameterization that has nice properties is to assume $\sum n_i \alpha_i = 0$.

Of interest to the experimenter is whether or not there is a **treatment effect**, that is do any of the levels of the treatment provide higher (lower) mean response than other levels. This can be hypothesized symbolically as $H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ (no treatment effect) against the alternative $H_A :$ Not all $\alpha_i = 0$ (treatment effects exist). Note that if $\alpha_1 = \alpha_2 = \cdots = \alpha_k = 0$ then $\mu_1 = \cdots = \mu_k$.

As with the case where there are two treatments to compare, tests based on the assumption that the k populations are normal (mound–shaped) will be used, either assuming equal or unequal variances. Also, an alternative test (based on ranks) that does not assume that the k populations are normal is used to compare poulation medians.

7.1.1 Tests Based on Normally Distributed Data

When the underlying populations of measurements that are to be compared are approximately normal, with equal variances, the F–test is appropriate. To conduct this test, partition the total variation in the sample data to variation **within** and **among** treatments. This partitioning is referred to as the **Analysis of Variance** and is an important tool in many statistical procedures. First, define the following items, based on random outcomes Y_{ij} where i indexes treatment and j represents the replicate number, with n_i observations for treatment i and $n_\cdot = n_1 + \cdots + n_k$.

$$Y_{ij} \sim N(\mu_i, \sigma) \quad \Rightarrow \quad E\left\{Y_{ij}^2\right\} = \mu_i^2 + \sigma^2$$

$$\overline{Y}_{i\cdot} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i} \qquad \overline{Y}_{i\cdot} \sim N\left(\mu_i, \frac{\sigma}{\sqrt{n_i}}\right) \quad \Rightarrow \quad E\left\{\overline{Y}_{i\cdot}^2\right\} = \mu_i^2 + \frac{\sigma^2}{n_i}$$

$$\overline{Y}_{\cdot\cdot} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} Y_{ij}}{n_\cdot} = \frac{\sum_{i=1}^{k} n_i \overline{Y}_{i\cdot}}{n_\cdot}$$

$$\text{Total (Corrected) Sum of Squares: } TSS = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad df_{Total} = n_\cdot - 1$$

$$\text{Between Treatment Sum of Squares: } SST = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 = \sum_{i=1}^{k} n_i\left(\overline{Y}_{i\cdot} - \overline{Y}_{\cdot\cdot}\right)^2 \qquad df_T = k - 1$$

Within Treatment (Error) Sum of Squares: $SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2 = \sum_{i=1}^{k}\left(n_i - 1\right)S_i^2 \qquad df_E = n_. - k$

Under the null hypothesis of no treatment effects ($\mu_1 = \cdots = \mu_k = \mu$), or equivalently ($\tau_1 = \cdots = \tau_k = 0$) the following results are obtained, where MST and MSE are mean squares for treatments and error, respectively.

$$E\left\{SST\right\} = E\left\{\sum_{i=1}^{k} n_i\left(\overline{Y}_{i.} - \overline{Y}_{..}\right)^2\right\} = E\left\{\sum_{i=1}^{k} n_i\left(\overline{Y}_{i.}^2 - \overline{Y}_{..}^2\right)\right\} =$$

$$= \sum_{i=1}^{k} n_i\left[\left(\mu^2 + \frac{\sigma^2}{n_i}\right) - \left(\mu^2 + \frac{\sigma^2}{n_.}\right)\right] = (k-1)\sigma^2 \quad \Rightarrow \quad E\left\{MST\right\} = E\left\{\frac{SST}{k-1}\right\} = \sigma^2$$

$$E\left\{SSE\right\} = E\left\{\sum_{i=1}^{k}\sum_{k=1}^{n_i}\left(Y_{ij} - \overline{Y}_{i.}\right)^2\right\} = E\left\{\sum_{i=1}^{k}\left(n_i - 1\right)S_i^2\right\} = (n_. - k)\sigma^2 \quad \Rightarrow \quad E\left\{MSE\right\} = E\left\{\frac{SSE}{n_. - k}\right\} = \sigma^2$$

Under the null hypothesis of no treatment effects, $E\left\{MST\right\} = E\left\{MSE\right\} = \sigma^2$ and the ratio MST/MSE follows the F-distribution with $k-1$ numerator and $n_. - k$ denominator degrees of freedom. When the null is not true and not all $\alpha_i = 0$, then the ratio follows the non-central F-distribution with parameter λ given below.

$$\frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda} \qquad \lambda = \frac{\sum_{i=1}^{k} n_i\alpha_i^2}{\sigma^2} \qquad \nu_1 = k-1 \qquad \nu_2 = n_. - k$$

Once samples have been obtained and the y_{ij} are observed, the F-test is conducted as follows.

$$\begin{aligned}
\overline{y}_{i.} &= \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \\
s_i &= \sqrt{\frac{\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_{i.})^2}{n_i - 1}} \\
n_. &= n_1 + \cdots + n_k \\
\overline{y}_{..} &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} y_{ij}}{n_.} \\
TSS &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_{..}\right)^2
\end{aligned}$$

$$SST = \sum_{i=1}^{k} n_i \left(\overline{y}_{i.} - \overline{y}_{..}\right)^2$$
$$SSE = \sum_{i=1}^{k} (n_i - 1) s_i^2$$

Here, $\overline{y}_{i.}$ and s_i are the mean and standard deviation of measurements in the i^{th} treatment group, and $\overline{y}_{..}$ and $n_.$ are the overall mean and total number of all measurements. TSS is the total variability in the data (ignoring treatments), SST measures the variability in the sample means among the treatments (weighted by the sample sizes), and SSE measures the variability within the treatments.

Note that the goal is to determine whether or not the population means differ. If they do, we would expect SST to be large, since that sum of squares is measuring differences in the sample means. A test for treatment effects is conducted after constructing an Analysis of Variance table, as shown in Table 7.1. In that table, there are *sums of squares* for treatments (SST), for error (SSE), and total (TSS). Also, there are *degrees of freedom*, which represent the number of "independent" terms in the sum of squares. Then, the *mean squares*, are sums of squares divided by their degrees of freedom. Finally, the F–statistic is computed as $F = MST/MSE$. This will serve as the test statistic. Note that MSE is an extension of the pooled variance computed in Chapter 5 for two groups, and often it is written as $MSE = s^2$.

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	$SST = \sum_{i=1}^{k} n_i \left(\overline{y}_{i.} - \overline{y}_{..}\right)^2$	$k-1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
ERROR	$SSE = \sum_{i=1}^{k} \left(n_i - 1\right) s_i^2$	$n_. - k$	$MSE = \frac{SSE}{n_. - k}$	
TOTAL	$TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{..}\right)^2$	$n_. - 1$		

Table 7.1: The Analysis of Variance Table for the Completely Randomized (Parallel Groups) Design

The formal method of testing this hypothesis is as follows.

1. $H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad (\mu_1 = \cdots = \mu_k)$ (No treatment effect)
2. H_A : Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha, k-1, n_.-k}$
5. p-value: $P(F_{k-1,n_.-k} \geq F_{obs})$

Example 7.1: Body Mass Indices of NHL, NBA, and EPL, Players

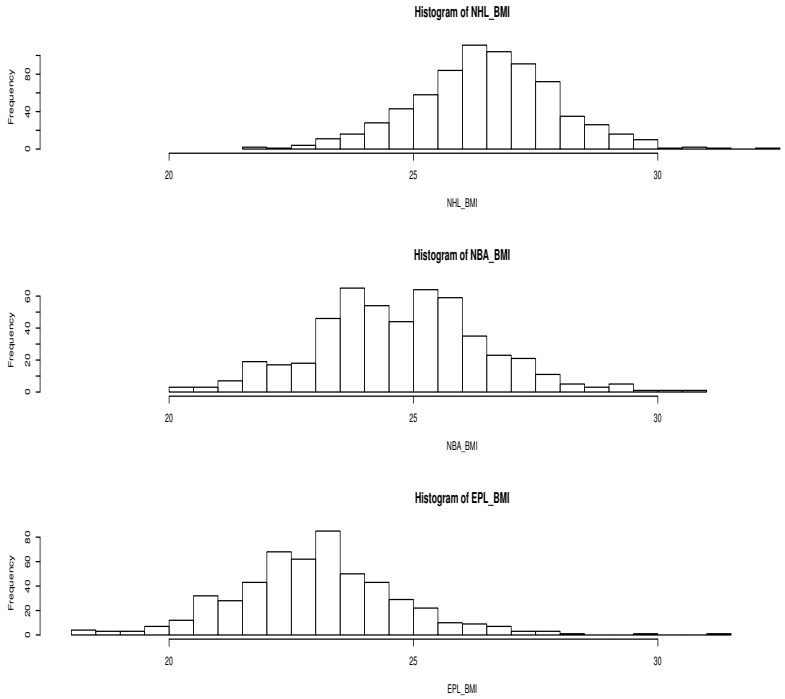

Figure 7.1: Histograms of NHL, NBA, and EPL Body Mass Indices

Consider an extension of the Body Mass Index analysis to include National Basketball Association players. The populations are NHL ($i = 1$), NBA ($i = 2$), and EPL ($i = 3$). Histograms for the three populations are given in Figure 7.1. The population sizes, means, and standard deviations are given below.

$$N_1 = 707 \quad N_2 = 505 \quad N_3 = 526 \qquad \mu_1 = 26.50 \quad \mu_2 = 24.74 \quad \mu_3 = 23.02 \qquad \sigma_1 = 1.45 \quad \sigma_2 = 1.72 \quad \sigma_3 = 1.71$$

While the population standard deviations (and thus variances) are not all equal, a "pooled" variance is used for computational purposes. Also, μ and α_i are computed.

$$\sigma^2 = \frac{717\left(1.45^2\right) + 505\left(1.72^2\right) + 526\left(1.71^2\right)^2}{717 + 505 + 526} = 2.60 \qquad \mu = \frac{717(26.50) + 505(24.74) + 526(23.02)}{717 + 505 + 526} = 24.94$$

$$\alpha_1 = 26.50 - 24.94 = 1.56 \qquad \alpha_2 = 24.74 - 24.94 = -0.20 \qquad \alpha_3 = 23.02 - 24.94 = -1.92$$

Note that these α_i are obtained under the assumption $\sum N_i\alpha_i = 0$. If samples of sizes $n_1 = n_2 = n_3 = 4$ and $n_1 = n_2 = n_3 = 12$ are taken, the following F-distributions for the ratio MST/MSE are obtained.

$$n_i = 4: \quad \frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda_1} \quad \lambda_1 = \frac{4\left(1.56^2 + (-0.20)^2 + (-1.92)^2\right)}{2.60} = 9.48 \quad \nu_1 = 3{-}1 = 2 \quad \nu_2 = 12{-}3 = 9$$

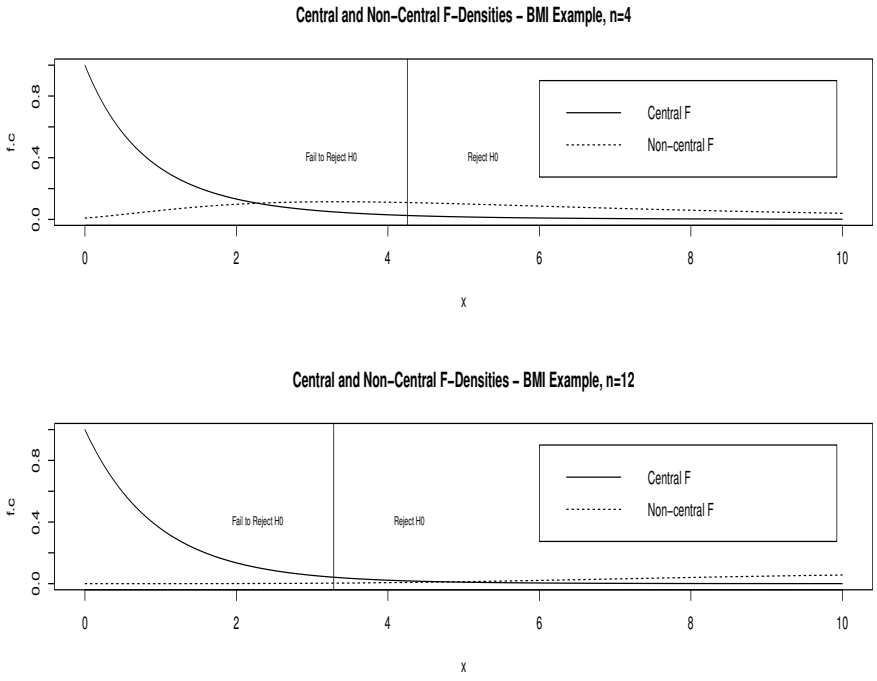

Figure 7.2: Central and non-central F-distributions for Body Mass Index example

$$n_i = 12: \quad \frac{MST}{MSE} \sim F_{\nu_1,\nu_2,\lambda_2} \quad \lambda_2 = \frac{12\left(1.56^2 + (-0.20)^2 + (-1.92)^2\right)}{2.60} = 28.43 \quad \nu_1 = 3-1 = 2 \quad \nu_2 = 36-3 = 33$$

When $n_1 = n_2 = n_3 = 4$, the critical value for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ at $\alpha = 0.05$ significance level is $F_{.05,2,9} = 4.256$. The power of the F-test under this configuration is $\pi_1 = .636$. When $n_1 = n_2 = n_3 = 12$, the critical value for testing $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0$ at $\alpha = 0.05$ significance level is $F_{.05,2,33} = 3.285$. The power of the F-test under this configuration is $\pi_2 = .997$. The central F-densities and the non-central F-densities with $\lambda_1 = 9.48$ and $\lambda_2 = 28.43$ for the denominator degrees of freedom of 9 and 33 are given in Figure 7.2.

Based on 100000 random sample of size $n_i = 4$ from each league, the F-test rejected the null hypothesis of no league differences in 63.4% of the samples. With samples of size $n_i = 12$, 99.7% of the F-tests rejected the null hypothesis. Despite the fact that the populations of measurements are not exactly normally distributed with equal variances, the test performs as expected. Computations for the first samples of size $n_1 = n_2 = n_3 = 12$ are given below.

$$\overline{y}_{1.} = 26.666 \quad \overline{y}_{2.} = 24.986 \quad \overline{y}_{3.} = 22.449 \quad \overline{y}_{..} = 24.701 \qquad s_1 = 1.968 \quad s_2 = 1.762 \quad s_3 = 1.149$$

$$SST = 12\left[(26.666 - 24.701)^2 + (24.986 - 24.701)^2 + (22.449 - 24.701)^2\right] = 108.167$$

$$df_T = 3 - 1 = 2 \qquad MST = \frac{108.167}{2} = 54.084$$

$$SSE = (12-1)\left[1.968^2 + 1.762^2 + 1.149^2\right] = 91.277 \quad df_E = 3(12) - 3 = 33 \quad MSE = \frac{91.277}{33} = 2.766$$

$$H_0 : \mu_1 = \mu_2 = \mu_3 \quad TS : F_{obs} = \frac{54.084}{2.766} = 19.55 \quad RR : F_{obs} \geq 3.285 \quad P = P\left(F_{2,33} \geq 19.55\right) < .0001$$

R Output

```
### Output

> round(ftest.out, 4)
     df_T df_E F(>05) P(F_obs>F(.05))
[1,]    2   33 3.2849          0.9942
> F[1]
[1] 19.55004
> cbind(ybar1[1], ybar2[1], ybar3[1], ybar[1], sd1[1], sd2[1], sd3[1])
         [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
[1,] 26.66637 24.98606 22.44932 24.70058 1.968428 1.762007 1.148883
```

∇

Example 7.2: Comparison of 5 Mosquito Repellents

A study compared $k = 5$ mosquito repellent patches on fabric for soldiers in military operations (Bhatnagar and Mehta (2007), [8]). The 5 treatments were: Odomos (1), Deltamethrin (2), Cyfluthrin (3), Deltamethrin+Odomos (4), and Cyfluthrin+Odomos (5), with $n_i = 30$ subjects per treatment, and a total of $n_{\cdot} = 150$ measurements. The response observed was the "Per Man-Hour Mosquito Catch." Sample statistics are given in Table 7.2, and the Analysis of Variance is given in Table 7.3. Data that have been generated to match the means and standard deviations are plotted in Figure 7.3. The overall mean (long line) and individual treatment means (short lines) are included.

Treatment	n_i	$\overline{y}_{i\cdot}$	s_i
Odomos (1)	30	7.900	3.367
Deltamethrin (2)	30	8.133	3.461
Cyfluthrin (3)	30	8.033	3.011
D+O(4)	30	6.333	3.122
C+O (5)	30	5.367	3.068

Table 7.2: Sample statistics for Mosquito Repellent study

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0 \quad (\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$ (No treatment effect)

ANOVA

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square	F_{obs}	$F_{.05}$	P
TREATMENTS	4	184.650	46.163	4.478	2.434	.0019
ERROR	145	1494.680	10.308			
TOTAL	149	1679.334				

Table 7.3: The Analysis of Variance table for the Mosquito Repellent study

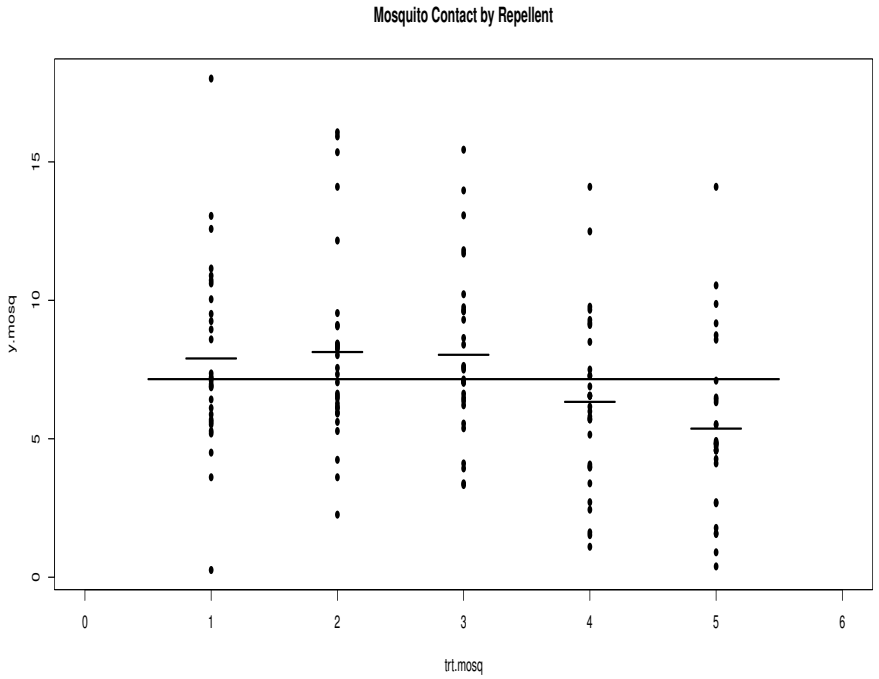

Figure 7.3: Mosquito catch by repellent treatment - data generated to match treatment means and standard deviations

2. H_A : Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE} = 4.478$
4. R.R.: $F_{obs} > F_{\alpha,k-1,n-k} = F_{0.05,4,145} = 2.434$
5. P-value: $P(F_{k-1,n.-k} \geq F_{obs}) = P(F_{4,145} \geq 4.478) = .0019$

The following R output gives the Analysis of Variance and the F-test.

R Output

```
### Output

> round(aov.out, 4)
           df        SS      MS      F F(.05)  P(>F)
Treatment   4  184.6501 46.1625 4.4782 2.4341 0.0019
Error     145 1494.6843 10.3082     NA     NA     NA
Total     149 1679.3345      NA     NA     NA     NA
```

The following R commands use the **aov** function to obtain the Analysis of Variance based on the raw data (not summary statistics).

R Commands and Output

```
## Commands
mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

trt.mosq <- factor(trt.mosq)
mosq.mod <- aov(y.mosq ~ trt.mosq)
summary(mosq.mod)

## Output

> summary(mosq.mod)
             Df Sum Sq Mean Sq F value  Pr(>F)
trt.mosq      4  184.6   46.16    4.48 0.00192 **
Residuals   145 1494.1   10.30
```

Since the F-statistic is sufficiently large, conclude that the means differ, then the following methods are used to make comparisons among treatments.

∇

Comparisons among Treatment Means

Assuming that it has been concluded that treatment means differ, we generally would like to know which means are significantly different. This is generally done by making contrasts among treatments. Special cases of contrasts include pre–planned or all pairwise comparisons between pairs of treatments.

A **contrast** is a linear function of treatment means, where the coefficients sum to 0. A contrast among population means can be estimated with the same contrast among sample means, and inferences can be made based on the sampling distribution of the contrast. Let C be the contrast among the population means, and $\hat{C}$ be its estimator based on means of the independent random samples.

$$C = a_1\mu_1 + \cdots + a_k\mu_k = \sum_{i=1}^{k} a_i\mu_i \text{ where } \sum_{i=1}^{k} a_i = 0 \qquad \hat{C} = a_1\overline{Y}_{1.} + \cdots + a_k\overline{Y}_{k.} = \sum_{i=1}^{k} a_i\overline{Y}_{i}$$

$$V\{\hat{C}\} = \sigma^2\left[\frac{a_1^2}{n_1} + \cdots + \frac{a_k^2}{n_k}\right] = \sigma^2\sum_{i=1}^{k}\frac{a_i^2}{n_i}$$

When the sample sizes are balanced (all n_i are equal), the formula for the variance clearly simplifies. Contrasts are specific to particular research questions, so the general rules for tests and Confidence Intervals are given here, followed by an application to the Mosquito Repellent study. Since the coefficients sum to 0, we are virtually always testing $H_0 : C = 0$ in practice.

Once samples are obtained, obtain $\hat{c}$, the contrast based on the observed sample means among the treatments.

$$\hat{c} = a_1\overline{y}_{1.} + \cdots + a_k\overline{y}_{k.} = \sum_{i=1}^{k} a_i\overline{y}_{i.} \qquad \hat{SE}\{\hat{C}\} = \sqrt{MSE\sum_{i=1}^{k}\frac{a_i^2}{n_i}}$$

Testing whether a contrast is equal to 0 and obtaining a $(1-\alpha)100\%$ Confidence Interval for C are done as follow.

$$H_0 : C = 0 \quad H_A : C \neq 0 \quad TS : t_C = \frac{\hat{c}}{\hat{SE}\{\hat{C}\}} \quad RR : |t_C| \geq t_{\alpha/2,n_.-k} \quad P = 2P\left(t_{n_.-k} \geq |t_C|\right)$$

$$(1-\alpha)100\% \text{ Confidence Interval for } C : \hat{c} \pm t_{\alpha/2,n_.-k}\hat{SE}\{\hat{C}\}$$

The test can be conducted as upper or lower-tailed with obvious adjustments. An alternative approach is to compute the sums of squares for the contrast SSC, and use an F-test, comparing its Mean Square to MSE.

$$SSC = \frac{(\hat{c})^2}{\sum_{i=1}^{k}\frac{a_i^2}{n_i}} \qquad df_C = 1 \qquad MSC = \frac{SSC}{1} = SSC$$

$$H_0 : C = 0 \quad H_A : C \neq 0 \quad TS : F_C = \frac{MSC}{MSE} \quad RR : F_C \geq F_{\alpha,1,n_.-k} \quad P = P\left(F_{1,n_.-k} \geq F_C\right)$$

Example 7.3: Comparison of 5 Mosquito Repellents

Suppose the researchers are interested in comparing the two treatments that use Deltamethrin (2 and 4) with the two treatments that use Cyfluthrin (3 and 5). Then, the following calculations are made.

$$C_1 = (\mu_2 + \mu_4) - (\mu_3 + \mu_5) \qquad a_1 = 0 \quad a_2 = a_4 = 1 \quad a_3 = a_5 = -1 \qquad n_i = 30 \qquad MSE = 10.308$$

$$\overline{y}_{2\cdot} = 8.133 \quad \overline{y}_{4\cdot} = 6.333 \quad \overline{y}_{3\cdot} = 8.033 \quad \overline{y}_{5\cdot} = 5.367 \qquad \hat{c}_1 = (8.133 + 6.333) - (8.033 + 5.367) = 1.066$$

$$\hat{SE}\{\hat{C}_1\} = \sqrt{10.308\left(\frac{0^2 + 1^2 + (-1)^2 + 1^2 + (-1)^2}{30}\right)} = 1.172$$

For a test ($\alpha = 0.05$) of $H_0 : C_1 = 0$, the test statistic, rejection region and P-value, along with a 95% Confidence Interval for C are given below.

$$TS : t_{C_1} = \frac{1.066}{1.172} = 0.910 \qquad RR : |t_{C_1}| \geq t_{.025,145} = 1.976 \qquad P = 2P\left(t_{145} \geq |0.910|\right) = .364$$

$$95\% \text{ CI for } C : 1.066 \pm 1.976(1.172) \equiv 1.066 \pm 2.316 \equiv (-1.250, 3.382)$$

There is no evidence of any difference between the effects of these two types of repellents. Next, we conduct the F-test, knowing in advance that its conclusion and P-value will be equivalent to 2-tailed t-test performed above (the only difference due to rounding is in third decimal place).

$$SSC_1 = \frac{1.066^2}{\frac{4}{30}} = 8.523 = MSC_1 \quad TS : F_{C_1} = \frac{8.523}{10.308} = 0.827 \quad RR : F_{C_1} \geq F_{.05,1,145} = 3.906$$

$$P = P\left(F_{1,145} \geq 0.827\right) = .365$$

For a second contrast (C_2), without going through all calculations, consider comparing Deltamethrin and Cyfluthrin (each without Odomos: 2 and 3) with Deltamethrin and Cyfluthrin (each with Odomos: 4 and 5). This involves: $a_1 = 0, a_2 = a_3 = 1, a_4 = a_5 = -1$. Note that the standard error of the contrast will be exactly the same as that for contrast $\hat{c}_1$.

$$\hat{c}_2 = 4.466 \quad TS : t_{C_2} = \frac{4.466}{1.172} = 3.811 \quad P = 2P\left(t_{145} \geq |3.811|\right) = .0002 \qquad 95\% \text{ CI: } 4.466 \pm 2.316 \equiv (2.150, 6.782)$$

There is evidence that the combined mean is higher without Odomos than with Odomos. Since low values are better (mosquito contacts), Odomos as an additive to the two chemicals (individually) is better

than no additive to the two chemicals individually. The F-test is given below. The R output that follows extends the calculations made in Example 7.2.

$$SSC_2 = \frac{4.466^2}{\frac{4}{30}} = 149.59 = MSC_2 \quad TS: F_{C_2} = \frac{149.59}{10.308} = 14.51 \quad P = P\left(F_{1,145} \geq 14.51\right) = .0005$$

R Output

```
## Output

> round(contrast.out, 4)
     Estimate Std Err      t 2P(>|t|)      LB     UB Sum Sq      F  P(>F)
[1,]    1.066  1.1724 0.9093   0.3647 -1.2511 3.3831 8.5227 0.8268 0.3647
> round(contrast.out, 4)
     Estimate Std Err      t 2P(>|t|)     LB     UB   Sum Sq       F P(>F)
[1,]    4.466  1.1724 3.8094    2e-04 2.1489 6.7831 149.5887 14.5117 2e-04
```

∇

A special class of contrasts are **orthogonal contrasts**. Two contrasts are orthogonal if the sum of the products of their a_i coefficients, divided by the sample sizes n_i, is 0. This concept is shown below.

$$C_1 = \sum_{i=1}^{k} a_{1i}\mu_i \qquad C_2 = \sum_{i=1}^{k} a_{2i}\mu_i \quad C_1 \text{ and } C_2 \text{ are orthogonal if } \quad \sum_{i=1}^{k} \frac{a_{1i}a_{2i}}{n_i} = 0$$

Note that if the sample sizes are all equal (balanced design), this simplifies to $\sum_{i=1}^{k} a_{1i}a_{2i} = 0$. The two contrasts in Example 7.3 are orthogonal (check this). If there are k treatments, and $k-1$ degrees of freedom for Treatments, any $k-1$ pairwise orthogonal contrasts' sums of squares will add up to the Treatment sum of squares. That is, SST can be decomposed into the sums of squares for the $k-1$ contrasts. The decomposition is not unique, there may be various sets of orthogonal contrasts.

Example 7.4: Comparison of 5 Mosquito Repellents

Consider these two other contrasts.

- (D versus C without O) vs (D versus C with Odomos): $C_3 = (\mu_2 - \mu_3) - (\mu_4 - \mu_5) = \mu_2 - \mu_3 - \mu_4 + \mu_5$
- Odomos only versus the four other treatments: $C_4 = 4\mu_1 - \mu_2 - \mu_3 - \mu_4 - \mu_5$

Table 7.4 gives the contrast coefficients for these four contrasts. For all six pairs, $\sum_{i=1}^{k} a_{ji}a_{j'i} = 0$, $j \neq j'$. Also given are the estimates, standard errors, t-tests, 95% Confidence Intervals, Sums of Squares and F-statistics.

Treatment (i)	C_1 ($j=1$)	C_2 ($j=2$)	C_3 ($j=3$)	C_4 ($j=4$)	$\overline{y}_{i.}$
1	0	0	0	4	7.900
2	1	1	1	-1	8.133
3	-1	1	-1	-1	8.033
4	1	-1	-1	-1	6.333
5	-1	-1	1	-1	5.367
$\hat{c}_j$	1.066	4.466	-0.866	3.734	
$\hat{SE}\{\hat{C}_j\}$	1.172	1.172	1.172	2.621	
t_{C_j}	0.909	3.809	-0.739	1.424	
P-value	.3642	.0002	.4613	.1565	
95% CI	(-1.251,3.383)	(2.149,6.783)	(-3.183,1.451)	(-1.447,8.915)	
SSC_j	8.523	149.589	5.625	20.914	
F_{C_j}	0.827	14.512	0.546	2.029	

Table 7.4: Orthogonal Contrasts for the Mosquito Repellent study

From Table 7.3, see that the Treatment sum of squares is $SST = 184.650$. As these four contrasts are pairwise orthogonal, their sums of squares add up to SST: $8.523 + 149.589 + 5.625 + 20.914 = 184.650$. Note that virtually all of the differences among the treatments (based on this set of contrasts) is contrast 2, comparing the average of D and C without O versus the average of D and C with O. The commands for Contrasts 3 and 4 are identical as that for Example 7.3 (with changes to the a^s), and are not included here.

∇

As the number of potential contrasts increases (as when k gets large), the chances of making false rejections of null hypotheses increases. Also, inferential methods are based on *a priori* contrasts being studied. A very conservative but widely used method for making simultaneous contrasts is Scheffe's method.

Scheffe's Method for All Possible Contrasts

This method makes use of the fact that all contrast sums of squares are smaller than the between treatments sum of squares (SST), since for any $k-1$ orthogonal contrasts, their sums of squares add up to SST. Thus, if any given contrast's sum of squares is large enough to reject $H_0 : \tau_1 = \cdots = \tau_k = 0$, when SSC replaces, SST in the F-test, that contrast is significantly different from 0 at the chosen significance level, which will be labeled α_E for the **experimentwise error rate**. This is the probability of making at least one incorrect conclusion with respect to the null hypotheses among all possible contrasts. The forms of the tests and the simultaneous $(1-\alpha_E)\,100\%$ Confidence Intervals for any set of contrasts are given here.

$$C = \sum_{i=1}^{k} a_i \mu_i \quad H_0 : C = 0 \quad H_A : C \neq 0 \qquad \text{Reject } H_0 \text{ if } : |\hat{c}| \geq \hat{SE}\{\hat{C}\}\sqrt{(k-1)F_{\alpha_E,k-1,n_\cdot-k}}$$

$$\text{Simultaneous } (1-\alpha_E)\,100\% \text{ Confidence Intervals: } \hat{c} \pm \hat{SE}\{\hat{C}\}\sqrt{(k-1)F_{\alpha_E,k-1,n_\cdot-k}}$$

Note that if the test rejects $H_0 : C = 0$, the corresponding Confidence Interval will be entirely positive or negative. If the test fails to reject H_0, the interval will contain 0.

Example 7.5: Scheffe's Method for Mosquito Repellent Contrasts

Considering the 4 contrasts described previously, the estimated standard errors were 1.172 for contrasts 1, 2, and 3, was 2.621 for contrast 4. Recall that $k = 5$ treatments were compared and the critical F-value for the Analysis of Variance was $F_{.05,5-1,150-5} = F_{.05,4,145} = 2.434$. For contrasts 1-3, the critical value and simultaneous Confidence Intervals are of the form given below.

$$\text{Conclude } C \neq 0 \text{ if: } |\hat{c}| \geq 1.172\sqrt{(5-1)2.434} = 3.657 \qquad \text{Simultaneous 95\% CI's: } \hat{c} \pm 3.657$$

Only contrast 2, with $\hat{c}_2 = 4.466$ meets that criteria. For contrast 4, the critical value is $2.621\sqrt{(5-1)2.434} = 8.178$ and that contrast, $\hat{c}_4 = 3.734$ does not exceed the critical value.

∇

Special cases of contrasts are **pairwise comparisons** among treatments. We will first consider making simultaneous comparisons for each "active" treatment with a control, and then how to make all possible comparisons. The methods are very similar.

Dunnett Method for Comparing Treatments With a Control

In many situations, researchers would like to compare each treatment with the control (when there is a natural control group). Here, the goal is to make all comparisons of treatment groups versus control ($k-1$, in all) with an overall confidence level of $(1-\alpha_E)\,100\%$, where α_E is the **experimentwise error rate**. This is the probability of making at least one incorrect conclusion with respect to the null hypotheses among the $k-1$ comparisons when all of the null hypotheses are true.

If the control group is labeled as treatment 1, the goal is to make inferences concerning $\mu_i - \mu_1$ for $i = 2, \ldots, k$. Special tables, containing Dunnett's critical values are available on the powerpoint slides corresponding to this chapter and values can also be obtained in certain packages in R. The key components are: (1) the number of comparisons $(k-1)$, (2) the error degrees of freedom $\nu = (n_{\cdot} - k)$ for the CRD, and (3) whether the test is 1- or 2-sided. In testing whether the individual treatments differ from the control, the tests can be 2-sided or 1-sided (if 1-sided, the direction must be chosen before observing the sample data). Note that these are similar to 2-sample t-tests with the exception that $k-1$ tests are being conducted simultaneously.

The tests (and simultaneous Confidence Intervals) are conducted making use of the following quantities.

$$\text{2-sided: } D_{\text{2-sided}} = d_{\alpha_E/2,k-1,\nu}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_1}\right)} \qquad \text{1-sided: } D_{\text{1-sided1}} = d_{\alpha_E,k-1,\nu}\sqrt{MSE\left(\frac{1}{n_i}+\frac{1}{n_1}\right)}$$

For 2-sided tests, conclude that treatment i mean is significantly different from the control mean if $|\overline{y}_{i\cdot} - \overline{y}_{1\cdot}| \geq D_{\text{2-sided}}$.

For 1-sided (Trt > Control) tests, conclude that treatment i mean is significantly higher than the control if $\overline{y}_{i\cdot} - \overline{y}_{1\cdot} \geq D_{\text{1-sided}}$.

For 1-sided (Trt < Control) tests, conclude that treatment i mean is significantly lower than the control if $\overline{y}_{i\cdot} - \overline{y}_{1\cdot} \leq -D_{\text{1-sided}}$.

Simultaneous Confidence Intervals can also be computed, and give the same conclusions as the corresponding tests.

$$\text{2-sided:} \quad (\overline{y}_{i\cdot} - \overline{y}_{1\cdot}) \pm D_{\text{2-sided}}$$

$$\text{Upper 1-sided:} \quad \left[(\overline{y}_{i\cdot} - \overline{y}_{1\cdot}) - D_{\text{1-sided}}, \infty\right] \qquad \text{Lower 1-sided:} \quad \left[-\infty, (\overline{y}_{i\cdot} - \overline{y}_{1\cdot}) + D_{\text{1-sided}}\right]$$

Based on each confidence interval, it can be determined whether the treatment differs from the control by determining whether or not 0 is included in the interval.

Example 7.6: Comparisons of 4 Mosquito Repellents with a Control

Treating Odomos as the control condition, $k-1 = 5-1 = 4$ and $\nu = n_{\cdot} - k = 150-5 = 145$, the 2-tailed critical value is 2.468 (for $\alpha_E = 0.05$ and $df = 150$). The 1-tailed critical value is 2.178. The standard error for the difference between means i and 1 for each treatment, as well as the form of the Confidence Intervals, are given below.

$$\hat{SE}\{\overline{Y}_{i\cdot} - \overline{Y}_{1\cdot}\} = \sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_1}\right)} = \sqrt{10.308\left(\frac{1}{30} + \frac{1}{30}\right)} = 0.829 \qquad D_{\text{2-sided}} = 2.468(0.829) = 2.046$$

$$95\%\text{ CI for } \mu_i - \mu_1\text{: } (\overline{y}_{i\cdot} - \overline{y}_{1\cdot}) \pm 2.46(0.829) \equiv (\overline{y}_{i\cdot} - \overline{y}_{1\cdot}) \pm 2.046$$

$$\overline{y}_{2\cdot} - \overline{y}_{1\cdot} = 8.133 - 7.900 = 0.233 \qquad \overline{y}_{3\cdot} - \overline{y}_{1\cdot} = 8.033 - 7.900 = 0.133$$

$$\overline{y}_{4\cdot} - \overline{y}_{1\cdot} = 6.333 - 7.900 = -1.567 \qquad \overline{y}_{5\cdot} - \overline{y}_{1\cdot} = 5.367 - 7.900 = -2.533$$

$$\mu_2 - \mu_1 : 0.233 \pm 2.046 \equiv (-1.813, 2.279) \quad \mu_3 - \mu_1 : 0.133 \pm 2.046 \equiv (-1.913, 2.179)$$

$$\mu_4 - \mu_1 : -1.567 \pm 2.046 \equiv (-3.613, 0.479) \quad \mu_5 - \mu_1 : -2.533 \pm 2.046 \equiv (-4.579, -0.483)$$

Based on the 2-sided Confidence Intervals, only treatment 5 is significantly different from the "control" treatment. The form of the tests for (lower) 1-tailed tests showing reduction of mosquito contacts would be as follows, where $D_{\text{1-sided}} = 2.178(0.829) = 1.806$.

$$\text{Conclude } H_A : \mu_i - \mu_1 < 0 \text{ if } \overline{y}_{i\cdot} - \overline{y}_{1\cdot} \leq -D_{\text{1-sided}} = -1.806$$

The same conclusions hold as did for the 2-tailed Confidence Intervals (this will not generally be the case).

The R commands below make use of the **multcomp** package and extend the program from Example 7.4.

R Commands and Output

```
## Commands
trt.mosq <- factor(trt.mosq)

mosq.mod1 <- aov(y ~ trt.mosq)
anova(mosq.mod1)

library(multcomp)
mosq.dunnett <- glht(mosq.mod1, linfct=mcp(trt.mosq="Dunnett"))
summary(mosq.dunnett)
confint(mosq.dunnett)

## Output
> summary(mosq.dunnett)

         Simultaneous Tests for General Linear Hypotheses

Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = y ~ trt.mosq)

Linear Hypotheses:
           Estimate Std. Error t value Pr(>|t|)
2 - 1 == 0    0.233      0.829   0.281  0.99579
3 - 1 == 0    0.133      0.829   0.160  0.99953
4 - 1 == 0   -1.567      0.829  -1.890  0.18449
5 - 1 == 0   -2.533      0.829  -3.056  0.00971 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
(Adjusted p values reported -- single-step method)

> confint(mosq.dunnett)

         Simultaneous Confidence Intervals
Multiple Comparisons of Means: Dunnett Contrasts
Fit: aov(formula = y ~ trt.mosq)

Quantile = 2.4695
95\% family-wise confidence level
 Linear Hypotheses:
           Estimate lwr     upr
2 - 1 == 0  0.2330  -1.8142  2.2802
3 - 1 == 0  0.1330  -1.9142  2.1802
4 - 1 == 0 -1.5670  -3.6142  0.4802
5 - 1 == 0 -2.5330  -4.5802 -0.4858
```

∇

Bonferroni Method of Multiple Comparisons

The Bonferroni method is used in many situations and is based on the following premise: If there are c comparisons to be made simultaneously, and desire to be $(1 - \alpha_E)\,100\%$ confident that all are correct, each comparison should be made at a higher level of confidence (lower probability of type I error). If individual

comparisons are made at $\alpha_I = \alpha_E/c$ level of significance, there is an overall error rate no larger than α_E. This method is conservative and can run into difficulties (low power) as the number of comparisons gets very large. The general procedures for simultaneous tests and Confidence Intervals are as follow in terms of comparing pairs of treatment means.

$$\text{Define:} \quad B_{ii'} = t_{\alpha_E/(2c),\nu}\hat{SE}\{\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}\} = t_{\alpha_E/(2c),\nu}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)} \quad i < i'$$

Conclude $\mu_i \neq \mu_{i'}$ if $|\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}| \geq B_{ii'}$ Simultaneous $(1-\alpha_E)\,100\%$ CI's for $\mu_i - \mu_{i'}$: $(\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}) \pm B_{ii'}$

where $t_{\alpha_E/(2c),\nu}$, with ν being the error degrees of freedom, $\nu = n_{\cdot} - k$ for the Completely Randomized Design, is obtained from the Bonferroni t–table (see chapter powerpoint slides) or from statistical packages or EXCEL.

Tukey Method for All Pairwise Comparisons

Various methods have been developed to handle all possible comparisons and keep the overall error rate at α_E, including the widely reported Bonferroni procedure described above. Another commonly used procedure is Tukey's Honest Significant Difference method, which is more powerful than the Bonferroni method (but more limited in its applicability). Statistical computer packages can make these comparisons automatically. Tukey's method can be used for tests and confidence intervals for all pairs of treatment means simultaneously. If there are k treatments, their will be $\frac{k(k-1)}{2}$ such tests or intervals. The general forms, allowing for different sample sizes for treatments i and i' are as follow (the unequal sample size case is referred to as the "Tukey-Kramer" method). The procedure makes use of the **Studentized Range Distribution** with critical values, $q_{\alpha_E,k,\nu}$, indexed by the number of treatments (k) and error degrees of freedom $\nu = n_{\cdot} - k$ for the Completely Randomized Design. The R functions **qtukey** and **ptukey** in R give quantiles and probabilities for the distribution. A table of critical values for $\alpha_E = 0.05$ is given in this chapter's powerpoint slides.

$$\text{Define:} \quad HSD_{ii'} = \frac{q_{\alpha_E,k,\nu}}{2}\hat{SE}\{\overline{Y}_{i\cdot} - \overline{Y}_{i'\cdot}\} = \frac{q_{\alpha_E,k,\nu}}{2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_{i'}}\right)} \quad i < i'$$

Conclude $\mu_i \neq \mu_{i'}$ if $|\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}| \geq HSD_{ii'}$ Simultaneous $(1-\alpha_E)\,100\%$ CI's for $\mu_i - \mu_{i'}$: $(\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}) \pm HSD_{ii'}$

When the sample sizes are equal ($n_i = n_{i'}$), the formula for $HSD_{ii'}$ can be simplified as follows.

$$HSD_{ii'} = q_{\alpha_E,k,\nu}\sqrt{MSE\left(\frac{1}{n_i}\right)} \quad i < i'$$

Example 7.7: Comparison of 5 Mosquito Repellents

The Bonferroni and Tukey methods are used to obtain simultaneous 95% CI's for each difference in mean mosquito contacts. The general form for the Bonferroni simultaneous 95% CI's (with $c = 5(4)/2 = 10$ and $\nu = 150 - 4 = 145$)) is given below. Recall that $MSE = 10.308$ and $n_i = 30$ for each treatment.

$$B_{ii'} = t_{.05/(2(10)),145}\sqrt{10.38\left(\frac{1}{30} + \frac{1}{30}\right)} = 2.851(0.829) = 2.363 \qquad \text{Simultaneous 95\% CIs: } (\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}) \pm 2.363$$

For Tukey's method, the confidence intervals are of the following form.

$$HSD_{ii'} = q_{0.05,5,145}\sqrt{10.308\left(\frac{1}{30}\right)} = 3.907(0.586) = 2.290 \qquad \text{Simultaneous 95\% CIs: } (\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}) \pm 2.290$$

The corresponding confidence intervals are given in Table 7.5.

		Simultaneous 95% CI's	
Comparison	$\overline{y}_{i\cdot} - \overline{y}_{i'\cdot}$	Bonferroni	Tukey
1 vs 2	$7.900 - 8.133 = -0.233$	$(-2.596, 2.130)$	$(-2.523, 2.057)$
1 vs 3	$7.900 - 8.033 = -0.133$	$(-2.496, 2, 230)$	$(-2.423, 2.157)$
1 vs 4	$7.900 - 6.333 = 1.567$	$(-0.796, 3.930)$	$(-0.723, 3.857)$
1 vs 5	$7.900 - 5.367 = 2.533$	$(0.170, 4.896)$	$(0.243, 4.823)$
2 vs 3	$8.133 - 8.033 = 0.100$	$(-2.263, 2.463)$	$(-2.190, 2.390)$
2 vs 4	$8.133 - 6.333 = 1.800$	$(-0.563, 4.163)$	$(-0.490, 4.090)$
2 vs 5	$8.133 - 5.367 = 2.766$	$(0.403, 5.129)$	$(0.476, 5.056)$
3 vs 4	$8.033 - 6.333 = 1.700$	$(-0.663, 4.063)$	$(-0.590, 3.990)$
3 vs 5	$8.033 - 5.367 = 2.666$	$(0.303, 5.029)$	$(0.376, 4.956)$
4 vs 5	$6.333 - 5.367 = 0.966$	$(-1.397, 3.329)$	$(-1.324, 3.256)$

Table 7.5: Bonferroni and Tukey multiple comparisons for the mosquito repellent study

Based on the intervals in Table 7.5, it can be concluded that treatments 1 (Odomos) and 5 (Cyfluthrin + Odomos) are significantly different, as are treatments 2 (Deltamethrin) and 5, and treatments 3 (Cyfluthrin) and 5.

While it is easy to write a function in R to conduct the Bonferroni method, there does not appear an easy "follow up" to the ANOVA. There is an easy one for Tukey's honest significant difference method, the **TukeyHSD** function. Note that R takes the mean with the higher subscript minus the mean with the lower subscript.

R Commands and Output

```
## Commands

### Tukey follow-up to 1-Way ANOVA
mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

trt.mosq <- factor(trt.mosq)
mosq.mod1 <- aov(y.mosq ~ trt.mosq)
anova(mosq.mod1)
TukeyHSD(mosq.mod1, "trt.mosq")

### Output
> TukeyHSD(mosq.mod1, "trt.mosq")
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = y ~ trt.mosq)
```

```
$trt.mosq
      diff       lwr        upr     p adj
2-1  0.233 -2.056985  2.5229848 0.9986197
3-1  0.133 -2.156985  2.4229848 0.9998497
4-1 -1.567 -3.856985  0.7229848 0.3272067
5-1 -2.533 -4.822985 -0.2430152 0.0221285
3-2 -0.100 -2.389985  2.1899848 0.9999517
4-2 -1.800 -4.089985  0.4899848 0.1965250
5-2 -2.766 -5.055985 -0.4760152 0.0093768
4-3 -1.700 -3.989985  0.5899848 0.2474716
5-3 -2.666 -4.955985 -0.3760152 0.0136760
5-4 -0.966 -3.255985  1.3239848 0.7710691

> bon.ci(0.05, y.mosq, trt.mosq)
      Trt i Trt i'   Diff Lower Bound Upper Bound p adjusted
 [1,]     1      2 -0.232      -2.595       2.130      1.000
 [2,]     1      3 -0.132      -2.495       2.231      1.000
 [3,]     1      4  1.567      -0.795       3.930      0.606
 [4,]     1      5  2.534       0.171       4.896      0.027
 [5,]     2      3  0.100      -2.262       2.463      1.000
 [6,]     2      4  1.800      -0.563       4.162      0.315
 [7,]     2      5  2.766       0.403       5.129      0.011
 [8,]     3      4  1.699      -0.663       4.062      0.421
 [9,]     3      5  2.666       0.303       5.028      0.016
[10,]     4      5  0.966      -1.396       3.329      1.000
```

∇

Unequal Variances - Welch's Test

When the variances are unequal among the populations, Welch's test, considered in the 2-sample case in Chapter 5 can be extended to $k > 2$ groups. The test involves making an adjustment to a weighted between treatments sum of squares, and the error degrees of freedom, for an approximate F-test for differences among the treatment means. Let n_i be the sample size for treatment i, $\overline{y}_{i\cdot}$ be the sample mean, and s_i be the sample standard deviation.

$$w_i = \frac{n_i}{s_i^2} \quad w_\cdot = \sum_{i=1}^{k} w_i \qquad F^* = \frac{1}{k-1}\left[\sum_{i=1}^{k} w_i \overline{y}_{i\cdot}^2 - \frac{\left(\sum_{i=1}^{k} w_i \overline{y}_{i\cdot}\right)^2}{w_\cdot}\right]$$

$$C_W = \sum_{i=1}^{k}\left[\frac{1}{n_i - 1}\left(1 - \frac{w_i}{w_\cdot}\right)^2\right] \qquad m_W = \left[1 + \frac{2(k-2)}{k^2-1} C_W\right]^{-1} \qquad \nu_W = \left[\frac{3}{k^2-1} C_W\right]^{-1}$$

$$\Rightarrow \quad F_W = m_W F^* \dot{\sim} F_{k-1,\nu_W}$$

The null hypothesis of no treatment effects is rejected if $F_W \geq F_{\alpha,k-1,\nu_W}$.

When the variances are not all equal, one approach to making all pairwise comparisons is the **Games-Howell** procedure. It combines Tukey's method based on critical values from the studentized range distribution with Satterthwaite's approximation for degrees of freedom among pairs of treatment means.

$$\nu_{ii'} = \frac{\left[\frac{s_i^2}{n_i} + \frac{s_{i'}^2}{n_{i'}}\right]^2}{\left[\frac{(s_i^2/n_i)^2}{n_i - 1} + \frac{(s_{i'}^2/n_{i'})^2}{n_{i'} - 1}\right]}$$

$$GH_{ii'} = \frac{q_{\alpha_E,k,\nu_{ii'}}}{\sqrt{2}}\hat{SE}\{\overline{Y}_{i.} - \overline{Y}_{i'.}\} = \frac{q_{\alpha_E,k,\nu_{ii'}}}{\sqrt{2}}\sqrt{\frac{s_i^2}{n_i} + \frac{s_{i'}^2}{n_{i'}}}$$

Conclude $\mu_i \neq \mu_{i'}$ if $|\overline{y}_i - \overline{y}_{i'}| \geq GH_{ii'}$ Simultaneous $(1-\alpha_E)\,100\%$ CI's for $\mu_i - \mu_{i'}$: $(\overline{y}_{i.} - \overline{y}_{i'.}) \pm GH_{ii'}$

Example 7.8: Geographical Differences in Sea Lion Barking Acoustics

A study compared sea lion barking acoustics for $k = 7$ locations in Australia (Ahonen, et al. (2014), [4]). The summary data, as well as calculations needed to compute Welch's F-statistic are given in Table 7.6. The authors reported data for 10 acoustic variables, we consider duration (ms). There were 7-10 sea lions at the various locations, and each was observed for 20 barks. No random sea lion effects were considered for the model, had they been, this could be analyzed as a nested design. Bartlett's test for equal variances (computations not shown here) gives a very large chi-square statistic $X_B^2 = 348.4$ with $7 - 1 = 6$ degrees of freedom. There is strong evidence of unequal variances among the locations.

Colony (i)	n_i	$\overline{y}_{i.}$	s_i	w_i	$w_i\overline{y}_{i.}$	$w_i\overline{y}_{i.}^2$	C_{Wi}
Lewis Island (1)	200	55	19	0.5540	30.4709	1675.9003	0.003297
Liguanea Island (2)	160	62	30	0.1778	11.0222	683.3778	0.005546
Olive Island (3)	200	69	17	0.6920	47.7509	3294.8097	0.002923
Blefuscu Island (4)	200	51	14	1.0204	52.0408	2654.0816	0.002124
Lilliput Island (5)	160	61	44	0.0826	5.0413	307.5207	0.005938
North Fisherman Island (6)	140	73	25	0.2240	16.3520	1193.6960	0.006131
Beagle Island (7)	180	77	33	0.1653	12.7273	980.0000	0.004971
Sum	1240	#N/A	#N/A	2.9162	175.4054	10789.3860	0.03093

Table 7.6: Sea Lion barking acoustic duration data and calculations for Welch's F-test

$$F^* = \frac{1}{7-1}\left[10789.3860 - \frac{(175.4054)^2}{2.9162}\right] = 39.832 \quad C_W = 0.03093 \quad m_W = \left[1 + \frac{2(7-2)}{7^2-1}(0.03093)\right]^{-1} = 0.9936$$

$$\nu_W = \left[\frac{3}{7^2-1}(0.03093)\right]^{-1} = 517.3$$

$$TS: F_W = 0.9936(39.832) = 39.577 \qquad RR: F_W \geq F_{.05,6,517.3} = 2.116 \qquad P = P\left(F_{6,517.3} \geq 39.577\right) = .0000$$

The following R commands conduct Bartlett's test and Welch's test directly on data that was generated to match the location means and SDs and allowed no negative responses.

R Commands and Output

```
## Commands

sb <- read.csv("http://www.stat.ufl.edu/~winner/data/sealion_bark.csv")
attach(sb); names(sb)

location <- factor(location)

bartlett.test(duration ~ location)
oneway.test(duration ~ location, var.equal=F)


### Output

> bartlett.test(duration ~ location)
        Bartlett test of homogeneity of variances
data:  duration by location
Bartlett's K-squared = 348.4, df = 6, p-value < 2.2e-16

> oneway.test(duration ~ location, var.equal=F)
        One-way analysis of means (not assuming equal variances)
data:  duration and location
F = 39.562, num df = 6.00, denom df = 517.29, p-value < 2.2e-16
```

The calculation for comparing the first two locations (Lewis and Liguanea Islands) is given here. R Commands for all 7(6)/2=21 pairs is given below.

$$\overline{y}_{1.} - \overline{y}_{2.} = 55 - 62 = -7 \qquad \hat{SE}\left\{\overline{Y}_{1.} - \overline{Y}_{2.}\right\} = \sqrt{\frac{19^2}{200} + \frac{30^2}{160}} = \sqrt{7.43} = 2.726$$

$$\nu_{12} = \frac{\left[\frac{19^2}{200} + \frac{30^2}{160}\right]^2}{\left[\frac{(19^2/200)^2}{200-1} + \frac{(30^2/160)^2}{160-1}\right]} = \frac{7.43^2}{0.2154} = 256.33 \qquad \frac{q_{.05,7,256.33}}{\sqrt{2}} = \frac{4.203}{\sqrt{2}} = 2.972$$

$$95\% \text{ CI for } \mu_1 - \mu_2 : -7 \pm 2.972(2.726) \equiv -7 \pm 8.10 \equiv (-15.10, 1.10)$$

R Output

```
## Output

> (loc_mean <- as.vector(tapply(duration,location,mean)))
[1] 55.00015 62.00025 68.99985 51.00020 61.00006 72.99957 76.99972
> (loc_var <- as.vector(tapply(duration,location,var)))
[1]  361.0088  899.9931  289.0005  196.0056 1935.9785  625.0147 1088.9722
> (loc_n <- as.vector(tapply(duration,location,length)))
[1] 200 160 200 200 160 140 180

> round(gh.out,3)
     Trt i Trt j    Diff    SE      DF Lower Bound Upper Bound
```

```
 [1,]    1    2  -7.000 2.726 256.329    -15.101      1.101
 [2,]    1    3 -14.000 1.803 393.175    -19.343     -8.657
 [3,]    1    4   4.000 1.669 365.893     -0.948      8.948
 [4,]    1    5  -6.000 3.729 206.309    -17.103      5.104
 [5,]    1    6 -17.999 2.504 246.031    -25.443    -10.556
 [6,]    1    7 -22.000 2.803 279.378    -30.323    -13.676
 [7,]    2    3  -7.000 2.659 238.603    -14.906      0.907
 [8,]    2    4  11.000 2.570 214.040      3.350     18.650
 [9,]    2    5   1.000 4.210 280.561    -11.503     13.504
[10,]    2    6 -10.999 3.176 297.314    -20.429     -1.570
[11,]    2    7 -14.999 3.417 337.824    -25.135     -4.864
[12,]    3    4  18.000 1.557 383.887     13.384     22.615
[13,]    3    5   8.000 3.680 196.999     -2.964     18.964
[14,]    3    6  -4.000 2.431 226.937    -11.232      3.232
[15,]    3    7  -8.000 2.738 261.311    -16.135      0.135
[16,]    4    5 -10.000 3.617 184.831    -20.781      0.782
[17,]    4    6 -21.999 2.333 199.993    -28.949    -15.049
[18,]    4    7 -26.000 2.651 236.117    -33.885    -18.114
[19,]    5    6 -12.000 4.070 257.826    -24.094      0.095
[20,]    5    7 -16.000 4.260 292.740    -28.648     -3.351
[21,]    6    7  -4.000 3.243 317.798    -13.622      5.622
```

∇

7.1.2 Test Based on Non–Normal Data

A nonparametric test for the Completely Randomized Design (CRD), where each experimental unit receives only one treatment, is the **Kruskal-Wallis Test**. The idea behind the test is similar to that of the Wilcoxon Rank Sum test. The main difference is that instead of comparing 2 population distributions, we are comparing $k \geq 2$ distributions. Sample measurements are ranked from 1 (smallest) to $n_\cdot = n_1 + \cdots + n_k$ (largest), with ties being replaced with the means of the ranks the tied subjects would have received had they not tied. For each treatment, the sum of the ranks of the sample measurements are computed, and labeled T_i. The sample size from the i^{th} treatment is n_i, and the total sample size is $n_\cdot = n_1 + \cdots + n_k$.

The hypothesis we wish to test is whether the k population distributions are identical (or that all medians are equal) against the alternative that some distribution(s) is (are) shifted to the right of other(s). This is similar to the hypothesis of no treatment effect that we tested in the previous section. The test is conducted as follows.

1. H_0 : The k population distributions are identical ($M_1 = M_2 = \cdots = M_k$)
2. H_A : Not all k distributions are identical (Not all M_i are equal)
3. T.S.: $H = \frac{12}{n_\cdot(n_\cdot+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n_\cdot + 1)$.
4. R.R.: $H > \chi^2_{\alpha,k-1}$
5. P-value: $P\left(\chi^2_{k-1} \geq H\right)$

If we do reject H_0, and conclude treatment differences exist, we could run the Wilcoxon Rank Sum test on all pairs of treatments, adjusting the individual α levels by taking α_E/c where c is the number of

comparisons, so that the overall test (on all pairs) has a significance level of α_E. This is an example of the Bonferroni procedure.

An alternative approach is to use the rank sums from the Kruskal-Wallis test directly. If the Kruskal-Wallis test does not reject the null hypothesis, stop and do not make any pairwise comparisons. Otherwise, compare the differences in the rank averages as follows, where $q_{\alpha_E,k,\infty}$ is the critical value from the studentized range distribution with k groups and degrees of freedom equal to infinity.

$$\text{Conclude Medians differ if: } |\overline{T}_i - \overline{T}_{i'}| \geq \frac{q_{\alpha_E,k,\infty}}{\sqrt{2}} \sqrt{\frac{n_\cdot (n_\cdot + 1)}{12} \left(\frac{1}{n_i} + \frac{1}{n_{i'}} \right)}$$

Example 7.9: Mechanical Properties and Anthocyanins Extractability of Grape Berries

A study was conducted to compare $k = 6$ treatments based on sugar content of berries with respect to various physical and mechanical properties (Zouid, et al (2013), [54]. The six treatments were sugar equivalents (g/L) of 176.5, 192.6, 209.3, 225.0, 242.1, and 258.5. One response observed was anthocyanin extractability (in percent). Data have been generated to match the authors means and standard deviation, with $n_i = 15$ replicates per treatment. Data and ranks are given in Table 7.7.

j	Trt1	Trt2	Trt3	Trt4	Trt5	Trt6	Rank1	Rank2	Rank3	Rank4	Rank5	Rank6
1	91.1514	90.4136	92.7392	94.7763	94.4086	94.5816	14	8	31	74	63	66
2	91.2250	92.9450	94.5272	94.6568	94.0188	94.6527	15	35	65	68	57	67
3	91.2764	91.4339	92.3321	94.7164	95.7081	93.3328	16	18	27	71	84	43
4	90.2144	91.8134	93.3022	93.3512	94.1216	93.1514	7	20	41	44	59	38
5	90.0281	89.1463	91.8391	93.7673	93.7188	94.5271	5	2	21	51	50	64
6	89.5134	93.2797	93.6013	91.9625	96.6605	95.0653	3	40	47	22	89	77
7	91.4120	93.8751	92.0830	92.0927	94.8797	94.1368	17	55	24	25	75	60
8	94.2134	90.6342	95.5857	92.8051	95.2221	96.9886	61	10	83	33	78	90
9	92.0028	96.2767	93.5188	92.5493	93.8019	95.8473	23	87	46	30	53	85
10	93.9741	95.4191	93.1780	90.9968	95.4145	96.6470	56	82	39	13	81	88
11	90.1335	94.7650	94.7178	94.3609	95.3483	93.1055	6	73	72	62	80	37
12	89.6315	90.9415	90.9408	93.7884	92.4709	93.8121	4	12	11	52	29	54
13	90.5289	92.8002	93.4466	93.3178	92.8295	94.7117	9	32	45	42	34	70
14	92.1393	94.0834	92.9888	95.3392	94.6859	95.0527	26	58	36	79	69	76
15	88.6058	91.7731	93.6493	92.3693	96.0107	93.6875	1	19	48	28	86	49
Mean	91.07	92.64	93.23	93.39	94.62	94.62	17.53	36.73	42.40	46.27	65.80	64.27
SD	1.56	2.01	1.18	1.25	1.16	1.18	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A
Sum	#N/A	#N/A	#N/A	#N/A	#N/A	#N/A	263	551	636	694	987	964

Table 7.7: Anthocyanin Extraction of Berries for $k = 6$ sugar contents

The Kruskal-Wallis test is conducted to determine whether the extraction distributions are significantly different among the six groups. The test is conducted at the $\alpha = 0.05$ significance level.

1. H_0 : The 6 population medians are identical ($M_1 = \cdots = M_6$)
2. H_A : Not all 6 medians are identical (Not all M_i are equal)

3. T.S.: $H = \frac{12}{n_\cdot(n_\cdot+1)}\sum_{i=1}^{k}\frac{T_i^2}{n_i} - 3(n_\cdot+1) = \frac{12}{90(91)}\left(\frac{(263)^2}{15}+\frac{(551)^2}{15}+\frac{(636)^2}{15}+\frac{(694)^2}{15}+\frac{(987)^2}{15}+\frac{(964)^2}{15}\right) - 3(91) = 308.90 - 273 = 35.90.$
4. R.R.: $H \geq \chi^2_{\alpha,k-1} = \chi^2_{.05,5} = 11.071$
5. P–value: $P(\chi^2_5 \geq 35.90) = .0000$

Reject H_0, and conclude differences exist. Based on the high rank sums for the two highest sugar content treatments, they appear to have higher anthocyanin extractability than the other treatments. The critical value for comparing mean ranks is computed below, with $q_{.05,6,\infty} = 4.030$.

$$\frac{q_{\alpha_E,\infty}}{\sqrt{2}}\sqrt{\frac{n_\cdot\,(n_\cdot+1)}{12}\left(\frac{1}{n_i}+\frac{1}{n_{i'}}\right)} = \frac{4.030}{\sqrt{2}}\sqrt{\frac{90\,(91)}{12}\left(\frac{1}{15}+\frac{1}{15}\right)} = 2.850\sqrt{91} = 27.19$$

Treatments 4, 5, and 6 have significantly higher medians than treatment 1, and treatments 5 and 6 have significantly higher medians than treatment 2.

R Commands and Output

```
## Commands

bt <- read.csv("http://www.stat.ufl.edu/~winner/data/berry_texture.csv")
attach(bt); names(bt)

sugar <- factor(sugar)

kruskal.test(anthExt ~ sugar)

## Output

> kruskal.test(anthExt ~ sugar)
        Kruskal-Wallis rank sum test
data:  anthExt by sugar
Kruskal-Wallis chi-squared = 35.9, df = 5, p-value = 9.944e-07
```

∇

7.2 Randomized Block Design (RBD) For Studies Based on Matched Units

In crossover designs (aka within subjects designs), each unit or subject receives each treatment. In these cases, units are referred to as **blocks**. In other studies, units or subjects may be matched based on external criteria. The notation for the Randomized Block Design (RBD) is very similar to that of the CRD, with a additional elements. The model we are assuming here is written as follows.

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} = \mu_i + \beta_j + \epsilon_{ij} \quad i = 1,\ldots,k; \quad j = 1,\ldots,b$$

Here, μ represents the overall mean measurement, α_i is the (**fixed**) effect of the i^{th} treatment, β_j is the (typically **random**) effect of the j^{th} block, and ϵ_{ij} is a random error component that can be thought of as the variation in measurements if the same experimental unit received the same treatment repeatedly. Note that just as before, μ_i represents the mean measurement for the i^{th} treatment (across all blocks). The general situation will consist of an experiment with k treatments being received by each of b blocks. Blocks can be fixed or random, typically they are random.

7.2.1 Test Based on Normally Distributed Data

When the (random) block effects (β_j) and random error terms (ϵ_{ij}) are independent and normally distributed, an F–test is conducted that is similar to that described for the Completely Randomized Design, but with an extra source of variation. If blocks are fixed, the analysis is the same. The notation used is as follows.

$$\begin{aligned}
\overline{y}_{i.} &= \frac{\sum_{j=1}^{b} y_{ij}}{b} \\
\overline{y}_{.j} &= \frac{\sum_{i=1}^{k} y_{ij}}{k} \\
n_{.} &= b \cdot k \\
\overline{y}_{..} &= \frac{\sum_{i=1}^{k}\sum_{j=1}^{b} y_{ij}}{bk} \\
TSS &= \sum_{i=1}^{k}\sum_{j=1}^{b} \left(y_{ij} - \overline{y}_{..}\right)^2 \\
SST &= \sum_{i=1}^{k} b\left(\overline{y}_{i.} - \overline{y}_{..}\right)^2 \\
SSB &= \sum_{j=1}^{b} k\left(\overline{y}_{.j} - \overline{y}_{..}\right)^2 \\
SSE &= \sum_{i=1}^{k}\sum_{j=1}^{b} \left(y_{ij} - \overline{y}_{i.} - \overline{y}_{.j} + \overline{y}_{..}\right)^2
\end{aligned}$$

Note that the Analysis of Variance simply has added items representing the block means $\left(\overline{y}_{.j}\right)$ and variation among the block means (SSB). We can further think of this as decomposing the total variation into differences among the treatment means (SST), differences among the block means (SSB), and random variation not explained by either differences among treatment or block means (SSE). Also, note that $SSE = TSS - SST - SSB$.

Once again, the main purpose for conducting this type of experiment is to detect differences among the treatment means (treatment effects). The test is very similar to that of the CRD, with only minor adjustments. We often are not interested in testing for differences among blocks, since we expect there to be differences among them (that's why the design was set up this way), and they were just a random sample from a population of such experimental units. However, in some cases, estimating the unit to unit (subject to subject) variance component is of interest. The testing procedure can be described as follows.

1. $H_0 : \alpha_1 = \cdots = \alpha_k = 0 \quad (\mu_1 = \cdots = \mu_k)$ (No treatment effect)

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	SST	$k-1$	$MST = \frac{SST}{k-1}$	$F = \frac{MST}{MSE}$
BLOCKS	SSB	$b-1$	$MSB = \frac{SSB}{b-1}$	
ERROR	SSE	$(b-1)(k-1)$	$MSE = \frac{SSE}{(b-1)(k-1)}$	
TOTAL	TSS	$bk-1$		

Table 7.8: The Analysis of Variance Table for the Randomized Block Design

2. H_A : Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,k-1,(b-1)(k-1)}$
5. p-value: $P(F_{k-1,(b-1)(k-1)} \geq F_{obs})$

The procedures to make comparisons among means are very similar to the methods used for the CRD. In each formula described previously for Scheffe's, Dunnett's, Bonferroni's, and Tukey's methods, n_i is replaced by b, when making comparisons among treatment means, and $\nu = (b-1)(k-1)$ is the error degrees of freedom.

The **Relative Efficiency** of conducting the Randomized Block Design, as opposed to the Completely Randomized Design is:

$$RE(RB, CR) = \frac{MSE_{CR}}{MSE_{RB}} = \frac{(b-1)MSB + b(t-1)MSE}{(bt-1)MSE}.$$

This represents the number of times as many replicates would be needed for each treatment in a CRD to obtain as precise of estimates of differences between two treatment means as were obtained by using b experimental units per treatment in the RBD. It measures reduction in experimental error due to using the block design.

Example 7.10: Comparison of 3 Methods for Estimating Value of Wood Logs

A study compared 3 methods of assessing the lumber value of logs (Lin and Wang (2012), [36]). The $k = 3$ treatments the authors compared was the actual sawmill value of the log, a value based on a heuristic programming algorithm, and a value based on a dynamic programming algorithm. Each "treatment" was measured on $b = 30$ logs (acting as the blocks). The goal was to compare the 3 treatments at valuating the

logs. Data are given in Table 7.9. A crude interaction plot is given in Figure 7.4, which plots the valuation versus log ID, with seperate lines for the three methods.

Log ID	Actual	Heuristic	Dynamic	LogMean
1	17.67	20.83	21.03	19.8433
2	31.76	35.05	34.24	33.6833
3	30.77	33.60	34.87	33.0800
4	40.27	42.52	42.89	41.8933
5	33.51	35.06	36.48	35.0167
6	23.07	25.37	26.34	24.9267
7	21.33	21.95	23.00	22.0933
8	26.28	28.07	28.69	27.6800
9	28.89	31.94	32.49	31.1067
10	18.46	19.14	21.76	19.7867
11	35.61	38.18	39.87	37.8867
12	23.15	25.67	27.22	25.3467
13	18.03	19.58	20.70	19.4367
14	28.22	30.89	30.05	29.7200
15	20.33	21.36	21.62	21.1033
16	12.42	13.01	14.02	13.1500
17	21.90	24.52	25.06	23.8267
18	36.16	38.12	38.86	37.7133
19	13.73	14.74	15.12	14.5300
20	15.74	17.96	18.00	17.2333
21	19.22	20.69	20.83	20.2467
22	17.12	19.12	19.31	18.5167
23	15.21	16.42	16.63	16.0867
24	22.03	23.58	24.24	23.2833
25	31.22	32.66	32.90	32.2600
26	25.69	28.39	28.81	27.6300
27	29.25	31.63	30.72	30.5333
28	32.77	33.29	35.87	33.9767
29	31.88	34.79	34.82	33.8300
30	24.54	26.23	26.54	25.7700
Trt Mean	24.8743	26.8120	27.4327	26.3730

Table 7.9: Log Values for 3 Methods of Valuation

$$TSS = (17.67 - 26.3730)^2 + \cdots + (26.54 - 26.3730)^2 = 5170.073 \qquad df = 30(3) - 1 = 89$$

$$SST = 30\left[(24.8743 - 26.3730)^2 + (26.8120 - 26.3730)^2 + (27.4327 - 26.3730)^2\right] = 106.8536 \qquad df_T = 3{-}1 = 2$$

$$SSB = 3\left[(19.8433 - 26.3730)^2 + \cdots + (25.7700 - 26.3730)^2\right] = 5042.772 \qquad df_B = 30 - 1 = 29$$

$$SSE = (17.67 - 19.8433 - 24.8743 + 26.3730)^2 + \cdots + (26.54 - 25.7700 - 26.4327 + 26.3730)^2 =$$

$$5170.073 - 106.8536 - 5042.772 = 20.448 \qquad df_E = (30 - 1)(3 - 1) = 58$$

We can now test for treatment effects, and if necessary use Tukey's method to make pairwise comparisons among the three methods ($\alpha_E = 0.05$ significance level).

1. $H_0 : \alpha_1 = \alpha_2 = \alpha_3 = 0 \quad (\mu_1 = \mu_2 = \mu_3)$ (No differences among valuation method means)

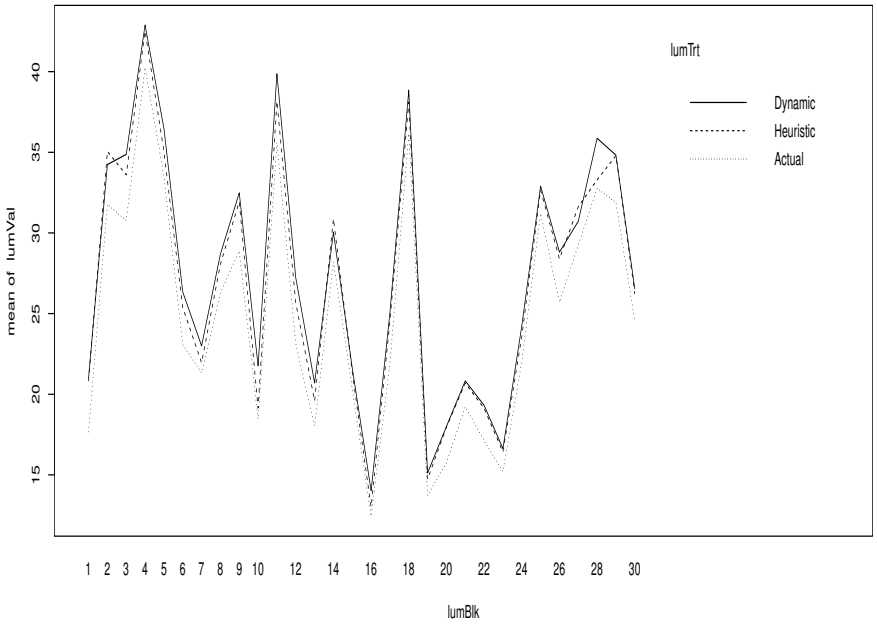

Figure 7.4: Plot of valuation versus log ID, with separate lines for valuation method

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_{obs}
TREATMENTS	106.854	2	53.427	151.546
BLOCKS	5042.772	29	173.889	
ERROR	20.448	58	0.353	
TOTAL	5170.073	89		

Table 7.10: Analysis of Variance table for log valuation data (RBD)

2. H_A : Not all α_i are 0 (Valuation differences exist)
3. T.S. $F_{obs} = \frac{MST}{MSE} = 151.546$
4. R.R.: $F_{obs} \geq F_{\alpha,k-1,(b-1)(k-1)} = F_{0.05,2,58} = 3.156$
5. P-value: $P(F_{2,58} \geq F_{obs}) = P(F_{2,58} \geq 151.546) = 0.0000$

Tukey's method to is used determine which valuations differ significantly. Recall that for Tukey's method, simultaneous confidence intervals of the form given below are computed, with k being the number of treatments (k=3), b being the number of blocks, and n_i the number of measurements per valuation method ($n_i = b = 30$).

$$(\overline{y}_{i.} - \overline{y}_{i'.}) \pm q_{\alpha,k,(b-1)(k-1)}\sqrt{MSE\left(\frac{1}{n_i}\right)} \quad \Longrightarrow \quad (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 3.402\sqrt{0.353\left(\frac{1}{30}\right)} \quad \Longrightarrow (\overline{y}_{i.} - \overline{y}_{i'.}) \pm 0.369$$

The corresponding simultaneous 95% confidence intervals and conclusions are given in Table 7.11. Conclude

Comparison	$\overline{y}_{i.} - \overline{y}_{i'.}$	CI	Conclusion
Actual vs Heuristic	$24.874 - 26.812 = -1.938$	$(-2.307, -1.569)$	$\mu_A < \mu_H$
Actual vs Dynamic	$24.874 - 27.433 = -2.559$	$(-2.928, -2.190)$	$\mu_A < \mu_D$
Heuristic vs Dynamic	$26.812 - 27.433 = -0.621$	$(-0.990, -0.252)$	$\mu_H < \mu_D$

Table 7.11: Tukey's simultaneous 95% CI's for wood log valuation data (RBD)

that Actual sawmill valuation is significantly lower than Heuristic, which is significantly lower than Dynamic.

The relative efficiency of using this design as opposed to a Completely Randomized Design is obtained below

$$RE(RB, CR) = \frac{MSE_{CR}}{MSE_{RB}} = \frac{(30-1)(173.889) + 30(3-1)(0.353)}{(30(3)-1)(0.353)} = \frac{5063.96}{31.42} = 161$$

A total of 161(30)=4830 wood logs per treatment would be needed to have as precise of comparisons between treatment means if this had been conducted as a CRD (independent samples design). Blocking was very effective in this study.

Note that to run this in R, it is necessary to have a separate row for each observation, along with a treatment ID and block ID.

R Commands and Output

```
## Commands

saw <- read.csv("http://www.stat.ufl.edu/~winner/data/sawmill1.csv")
attach(saw); names(saw)
lumTrt <- factor(lumTrt)
lumBlk <- factor(lumBlk)
```

```
levels(lumTrt) <- c("Actual", "Heuristic", "Dynamic")

saw.mod1 <- aov(lumVal ~ lumTrt + lumBlk)
anova(saw.mod1)
TukeyHSD(saw.mod1, "lumTrt")

interaction.plot(lumBlk, lumTrt, lumVal)

## Output
> anova(saw.mod1)
Analysis of Variance Table
Response: lumVal
          Df Sum Sq Mean Sq F value    Pr(>F)
lumTrt     2  106.8  53.424  151.53 < 2.2e-16 ***
lumBlk    29 5042.8 173.889  493.21 < 2.2e-16 ***
Residuals 58   20.4   0.353
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> TukeyHSD(saw.mod1, "lumTrt")
  Tukey multiple comparisons of means
    95% family-wise confidence level
Fit: aov(formula = lumVal ~ lumTrt + lumBlk)
$lumTrt
                      diff       lwr       upr    p adj
Heuristic-Actual  1.9376667 1.5689056 2.3064277 0.000000
Dynamic-Actual    2.5583333 2.1895723 2.9270944 0.000000
Dynamic-Heuristic 0.6206667 0.2519056 0.9894277 0.000449
```

7.2.2 Test Based on Non-Normal Data

A nonparametric procedure that can be used to analyze data from the Randomized Block Design (RBD), where each subject or block receives each treatment is Friedman's Test. The idea behind Friedman's Test is to rank the measurements corresponding to the k treatments within each block. Then the rank sum corresponding to each treatment is obtained. This test can also be used when the data consists of preferences (ranks) by raters among k competing items.

Once the measurements are ranked within each block from 1 (smallest) to k (largest), and the rank sums $T_1, T_2, \ldots, T_k$ are computed for each treatment, the test is conducted as follows (assume b blocks are used in the experiment).

1. H_0 : The k population distributions have equal medians ($M_1 = M_2 = \cdots = M_k$)
2. H_A : Not all k distributions have identical medians (Not all M_i are equal)
3. T.S.: $F_r = \frac{12}{bk(k+1)} \sum_{i=1}^{k} T_i^2 - 3b(k+1)$.
4. R.R.: $F_r \geq \chi^2_{\alpha,k-1}$.
5. P–value:$P\left(\chi^2_{k-1} \geq F_r\right)$

Either k (the number of treatments) or b (the number of blocks) should be larger than 5 for this test to be appropriate.

When there are ties among treatments within blocks, an adjusted statistic can be computed (Hollander and Wolfe (1999), [27]).

$$F_r' = \frac{12\sum_{i=1}^{k}\left(T_i - \overline{T}\right)^2}{bk(k+1) - \frac{1}{k-1}\sum_{j=1}^{b}\left[\sum_{i=1}^{k} t_{ij}^3 - k\right]}$$

where t_{ij} is the size of the i^{th} group in j^{th} block. When there are no ties within blocks, all group sizes within blocks are 1, and the last term in the denominator is 0, and $F_r' = F_r$.

If we reject H_0, and conclude treatment effects exist, we can conduct Wilcoxon's Signed–Rank Test on all pairs of treatments (adjusting α_I for the number of comparisons being made, as in Bonferroni's method), to determine which pairs differ significantly.

Example 7.11: Comparison of 3 Methods for Estimating Volume of Wood Logs

A study compared 3 methods of measuring lumber volume of logs (Lin and Wang (2012), [36]). The $k = 3$ treatments the authors compared was the actual sawmill volume of the log, a volume based on heuristic programming algorithm, and a volume based on a dynamic programming algorithm. Each "treatment" was measured on $b = 30$ logs (acting as the blocks). The goal was to compare the 3 treatments at valuating the logs. Data are given in Table 7.12. We will test for treatment effects using Friedman's test ($\alpha = 0.05$).

1. H_0 : The 3 distributions of volumes have equal medians for the three methods ($M_1 = M_2 = M_3$)
2. H_A : The 3 distributions of volumes do not have equal medians (Not all M_i are equal)
3. T.S.: $F_r = \frac{12}{bk(k+1)}\sum_{i=1}^{k} T_i^2 - 3b(k+1) = \frac{12}{30(3)(4)}[(42.0)^2+(63.5)^2+(74.5)^2] - 3(30)(4) = 378.22 - 360 = 18.22$.
4. R.R.: $F_r \geq \chi^2_{\alpha,k-1} = \chi^2_{.05,2} = 5.99$.
5. P-value: $P\left(\chi_2^2 \geq 18.22\right) = .0001$

We reject H_0, and conclude that volume assessment method differences exist. Based on the rank sums, it appears that the Actual method provides lower volume assessments than the Heuristic and Dynamic methods. Note that there are ties in 9 of the 30 blocks. To obtain the adjusted Friedman's test statistic, make the following computations.

$$\overline{T} = 60 \qquad \sum_{j=1}^{b}\left[\sum_{i=1}^{k} t_{ij}^3 - k\right] = 21(1^3+1^3+1^3-3) + 9(1^3+2^3+0^3-3) = 21(0) + 9(6) = 54$$

$$F_r' = \frac{12\left[(42.0-60)^2 + (63.5-60)^2 + (74.5-60)^2\right]}{30(3)(4) - \frac{1}{3-1}(54)} = \frac{6558}{333} = 19.694$$

LogID	Actual	Heuristic	Dynamic	Rank1	Rank2	Rank3
1	0.090	0.092	0.093	1	2	3
2	0.123	0.121	0.121	3	1.5	1.5
3	0.135	0.140	0.141	1	2	3
4	0.152	0.156	0.153	1	3	2
5	0.126	0.131	0.133	1	2	3
6	0.101	0.105	0.107	1	2	3
7	0.091	0.092	0.093	1	2	3
8	0.115	0.117	0.120	1	2	3
9	0.121	0.123	0.129	1	2	3
10	0.094	0.092	0.093	3	1	2
11	0.174	0.175	0.173	2	3	1
12	0.139	0.139	0.140	1.5	1.5	3
13	0.135	0.135	0.145	1.5	1.5	3
14	0.153	0.154	0.156	1	2	3
15	0.123	0.118	0.118	3	1.5	1.5
16	0.092	0.100	0.099	1	3	2
17	0.146	0.143	0.141	3	2	1
18	0.163	0.162	0.168	2	1	3
19	0.096	0.100	0.100	1	2.5	2.5
20	0.092	0.096	0.097	1	2	3
21	0.127	0.128	0.135	1	2	3
22	0.120	0.132	0.131	1	3	2
23	0.111	0.117	0.116	1	3	2
24	0.138	0.135	0.145	2	1	3
25	0.176	0.183	0.182	1	3	2
26	0.096	0.102	0.102	1	2.5	2.5
27	0.115	0.116	0.116	1	2.5	2.5
28	0.137	0.140	0.141	1	2	3
29	0.110	0.116	0.116	1	2.5	2.5
30	0.103	0.110	0.110	1	2.5	2.5
Sum	#N/A	#N/A	#N/A	42	63.5	74.5

Table 7.12: Volume assessments of wood logs and ranks by 3 Methods - RBD

Based on the pairwise Wilcoxon signed-rank tests given in the R output below, there are $c = k(k-1)/2 = 3$ comparisons and the Bonferroni adjusted P-values are the minimum of 1 and and c times the individual P-values. Thus, for the comparison between Actual and Heuristic, $P_{\text{adj}} = min(1, 3(.002063)) = .0062$; for Actual versus Dynamic, $P_{\text{adj}} = min(1, 3(9.51e-05)) = .0003$; and for Heuristic versus Dynamic, $P_{\text{adj}} = min(1, 3(.05147)) = .1544$. Both Heuristic and Dynamic have significantly higher medians than Actual, Heuristic and Dynamic are not significantly different.

R Commands and Output

```
## Commands
saw <- read.csv("http://www.stat.ufl.edu/~winner/data/sawmill1.csv")
attach(saw); names(saw)

lumTrt <- factor(lumTrt)
lumBlk <- factor(lumBlk)

friedman.test(lumVol ~ lumTrt | lumBlk)

actual <- lumVol[lumTrt==1]
heuristic <- lumVol[lumTrt==2]
dynamic <- lumVol[lumTrt==3]

wilcox.test(actual, heuristic, paired=T)
wilcox.test(actual, dynamic, paired=T)
wilcox.test(heuristic, dynamic, paired=T)

## Output

> friedman.test(lumVol ~ lumTrt | lumBlk)
        Friedman rank sum test
data:  lumVol and lumTrt and lumBlk
Friedman chi-squared = 19.694, df = 2, p-value = 5.291e-05

> wilcox.test(actual, heuristic, paired=T)
        Wilcoxon signed rank test with continuity correction
data:  actual and heuristic
V = 67.5, p-value = 0.002063
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(actual, dynamic, paired=T)
        Wilcoxon signed rank test with continuity correction
data:  actual and dynamic
V = 42.5, p-value = 9.51e-05
alternative hypothesis: true location shift is not equal to 0

> wilcox.test(heuristic, dynamic, paired=T)
        Wilcoxon signed rank test with continuity correction
data:  heuristic and dynamic
V = 74, p-value = 0.05147
alternative hypothesis: true location shift is not equal to 0
```

7.3 Latin Square Designs

The Latin Square Design is an extension of the Randomized Block Design to include two blocking factors, each with the same number of levels as the primary treatment factor. A classic example is to compare 4

brands of automobile tires (treatment factor) using 4 cars as one blocking factor (**random**) and tire position on car as the second blocking factor (**fixed**). The experimental design can be set up as in Table 7.13. Due to there being three factors (treatments, row, and columns) needing three index letters (i, j, k), t is used as the number of treatments.

	Tire Position			
Car	1	2	3	4
1	A	B	C	D
2	B	C	D	A
3	C	D	A	B
4	D	A	B	C

Table 7.13: Latin Square Design (4 Treatments)

The 4 test cars would be randomized to the 4 labels (1,2,3,4); the 4 tire positions (Right Front, Left Front, Right Rear, Left Rear) would be randomized to the 4 labels, and the 4 brands would be randomized to the 4 labels (A,B,C,D). In practice, this experiment would be replicated in multiple squares (sets of 4 cars each). The model and Analysis of Variance can be obtained as follows for the case of an experiment consisting of one latin square with t treatments and $n_{\cdot} = t^2$ observations.

$$Y_{ij}^{k} = \mu + \alpha_k + \beta_i + \gamma_j + \epsilon_{ij}$$

where α_k is the effect of treatment k, β_i is the effect of row i, and γ_j is the effect of column j. Note that we only use two subscripts, as each observation is indexed by its row and column, the superscript identifies the treatment. To obtain the analysis of variance, we obtain the row means $(\overline{y}_{i\cdot})$, the column means $(\overline{y}_{\cdot j})$, the treatment means $(\overline{y}_k)$, and the overall mean $(\overline{y}_{\cdot\cdot})$. Then, the following sums of squares and the Analysis of Variance are given in Table 7.14.

$$\begin{aligned}
TSS &= \sum_{i=1}^{t}\sum_{j=1}^{t}\left(y_{ij}^{k} - \overline{y}_{\cdot\cdot}\right)^2 \\
SST &= t\sum_{k=1}^{t}\left(\overline{y}_k - \overline{y}_{\cdot\cdot}\right)^2 \\
SSR &= t\sum_{i=1}^{t}\left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2 \\
SSC &= t\sum_{i=1}^{t}\left(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot}\right)^2 \\
SSE &= TSS - SST - SSR - SSC
\end{aligned}$$

Once again, the main purpose for conducting this type of experiment is to detect differences among the treatment means (treatment effects). The test is very similar to that of the CRD and RBD, with only minor adjustments. We are rarely interested in testing for differences among either blocking factor, since we expect there to be differences among them (that's why the experiment is designed this way). The treatments are the items chosen specifically to be compared in the experiment. The testing procedure can be described as follows.

1. $H_0 : \alpha_1 = \cdots = \alpha_t = 0$ (No treatment effect)

ANOVA

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F
TREATMENTS	SST	$t-1$	$MST = \frac{SST}{t-1}$	$F = \frac{MST}{MSE}$
ROWS	SSR	$t-1$	$MSR = \frac{SSR}{t-1}$	
COLUMNS	SSC	$t-1$	$MSC = \frac{SSC}{t-1}$	
ERROR	SSE	$(t-1)(t-2)$	$MSE = \frac{SSE}{(t-1)(t-2)}$	
TOTAL	TSS	t^2-1		

Table 7.14: The Analysis of Variance Table for the Latin Square Design

2. H_A : Not all α_i are 0 (Treatment effects exist)
3. T.S. $F_{obs} = \frac{MST}{MSE}$
4. R.R.: $F_{obs} \geq F_{\alpha,t-1,(t-1)(t-2)}$
5. p-value: $P\left(F_{t-1,(t-1)(t-2)} \geq F_{obs}\right)$

The procedures to make comparisons among means are also very similar to the methods used for the CRD and the RBD. In each formula described previously for Dunnett's, Bonferroni's, and Tukey's methods, n_i in the CRD and b in the RBD are replaced with t, when making comparisons among treatment means, and $\nu = (t-1)(t-2)$.

When t is small, multiple squares can be run, allowing for more error degrees of freedom. The same or new row/column block levels can be used. For instance in the car tire example, 8 or 12 (or any multiple of 4) cars could be used, while the tire positions would remain the same.

The **Relative Efficiency** of conducting the Latin Square Design, as opposed to the Completely Randomized Design is:

$$RE(LS, CR) = \frac{MSE_{CR}}{MSE_{LS}} = \frac{MSR + MSC + (t-1)MSE}{(t+1)MSE}$$

This represents the number of times as many replicates would be needed for each treatment in a CRD to obtain as precise of estimates of differences between two treatment means as we were able to obtain by using t experimental units per treatment in the Latin Square.

Example 7.12: Dye Decolorisation in Copper Alginate

An experiment was conducted to compare $t = 3$ CuSO4 molar concentrations (.075, .150, 0.225) with respect to dye decolorisation (Teerapatsakul, et al (2008), [49]). The row blocking factor was alginate type (percentage of alginate from 2 vendors (100:0, 50:50, and 0:100)). The column blocking factor was alginate concentration (1.5, 3.0, 4.5). The experiment was conducted as a 3x3 latin square. The design and data are given in Table 7.15.

	Alginate Concentration			
Alginate Type	1	2	3	Row Mean
1	A 53.9	B 63.7	C 58.2	58.60
2	B 50.3	C 52.8	A 57.4	53.50
3	C 53.0	A 45.8	B 46.1	48.30
Column Mean	52.40	54.10	53.90	53.47
Treatment Mean	A 52.37	B 53.37	C 54.67	

Table 7.15: Dye decolorisation (%) in Copper Alginate

$$\text{Total SS: } TSS = (53.9 - 53.4667)^2 + \cdots + (46.1 - 53.4667)^2 = 266.52 \qquad df = 3^2 - 1 = 8$$

$$\text{Treatment SS: } SST = 3\left[(52.3667 - 53.4667)^2 + (53.3667 - 53.4667)^2 + (54.667 - 53.4667)^2\right] = 7.98$$

$$df_T = 3 - 1 = 2 \qquad MST = 3.99$$

$$\text{Row SS: } SSR = 3\left[(58.60 - 53.4667)^2 + (53.50 - 53.4667)^2 + (48.30 - 53.4667)^2\right] = 159.14$$

$$df_R = 3 - 1 = 2 \qquad MSR = 79.57$$

$$\text{Column SS: } SSC = 3\left[(52.40 - 53.4667)^2 + (54.10 - 53.4667)^2 + (53.90 - 53.4667)^2\right] = 5.18$$

$$df_C = 3 - 1 = 2 \qquad MSC = 2.59$$

$$\text{Error SS: } SSE = 266.52 - 7.98 - 159.14 - 5.18 = 94.22 \qquad df_E = (3-1)(3-2) = 2 \qquad MSE = 47.11$$

The F-statistic used for testing for CuSO4 concentration effects is $F_T = 3.99/47.11 = 0.0847$, with critical F-value of $F_{2,2,.05} = 19$ and P-value=.9219. There is no evidence of CuSO4 concentration effects on dye decolorisation. The relative efficiency of this design is 1.36.

$$RE(LS, CR) = \frac{79.57 + 2.59 + (3-1)47.11}{(3+1)47.11} = 1.36$$

∇

7.4 R Code for Chapter 7

```
### Chapter 7

### Example 7.1
```

```
bmi.sim <- read.csv("http://www.stat.ufl.edu/~winner/data/nhl_nba_ebl_bmi.csv")
attach(bmi.sim); names(bmi.sim)
set.seed(13579)
N <- rep(0,3)
N[1] <- 717
N[2] <- 505
N[3] <- 526
NHL_BMI <- NHL_BMI[1:N[1]]
NBA_BMI <- NBA_BMI[1:N[2]]
EPL_BMI <- EPL_BMI[1:N[3]]

## Figure 7.1

par(mfrow=c(3,1))
hist(NHL_BMI,breaks=25,xlim=c(18,34))
hist(NBA_BMI,breaks=25,xlim=c(18,34))
hist(EPL_BMI,breaks=25,xlim=c(18,34))
par(mfrow=c(1,1))
## End Figure 7.1

# Obtain population means and SDs by league
sigma <- rep(0,3)
sigma[1] <- sd(NHL_BMI) * sqrt((N[1]-1)/N[1])
sigma[2] <- sd(NBA_BMI) * sqrt((N[2]-1)/N[2])
sigma[3] <- sd(EPL_BMI) * sqrt((N[3]-1)/N[3])
mu <- rep(0,3)
mu[1] <- mean(NHL_BMI)
mu[2] <- mean(NBA_BMI)
mu[3] <- mean(EPL_BMI)
sigma2.all <- sum(N * sigma^2) / sum(N)
mu.all <- sum(N * mu) / sum(N)
cbind(sigma2.all, mu.all)

# Set sample sizes and obtain number of groups=t
sampsz <- rep(3,0)
sampsz[1] <- 12
sampsz[2] <- 12
sampsz[3] <- 12
sumssz <- sum(sampsz)
num.trt <- length(sampsz)

# Power when n_i=4, n_i=12
(power.4 <- 1-pf(qf(.95,2,9),2,9,9.48))
(power.12 <- 1-pf(qf(.95,2,33),2,33,28.43))

# Set up the number of samples and holders for sample means and SDs
N.sim <- 100000
set.seed(6529)
ybar1 <- numeric(N.sim)
sd1 <- numeric(N.sim)
ybar2 <- numeric(N.sim)
sd2 <- numeric(N.sim)
ybar3 <- numeric(N.sim)
sd3 <- numeric(N.sim)
ybar <- numeric(N.sim)

# Loop through the samples
for (i in 1:N.sim) {
samp1 <- sample(1:N[1],sampsz[1],replace=F)
samp2 <- sample(1:N[2],sampsz[2],replace=F)
samp3 <- sample(1:N[3],sampsz[3],replace=F)
y1 <- NHL_BMI[samp1]
y2 <- NBA_BMI[samp2]
y3 <- EPL_BMI[samp3]
```

```
ybar1[i] <- mean(y1); sd1[i] <- sd(y1)
ybar2[i] <- mean(y2); sd2[i] <- sd(y2)
ybar3[i] <- mean(y3); sd3[i] <- sd(y3)
ybar[i] <- (mean(y1)+mean(y2)+mean(y3))/num.trt
}

# Obtain the ANOVA, F-test, Proportion of Samples Rejecting
SST <- sampsz[1]*(ybar1-ybar)^2 + sampsz[2]*(ybar2-ybar)^2 +
sampsz[3]*(ybar3-ybar)^2
SSE <- (sampsz[1]-1)*sd1^2 + (sampsz[2]-1)*sd2^2 + (sampsz[3]-1)*sd3^2
MST <- SST/(num.trt-1)
MSE <- SSE/(sumssz-num.trt)
F <- MST/MSE
f.alpha <- qf(.95,num.trt-1,sumssz-num.trt)
reject <- sum(F >= f.alpha)/N.sim

ftest.out <- cbind(num.trt-1,sumssz-num.trt, f.alpha, reject)
colnames(ftest.out) <- c("df_T", "df_E", "F(>05)", "P(F_obs>F(.05))")
round(ftest.out, 4)
F[1]
cbind(ybar1[1], ybar2[1], ybar3[1], ybar[1], sd1[1], sd2[1], sd3[1])

### Figure 7.2
## F and non-central F

## n_i=4

x <- seq(0,10,.01)
f.c <- df(x,2,9)
f.nc <- df(x,2,9,9.48)

par(mfrow=c(2,1))
plot(x,f.c,type="l",
  main="Central and Non-Central F-Densities - BMI Example, n=4")
lines(x,f.nc,lty=2)
legend(6,0.9,c("Central F", "Non-central F"),lty=c(1,2))
abline(v=qf(.95,2,9))
text(qf(.95,2,9)-1,0.4,"Fail to Reject H0",cex=0.7)
text(qf(.95,2,9)+1,0.4,"Reject H0",cex=0.7)

### n_i=12

x <- seq(0,10,.01)
f.c <- df(x,2,33)
f.nc <- df(x,2,33,28.42)
plot(x,f.c,type="l",
  main="Central and Non-Central F-Densities - BMI Example, n=12")
lines(x,f.nc,lty=2)
legend(6,0.9,c("Central F", "Non-central F"),lty=c(1,2))
abline(v=qf(.95,2,33))
text(qf(.95,2,33)-1,0.4,"Fail to Reject H0",cex=0.7)
text(qf(.95,2,33)+1,0.4,"Reject H0",cex=0.7)
### End of Figure 7.2

rm(list=ls(all=TRUE))

### Example 7.2

### Summary Stats
k <- 5
n <- rep(30, k)
ybar <- c(7.900, 8.133, 8.033, 6.333, 5.367)
sd <- c(3.367, 3.461, 3.011, 3.122, 3.068)
```

```
ybar.all <- sum(n*ybar) / sum(n)
### ANOVA Computations from Summary Stats
SST <- sum(n * (ybar-ybar.all)^2)
SSE <- sum((n-1) * sd^2)
dfT <- k-1
dfE <- sum(n)-k
TSS <- SST + SSE
dfTOT <- sum(n)-1
MST <- SST/dfT
MSE <- SSE/dfE
F_obs <- MST/MSE
F.05 <- qf(.95,dfT,dfE)
F_p <- 1-pf(F_obs,dfT,dfE)

df <- rbind(dfT, dfE, dfT+dfE)
SS <- rbind(SST, SSE, TSS)
MS <- rbind(MST, MSE, NA)
F <- rbind(F_obs, NA, NA)
F.a <- rbind(F.05, NA, NA)
F.p <- rbind(F_p, NA, NA)
aov.out <- cbind(df, SS, MS, F, F.a, F.p)
rownames(aov.out) <- c("Treatment", "Error", "Total")
colnames(aov.out) <- c("df", "SS", "MS", "F", "F(.05)", "P(>F)")
round(aov.out, 4)

mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)
trt.mosq <- factor(trt.mosq)
mosq.mod <- aov(y.mosq ~ trt.mosq)
summary(mosq.mod)

rm(list=ls(all=TRUE))

### Example 7.3

### Summary Stats
k <- 5
n <- rep(30, k)
ybar <- c(7.900, 8.133, 8.033, 6.333, 5.367)
sd <- c(3.367, 3.461, 3.011, 3.122, 3.068)
ybar.all <- sum(n*ybar) / sum(n)

### ANOVA Computations from Summary Stats
SST <- sum(n * (ybar-ybar.all)^2)
SSE <- sum((n-1) * sd^2)
dfT <- k-1
dfE <- sum(n)-k
TSS <- SST + SSE
dfTOT <- sum(n)-1
MST <- SST/dfT
MSE <- SSE/dfE
F_obs <- MST/MSE
F.05 <- qf(.95,dfT,dfE)
F_p <- 1-pf(F_obs,dfT,dfE)

df <- rbind(dfT, dfE, dfT+dfE)
SS <- rbind(SST, SSE, TSS)
MS <- rbind(MST, MSE, NA)
F <- rbind(F_obs, NA, NA)
F.a <- rbind(F.05, NA, NA)
F.p <- rbind(F_p, NA, NA)
aov.out <- cbind(df, SS, MS, F, F.a, F.p)
rownames(aov.out) <- c("Treatment", "Error", "Total")
colnames(aov.out) <- c("df", "SS", "MS", "F", "F(.05)", "P(>F)")
```

```
round(aov.out, 4)

### Example 7.4
### Contrasts
LC1 <- c(0, 1, -1, 1, -1)
c1 <- sum(LC1*ybar)
se.c1 <- sqrt(MSE*(sum(LC1^2/n)))
t.C1 <- c1/se.c1
t.C1.p <- 2 * (1-pt(abs(t.C1),dfE))
C1.CI <- c1 + qt(c(.025,.975),dfE) * se.c1
SSC1 <- c1^2 / sum(LC1^2/n)
F.C1 <- SSC1 / MSE
F.C1.p <- 1 - pf(F.C1,1,dfE)

contrast.out <- cbind(c1, se.c1, t.C1, t.C1.p, C1.CI[1], C1.CI[2],
      SSC1, F.C1, F.C1.p)
colnames(contrast.out) <- c("Estimate", "Std Err", "t", "2P(>|t|)",
    "LB", "UB", "Sum Sq", "F", "P(>F)")
round(contrast.out, 4)

LC2 <- c(0, 1, 1, -1, -1)
c2 <- sum(LC2*ybar)
se.c2 <- sqrt(MSE*(sum(LC2^2/n)))
t.C2 <- c2/se.c2
t.C2.p <- 2 * (1-pt(abs(t.C2),dfE))
C2.CI <- c2 + qt(c(.025,.975),dfE) * se.c2
SSC2 <- c2^2 / sum(LC2^2/n)
F.C2 <- SSC2 / MSE
F.C2.p <- 1 - pf(F.C2,1,dfE)

contrast.out <- cbind(c2, se.c2, t.C2, t.C2.p, C2.CI[1], C2.CI[2],
      SSC2, F.C2, F.C2.p)
colnames(contrast.out) <- c("Estimate", "Std Err", "t", "2P(>|t|)",
    "LB", "UB", "Sum Sq", "F", "P(>F)")
round(contrast.out, 4)

rm(list=ls(all=TRUE))

### Example 7.5

mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

## Figure 7.3 (New)
# win.graph(height=5.5, width=7.0)
par(mfrow=c(1,1))
plot(y.mosq ~ trt.mosq, main="Mosquito Contact by Repellent",
   xlim=c(0,6), pch=16)
lines(c(0.5, 5.5), c(mean(y.mosq),mean(y.mosq)), lwd=2)
for (i in 1:length(unique(trt.mosq))) {
  lines(c(i-0.2,i+0.2),
   c(mean(y.mosq[trt.mosq==i]),mean(y.mosq[trt.mosq==i])),
    lwd=2)
}
## End of Figure 7.3

trt.mosq <- factor(trt.mosq)
mosq.mod1 <- aov(y ~ trt.mosq)
anova(mosq.mod1)

# install.packages("multcomp")
library(multcomp)
mosq.dunnett <- glht(mosq.mod1, linfct=mcp(trt.mosq="Dunnett"))
summary(mosq.dunnett)
```

```
confint(mosq.dunnett)

rm(list=ls(all=TRUE))

### Example 7.6

### Summary Stats
k <- 5
n <- rep(30, k)
ybar <- c(7.900, 8.133, 8.033, 6.333, 5.367)
sd <- c(3.367, 3.461, 3.011, 3.122, 3.068)

### Tukey follow-up to 1-Way ANOVA
mp <- read.csv("http://www.stat.ufl.edu/~winner/data/mosquito_patch.csv")
attach(mp); names(mp)

trt.mosq <- factor(trt.mosq)
mosq.mod1 <- aov(y.mosq ~ trt.mosq)
anova(mosq.mod1)
TukeyHSD(mosq.mod1, "trt.mosq")

### Bonferroni method
bon.ci <- function(alpha_E, y, trt.y) {
ybar <- as.vector(tapply(y, trt.y, mean))       # Obtain Trt Means
sd <- as.vector(tapply(y, trt.y, sd))           # Obtain Trt SDs
n <- as.vector(tapply(y, trt.y, length))        # Obtain Trt n's
k <- length(ybar)                       # Obtain k = # trts
c <- k*(k-1)/2                          # Obtain c = # comparisons
SSE <- sum((n-1)*sd^2)                  # Error Sum of Squares
dfE <- sum(n)-k                         # Error df
MSE <- SSE/dfE                          # Error Mean Square
bon.t <- qt(1-alpha_E/(2*c),dfE)        # Critical value
bon.out <- matrix(rep(0,6*c),ncol=6)    # Matrix to hold results
bon.row <- 0
## Loop through all i, i' and compute B_{ii'} and CI's
for(i1 in 1:(k-1)) {
for (i2 in (i1+1):k) {
bon.row <- bon.row + 1
bon.out[bon.row,1] <- i1
bon.out[bon.row,2] <- i2
bon.out[bon.row,3] <- ybar[i1] - ybar[i2]
bon.out[bon.row,4] <-
(ybar[i1] - ybar[i2]) - bon.t*sqrt(MSE*(1/n[i1] + 1/n[i2]))
bon.out[bon.row,5] <-
(ybar[i1] - ybar[i2]) + bon.t*sqrt(MSE*(1/n[i1] + 1/n[i2]))
t <- (ybar[i1] - ybar[i2]) / sqrt(MSE*(1/n[i1] + 1/n[i2]))
bon.out[bon.row,6] <- min(1, c*2*(1-pt(abs(t),dfE)))
}}
colnames(bon.out) <- c("Trt i","Trt i'","Diff","Lower Bound","Upper Bound",
"p adjusted")
round(bon.out,3)
}

bon.ci(0.05, y.mosq, trt.mosq)

rm(list=ls(all=TRUE))

### Example 7.7

sb <- read.csv("http://www.stat.ufl.edu/~winner/data/sealion_bark.csv")
attach(sb); names(sb)

location <- factor(location)
bartlett.test(duration ~ location)
```

```
oneway.test(duration ~ location, var.equal=F)

(loc_mean <- as.vector(tapply(duration,location,mean))) ## Vector of means
(loc_var <- as.vector(tapply(duration,location,var))) ## Vector of vars
(loc_n <- as.vector(tapply(duration,location,length))) ## Vector of ns

### Games-Howell method
k <- length(loc_n)
gh.out <- matrix(rep(0,7*k*(k-1)/2),ncol=7)
gh.row <- 0
for(i1 in 1:(k-1)) {
for (i2 in (i1+1):k) {
gh.row <- gh.row + 1
gh.out[gh.row,1] <- i1
gh.out[gh.row,2] <- i2
gh.out[gh.row,3] <- loc_mean[i1] - loc_mean[i2]
diff_mean_var <- loc_var[i1]/loc_n[i1]+loc_var[i2]/loc_n[i2]
gh.out[gh.row,4] <- sqrt(diff_mean_var)
diff_df <- diff_mean_var^2 /
  ( ((loc_var[i1]/loc_n[i1])^2/(loc_n[i1]-1)) +
  ((loc_var[i2]/loc_n[i2])^2/(loc_n[i2]-1)) )
gh.out[gh.row,5] <- diff_df
gh.q <- qtukey(.95,k,diff_df)
gh.out[gh.row,6] <-
  (loc_mean[i1] - loc_mean[i2]) - (gh.q/sqrt(2)) * sqrt(diff_mean_var)
gh.out[gh.row,7] <-
  (loc_mean[i1] - loc_mean[i2]) + (gh.q/sqrt(2)) * sqrt(diff_mean_var)
}}
colnames(gh.out) <- c("Trt i","Trt j","Diff","SE", "DF",
   "Lower Bound","Upper Bound")
round(gh.out,3)

rm(list=ls(all=TRUE))

### Example 7.8

bt <- read.csv("http://www.stat.ufl.edu/~winner/data/berry_texture.csv")
attach(bt); names(bt)

sugar <- factor(sugar)
kruskal.test(anthExt ~ sugar)

rm(list=ls(all=TRUE))

### Example 7.9

saw <- read.csv("http://www.stat.ufl.edu/~winner/data/sawmill1.csv")
attach(saw); names(saw)
lumTrt <- factor(lumTrt)
lumBlk <- factor(lumBlk)
levels(lumTrt) <- c("Actual", "Heuristic", "Dynamic")

saw.mod1 <- aov(lumVal ~ lumTrt + lumBlk)
anova(saw.mod1)
TukeyHSD(saw.mod1, "lumTrt")

## Figure 7.4
interaction.plot(lumBlk, lumTrt, lumVal)

rm(list=ls(all=TRUE))

### Example 7.10

saw <- read.csv("http://www.stat.ufl.edu/~winner/data/sawmill1.csv")
```

```
attach(saw); names(saw)
lumTrt <- factor(lumTrt)
lumBlk <- factor(lumBlk)

friedman.test(lumVol ~ lumTrt | lumBlk)

actual <- lumVol[lumTrt==1]
heuristic <- lumVol[lumTrt==2]
dynamic <- lumVol[lumTrt==3]
wilcox.test(actual, heuristic, paired=T)
wilcox.test(actual, dynamic, paired=T)
wilcox.test(heuristic, dynamic, paired=T)

rm(list=ls(all=TRUE))
```

Chapter 8

Categorical Data Analysis

Recall that variables can be categorical or numeric. The past four chapters dealt with making inferences for quantitative responses. In this chapter, methods commonly used to analyze data when the response variable is categorical are introduced. First, consider estimating and testing proportions corresponding to a single binomial (2 possible outcomes) or multinomial ($k > 2$ possible outcomes) variable. Then, cases of testing for associations among two or more categorical variables are covered.

8.1 Inference Concerning a Single Variable

A single variable can have two levels, and counts are modeled by the Binomial distribution, or it can have $k > 2$ levels and counts are modeled by the Multinomial distribution. Note that the Binomial is a special case of the Multinomial, however there are many methods that apply strictly to binary outcomes.

8.1.1 Variables with Two Possible Outcomes

In the case of a binary variable, the goal is typically to estimate the true underlying probability of success, π. The sample proportion $\hat{\pi} = Y/n$ from a binomial experiment with n trials and Y successes has a sampling distribution with mean π and standard error $\sqrt{\pi(1-\pi)/n}$. In large samples, the sampling distribution is approximately normal. One commonly used rule of thumb is that $n\pi \geq 5$ and $n(1-\pi) \geq 5$. When estimating π, the estimated standard error must be used, where π is replaced with $\hat{\pi}$. Note that the standard error is maximized for a given n when $\pi = 1 - \pi = 0.5$, so a conservative case uses $\pi = 0.5$ in the standard error. The large-sample $(1-\alpha)100\%$ Confidence Interval for π and the sample size needed for a given margin of error, E, are given below.

$$(1-\alpha)100\%\text{CI for } \pi : \hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}\,(1-\hat{\pi})}{n}} \qquad E = z_{\alpha/2}\sqrt{\frac{\pi\,(1-\pi)}{n}} \quad \Rightarrow \quad n = \frac{z^2_{\alpha/2}\pi(1-\pi)}{E^2}$$

In small samples, the large-sample normal approximation does poorly in terms of coverage rates for π. It has been seen that making an adjustment to the success count and the sample size performs well. This is referred to as the **Wilson-Agresti-Coull** method. Let y be the observed number of successes in the n trials, then the Confidence Interval is obtained as follows. Note that since $z_{.025} = 1.96 \approx 2$, for a 95% Confidence Interval, this can be thought of as adding 2 Successes and 2 Failures to the observed data (Agresti and Coull (1998), [2]).

$$\tilde{y} = y + 0.5z^2_{\alpha/2} \quad \tilde{n} = n + z^2_{\alpha/2} \quad \tilde{\pi} = \frac{\tilde{y}}{\tilde{n}} \qquad (1-\alpha)100\%\text{CI for } \pi : \tilde{\pi} \pm z_{\alpha/2}\sqrt{\frac{\tilde{\pi}\,(1-\tilde{\pi})}{\tilde{n}}}$$

Example 8.1: Estimating Shaquille O'Neal's Free Throw Success Probability

During Shaquille O'Neal's NBA regular season career, he took 11252 free throws, successfully making 5935, so that $\pi = 5935/11252 = .5275$. Stringing out his within game free throw attempts into a sequence of 1^s and 0^s, and taking 100000 random samples of size $n = 10$, the coverage rates for the two methods are 88.5% for the "traditional" large-sample method and 94.8% for the Wilson-Agresti-Coull method. For the small-sample case, the adjustment performs very well. When the samples are of size $n = 30$, the coverage rates are 93.2% and 95.9%, respectively. For samples of size $n = 100$, they are 94.3% and 95.3%, respectively.

R Output

```
## Output

> round(ft.out, 4)
          pi pi-hat cover pi-tilde cover pi-hat mean width pi-tilde mean width
n=10  0.5275       0.8836         0.9455            0.5842              0.5120
n=30  0.5275       0.9321         0.9588            0.3510              0.3319
n=100 0.5275       0.9436         0.9538            0.1946              0.1911
```

For the first sample of size $n = 10$, $y = 7$ free throws were successes and the following calculations are used to obtain the 95% Confidence Intervals for π.

$$\hat{\pi} = \frac{7}{10} = 0.7 \qquad \hat{SE}\{\hat{\pi}\} = \sqrt{\frac{0.7\,(1-0.7)}{10}} = 0.145 \qquad 0.70 \pm 1.96(0.145) \equiv 0.70 \pm 0.284 \equiv (0.416, 0.984)$$

$$\tilde{y} = 7 + 0.5(1.96)^2 = 8.92 \quad \tilde{n} = 10 + (1.96)^2 = 13.84 \quad \tilde{\pi} = \frac{8.92}{13.84} = 0.645 \quad \sqrt{\frac{0.645\,(1-0.645)}{13.84}} = 0.129$$

$$0.645 \pm 1.96(0.129) \equiv 0.645 \pm 0.253 \equiv (0.392, 0.898)$$

Both intervals contain $\pi = 0.5275$.

∇

A Large-sample test of whether $\pi = \pi_0$ can also be conducted. For instance, a test may be whether a majority of people favor a political candidate or referendum, or whether a defective rate is below some tolerance level.

2-tailed test: $H_0 : \pi = \pi_0 \quad H_A : \pi \neq \pi_0 \quad TS : z_{obs} = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : |z_{obs}| \geq z_{\alpha/2} \quad P = 2P\left(Z \geq |z_{obs}|\right)$

Upper-tailed test: $H_0 : \pi \leq \pi_0 \quad H_A : \pi > \pi_0 \quad TS : z_{obs} = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : z_{obs} \geq z_{\alpha} \quad P = P\left(Z \geq z_{obs}\right)$

Lower-tailed test: $H_0 : \pi \geq \pi_0 \quad H_A : \pi < \pi_0 \quad TS : z_{obs} = \dfrac{\hat{\pi} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \quad RR : z_{obs} \leq -z_{\alpha} \quad P = P\left(Z \leq z_{obs}\right)$

An exact test can be conducted by use of the binomial distribution and statistical packages by obtaining the exact probability that the count could be more extreme than the observed count y under the null hypothesis. See the examples below.

Example 8.2: NBA Point Spread and Over/Under Outcomes for 2014-2015 Regular Season

For each NBA game there is a "point spread" for bettors to wager on. If the home team is favored to win the game by 5 points, it must win by 6 or more points to "cover the spread," if it loses the game or wins by less than 5 points, the home team loses the bet, and if it wins by 5 points, the best is a tie or "push." For the 2014-2015 regular season games, the home team covered the spread in 588 games, failed to cover the spread in 615 games, and "tied" the spread in 27 games. We treat these games as a sample of the infinite population of games that could be played among NBA teams, and eliminate the 27 "pushes." The test is whether the true underlying probability that the home team covers is 0.50. Otherwise bettors could have an advantage over bookmakers. $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$.

$$y = 588 \qquad n = 615 + 588 = 1203 \qquad \hat{\pi} = \frac{588}{1203} = 0.4888 \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{0.5(1-0.5)}{1203}} = 0.0144$$

$$z_{obs} = \frac{0.4888 - 0.5}{0.0144} = -0.78 \qquad P = 2P(Z \geq 0.78) = 2(0.2177) = 0.4354$$

There is no evidence of a "bias" (positive or negative) in terms of the home team performance against the spread. An exact test is given here. Under the null hypothesis, the expected value of Y is $n\pi_0 = 1203(0.5) = 601.5$. The observed y is 588, which is 13.5 below its expected value. Had y been 615, it would have been 13.5 above its expected value. The exact 2-tailed P-value is obtained as follows.

$$P = P\left(Y \leq 588 | Y \sim Bin(1203, 0.5)\right) + P\left(Y \geq 615 | Y \sim Bin(1203, 0.5)\right) = 0.22675 + 0.22675 = .4535$$

A similar test can be done for the "Over/Under" bet. Bookmakers set a total score for the two teams, and if the combined points exceed this line the Over wins, if it falls short, the Under wins, and if it ties, it is a "Push." For the Over/Under bet for that season, Under won 633 times, Over won 583 times, and there were 14 Pushes. Again, we eliminate the Pushes, and test $H_0 : \pi = 0.50$ versus $H_A : \pi \neq 0.50$, where π is the probability Over wins.

$$y = 583 \qquad n = 633 + 583 = 1216 \qquad \hat{\pi} = \frac{583}{1216} = 0.4794 \qquad SE\left\{\hat{\pi}\right\}_{H_0} = \sqrt{\frac{0.5(1-0.5)}{1216}} = 0.0143$$

$$z_{obs} = \frac{0.4794 - 0.5}{0.0143} = -1.44 \qquad P = 2P(Z \geq 1.44) = 2(.0749) = 0.1498$$

Again there is no evidence of a bias. An exact P-value is obtained below.

$$P = P\left(Y \leq 583 | Y \sim Bin(1216, 0.5)\right) + P\left(Y \geq 633 | Y \sim Bin(1216, 0.5)\right) = 0.07997 + 0.07997 = 0.1599$$

R Output

```
### Output
> round(cov.out, 4)
     pi(H0)   y    n  pihat SE{H0}       Z   P(Z) P(Exact) SE{pihat}  Lower Upper
[1,]    0.5 588 1203 0.4888 0.0144 -0.7785 0.4363   0.4535    0.0144 0.4605 0.517
> ### Exact Tests
> binom.test(Y.Cov,n.Cov,p=0.5,alternative="two.sided")

        Exact binomial test

data:  Y.Cov and n.Cov
number of successes = 588, number of trials = 1203, p-value = 0.4535
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.4601721 0.5174390
sample estimates:
probability of success
             0.4887781
```

∇

8.1.2 Variables with $k > 2$ Possible Outcomes

A Multinomial experiment is an extension of the Binomial experiment with the caveat that each of n trials can end in one of k possible outcomes or categories. The probability of outcome i is π_i, and the count of the number of trials falling in category i is y_i. The following restrictions must hold.

$$\pi_1 + \cdots + \pi_k = 1 \qquad y_1 + \cdots + y_k = n$$

Category (i)	π_{i0}	Expected #	Observed #	X^2
Black (1)	1/6	344	306	4.198
White (2)	1/6	344	338	0.105
Red (3)	1/6	344	432	22.512
Yellow (4)	1/6	344	348	0.047
Blue (5)	1/6	344	331	0.491
Green (6)	1/6	344	309	3.561
Total	1	2064	2064	30.913

Table 8.1: Numbers of e-mails received by shirt color - Internet personal ad experiment

Note that for the Binomial case, we have previously labeled $y_1 = y$ and $y_2 = n - y$ where category 1 represents "Success" and category 2 represents "Failure." The probability of the experiment resulting in the observed counts $(y_1, \ldots, y_k)$ is as follows.

$$p(y_1, \ldots, y_k) = \frac{n!}{y_1! \cdots y_k!} \pi_1^{y_1} \cdots \pi_k^{y_k} \qquad \pi_1 + \cdots + \pi_k = 1 \quad y_1 + \cdots + y_k = n \quad \pi_i \geq 0 \quad y_i \geq 0$$

A test can be conducted for $H_0 : \pi_1 = \pi_{10}, \ldots, \pi_k = \pi_{k0}$, for some specified set of k probabilities $\pi_{10}, \ldots, \pi_{k0}$ that sum to 1. Again this simply extends the binomial test of $H_0 : \pi = \pi_0$. Once the observed counts $y_1, \ldots, y_k$ are obtained, the test is conducted as follows, where E_i is the expected count for category i under the null hypothesis ($\pi_i = \pi_{i0}$) and observed sample size (n).

$$E_i = n\pi_{i0} \qquad TS : X^2_{obs} = \sum_{i=1}^{k} \frac{(y_i - E_i)^2}{E_i} \qquad RR : X^2_{obs} \geq \chi^2_{\alpha,k-1} \qquad P = P\left(\chi^2_{k-1} \geq X^2_{obs}\right)$$

Example 8.3: Color Preferences in Physical Attraction

An experiment was conducted to determine whether there is evidence of differences in attraction to various colors (Gueguen and Jacob (2013), [24]). Women registered on an internet personal ad site were photographed in shirts of $k = 6$ colors: red, black, white, yellow, blue, and green which were shown on the web site. The total numbers of e-mails received from the various shirt colors were obtained and given in Table 8.1. The authors tested the hypothesis that the $k = 6$ colors would be equally responded to in the general population. That is $H_0 : \pi_1 = \cdots = \pi_6 = 1/6$. There were a total of $n = 2064$ e-mail messages received during the study period. The test statistic is $X^2_{obs} = 30.912$ with $k - 1 = 6 - 1 = 5$ degrees of freedom. The critical Chi-square value is $\chi^2_{.05,5} = 11.071$, and the P-value for the test is .0000. There is strong evidence for differences among the preferences of the colors.

R Commands and Output

```
### Commands

### Default probs are 1/#categories
chisq.test(c(306,338,432,348,331,309),p=c(1/6,1/6,1/6,1/6,1/6,1/6))

### Output
```

y	$p(y)$	Expected #	Observed #	X^2
0	.3937	226.74	229	0.0225
1	.3670	211.39	211	0.0007
2	.1711	98.54	93	0.3115
3	.0532	30.62	35	0.6265
≥ 4	.0151	8.71	8	0.0579
Total	1	576	576	1.0191

Table 8.2: Probability Distribution for Number of bombs hitting within 576 areas on a grid in the south of London during World War II

```
> chisq.test(c(306,338,432,348,331,309),p=c(1/6,1/6,1/6,1/6,1/6,1/6))
        Chi-squared test for given probabilities
data:  c(306, 338, 432, 348, 331, 309)
X-squared = 30.913, df = 5, p-value = 9.746e-06
```

∇

This test is often used when testing whether data come from a particular family of probability distributions.

- A family of distributions (e.g. Poisson, Negative Binomial, Normal, Gamma, Beta) is considered for the data.
- Data are sampled and used to estimate the m parameter(s) of the distribution.
- The data are placed in k mutually exclusive and exhaustive ranges of values.
- The observed counts n_i and the expected counts under the hypothesized distribution are obtained (expected counts should be ≥ 5)
- The chi-square statistic is computed and has $k-1-m$ degrees of freedom under the null hypothesis.

Example 8.4: Bombings in London During World War II

A widely reported application of the Poisson Distribution involves the counts of the number of bombs hitting among 576 areas of $0.5km^2$ in south London during WWII (Clarke (1946), [15], also reported in Feller (1950), [22]). There were a total of 537 bombs hit with a mean of $537/576 = .9323$. Table 8.2 gives the counts, and their expected counts ($576p(y)$) under the Poisson distribution with $\lambda = 0.9323$ for the occurrences of 0 bombs, 1 bomb, ..., ≥ 4 bombs.

For a test of whether these data are modeled by a Poisson distribution with mean 0.9323, the test statistic is $X^2_{obs} = 1.0191$ with degrees of freedom $k-1-m = 5-1-1 = 3$. The critical value for $\alpha = 0.05$ is $\chi^2_{.05,3} = 7.815$, and the P-value is $P\left(\chi^2_3 \geq 1.0191\right) = .7966$.

R Output

Range	Observed #	Expected #	X^2
0-4.75	74	81.142	0.629
4.75-5.25	157	136.891	2.954
5.25-5.75	235	221.927	0.770
5.75-6.25	251	271.450	1.541
6.25-6.75	256	261.602	0.120
6.75-7.25	201	205.546	0.101
7.25-7.75	138	135.332	0.053
7.75-8.25	79	76.350	0.092
8.25-8.75	32	37.593	0.832
8.75-9.25	18	16.404	0.155
9.25-∞	13	9.763	1.073
Total	1454	1454	8.319

Table 8.3: Goodness-of-fit test for Male Rock and Roll marathon speeds as Gamma distribution

```
### Output

> (mean.bomb <- 537/576)
[1] 0.9322917
> (exp.bomb <- sum(obs.bomb)*c(p0,p1,p2,p3,p4))
[1] 226.742723 211.390351  98.538731  30.622279   8.705916
> round(bomb.out, 4)
     X2 stat DF X2(.05) P-value
[1,]  1.0176  3  7.8147   0.797
```

∇

Example 8.5: Male Rock and Roll Marathon Velocities

Previously the parameters of the Gamma distribution to model male Rock and Roll marathon speeds were estimated by the method of moments as $\alpha = 35.896$ and $\beta = 5.665$, treating these speeds as a sample from a conceptual population. Note that other methods of estimation involve maximizing the likelihood function and minimizing the chi-square goodness of fit statistic. The range of velocities is broken into the following $k = 11$ categories: $(0, 4.75], (4.75, 5.25], \ldots, (8.75, 9.25], (9.25, \infty)$. Table 8.3 gives the observed and expected counts, and computations for the chi-square goodness-of-fit test. The degrees of freedom are 11-1-2=8, with critical chi-square value of $\chi^2_{.05,8} = 15.507$ and P-value $P\left(\chi^2_8 \geq 8.319\right) = .4029$.

R Output

```
### Output

> round(X2.out,4)
       alpha   beta Test Stat DF X2(.05) P-value
[1,] 35.8964 5.6646    8.3191  8 15.5073  0.4029
> round(cbind(cell.top,n.cells,tot.n.cells,exp.cells,X2.ab.cell),3)
       cell.top n.cells tot.n.cells exp.cells X2.ab.cell
 [1,]     4.75      74          74    81.142      0.629
 [2,]     5.25     157         231   136.891      2.954
 [3,]     5.75     235         466   221.927      0.770
 [4,]     6.25     251         717   271.450      1.541
```

```
 [5,]       6.75     256         973   261.602      0.120
 [6,]       7.25     201        1174   205.546      0.101
 [7,]       7.75     138        1312   135.332      0.053
 [8,]       8.25      79        1391    76.350      0.092
 [9,]       8.75      32        1423    37.593      0.832
[10,]       9.25      18        1441    16.404      0.155
[11,] 9999999.00      13        1454     9.763      1.073
```

∇

8.2 Introduction to Tests for Association for Two Categorical Variables

The data are generally counts of individuals or units, and are given in the form of an $r \times c$ **contingency table**. Throughout these notes, the rows of the table will represent the r levels of the explanatory variable, and the columns will represent the c levels of the response variable. The numbers within the table are the counts of the numbers of individuals falling in that cell's combination of levels of the explanatory and response variables. The general set–up of an $r \times c$ contingency table is given in Table 8.4.

		Response Variable				
		1	2	$\cdots$	c	
Explanatory Variable	1	n_{11}	n_{12}	$\cdots$	n_{1c}	$n_{1.}$
	2	n_{21}	n_{22}	$\cdots$	n_{2c}	$n_{2.}$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}	$n_{r.}$
		$n_{.1}$	$n_{.2}$	$\cdots$	$n_{.c}$	$n_{..}$

Table 8.4: An $r \times c$ Contingency Table

Recall that categorical variables can be **nominal** or **ordinal**. Nominal variables have levels that have no inherent ordering, such as gender (male, female) or hair color (black, blonde, brown, red). Ordinal variables have levels that do have a distinct ordering such as reviewer's assessment of a movie (negative opinion, mixed opinion, positive opinion).

In this chapter, the following cases are covered.

- 2×2 tables (both variables have two levels)
- Both variables are nominal.
- Both variables are ordinal.
- Explanatory variable is nominal, response variable is ordinal.
- Tables are nominal or ordinal ratings of the same objects by two raters.

8.3 2×2 Tables

There are many situations where both the independent and dependent variables have two levels. One example is efficacy studies for drugs, where subjects are assigned at random to active drug or placebo (explanatory variable) and the outcome measure is whether or not the patient is cured (response variable). A second example is epidemiological studies where disease state is observed (response variable), as well as exposure to risk factor (explanatory variable). Drug efficacy studies are generally conducted as randomized clinical trials, while epidemiological studies are generally conducted in cohort (prospective) and case–control (retrospective) settings.

For this particular case, we will generalize the explanatory variable's levels to exposed (E) and not exposed ($\overline{E}$), and the response variable's levels as disease (D) and no disease ($\overline{D}$). These interpretations can be applied in either of the two settings described above and can be generalized to virtually any application. The data for this case will be of the form of Table 8.5.

		Disease State		
		D (Present)	$\overline{D}$ (Absent)	Total
Exposure State	E (Present)	$n_{11} = y_1$	$n_{12} = n_1 - y_1$	$n_{1.} = n_1$
	$\overline{E}$ (Absent)	$n_{21} = y_2$	$n_{22} = n_2 - y_2$	$n_{2.} = n_2$
Total		$n_{.1} = y_1 + y_2$	$n_{.2} = (n_1 - y_1) + (n_2 - y_2)$	$n_{..} = n_1 + n_2$

Table 8.5: A 2×2 Contingency Table

In the case of drug efficacy studies, the exposure state can be thought of as the drug the subject is randomly assigned to. Exposure could imply that a subject was given the active drug, while non–exposure could imply having received placebo. In either type study, there are three measures of association commonly estimated and reported. These are the **absolute risk** (aka difference in proportions), the **relative risk** and the **odds ratio**.

These methods are also used when the explanatory variable has more than two levels, and the response variable has two levels. The methods described below are computed within pairs of levels of the explanatory variables, with one level forming the "baseline" group in comparisons.

8.3.1 Difference in Proportions: $\pi_1 - \pi_2$

In many studies, the goal is to compare the Success probabilities for two groups. These studies can be based on large samples or small samples, and can be based on independent or paired samples.

For the large sample case, based on independent samples, the estimators $\hat{\pi}_1 = Y_1/n_1$ and $\hat{\pi}_2 = Y_2/n_2$ for the two groups are independent and have sampling distributions that are approximately normal. The relevant results are given below.

$$E\{\hat{\pi}_1 - \hat{\pi}_2\} = \pi_1 - \pi_2 \quad SE\{\hat{\pi}_1 - \hat{\pi}_2\} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}} \quad \hat{SE}\{\hat{\pi}_1 - \hat{\pi}_2\} = \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

$$(1-\alpha)100\% \text{ CI for } \pi_1-\pi_2 : (\hat{\pi}_1-\hat{\pi}_2)\pm z_{\alpha/2}\hat{SE}\{\hat{\pi}_1-\hat{\pi}_2\} \quad \equiv \quad (\hat{\pi}_1-\hat{\pi}_2)\pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1}+\frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

In terms of testing the hypothesis $H_0 : \pi_1 - \pi_2 = 0$, an adjustment is made to the standard error of $\hat{\pi}_1 - \hat{\pi}_2$. In this case the overall combined proportion of successes is obtained and used in the "pooled" standard error.

$$\hat{\pi} = \frac{y_1+y_2}{n_1+n_2} \qquad \hat{SE}_p\{\hat{\pi}_1-\hat{\pi}_2\} = \sqrt{\hat{\pi}(1-\hat{\pi})\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}$$

The test statistic for testing $H_0 : \pi_1 - \pi_2 = 0$ is given below with the usual rules for rejection regions and P-values for 2-tailed and 1-tailed tests.

$$TS : z_{obs} = \frac{\hat{\pi}_1-\hat{\pi}_2}{\hat{SE}_p\{\hat{\pi}_1-\hat{\pi}_2\}} = \frac{\hat{\pi}_1-\hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\left[\frac{1}{n_1}+\frac{1}{n_2}\right]}}$$

Example 8.6: Risk Taking After Large Financial Losses

An Australian natural experiment considered the effect of large losses on subsequent risk taking behavior (Page, Savage, and Torgler (2014), [40]). The study included a sample of $n_1 = 94$ people who had been affected by the flood in Brisbane during 2011 and a sample of $n_2 = 107$ people who had not been affected. The subjects in the experiment were given the choice between a certain \$10 and a scratch card valued at \$10, but with a maximum prize of \$500,000. The scratch card is considered the "high risk" choice. Of the affected participants, $y_1 = 75$ chose the scratch card, of the unaffected, $y_2 = 53$ chose the scratch card.

$$\hat{\pi}_1 = \frac{75}{94} = .7979 \quad \hat{\pi}_2 = \frac{53}{107} = 0.4953 \quad \hat{\pi}_1-\hat{\pi}_2 = .7979-.4953 = .3026 \qquad \hat{\pi} = \frac{75+53}{94+107} = \frac{128}{201} = 0.6368$$

$$\hat{SE}\{\hat{\pi}_1-\hat{\pi}_2\} = \sqrt{\frac{.7979(.2021)}{94}+\frac{.4953(.5047)}{107}} = .0637 \qquad .3026\pm 1.96(.0637) \equiv .3026\pm .1248 \equiv (.1778, .4274)$$

$$\hat{SE}_p\{\hat{\pi}_1-\hat{\pi}_2\} = \sqrt{.6368(.3632)\left[\frac{1}{94}+\frac{1}{107}\right]} = .0680 \qquad z_{obs} = \frac{.3026}{.0680} = 4.451 \quad P = 2P(Z \geq 4.451) \approx 0$$

This provides empirical evidence consistent with prospect theory that states that people adopt risk taking attitudes after losses.

R Commands and Output

```
### Commands

y1 <- 75; n1 <- 94 ## Successes and Total for Group 1 (Affected by Flood)
y2 <- 53; n2 <- 107 ## Successes and Total for Group 2 (Unaffected)

pihat.1 <- y1/n1
pihat.2 <- y2/n2
pihat <- (y1+y2)/(n1+n2)
se.pihat.12 <- sqrt((pihat.1*(1-pihat.1)/n1)+(pihat.2*(1-pihat.2)/n2))
se.pihat.12p <- sqrt(pihat*(1-pihat)*(1/n1+1/n2))
z025 <- qnorm(.975,0,1)

pi12.ci <- (pihat.1-pihat.2) + c(-z025,z025)*se.pihat.12    # 95%CI for pi1-pi2
pi12.z <- (pihat.1-pihat.2)/se.pihat.12p                    # Z_obs for H0:pi1-pi2=0
pi12.p <- 2 * (1-pnorm(abs(pi12.z)))                        # 2-sided P-value

pi12.out <- cbind(y1, y2, n1, n2, pihat.1, pihat.2, pihat, se.pihat.12, pi12.ci[1],
     pi12.ci[2], se.pihat.12p, pi12.z, pi12.p)
colnames(pi12.out) <- c("y1", "y2", "n1", "n2", "pihat1", "pihat2", "pooled",
    "SE{Diff}", "Lower",  "Upper", "SE{(H0)}", "Z", "P-value")
round(pi12.out, 4)

prop.test(c(y1,y2),c(n1,n2),correct=F)

### Output

> round(pi12.out, 4)
     y1 y2 n1  n2 pihat1 pihat2 pooled SE{Diff}  Lower  Upper SE{(H0)}      Z P-value
[1,] 75 53 94 107 0.7979 0.4953 0.6368   0.0637 0.1778 0.4273    0.068 4.4502       0
>
> prop.test(c(y1,y2),c(n1,n2),correct=F)

        2-sample test for equality of proportions without continuity correction

data:  c(y1, y2) out of c(n1, n2)
X-squared = 19.804, df = 1, p-value = 8.58e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.1777846 0.4273059
sample estimates:
   prop 1    prop 2
0.7978723 0.4953271
```

Note that R presents the "Z-test" as a chi-square test (with 1 degree of freedom), $z_{obs}^2 = 4.4502^2 = 19.804$. The P-values are identical for a 2-tailed test.

∇

8.3.2 Relative Risk and Odds Ratio

Two other measures that can be used to compare two proportions are the **Relative Risk** and the **Odds Ratio**. These are generally reported in medical and epidemiology studies, particularly when the probabilities are small. Relative risk is a ratio of the two probabilities. In epidemiology studies it is the ratio of the probability of obtaining the disease among those exposed to some risk factor to the probability of obtaining disease among those not exposed.

$$\text{Relative Risk:} \quad RR = \frac{P(D|E)}{P(D|\overline{E})} = \frac{\pi_1}{\pi_2}$$

Based on this definition:

- A relative risk greater than 1.0 implies the exposed group have a higher probability of contracting disease than the unexposed group.
- A relative risk less than 1.0 implies that the exposed group has a lower chance of contracting disease than unexposed group (we might expect this to be the case in drug efficacy studies).
- A relative risk of 1.0 implies that the risk of disease is the same in both exposure groups (no association between exposure state and disease state).

Note that the Relative Risk is a population parameter that must be estimated based on sample data. We will be able to calculate confidence intervals for the relative risk, allowing inferences to be made concerning this population parameter, based on the range of values of RR within the $(1-\alpha)100\%$ confidence interval. The procedure to compute a $(1-\alpha)100\%$ confidence interval for the population relative risk is as follows.

1. Obtain the sample proportions of exposed and unexposed subjects who contract disease. These values are: $\hat{\pi}_E = \hat{\pi}_1 = \frac{y_1}{n_1}$ and $\hat{\pi}_{\overline{E}} = \hat{\pi}_2 = \frac{y_2}{n_2}$, respectively.
2. Compute the estimated relative risk: $\hat{RR} = \frac{\hat{\pi}_E}{\hat{\pi}_{\overline{E}}} = \frac{\hat{\pi}_1}{\hat{\pi}_2}$.
3. Compute $v_{RR} = \frac{(1-\hat{\pi}_E)}{y_1} + \frac{(1-\hat{\pi}_{\overline{E}})}{y_2} = \frac{1-\hat{\pi}_1}{y_1} + \frac{1-\hat{\pi}_2}{y_2}$ This is the estimated variance of $\log\left(\hat{RR}\right)$.
4. The confidence interval can be computed as: $\left(\hat{RR}e^{-z_{\alpha/2}\sqrt{v_{RR}}}, \hat{RR}e^{z_{\alpha/2}\sqrt{v_{RR}}}\right)$.

Example 8.7: Pamidronate for Skeletal Events in Myeloma Patients

An efficacy study was conducted for the drug pamidronate in patients with stage III multiple myeloma and at least one lytic lesion (Berenson, et al.,(1996), [7]). In this randomized clinical trial, patients were assigned at random to receive either pamidronate (E) or placebo ($\overline{E}$). One endpoint reported was the occurrence of any skeletal events after 9 cycles of treatment (D) or non–occurrence ($\overline{D}$). The results are given in Table 8.6. We will use the data to compute a 95% confidence interval for the relative risk of suffering skeletal events (in a time period of this length) for patients on pamidronate relative to patients not on the drug.

		Occurrence of Skeletal Event		
		Yes (D)	No ($\overline{D}$)	
Treatment Group	Pamidronate (E)	47	149	196
	Placebo ($\overline{E}$)	74	107	181
		121	256	377

Table 8.6: Observed cell counts for pamidronate data

First, obtain the proportions of patients suffering skeletal events among those receiving the active drug, and among those receiving the placebo

$$\hat{\pi}_E = \frac{n_{11}}{n_{1.}} = \frac{y_1}{n_1} = \frac{47}{196} = 0.240 \qquad \hat{\pi}_{\overline{E}} = \frac{n_{21}}{n_{2.}} = \frac{y_2}{n_2} = \frac{74}{181} = 0.409$$

Then compute the estimated relative risk ($\hat{RR}$) and the estimated variance of its natural log (v).

$$\hat{RR} = \frac{\hat{\pi}_E}{\hat{\pi}_{\overline{E}}} = \frac{.240}{.409} = 0.587 \qquad v_{RR} = \frac{(1-\hat{\pi}_E)}{n_{11}} + \frac{(1-\hat{\pi}_{\overline{E}})}{n_{21}} = \frac{(1-.240)}{47} + \frac{(1-.409)}{74} = .016 + .008 = .024$$

Finally, compute a 95% confidence interval for the population relative risk (recall that $z_{.025} = 1.96$).

$$\left(\hat{RR}e^{-z_{.05/2}\sqrt{v_{RR}}}, \hat{RR}e^{z_{.05/2}\sqrt{v_{RR}}}\right) \equiv \left(0.587e^{-1.96\sqrt{.024}}, 0.587e^{1.96\sqrt{.024}}\right)$$

$$\equiv (0.587(0.738), 0.587(1.355)) \equiv (0.433, 0.795)$$

Thus, we can be confident that the relative risk of suffering a skeletal event (in this time period) for patients on pamidronate (relative to patients not on pamidronate) is between 0.433 and 0.795. Since this entire interval is below 1.0, there is evidence that pamidronate is effective at reducing the risk of skeletal events. Further, an estimate is obtained that pamidronate changes the risk by $\left(\hat{RR} - 1\right)100\% = (0.587 - 1)100\% = -41.3\%$.

∇

The **Odds Ratio** is the ratio of odds of success for the two groups. First define the **odds** of an event occurring. If π is the probability that an event occurs, the odds o that it occurs is $o = \pi/(1-\pi)$. The odds can be interpreted as the number of times the event will occur for every time it will not occur if the process were repeated many times. For example, if you toss a coin, the probability it lands heads is $\pi = 0.5$. The corresponding odds of a head are $o = 0.5/(1-0.5) = 1.0$. Thus if you toss a coin many the times, the odds of a head are 1.0 (or 1-to-1 if you've ever been to a horse or dog track). Note that while odds are not probabilities, they are very much related to them: high probabilities are associated with high odds, and low probabilities are associated with low odds. In fact, for events with very low probabilities, the odds are very close to the probability of the event.

The ratio of the two odds is called the **odds ratio**. The odds ratio (OR) is similar to the relative risk, and is virtually equivalent to it when the prevalence of the disease ($P(D)$) is low. The odds ratio is computed as follows.

$$OR = \frac{\text{odds of disease given exposed}}{\text{odds of disease given unexposed}} = \frac{\text{odds of exposure given diseased}}{\text{odds of exposure given not diseased}} = \frac{n_{11}/n_{21}}{n_{12}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The odds ratio is similar to relative risk in the sense that it is a population parameter that must be estimated, as well as the interpretations associated with it in terms of whether its value is above, below, or equal to 1.0.

- If the odds ratio is greater than 1.0, the odds (and thus probability) of disease is higher among exposed (group 1) than unexposed (group 2).
- If the odds ratio is less than 1.0, the odds (and thus probability) of disease is lower among exposed (group 1) than unexposed (group 2).

- If the odds ratio is 1.0, the odds (and thus probability) of disease is the same for both groups (no association between exposure to risk factor and disease state).

The procedure to compute a $(1-\alpha)100\%$ confidence interval for the population odds ratio is as follows.

1. Obtain the estimated odds ratio: $\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$.
2. Compute $v_{OR} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}$ (this is the variance of $\log\left(\hat{OR}\right)$).
3. The confidence interval can be computed as: $\left(\hat{OR}e^{-z_{\alpha/2}\sqrt{v_{OR}}}, \hat{OR}e^{z_{\alpha/2}\sqrt{v_{OR}}}\right)$.

Example 8.8: Case-Control Study of Lip Cancer

An epidemiological case-control study was reported, with cases being 537 people diagnosed with lip cancer (D) and controls being made up of 500 people without lip cancer ($\overline{D}$) where all were patients at the Mayo Clinic (Broders (1920), [9]). One risk factor measured was whether or not the subject had smoked a pipe (pipe smoker E, non-pipe smoker $\overline{E}$). Table 8.7 gives the numbers of subjects falling in each lip cancer/pipe smoking combination. We would like to compute a 95% confidence interval for the population odds ratio, and determine whether or not pipe smoking is associated with higher (or possibly lower) odds (and probability) of contracting lip cancer.

		Occurrence of Lip Cancer		
		Yes (D)	No ($\overline{D}$)	
Pipe Smoking Status	Yes (E)	339	149	488
	No ($\overline{E}$)	198	351	549
		537	500	1037

Table 8.7: Observed cell counts for lip cancer/pipe smoking data

We compute the confidence interval as described above, again recalling that $z_{\alpha/2} = z_{0.025} = 1.96$:

1. $\hat{OR} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{339(351)}{149(198)} = 4.03$.
2. $v_{OR} = \frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}} = \frac{1}{339} + \frac{1}{149} + \frac{1}{198} + \frac{1}{351} = 0.0176$
3. 95% CI: $\left(\hat{OR}e^{-z_{\alpha/2}\sqrt{v_{OR}}}, \hat{OR}e^{z_{\alpha/2}\sqrt{v_{OR}}}\right) = \left(4.03e^{-1.96\sqrt{.0176}}, 4.03e^{1.96\sqrt{.0176}}\right) = (3.11, 5.23)$.

We can be 95% confident that the population odds ratio is between 3.11 and 5.23. That is the odds of contracting lip cancer is between 3.1 and 5.2 times higher among pipe smokers than non–pipe smokers. Note that in making the inference that pipe smoking *causes* lip cancer, it would need to demonstrated this association is present after controlling for other potential risk factors. Methods for controlling for other factors include the Mantel-Haenszel test given below.

∇

To understand why inference for the Relative Risk and Odds Ratio are based on the log of the measures as opposed to the estimate itself, consider the following example.

Example 8.9: Seat Belt Violations Among Traffic Stops by Gender

For the Charlotte, NC traffic stop data, there were $N_m = 46294$ traffic stops of male drivers and $N_f = 33590$ stops of female drivers. Among male drivers, there were 437 seat belt violations; among female drivers, there were 194. For this population, the following parameters are obtained.

$$\pi_m = \frac{437}{46294} = .00944 \quad \pi_f = \frac{194}{33590} = .00578 \quad RR = \frac{.00944}{.00578} = 1.6344 \quad OR = \frac{.00944/(1-.00944)}{.00578(1-.00578)} = 1.6405$$

A set of 10000 samples were obtained with sample sizes $n_m = n_f = 4000$, where estimated Relative Risks $\left(\hat{RR}\right)$, Odds Ratios $\left(\hat{OR}\right)$, and the estimated standard errors of their logs were computed and saved. Histograms of $\hat{RR}$, $\hat{OR}$, $\log\left(\hat{RR}\right)$ and $\log\left(\hat{OR}\right)$ given in Figure 8.1. Clearly, the "log" versions have sampling distributions that are approximately normal, while the untransformed versions do not. That is the basis for computing the Confidence Intervals as shown previously.

$$(1-\alpha)100\% \text{ CI for } log(RR) : \log\left(\hat{RR}\right) \pm z_{\alpha/2}\sqrt{v_{RR}} \equiv (A, B) \qquad (1-\alpha)100\% \text{ CI for } RR : \left(e^A, e^B\right)$$

The same logic applies for the Odds Ratio. Based on the 10000 samples, the Confidence Interval for the Relative Risk covered 1.6344 in 96.4% of the samples. The Odds Ratio (1.6405) was covered in 96.4% of the samples.

R Output

```
## Output

> round(RROR.out, 4)
       pi.m   pi.f  RR.mf  OR.mf RR.cov OR.cov
[1,] 0.0094 0.0058 1.6344 1.6405 0.9636 0.9636
```

∇

8.3.3 Small–Sample Inference — Fisher's Exact Test

The tests for association described previously all assume that the samples are sufficiently large so that the estimators have sampling distributions that are approximately normal. However, in many instances studies are based on small samples. This may arise due to cost or ethical reasons. A test due to R.A. Fisher, **Fisher's exact test**, is widely used in this particular situation. The logic of the test goes as follows.

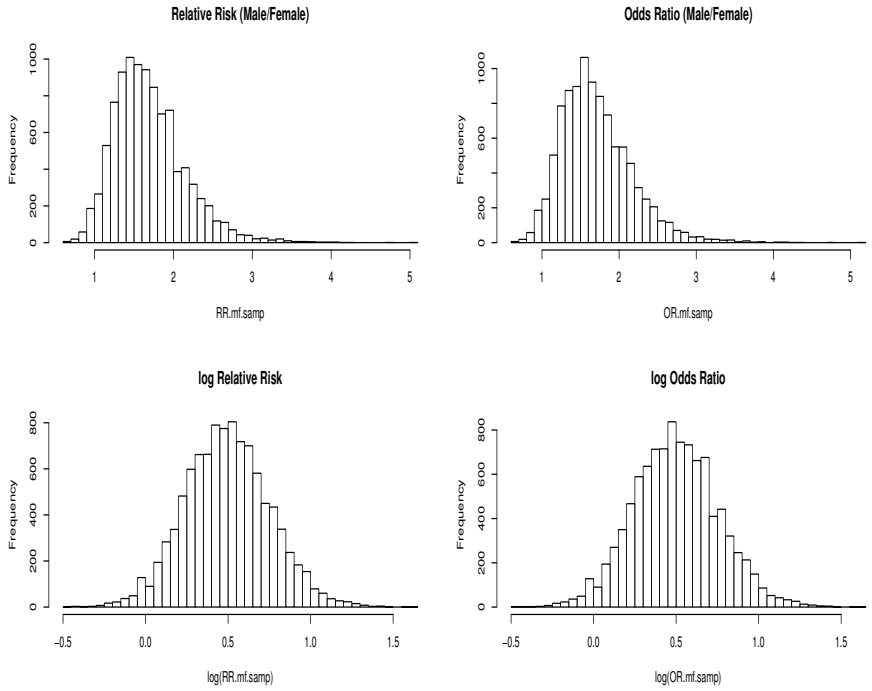

Figure 8.1: Seat Belt Violations Proportions by Gender: Relative Risk, Odds Ratio, log(RR), log(OR)

Take a sample with $n_{1\cdot}$ people (or experimental units) that are from group 1 and $n_{2\cdot}$ that are from group 2. Further observe $n_{\cdot 1}$ individuals that are "Successes," of which n_{11} were from group 1. The question is, conditional on the number from group 1 and the number of Successes, what is the probability that as many or more (fewer) of the events could have been from group 1 (under the assumption that there is no difference in the population). The test makes use of the **hypergeometric distribution**, and results in computing a probability of as strong or stronger evidence in favor of the alternative hypothesis than was observed (P-value). For a 1-tailed test $H_0 : \pi_1 \leq \pi_2$ versus $H_A : \pi_1 > \pi_2$, the P-value is obtained as follows.

$$P = \sum_{n=n_{11}}^{\min(n_{1\cdot},n_{\cdot 1})} \frac{\binom{n_{1\cdot}}{n}\binom{n_{2\cdot}}{n_{\cdot 1}-n}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}} \qquad \binom{a}{b} = \frac{a!}{b!(a-b)!} \qquad a \geq b$$

For a 1-tailed test $H_0 : \pi_1 \geq \pi_2$ versus $H_A : \pi_1 < \pi_2$, the P-value is obtained as follows.

$$P = \sum_{n=\min(0,n_{\cdot 1}-n_{2\cdot})}^{n_{11}} \frac{\binom{n_{1\cdot}}{n}\binom{n_{2\cdot}}{n_{\cdot 1}-n}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

For a 2-sided test, all cases where the absolute value of the difference in proportions is as large or larger than the observed difference are used in the calculation of the P-value.

Example 8.10: Early Use of Antiseptic in Amputations

A study was reported on the effects of antiseptic treatment among amputations in a British surgical hospital (Lister (1870), [37]). Tragically for Dr. Lister, he lived before Fisher, so he felt unable to make an inference based on statistical methodology, although he saw the effect was certainly there. Fisher's exact test can be used to make the inference. The study had two groups: one group based on amputation without antiseptic (years 1864-1866), and a group based on amputation with antiseptic (years 1867-1869). All surgeries were in the same hospital. We will consider the patients with antiseptic as the exposed (group 1). The endpoint reported was death (apparently due to the surgery and disease that was associated with it). The results are given in Table 8.8.

		Surgical Outcome		
		Death	No Death	
Treatment Group	Antiseptic (E)	6	34	40
	Control ($\overline{E}$)	16	19	35
		22	53	75

Table 8.8: Observed cell counts for antiseptic data

Note that this study is based on historical, as opposed to concurrent controls. There were 40 patients exposed to the antiseptic and 22 deaths, of which 6 were treated with antiseptic, and 16 in the untreated group. Now if the treatment is effective, it should reduce deaths, so we have to ask what is the probability that 6 or fewer of the 22 deaths could have been in the antiseptic group, given there were 40 patients in that group. More extreme cases would have been 0 deaths in group 1 (all 22 in group 2), up through 5 deaths in group 1 (17 in group 2). For a lower-tailed test (showing antiseptic reduces risk of death), the P-value is computed as follows.

$$\frac{\binom{40}{6}\binom{35}{16}}{\binom{75}{22}} + \frac{\binom{40}{5}\binom{35}{17}}{\binom{75}{22}} + \cdots + \frac{\binom{40}{0}\binom{35}{22}}{\binom{75}{22}} = .0037$$

That is, under the assumption of no treatment effect, the probability that based on a sample of this size, and this number of deaths, it is very unlikely that the sample results would have been this strong or stronger in favor of the antiseptic group. If we conduct the test with $\alpha = 0.05$, the p–value (.0037) is smaller than α, and we conclude that the antiseptic was associated with a lower probability of death.

For a 2-tailed test, the following additional computations are needed.

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{6}{40} - \frac{16}{35} = -.3071 \quad n_{11} = 17 \Rightarrow n_{21} = 5 \Rightarrow \hat{\pi}_1 - \hat{\pi}_2 = .2821 \quad n_{11} = 18 \Rightarrow n_{21} = 4 \Rightarrow \hat{\pi}_1 - \hat{\pi}_2 = .3357$$

$$\Rightarrow \quad P = .0037 + \frac{\binom{40}{18}\binom{35}{4}}{\binom{75}{22}} + \cdots + \frac{\binom{40}{22}\binom{35}{0}}{\binom{75}{22}} = .0037 + .0013 = .0050$$

R Commands and Output

```
## Commands
(lister <- matrix(c(6,34,16,19),byrow=T,ncol=2))

fisher.test(lister,alt="less")
```

```
fisher.test(lister,alt="two.sided")

## Output
> (lister <- matrix(c(6,34,16,19),byrow=T,ncol=2))
     [,1] [,2]
[1,]    6   34
[2,]   16   19
> fisher.test(lister,alt="less")
        Fisher's Exact Test for Count Data
data:  lister
p-value = 0.003685
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.0000000 0.5927603
sample estimates:
odds ratio
 0.2142773

> fisher.test(lister,alt="two.sided")
        Fisher's Exact Test for Count Data
data:  lister
p-value = 0.005018
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.05825328 0.69559373
sample estimates:
odds ratio
 0.2142773
```

∇

8.3.4 McNemar's Test for Paired Designs

When the same units are being observed under both experimental treatments (or units have been matched based on some criteria), McNemar's test can be used to test for treatment effects. The relevant subjects (pairs) are the ones who respond differently under the two conditions. Counts will appear as in Table 8.9.

		Trt 2 Outcome: Present	Trt 2 Outcome: Absent	
Trt 1 Outcome	Present	n_{11}	n_{12}	$n_{1.}$
	Absent	n_{21}	n_{22}	$n_{2.}$
		$n_{.1}$	$n_{.2}$	$n_{..}$

Table 8.9: Notation for McNemar's Test

Note that n_{11} is the number of units that have the outcome characteristic present under both treatments, while n_{22} is the number having the outcome characteristic absent under both treatments. None of these subjects offer any information regarding the difference in treatment effects. The units that provide information are the n_{12} cases that have the outcome present under treatment 1, and absent under treatment 2; and the n_{21} units that have the outcome absent under treatment 1, and present under treatment 2. Note that treatment 1 and treatment 2 can also be "Before" and "After" treatment, or any two conditions.

A large-sample test for treatment effects can be conducted as follows.

- H_0 : Pr(Outcome Present|Trt 1)=Pr(Outcome Present|Trt 2) $\Rightarrow$ No Trt effect
- H_A : The probabilities differ (Trt effects - This can be 1-sided also)
- $TS : z_{obs} = \frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}}$
- $RR : |z_{obs}| \geq z_{\alpha/2}$ (For 2-sided test)
- P-value: $2P(Z \geq |z_{obs}|)$ (For 2-sided test)

Often this test is reported as a chi-square test. The statistic is the square of the z-statistic above, and its treated as a chi-square random variable with one degree of freedom. The 2-sided z-test, and the chi-square test are mathematically equivalent.

An exact test is based on the binomial distribution. Under the null hypothesis of no treatment effect, the count n_{12} is distributed binomial with $n = n_{12} + n_{21}$ and $\pi = 0.5$. The P-value is computed as follows.

$$H_0 : \pi_1 = \pi_2 \quad H_A : \pi_1 \neq \pi_2 \quad P = 2 \min \left[P\left(Y \leq n_{12}\right), P\left(Y \geq n_{12}\right) \right] \quad Y \sim Bin\left(n = n_{12} + n_{21}, \pi = 0.5\right)$$

If trying to demonstrate that $\pi_1 > \pi_2$, we would expect $n_{12} > n_{21}$ and $P = P\left(Y \geq n_{12}\right)$. If the goal is to demonstrate that $\pi_1 < \pi_2$, we would expect $n_{12} < n_{21}$ and $P = P\left(Y \leq n_{12}\right)$.

Example 8.11: Framing of Risky Outcomes

In one of many studies testing prospect theory, subjects were asked to make two decisions regarding risky gambles (Kahneman and Tversky (1984), [31]). The decision choices are given below.

- Decision 1: Choose between (A): a sure gain of \$240 and (B): a 25% chance of winning \$1000 and 75% chance of winning \$0.
- Decision 2: Choose between (C): a sure loss of \$750 and (D): a 75% chance of losing \$1000 and a 25% chance of losing \$0.

The results are given below. Decision 1 is a Positive frame, Decision 2 is Negative. Choices A and C are "sure thing" selections, B and D are "risky."

- In 16 subjects, both sure things (A and C) were chosen.
- In 110 subjects, the Positive sure thing (A) and Negative risky bet (D) were chosen.
- In 4 subjects, the Positive risky bet (B) and Negative sure thing (C) were chosen.
- In 20 subjects, both risky bets (B and D) were chosen.

The data are summarized in Table 8.10.

We can test whether the tendency to choose between a sure thing and risky bet depends on whether the choice is framed positive (gain) or negative (loss) based on McNemar's test, since both outcomes are being observed on the same subjects.

		Negative Frame: Sure Thing	Negative Frame: Risky Bet	
Positive Frame	Sure Thing	16	110	126
	Risky Bet	4	20	24
		20	130	150

Table 8.10: Positive and Negative frames and subjects' selections between sure thing and risky bet

- H_0 : No differences in tendency to choose between sure thing and risky bet under the two frames
- H_A : The probabilities differ
- $TS : z_{obs} = \frac{110-4}{\sqrt{110+4}} = \frac{106}{10.6771} = 9.9278$
- $RR : |z_{obs}| \geq z_{.025} = 1.96$ (For 2-sided test, with $\alpha = 0.05$)
- P-value: $2P(Z \geq 9.9278) \approx 0$ (For 2-sided test)

Thus, we conclude that the tendencies differ. People tend to choose the sure thing when posed as a gain, and the risky bet when posed as a loss. The exact P-value is set-up below.

$$P = 2P\left(Y \geq 110 | Y \sim Bin(n = 114, \pi = 0.5)\right) \approx 0$$

R Commands and Output

```
## Commands

(bet <- matrix(c(16,110,4,20),byrow=T,,ncol=2))

mcnemar.test(bet,correct=F)
z.stat <- (bet[1,2]-bet[2,1])/sqrt(bet[1,2]+bet[2,1])
z.p <- 2*(1-pnorm(abs(z.stat),0,1))
binom.p <- 2*(1-pbinom(max(bet[1,2],bet[2,1])-1,bet[1,2]+bet[2,1],0.5))

bet.out <- cbind(bet[1,2], bet[2,1], z.stat, z.stat^2, z.p, binom.p)
colnames(bet.out) <- c("n12=+R/-S", "n21=+S/-R", "z", "z^2", "P(z)", "P(exact)")
round(bet.out, 4)

### Output

> (bet <- matrix(c(16,110,4,20),byrow=T,,ncol=2))
     [,1] [,2]
[1,]   16  110
[2,]    4   20
>
>> mcnemar.test(bet,correct=F)

        McNemar's Chi-squared test

data:  bet
McNemar's chi-squared = 98.561, df = 1, p-value < 2.2e-16

> round(bet.out, 4)
     n12=+R/-S n21=+S/-R      z     z^2 P(z) P(exact)
[1,]       110         4 9.9278 98.5614    0        0
```

The chi-square statistic from **mcnemar.test** is the square of the z-statistic. They give identical P-values for a 2-tailed test.

∇

8.3.5 Mantel–Haenszel Estimate for Stratified Samples

In some situations, the subjects in the study may come from one of several populations (strata). For instance, an efficacy study may have been run at multiple centers, and there may be some "center" effect that is related to the response. Another example is if race is related to the outcomes, and we may wish to adjust for race by computing odds ratios separately for each race, then combine them.

This is a situation where we would like to determine if there is an association between the explanatory and response variables, after *controlling* for a second explanatory variable. If there are k populations, then we can arrange the data (in a different notation than in the previous sections) as displayed in Table 8.11. Note that for each table, n_i is the sample size for that strata ($n_i = A_i + B_i + C_i + D_i$).

		Strata 1 Outcome				Strata k Outcome		
		Success	Failure)	Total		Success	Failure	Total
Group	1	A_1	B_1		... 1	A_k	B_k	
	2	C_1	D_1		... 2	C_k	D_k	
Total				n_1				n_k

Table 8.11: Contingency Tables for Mantel–Haenszel Estimator

The estimator of the odds ratio is computed as:

$$OR_{MH} = \frac{R}{S} = \frac{\sum_{i=1}^{k} R_i}{\sum_{i=1}^{k} S_i} = \frac{\sum_{i=1}^{k} A_i D_i / n_i}{\sum_{i=1}^{k} B_i C_i / n_i}$$

One estimate of the variance of the log of OR_{MH} is:

$$v = \hat{V}(ln(OR_{MH})) = \frac{1}{S^2} \sum_{i=1}^{k} S_i^2 \left(\frac{1}{A_i} + \frac{1}{B_i} + \frac{1}{C_i} + \frac{1}{D_i} \right)$$

As with the odds ratio, we can obtain a 95% CI for the population odds ratio as:

$$(OR_{MH} e^{-1.96\sqrt{v}}, OR_{MH} e^{1.96\sqrt{v}})$$

.

Example 8.12: Relationship Between Smoking and Death

A large study relating smoking habits and death rates reported that cigarette smoking was related to higher death rate (Hammond and Horn, (1954), [25]). Men were classified as regular cigarette smokers (E) and noncigarette smokers ($\overline{E}$). The nonsmokers had never smoked cigarettes regularly. There were a total

of 187,766 men who were successfully traced from the early 1952 start of study through October 31,1953. Of that group, 4854 (2.6%) had died.

A second variable that would clearly be related to death was age. In this study, all men were 50–69 at entry. The investigators then broke these ages down into four strata (50–54,55–59,60–64,65–69). The overall outcomes (disregarding age) are given in Table 8.12. Note that the overall odds ratio is $OR = (3002(78092))/(104820(1852)) = 1.21$.

		Occurrence of Death		
		Yes (D)	No ($\overline{D}$)	
Cigarette Smoking	Yes (E)	3002	104280	107822
Status	No ($\overline{E}$)	1852	78092	79944
		4854	182912	187766

Table 8.12: Observed cell counts for cigarette smoking/death data

The data, stratified by age group, are given in Table 8.13. Also, the odds ratios, proportion deaths ($P(D)$), and proportion smokers ($P(E)$) are given.

Age Group (i)	A_i	B_i	C_i	D_i	n_i	R_i	S_i	OR	$P(D)$	$P(E)$
50–54 (1)	647	39990	204	20132	60973	213.6	133.8	1.60	.0140	.6665
55–59 (2)	857	32894	394	21671	55816	332.7	232.2	1.43	.0224	.6047
60–64 (3)	855	20739	488	19790	41872	404.1	241.7	1.67	.0321	.5157
65–69 (4)	643	11197	766	16499	29105	364.5	294.7	1.24	.0484	.4068

Table 8.13: Observed cell counts and odds ratio calculations (by age group) for cigarette smoking/death data

Note that the odds ratio is higher within each group than it is for the overall group. This is referred to as *Simpson's Paradox*. In this case it can be explained as follows:

- Mortality increases with age from 1.40% for 50–54 to 4.84% for 65–69.
- As age increases, the proportion of smokers decreases from 66.65% to 40.68%
- A higher proportion of nonsmokers are in the higher risk (age) groups than are smokers. Thus, the nonsmokers are at a "disadvantage" because more of them are in the higher age groups (many smokers in the population have already died before reaching that age group).

This leads to an estimate of the odds ratio *adjusted for age*. That is what the Mantel–Haenszel estimator provides. It is computed as described above.

$$R = \sum_{i=1}^{4} R_i = 213.6 + 332.7 + 404.1 + 364.5 = 1314.9 \quad S = \sum_{i=1}^{4} S_i = 133.8 + 232.2 + 241.7 + 294.7 = 902.4$$

$$OR_{MH} = \frac{R}{S} = \frac{1314.9}{902.4} = 1.46$$

The estimated variance of $ln(OR_{MH})$ is 0.00095 (trust me). Then the following 95%CI for the odds ratio in the population of males in the age group 50–69 (adjusted for age) is obtained.

$$(OR_{MH}e^{-1.96\sqrt{v}}, OR_{MH}e^{1.96\sqrt{v}}) \equiv (1.46e^{-1.96\sqrt{.00095}}, 1.46e^{1.96\sqrt{.00095}}) \equiv (1.37, 1.55).$$

We can be very confident that the odds of death (during the length of time of the study – 20 months) is between 37% and 55% higher for smokers than nonsmokers, after controlling for age (among males in the 50–69 age group).

R Commands and Output

```
### Commands

### Enter data by COLUMNS within strata
(smoke <- array(c(647,204,39990,20132, 857,394,32894,21671,
   855,488,20739,19790, 643,766,11197,16499),dim=c(2,2,4)))

mantelhaen.test(smoke,exact=F,correct=F,alternative="two.sided")

### Output
> mantelhaen.test(smoke,exact=F,correct=F,alternative="two.sided")

        Mantel-Haenszel chi-squared test without continuity correction

data:  smoke
Mantel-Haenszel X-squared = 151.03, df = 1, p-value < 2.2e-16
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.372101 1.547602
sample estimates:
common odds ratio
         1.457212
```

∇

8.4 Nominal Explanatory and Response Variables

In cases where both the explanatory and response variables are nominal, the most commonly used method of testing for association between the variables is the **Pearson Chi–Squared Test**. In these situations, we are interested if the probability distributions of the response variable are the same at each level of the explanatory variable.

As we have seen before, the data represent counts, and appear as in Table 8.4. The n_{ij} values are referred to as the **observed** counts. If the variables are independent (not associated), then the population probability distributions for the response variable will be identical within each level of the explanatory variable, as in Table 8.14.

The special case of 2×2 tables has already been covered. Now generalize to r groups (treatments) and c possible outcomes. To perform Pearson's Chi–square test, compute the **expected** values for each cell count

		Response Variable				
		1	2	$\cdots$	c	
Explanatory Variable	1	p_1	p_2	$\cdots$	p_c	1.0
	2	p_1	p_2	$\cdots$	p_c	1.0
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	r	p_1	p_2	$\cdots$	p_c	1.0

Table 8.14: Probability distributions of response variable within levels of explanatory variable under condition of no association between the two variables.

under the hypothesis of independence, and obtain a statistic based on discrepancies between the observed and expected values.

$$\text{observed} = n_{ij} \qquad \text{expected} = E_{ij} = \frac{n_{i.}n_{.j}}{n_{..}}$$

The expected values represent how many individuals would have fallen in cell (i, j) if the probability distributions of the response variable were the same for each level of the explanatory (grouping) variable. They apply the marginal proportion of cases in column j, $n_{.j}/n_{..}$ to the number of units in row i, $n_{i.}$. The test is conducted as follows:

1. H_0 : No association between the explanatory and response variables (see Table 8.14).
2. H_A : Explanatory and response variables are associated
3. T.S.: $X^2_{obs} = \sum_{\text{all cells}} \frac{(\text{observed}-\text{expected})^2}{\text{expected}} = \sum_{i,j} \frac{(n_{ij}-E_{ij})^2}{E_{ij}}$
4. RR: $X^2_{obs} > \chi^2_{\alpha,(r-1)(c-1)}$
5. P–value: $P(\chi^2_{(r-1)(c-1)} \geq X^2_{obs})$

If the chi-square test rejects the null hypothesis, **standardized (adjusted) residuals** can be used to determine which cells are the "cause" of the association between the variables. These are like Z-statistics. Generally, standardized residuals larger than 2 or 3 in absolute values are considered to be evidence against independence in that cell.

$$R_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}\left(1 - \frac{n_{i.}}{n_{..}}\right)\left(1 - \frac{n_{.j}}{n_{..}}\right)}}$$

Example 8.13: Jury Decisions in Product Liability Cases

An experiment was conducted regarding jurors' decisions to award plaintiffs in product liability trials (Culp and Pollage (2002) [17]). The observed and expected values are given in Table 8.15. There were $r = 5$ treatments and $c = 2$ outcomes (award in favor of plaintiff, or not). The five conditions were as follows (all conditions included the jurors hearing the facts of the case).

1. Judge's instruction on strict liability and lawyer's oral arguments
2. Judge's instruction on negligence and lawyer's oral arguments
3. No judge's instruction or lawyer's oral arguments (Control)
4. Judge's instruction on strict liability but no lawyer's oral arguments
5. Judge's instruction on negligence but no lawyer's oral arguments

Jury Condition (i)	Award	No Award	Total
Strict Liability/Oral Argument (1)	15 (21.80)	43 (36.20)	58
Negligence/Oral Argument (2)	18 (17.66)	29 (29.34)	47
Control (3)	7 (14.66)	32 (24.34)	39
Strict Liability/No Oral Argument (4)	37 (28.19)	38 (46.81)	75
Negligence/No Oral Argument (5)	38 (32.70)	49 (54.30)	87
Total	115	191	306

Table 8.15: Observed (expected) values of numbers of jurors voting to award or not award plaintiff in product liability trial)

Overall, the proportion of jurors voting to award the plaintiff is $115/206 = .3758$, and the proportion voting no award is .6242. These proportions are applied to the row totals to obtain the expected counts under the hypothesis of no association between juror condition and voting outcome.

$$E_{11} = \left(\frac{115}{306}\right)(58) = 21.80 \quad E_{12} = \left(\frac{191}{306}\right)(58) = 36.20 \cdots E_{51} = \left(\frac{115}{306}\right)(87) = 32.70 \quad E_{52} = \left(\frac{191}{306}\right)(87) = 54.30$$

The test of whether there is an association between jury condition and vote outcome is conducted below.

H_0:Jury condition and voting outcome are independent vs H_A: Jury condition and voting outcome are associated.

$$TS: X^2_{obs} = \sum_{i=1}^{5}\sum_{j=1}^{2}\frac{(n_{ij}-E_{ij})^2}{E_{ij}} = \frac{(15-21.80)^2}{21.80} + \cdots + \frac{(49-54.30)^2}{54.30} = 2.121 + \cdots + 0.517 = 15.609$$

$$RR: X^2_{obs} \geq \chi^2_{.05,(5-1)(2-1)} = 9.488 \quad P = P\left(\chi^2_4 \geq 15.609\right) = .0036$$

The standardized residuals for the control treatment (Jury Condition 3) are -2.71 for Award and $+2.71$ for No Award, while those for Jury Condition 4 are $+2.42$ and -2.42, respectively. While these do not exceed 3 in absolute value, they are well above 2. Fewer jurors in the Control Group voted to award the plaintiff than expected under independence, and more voted to award the plaintiff in Jury Condition 4. The calculations for the Control Group are given below.

$$R_{31} = \frac{7-14.66}{\sqrt{14.66(1-39/306)(1-115/306)}} = \frac{-7.66}{2.83} = -2.71 \quad R_{32} = \frac{32-24.34}{\sqrt{24.34(1-39/306)(1-191/306)}} = \frac{7.66}{2.83} = 2.71$$

R Commands and Output

```
## Commands

pla <- read.csv("http://www.stat.ufl.edu/~winner/data/productliability_award.csv")
attach(pla); names(pla)

(jury_award <- table(jury,award))

X2_ja <- chisq.test(jury_award, correct=F)
X2_ja
X2_ja$stdres

## Output

> (jury_award <- table(jury,award))
    award
jury  0  1
   1 43 15
   2 29 18
   3 32  7
   4 38 37
   5 49 38
> X2_ja
        Pearson's Chi-squared test
data:  jury_award
X-squared = 15.608, df = 4, p-value = 0.003592
> X2_ja$stdres
    award
jury          0          1
   1  2.0470036 -2.0470036
   2 -0.1101878  0.1101878
   3  2.7100635 -2.7100635
   4 -2.4184629  2.4184629
   5 -1.3878162  1.3878162
```

∇

8.5 Ordinal Explanatory and Response Variables

In situations where both the explanatory and response variables are ordinal, we would like to take advantage of the fact that the levels of the variables have distinct orderings. We can ask questions such as: Do individuals with high levels of the explanatory variable tend to have high (low) levels of the corresponding response variable. For instance, suppose that the explanatory variable is dose, with increasing (possibly numeric) levels of amount of drug given to a subject, and the response variable is an ordinal measure (possibly subjective) of degree of improvement. Then, we may be interested in seeing if as dose increases, the degree of improvement increases (this is called a dose–response relationship).

Various measures have been developed for this type of experimental setting. Most are based on **concordant** and **discordant** pairs. Concordant pairs involve pairs where one unit scores higher on both variables than the other unit. Discordant pairs are pairs where one unit scores higher on one variable, but lower on the other variable, than the other unit.

In cases where there is a **positive association** between the two variables, we would expect more concordant than discordant pairs. That is, there should be many units that score high on both variables,

and many that score low on both, with fewer units scoring high one variable and low on the other. On the other hand, if there is a **negative association**, we would expect more discordant pairs than concordant pairs. That is, units will tend to score high on one variable, but lower on the other.

Two commonly reported measures of ordinal association are **gamma** and **Kendall's** τ_b. Both of these measures lie between -1 and 1. Negative values correspond to negative association, and positive values correspond to positive association. These types of association were described previously. A value of 0 implies no association between the two variables. Here, we give the formulas for the point estimates, their standard errors are better left to computers to handle. Tests of hypothesis and confidence intervals for the population measure are easily obtained from large–samples.

The point estimators for **gamma** and **Kendall's** τ_b are given below, where C is the number of concordant pairs and D is the number of discordant pairs.

$$\hat{\gamma} = \frac{C-D}{C+D} \qquad \hat{\tau}_b = \frac{C-D}{0.5\sqrt{(n_{\cdot\cdot}^2 - \sum n_{i\cdot}^2)(n_{\cdot\cdot}^2 - \sum n_{\cdot j}^2)}}$$

To conduct a large–sample test of whether or not the population parameter is 0 (that is, a test of association between the explanatory and response variables), we complete the following steps:

1. $H_0 : \gamma = 0$ (No association)
2. $H_A : \gamma \neq 0$ (Association exists)
3. T.S.: $z_{obs} = \frac{\hat{\gamma}}{\hat{SE}\{\hat{\gamma}\}}$
4. R.R.: $|z_{obs}| \geq z_{\alpha/2}$
5. p–value:$2P(z \geq |z_{obs}|)$

For a test concerning Kendall's τ_b, replace γ with τ_b. For a $(1-\alpha)100\%$ CI for the population parameter, simply compute the following (this time we use τ_b).

$$\hat{\tau}_b \pm z_{\alpha/2}\hat{SE}\{\hat{\tau}_b\}$$

Example 8.14: Jurors' Vote on Capital Trial

A study considered jurors' first vote in a capital trial and their view of the defendant's level of remorse (Eisenberg, Garvey, and Wells, 2001, [21]). The variables are described below.

- Level of defendant's remorse was classified by the answer to: How well does 'sorry' describe the defendant? with levels: Not at all, Not so well, Fairly well, Very Well.
- The juror's first vote was classified as: Life Imprisonment, Undecided between Life and Death, Death Penalty.

	Juror's First Vote			
Remorse	Life	Undecided	Death	Total
Not at all	12	6	56	74
Not so well	15	7	32	54
Fairly well	12	7	9	28
Very well	16	2	8	26
Total	55	22	105	182

Table 8.16: Numbers of subjects within each defendant's remorse and juror's first vote status combination

Observed counts are given in Table 8.16.

Concordant pairs are pairs where one subject scores higher on each variable than the other subject. Thus, all subjects in the "Very well" remorse group who voted "Death" are concordant with all subjects who had a lower remorse score and voted for "Life" or "Undecided." Similarly, all subjects in the "Not so well" remorse group and voted "Undecided" are concordant with all subjects in the "Not at all" remorse group and voted for "Life." The total number of concordant pairs (C) is:

$$C = 8(12+6+15+7+12+7) + 2(12+15+12) + 9(12+6+15+7) + 7(12+15) + 32(12+6) + 7(12) =$$

$$472 + 78 + 360 + 189 + 576 + 84 = 1759$$

Discordant pairs are pairs where one subject scores higher on one variable, but lower on the other variable than the other subject. Thus, all subjects in the "Very well" remorse group who voted "Life" are discordant with all subjects who had a lower remorse score and voted for "Undecided" or "Death." Similarly, all subjects in the "Not so well" remorse group and voted "Undecided" are discordant with all subjects in the "Not at all" remorse group and voted for "Death." Thus, the total number of discordant pairs (D) is:

$$D = 16(56+6+32+7+9+7) + 2(56+32+9) + 12(56+6+32+7) + 7(56+32) + 15(56+6) + 7(56) =$$

$$1872 + 194 + 1212 + 616 + 930 + 392 = 5216$$

Notice that there are more discordant pairs than concordant pairs. This is consistent with tougher judgments for defendants displaying lower levels of remorse.

$$\hat{\gamma} = \frac{C-D}{C+D} = \frac{1759-5216}{1759+5216} = \frac{-3457}{6975} = -0.496$$

$$\hat{\tau}_b = \frac{C-D}{0.5\sqrt{(n_{..}^2 - \sum n_{i.}^2)(n_{..}^2 - \sum n_{.j}^2)}}$$

$$= \frac{1759-5216}{0.5\sqrt{\left[182^2 - (74^2+54^2+28^2+26^2)\right]\left[182^2 - (55^2+22^2+105^2)\right]}}$$

$$\frac{-3457}{0.5\sqrt{(23272)(18590)}} = \frac{-3457}{10400} = -0.332$$

R Commands and Output

```
## Commands

rd1 <- read.csv("http://www.stat.ufl.edu/~winner/data/remorse_death.csv")
attach(rd1); names(rd1)

install.packages("vcdExtra")
library(vcdExtra)

(rd.table <- table(remorse,jurVote))
GKgamma(rd.table)
cor.test(remorse,jurVote, method="kendall")

## Output

> (rd.table <- table(remorse,jurVote))
       jurVote
remorse  1  2  3
      1 12  6 56
      2 15  7 32
      3 12  7  9
      4 16  2  8
> GKgamma(rd.table)
gamma        : -0.496
std. error   : 0.083
CI           : -0.659 -0.332
> cor.test(remorse,jurVote, method="kendall")
        Kendall's rank correlation tau
data:  remorse and jurVote
z = -5.0313, p-value = 4.873e-07
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
-0.332409
```

The 95% Confidence Interval for the population value of γ is well away from 0, and the test for Kendall's τ_B rejects the null hypothesis of no association. There is strong evidence for a negative association between defendant's remorse and juror's vote.

∇

8.6 Nominal Explanatory and Ordinal Response Variable

In the case where the explanatory variable is nominal and the response variable is ordinal, the **Kruskal–Wallis Test** can be used, which was described in Chapter 7.

1. H_0 : The probability distributions of the ordinal response variable are the same for each level of the explanatory variable (treatment group). (No association)
2. H_A : The probability distributions of the response variable are the not same for each level of the explanatory variable. (Association).
3. T.S.: $H = \frac{12}{n_{..}(n_{..}+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_{i.}} - 3(n_{..}+1)$

4. R.R.: $H > \chi^2_{\alpha,r-1}$
5. P-value: $P(\chi^2_{r-1} \geq H)$

The adjustment made for ties is given below. In this setting, there will generally be many ties.

$$H' = \frac{H}{\left[1 - \frac{\sum(t_j^3 - t_j)}{n_{..}^3 - n_{..}}\right]} \qquad t_j \equiv \text{ number of observations in the } j^{th} \text{ group of tied ranks}$$

Example 8.15: Soccer Game Outcomes Among European Premier Leagues - 2013/2014 Season

Soccer (football) games in premier leagues can end in one of three ordinal ways for the home team (Lose, Draw, Win). Treating the regular season games for the $r = 5$ leagues: England, France, Germany, Italy, and Spain as samples from conceptual populations of all possible games that could be played, the Kruskal-Wallis test is applied to determine whether the distributions differ. Table 8.17 contains the numbers of each outcome categories, the ranks, and the rank sums for each national league.

	Game Outcome			
League ($n_{i.}$)	Lose	Draw	Win	Sum (T_i)
England ($n_{1.} = 380$)	123	78	179	343978.5
France ($n_{2.} = 380$)	104	109	167	345573.0
Germany ($n_{3.} = 306$)	97	64	145	278539.5
Italy ($n_{4.} = 380$)	109	90	181	352081.5
Spain ($n_{5.} = 380$)	115	86	179	347878.5
# Matches	548	427	851	
Ranks	1–548	549–975	976–1826	
Avg. Rank	274.5	762	1401	

Table 8.17: Data and ranks for European Premier League Game Outcomes ($n_{..} = 1826$)

To obtain T_1, the rank sum for England, note that 123 of the games received the rank of 274.5 (the rank assigned to each loss), 78 received the rank of 762, and 179 received the rank of 1401.

$$T_1 = 123(274.5)+78(762)+179(1401) = 343978.5 \qquad \cdots \qquad T_5 = 115(274.5)+86(762)+179(1401) = 347878.5$$

Here, we will test whether (H_A) or not (H_0) the distributions of game outcomes differ among the five leagues. The test statistic is computed as follows.

$$H = \frac{12}{n_{..}(n_{..}+1)} \sum_{i=1}^{r} \frac{T_i^2}{n_i} - 3(n_{..}+1) =$$

$$\frac{12}{1826(1827)} \left(\frac{(343978.5)^2}{380} + \frac{(345573.0)^2}{380} + \frac{(278539.5)^2}{306} + \frac{(352081.5)^2}{380} + \frac{(347878.5)^2}{380} \right) - 3(1827) =$$

$$\frac{12}{1826(1827)}(311371601.2+314264995.6+253543310.7+326214164.8+318472238.8)-5481 = 5481.366-5481 = 0.366$$

There are many ties, so the adjustment is computed as follows.

$$t_1 = 548 \quad t_2 = 427 \quad t_3 = 851 \quad \sum_{j=1}^{3}\left(t_j^3 - t_j\right) = 858714300 \quad n_{\cdot\cdot}^3 - n_{\cdot\cdot} = 6088386150$$

$$H' = \frac{0.366}{1 - \frac{858714300}{6088386150}} = \frac{0.366}{0.859} = 0.426 \quad RR : H' \geq \chi^2_{.05,5-1} = 9.488 \quad P = P\left(\chi^2_4 \geq 0.426\right) = .9803$$

There is no evidence of differences among the leagues.

R Commands and Output

```
## Commands

euro13 <- read.csv("http://www.stat.ufl.edu/~winner/data/europesoccer2013.csv")
attach(euro13); names(euro13)

home.result <- ifelse(DiffGoal<0,0,ifelse(DiffGoal==0,1,2))   ## Assign 0 for loss, 1 for Tie, 2 for Win
League <- factor(League)

kruskal.test(home.result ~ League)

## Output

> kruskal.test(home.result ~ League)
        Kruskal-Wallis rank sum test
data:  home.result by League
Kruskal-Wallis chi-squared = 0.42598, df = 4, p-value = 0.9803
```

∇

8.7 Assessing Agreement Among Raters

As mentioned in Chapter 1, in many situations the response being measured is an assessment made by an investigator. For instance, in food or beverage tasting experiments, the response may be quality (color, taste, texture, smoothness), which would involve rating a product along some sort of Likert (ordinal) scale. Various varieties of soy sauce's color may be rated by judges on an ordinal scale of Light Brown, Brown, Intense Brown, Black. Unfortunately measurements such as these are much more subjective than mechanical measures such as viscosity or salt content. In many instances, a pair (or more) of raters may be used, and the level of their agreement is to be determined.

A measure of agreement that was developed in psychiatric diagnosis is **Cohen's** κ. It measures the proportion of agreement beyond chance agreement. It can take on negative values when the agreement is

worse than expected by chance, and the largest value it can take is 1.0, which occurs when there is perfect agreement. Suppose there are k categories. Let p_{ij} be the proportion of items rated as category i by rater 1 and category j by rater 2. Further, let $p_{i.}$ be the marginal proportion for category i by rater 1 and $p_{.j}$ be the marginal proportion for category j by rater 2 (see Table 8.18 below for a numeric example). Then the observed agreement, p_{obs}, and the expected (by chance) agreement, p_{exp}, are computed as follows, as well as Cohen's κ.

$$p_{\text{obs}} = p_{11} + \cdots + p_{kk} = \sum_{i=1}^{k} p_{ii} \qquad p_{\text{exp}} = p_{1.}p_{.1} + \cdots + p_{k.}p_{.k} = \sum_{i=1}^{k} p_{i.}p_{.i}$$

$$\hat{\kappa} = \frac{p_{\text{obs}} - p_{\text{exp}}}{1 - p_{\text{exp}}}$$

The standard error of $\hat{\kappa}$ is messy to compute, but can be obtained by various software packages.

While κ only detects disagreement, a modification, called **weighted** κ distinguishes among levels of disagreement when categories are ordered. That is, raters who disagree by one category are in stronger agreement than raters who differ by several categories. Weighted κ can use any weighting scheme, a very common one is to use is as follows.

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \qquad p^w_{\text{obs}} = \sum_{i=1}^{k}\sum_{i=1}^{k} w_{ij}p_{ij} \qquad p^w_{\text{exp}} = \sum_{i=1}^{k}\sum_{i=1}^{k} w_{ij}p_{i.}p_{.j}$$

$$\hat{\kappa}_w = \frac{p^w_{\text{obs}} - p^w_{\text{exp}}}{1 - p^w_{\text{exp}}}$$

Example 8.16: Agreement Among Professional Movie Reviewers

A study compared the level of agreement among popular movie critics (Agresti and Winner, 1997, [3]). The pairwise levels of agreement among 8 critics (Gene Siskel, Roger Ebert, Michael Medved, Jeffrey Lyons, Rex Reed, Peter Travers, Joel Siegel, and Gene Shalit) were computed. In this example, we will focus on Siskel and Ebert. There were 160 movies that both critics reviewed during the study period, the results are given in Table 8.18, which is written as a 3×3 contingency table. The ratings are from the trade publication *Variety* which evaluated critics' reviews as Con (low), Mixed (medium), and Pro (high).

If their ratings were independent (that is, knowledge of Siskel's rating gives no information as to Ebert's rating on the same movie), we would expect the following probabilities along the main diagonal (where the critics agree):

$$p_{11} = P(\text{Con}|\text{Siskel}) \cdot P(\text{Con}|\text{Ebert}) = (.281)(.263) = .074$$

$$p_{22} = P(\text{Mixed}|\text{Siskel}) \cdot P(\text{Mixed}|\text{Ebert}) = (.200)(.188) = .038$$

$$p_{33} = P(\text{Pro}|\text{Siskel}) \cdot P(\text{Pro}|\text{Ebert}) = (.281)(.263) = .285$$

So, even if their ratings were independent, we would expect the proportion of movies that they would agree on by chance to be $p_c = .074 + .038 + .285 = .397$. That is, we would expect them to agree about 40% of

Siskel Rating	Ebert Rating: Con	Mixed	Pro	Total
Con	24	8	13	45
	(.150)	(.050)	(.081)	(.281)
	(.074)	(.053)	(.155)	—
Mixed	8	13	11	32
	(.050)	(.081)	(.069)	(.200)
	(.053)	(.038)	(.110)	—
Pro	10	9	64	83
	(.063)	(.056)	(.400)	(.519)
	(.136)	(.098)	(.285)	—
Total	42	30	88	160
	.263	.188	.550	1.00

Table 8.18: Ratings on $n = 160$ movies by Gene Siskel and Roger Ebert – raw counts, observed proportions, and proportions expected under chance

the time, based on their marginal distributions. In fact, the observed proportion of movies for which they agree on is $p_o = .150 + .081 + .400 = .631$, so they agree on about 63% of the movies. We can now compute Cohen's κ:

$$\kappa = \frac{\text{observed agreement} - \text{chance agreement}}{1 - \text{chance agreement}} = \frac{.631 - .397}{1 - .397} = \frac{.234}{.603} = .388$$

This would be considered a moderate level of agreement. The sample difference between the observed agreement and the agreement expected under independence is 39% of the maximum possible difference.

R Commands and Output

```
### Commands

(siskel_ebert <- matrix(c(24,8,13,8,13,11,10,9,64),byrow=T,ncol=3))

install.packages("psych")
library(psych)
cohen.kappa(siskel_ebert)

### Output

> cohen.kappa(siskel_ebert)
Call: cohen.kappa1(x = x, w = w, n.obs = n.obs, alpha = alpha, levels = levels)

Cohen Kappa and Weighted Kappa correlation coefficients and confidence boundaries
                 lower estimate upper
unweighted kappa  0.27     0.39  0.51
weighted kappa    0.32     0.46  0.60

 Number of subjects = 160
```

∇

8.8 R Code for Chapter 8

```
### Chapter 8

## Example 8.1

shaq_ft <- read.csv("http://www.stat.ufl.edu/~winner/data/shaqoneal_gamestats.csv")
attach(shaq_ft); names(shaq_ft)
N.game <- length(FTA)
N.FT <- sum(FTA)
FT_01 <- rep(0,N.FT)
FT_count <- 0
for (i in 1:N.game) {
if (FTA[i] > 0) {
FT_01[(FT_count+1):(FT_count+FT[i])] <- 1
FT_count <- FT_count + FTA[i]
}
}
FT_pi <- sum(FT)/sum(FTA)
num.samp <- 100000
n.samp <- c(10, 30, 100)
set.seed(24680)
FT.y <- matrix(rep(0,3*num.samp), ncol=3)
CI1 <- matrix(rep(0,6*num.samp),ncol=6)
CI2 <- matrix(rep(0,6*num.samp),ncol=6)
z025 <- qnorm(.975,0,1)

for(i1 in 1:3) {
for (i2 in 1:num.samp) {
FT.y[i2,i1] <- sum(sample(FT_01, n.samp[i1], replace=F))
FT.pihat <- FT.y[i2,i1]/n.samp[i1]
CI1[i2,((i1-1)*2+1):((i1-1)*2+2)] <- FT.pihat +
      c(-1,1) * z025 * sqrt(FT.pihat*(1-FT.pihat)/n.samp[i1])
FT.y.tilde <- FT.y[i2,i1] + 0.5*z025^2
FT.n.tilde <- n.samp[i1] + z025^2
FT.pitil <- FT.y.tilde / FT.n.tilde
CI2[i2,((i1-1)*2+1):((i1-1)*2+2)] <- (FT.pitil) +
   c(-1,1) * z025 * sqrt(FT.pitil*(1-FT.pitil)/FT.n.tilde)
}}

wald1 <- sum(CI1[,1] <= FT_pi & CI1[,2] >= FT_pi) / num.samp
wald2 <- sum(CI1[,3] <= FT_pi & CI1[,4] >= FT_pi) / num.samp
wald3 <- sum(CI1[,5] <= FT_pi & CI1[,6] >= FT_pi) / num.samp

wac1 <- sum(CI2[,1] <= FT_pi & CI2[,2] >= FT_pi) / num.samp
wac2 <- sum(CI2[,3] <= FT_pi & CI2[,4] >= FT_pi) / num.samp
wac3 <- sum(CI2[,5] <= FT_pi & CI2[,6] >= FT_pi) / num.samp

ft.out <- rbind(cbind(FT_pi, wald1, wac1,
                 mean(CI1[,2]-CI1[,1]), mean(CI2[,2]-CI2[,1])),
                cbind(FT_pi, wald2, wac2,
                 mean(CI1[,4]-CI1[,3]), mean(CI2[,4]-CI2[,3])),
                cbind(FT_pi, wald3, wac3,
                 mean(CI1[,6]-CI1[,5]), mean(CI2[,6]-CI2[,5])))
rownames(ft.out) <- c("n=10", "n=30", "n=100")
colnames(ft.out) <- c("pi", "pi-hat cover", "pi-tilde cover",
                      "pi-hat mean width", "pi-tilde mean width")
round(ft.out, 4)

rm(list=ls(all=TRUE))

### Example 8.2
```

```
nbaou <- read.csv("http://www.stat.ufl.edu/~winner/data/nbaodds201415.csv")
attach(nbaou); names(nbaou)

#### Point Spread Analysis
table(TeamCov)
TeamCov01 <- subset(TeamCov, TeamCov != 0)   # Remove Pushes
table(TeamCov01)
Y.Cov <- sum(TeamCov01[TeamCov01 == 1])      # Games Home Team Covers
n.Cov <- length(TeamCov01)                   # Number of Games
pi.H0 <- 0.50                                # Null value for pi
pihat.Cov <- Y.Cov / n.Cov                   # Point estimate
se.pihat.Cov.CI <- sqrt(pihat.Cov * (1-pihat.Cov) / n.Cov)   # Std Error for CI
se.pihat.Cov.H0 <- sqrt(pi.H0 * (1-pi.H0) / n.Cov)           # Std Error for Z-test
Z.Cov.H0 <- (pihat.Cov - pi.H0) / se.pihat.Cov.H0            # Z-statistic
p.Cov.H0 <- 2*(1-pnorm(abs(Z.Cov.H0),0,1))                   # P-value for Z-test
pihat.Cov.CI <- pihat.Cov + c(-1.96, 1.96) * se.pihat.Cov.CI # Large-sample 95% CI
p.Cov.H0.exact <- pbinom(min(Y.Cov, n.Cov-Y.Cov),n.Cov, pi.H0) +
                1-pbinom(max(Y.Cov, n.Cov-Y.Cov)-1,n.Cov, pi.H0)  # Exact P-value

cov.out <- cbind(pi.H0, Y.Cov, n.Cov, pihat.Cov, se.pihat.Cov.H0, Z.Cov.H0,
    p.Cov.H0, p.Cov.H0.exact, se.pihat.Cov.CI, pihat.Cov.CI[1],
    pihat.Cov.CI[2])
colnames(cov.out) <- c("pi(H0)", "y", "n", "pihat", "SE{H0}", "Z", "P(Z)",
      "P(Exact})", "SE{pihat}", "Lower", "Upper")
round(cov.out, 4)
### Exact Tests
binom.test(Y.Cov,n.Cov,p=0.5,alternative="two.sided")

rm(list=ls(all=TRUE))

### Example 8.3

### Default probs are 1/#categories
chisq.test(c(306,338,432,348,331,309),p=c(1/6,1/6,1/6,1/6,1/6,1/6))

rm(list=ls(all=TRUE))

### Example 8.4

(mean.bomb <- 537/576)
p0 <- dpois(0,mean.bomb)     ### p(0) for Poisson(mean.bomb)
p1 <- dpois(1,mean.bomb)
p2 <- dpois(2,mean.bomb)
p3 <- dpois(3,mean.bomb)
p4 <- 1-p0-p1-p2-p3

obs.bomb <- c(229,211,93,35,8)
(exp.bomb <- sum(obs.bomb)*c(p0,p1,p2,p3,p4))

X2.stat <- sum((obs.bomb-exp.bomb)^2 / exp.bomb)
X2.df <- length(obs.bomb)-1-1
X2.05 <- qchisq(.95,X2.df)
X2.p <- 1-pchisq(X2.stat,X2.df)

bomb.out <- cbind(X2.stat, X2.df, X2.05, X2.p)
colnames(bomb.out) <- c("X2 stat", "DF", "X2(.05)", "P-value")
round(bomb.out, 4)

rm(list=ls(all=TRUE))

### Example 8.5

## Read data from website and attach data frame and obain variable names
rr.mar <- read.csv(
```

```
  "http://www.stat.ufl.edu/~winner/data/rocknroll_marathon_mf2015a.csv")
attach(rr.mar); names(rr.mar)

### Select only males and obtain Method of Moments Estimates of alpha, beta
m.mph <- mph[Gender=="M"]
(alpha.m <- mean(m.mph)^2 / var(m.mph))
(beta.m <- mean(m.mph) / var(m.mph))
n.mph <- length(m.mph)

### Assign bottom and top of cell ranges and obtain counts within cells
cell.bot <- c(0,seq(4.750000001,9.250000001,0.50))
cell.top <- c(seq(4.75,9.25,0.50),9999999)
n.cells <- rep(0,length(cell.bot))
tot.n.cells <- rep(0,length(cell.bot))
n.cells[1] <- sum(m.mph <= cell.top[1])
tot.n.cells[1] <- n.cells[1]
alpha1 <- alpha.m
beta1 <- beta.m
for (i in 2:length(cell.bot)) {
n.cells[i] <- sum(m.mph <= cell.top[i]) - tot.n.cells[i-1]
tot.n.cells[i] <- tot.n.cells[i-1] + n.cells[i]
}

### Set up computations by cell for chi-square statistic and compute it
X2.ab.cell <- rep(0,length(cell.bot))
X2.ab <- 0
exp.cells <- rep(0,length(cell.bot))
for (i in 1:length(cell.bot)) {
exp.cells[i] <- n.mph*
  (pgamma(cell.top[i],alpha1,beta1)-pgamma(cell.bot[i],alpha1,beta1))
X2.ab.cell[i] <- ((n.cells[i] - exp.cells[i])^2)/exp.cells[i]
X2.ab <- X2.ab + X2.ab.cell[i]
}

X2.df <- length(cell.bot)-1-2
X2.CV <- qchisq(.95,X2.df)
X2.pval <- 1-pchisq(X2.ab,X2.df)

X2.out <- cbind(alpha1,beta1,X2.ab,X2.df,X2.CV,X2.pval)
colnames(X2.out) <- c("alpha","beta","Test Stat","DF","X2(.05)","P-value")
round(X2.out,4)
round(cbind(cell.top,n.cells,tot.n.cells,exp.cells,X2.ab.cell),3)

rm(list=ls(all=TRUE))

### Example 8.6

y1 <- 75; n1 <- 94 ## Successes and Total for Group 1 (Affected by Flood)
y2 <- 53; n2 <- 107 ## Successes and Total for Group 2 (Unaffected)

pihat.1 <- y1/n1
pihat.2 <- y2/n2
pihat <- (y1+y2)/(n1+n2)
se.pihat.12 <- sqrt((pihat.1*(1-pihat.1)/n1)+(pihat.2*(1-pihat.2)/n2))
se.pihat.12p <- sqrt(pihat*(1-pihat)*(1/n1+1/n2))
z025 <- qnorm(.975,0,1)

pi12.ci <- (pihat.1-pihat.2) + c(-z025,z025)*se.pihat.12    # 95%CI for pi1-pi2
pi12.z <- (pihat.1-pihat.2)/se.pihat.12p                    # Z_obs for H0:pi1-pi2=0
pi12.p <- 2 * (1-pnorm(abs(pi12.z)))                        # 2-sided P-value

pi12.out <- cbind(y1, y2, n1, n2, pihat.1, pihat.2, pihat, se.pihat.12, pi12.ci[1],
     pi12.ci[2], se.pihat.12p, pi12.z, pi12.p)
colnames(pi12.out) <- c("y1", "y2", "n1", "n2", "pihat1", "pihat2", "pooled",
```

```
    "SE{Diff}", "Lower",  "Upper", "SE{(H0)}", "Z", "P-value")
round(pi12.out, 4)

prop.test(c(y1,y2),c(n1,n2),correct=F)

rm(list=ls(all=TRUE))

### Example 8.9

## Read data off web page, attach file as data frame, and list variable names
clt2016 <- read.csv("http://www.stat.ufl.edu/~winner/data/trafficstop.csv")
attach(clt2016); names(clt2016)

RsnStop.m <- RsnStop[DrvMale==1]
RsnStop.f <- RsnStop[DrvMale==0]
seatbelt.m <- ifelse(RsnStop.m==6,1,0)
seatbelt.f <- ifelse(RsnStop.f==6,1,0)
N.m <- length(seatbelt.m)
N.f <- length(seatbelt.f)

pi.m <- sum(seatbelt.m)/N.m
pi.f <- sum(seatbelt.f)/N.f
RR.mf <- pi.m/pi.f
OR.mf <- (pi.m*(1-pi.f)) / (pi.f*(1-pi.m))

n.samp <- 4000
num.sim <- 10000
set.seed(12345)
RR.mf.samp <- rep(0,num.sim)
OR.mf.samp <- rep(0,num.sim)
lnRR.se <- rep(0,num.sim)
lnOR.se <- rep(0,num.sim)

for (i in 1:num.sim) {
samp.m <- sample(seatbelt.m,n.samp,replace=F)
samp.f <- sample(seatbelt.f,n.samp,replace=F)
y.m <- sum(samp.m)
y.f <- sum(samp.f)
RR.mf.samp[i] <- y.m/y.f
OR.mf.samp[i] <- (y.m*(n.samp-y.f)) / (y.f*(n.samp-y.m))
lnRR.se[i] <- sqrt((1-y.m/n.samp)/y.m + (1-y.f/n.samp)/y.f)
lnOR.se[i] <- sqrt(1/y.m + 1/(n.samp-y.m) + 1/y.f + 1/(n.samp-y.f))
}

z.025 <- qnorm(.975,0,1)
RR.LB <- RR.mf.samp*exp(-z.025*lnRR.se)
RR.UB <- RR.mf.samp*exp(z.025*lnRR.se)
OR.LB <- OR.mf.samp*exp(-z.025*lnOR.se)
OR.UB <- OR.mf.samp*exp(z.025*lnOR.se)
RR.cov <- sum(RR.LB <= RR.mf & RR.UB >= RR.mf) / num.sim
OR.cov <- sum(OR.LB <= OR.mf & OR.UB >= OR.mf) / num.sim

RROR.out <- cbind(pi.m, pi.f, RR.mf, OR.mf, RR.cov, OR.cov)
colnames(RROR.out) <- cbind("pi.m", "pi.f", "RR.mf", "OR.mf", "RR.cov", "OR.cov")
round(RROR.out, 4)

## Figure 8.1
# win.graph(height=5.5, width=7.0)
par(mfrow=c(2,2))
hist(RR.mf.samp,breaks=50,main="Relative Risk (Male/Female)")
hist(OR.mf.samp,breaks=50,main="Odds Ratio (Male/Female)")
hist(log(RR.mf.samp),breaks=50, main="log Relative Risk")
hist(log(OR.mf.samp),breaks=50, main="log Odds Ratio")
## End Figure 8.1
```

```
rm(list=ls(all=TRUE))

### Example 8.10

(lister <- matrix(c(6,34,16,19),byrow=T,ncol=2))
fisher.test(lister,alt="less")
fisher.test(lister,alt="two.sided")

rm(list=ls(all=TRUE))

### Example 8.11

(bet <- matrix(c(16,110,4,20),byrow=T,,ncol=2))

mcnemar.test(bet,correct=F)
z.stat <- (bet[1,2]-bet[2,1])/sqrt(bet[1,2]+bet[2,1])
z.p <- 2*(1-pnorm(abs(z.stat),0,1))
binom.p <- 2*(1-pbinom(max(bet[1,2],bet[2,1])-1,bet[1,2]+bet[2,1],0.5))

bet.out <- cbind(bet[1,2], bet[2,1], z.stat, z.stat^2, z.p, binom.p)
colnames(bet.out) <- c("n12=+R/-S", "n21=+S/-R", "z", "z^2", "P(z)", "P(exact)")
round(bet.out, 4)

rm(list=ls(all=TRUE))

### Example 8.12

(smoke <- array(c(647,204,39990,20132, 857,394,32894,21671,
   855,488,20739,19790, 643,766,11197,16499),dim=c(2,2,4)))

mantelhaen.test(smoke,exact=F,correct=F,alternative="two.sided")

rm(list=ls(all=TRUE))

### Example 8.13

pla <- read.csv(
    "http://www.stat.ufl.edu/~winner/data/productliability_award.csv")
attach(pla); names(pla)

(jury_award <- table(jury,award))
X2_ja <- chisq.test(jury_award, correct=F)
X2_ja
X2_ja$stdres

rm(list=ls(all=TRUE))

### Example 8.14

rd1 <- read.csv("http://www.stat.ufl.edu/~winner/data/remorse_death.csv")
attach(rd1); names(rd1)
install.packages("vcdExtra")
library(vcdExtra)
(rd.table <- table(remorse,jurVote))
GKgamma(rd.table)
cor.test(remorse,jurVote, method="kendall")

rm(list=ls(all=TRUE))

### Example 8.15

euro13 <- read.csv("http://www.stat.ufl.edu/~winner/data/europesoccer2013.csv")
attach(euro13); names(euro13)
```

```
home.result <- ifelse(DiffGoal<0,0,ifelse(DiffGoal==0,1,2)) ## Assign 0 for loss, 1 for Tie, 2 for Win
League <- factor(League)
kruskal.test(home.result ~ League)

rm(list=ls(all=TRUE))

### Example 8.16

(siskel_ebert <- matrix(c(24,8,13,8,13,11,10,9,64),byrow=T,ncol=3))

install.packages("psych")
library(psych)
cohen.kappa(siskel_ebert)

rm(list=ls(all=TRUE))
```

Chapter 9

Linear Regression

Linear regression is used when there is a numeric response variable and numeric (and possibly categorical) predictor (explanatory) variable(s). The mean of the response variable is to be related to the predictor(s) with random error terms typically assumed to be independent and normally distributed with constant variance. The fitting of linear regression models is very flexible, allowing for fitting curvature, categorical predictors, and interactions between factors.

9.1 Simple Linear Regression

When there is a single numeric predictor, the model is referred to as **Simple Regression**. The response variable is denoted as Y and the predictor variable is denoted as X. The model is written as follows.

$$Y = \beta_0 + \beta_1 X + \epsilon \qquad \epsilon \sim N(0, \sigma) \text{ independent}$$

Here β_0 is the intercept (mean of Y when X=0) and β_1 is the slope (the change in the mean of Y when X increases by 1 unit). Of primary concern is whether $\beta_1 = 0$, which implies the mean of Y is constant (β_0), and thus Y and X are not associated.

9.1.1 Estimation of Model Parameters

A sample of pairs (X_i, Y_i) $i = 1, \ldots, n$ is observed. The goal is to choose estimators of β_0 and β_1 that minimize the error sum of squares: $Q = \sum_{i=1}^{n} \epsilon_i^2$. The resulting **ordinary least squares** estimators are given below (the formulas are derived making use of calculus).

$$Y_i = \beta_0 + \beta_1 X_+ \epsilon_i \quad i = 1, \ldots, n \qquad \epsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$$

Once estimates have been computed, **fitted values** and **residuals** are obtained for each observation. The **error sum of squares (SSE)** is obtained as the sum of the squared residuals from the regression fit.

Fitted Values: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ Residuals: $e_i = Y_i - \hat{Y}_i$ $SSE = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n (Y_i - \overline{Y})^2 - \hat{\beta}_1^2 \sum_{i=1}^n (X_i - \overline{X})^2$

The (unbiased) estimator of the error variance σ^2 is $s^2 = MSE = \frac{SSE}{n-2}$, where MSE is the **Mean Square Error**. The subtraction of 2 can be thought of as the fact two parameters have been estimated: β_0 and β_1.

The estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are linear functions of $Y_1, \ldots, Y_n$ and thus using basic rules of mathematical statistics, their sampling distributions are as follow, assuming the error terms are normal, independent, with constant variance.

$$\hat{\beta}_1 \sim N\left(\beta_1, \sqrt{\frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}}\right) \qquad \hat{\beta}_0 \sim N\left(\beta_0, \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}\right)$$

The estimated standard errors are the standard error with the unknown σ^2 replaced by MSE.

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{MSE}{\sum_{i=1}^n (X_i - \overline{X})^2}} \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{MSE \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}$$

Example 9.1: Bollywood Films' Revenues and Budgets 2013-2017

Box office data for $n = 190$ Bollywood films, as well as their approximate budgets (production and advertising) were obtained from bollywoodmoviereviewz.com. These films are being treated as a random sample of all movies that could have been made under similar conditions. Plots of gross revenues versus budget are given in Figure 9.1. As is often seen with this type of data, logarithmic transformations on Y and/or X can be helpful in linearizing the relationship. All four possibilities are considered.

Based on the plots, the model with both variables transformed to the logarithmic scale is fit. This is due to the linear relation with approximately constant variance. When both variables have been transformed this way, the slope can be interpreted as percent change in Y when X is increased by 1%. Calculations for the linear regression are given below.

$$n = 190 \qquad \overline{X} = 3.5049 \qquad \overline{Y} = 3.1846$$

$$\sum_{i=1}^n (X_i - \overline{X})^2 = 131.043 \qquad \sum_{i=1}^n (Y_i - \overline{Y})^2 = 381.436 \qquad \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = 172.9174$$

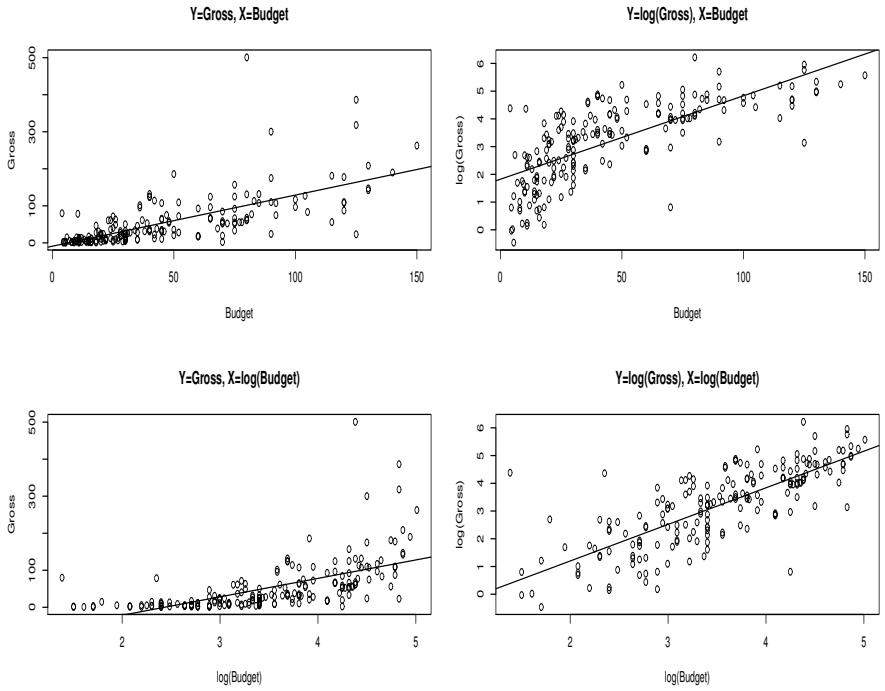

Figure 9.1: Bollywood Film Revenues and Budgets 2013-2017.

$$\hat{\beta}_1 = \frac{172.9174}{131.043} = 1.3195 \quad \hat{\beta}_0 = 3.1846 - 1.3195(3.5049) = -1.4401 \quad SSE = 381.436 - (1.3195)^2(131.043) = 153.2796$$

$$s^2 = MSE = \frac{SSE}{n-2} = \frac{153.2796}{190-2} = 0.8153$$

$$\hat{SE}\{\hat{\beta}_1\} = \sqrt{\frac{0.8153}{131.043}} = 0.0789 \qquad \hat{SE}\{\hat{\beta}_0\} = \sqrt{0.8153\left[\frac{1}{190} + \frac{3.5049^2}{131.043}\right]} = 0.2841$$

R Output

```
## Output

> round(ss.out, 4)
        SSYY    SSXX     SSXY      SSE    MSE beta1-hat  b0-hat SE{b1} SE{b0}
[1,] 381.436 131.043 172.9174 153.2636 0.8152    1.3195 -1.4402 0.0789 0.2841
```

∇

9.1.2 Inference Regarding β_1 and β_0

Primarily of interest are inferences regarding β_1. Note that if $\beta_1 = 0$, Y and X are not associated. We can test hypotheses and construct confidence intervals based on the estimate β_1 and its estimated standard error. The t-test is conducted as follows. Note that the null value β_{10} is almost always 0, and that software packages that report these tests always are treating β_{10} as 0.

$$H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10} \quad TS : t_{obs} = \frac{\hat{\beta}_1 - \beta_{10}}{\hat{SE}\{\hat{\beta}_1\}} \quad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \quad P = 2P\left(t_{n-2} \geq |t_{obs}|\right)$$

One-sided tests use the same test statistic, but the Rejection Region and P-value are changed to reflect the alternative hypothesis.

$$H_A^+ : \beta_1 > \beta_{10} \qquad RR : t_{obs} \geq t_{\alpha,n-2} \qquad P = P\left(t_{n-2} \geq t_{obs}\right)$$

$$H_A^- : \beta_1 < \beta_{10} \qquad RR : t_{obs} \leq -t_{\alpha,n-2} \qquad P = P\left(t_{n-2} \leq t_{obs}\right)$$

A $(1-\alpha)100\%$ confidence interval for β_1 is obtained as:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-2}\hat{SE}\{\hat{\beta}_1\}$$

Note that the confidence interval represents the values of β_{10} for which the two-sided test: $H_0 : \beta_1 = \beta_{10} \quad H_A : \beta_1 \neq \beta_{10}$ fails to reject the null hypothesis.

Inferences regarding β_0 are of less interest in practice, but can be conducted in analogous manner, using the estimate $\hat{\beta}_0$ and its estimated standard error $\hat{SE}\{\hat{\beta}_0\}$.

Example 9.2: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, a test of $H_0 : \beta_1 = 0$ and a 95% Confidence Interval for β_1 are obtained.

$$H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0 \quad TS : t_{obs} = \frac{1.3195}{0.0789} = 16.72 \quad RR : |t_{obs}| \geq 1.973 \quad P \approx 0$$

$$95\% \text{ Confidence Interval for } \beta_1 : 1.3195 \pm 1.973(0.0789) \equiv 1.3195 \pm 0.1557 \equiv (1.1638, 1.4752)$$

There is strong evidence of an association between log(Revenue) and log(Budget). Similarly, inference regarding the intercept β_0 can be made as well (although is of less interest as no movies had log(Budget)=0).

$$H_0 : \beta_0 = 0 \quad H_A : \beta_0 \neq 0 \quad TS : t_{obs} = \frac{-1.4402}{0.2841} = -5.069 \quad RR : |t_{obs}| \geq 1.973 \quad P \approx 0$$

95% Confidence Interval for $\beta_0 : -1.4402 \pm 1.973(0.2841) \equiv -1.4402 \pm 0.5605 \equiv (-2.0007, -0.8797)$

R Commands and Output

```
## Commands
## Analysis using lm (linear model) function in R

bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
confint(bolly.mod1)

## Output

> round(b.out, 4)
            Estimate Std. Error       t P-Value Lower Bound Upper Bound
Intercept    -1.4402     0.2841 -5.0695       0     -2.0007     -0.8798
log(Budget)   1.3195     0.0789 16.7298       0      1.1640      1.4751

> summary(bolly.mod1)
Call:
lm(formula = Y ~ X)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.44023    0.28410  -5.069 9.51e-07 ***
X            1.31955    0.07887  16.730  < 2e-16 ***

Residual standard error: 0.9029 on 188 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5961
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16

> confint(bolly.mod1)
                2.5 %    97.5 %
(Intercept) -2.000665 -0.879805
X            1.163955  1.475138
```

∇

9.1.3 Estimating a Mean and Predicting a New Observation @ $X = X^*$

There may be interest in estimating the mean response at a specific level X^*. The parameter of interest is $\mu^* = \beta_0 + \beta_1 X^*$. The point estimator, standard error, and $(1-\alpha)100\%$ Confidence Interval are given below.

$$\hat{Y}^* = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{MSE\left[\frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\%\text{ CI }: \hat{Y}^* \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*\right\}$$

To obtain a simultaneous $(1-\alpha)100\%$ Confidence Interval for the entire regression line (not just a single point), the Working-Hotelling method can be used.

$$\hat{Y}^* \pm \sqrt{2F_{\alpha,2,n-2}}\hat{SE}\left\{\hat{Y}^*\right\}$$

If the goal is to predict a new observation when $X = X^*$, uncertainty with respect to estimating the mean (as seen by the Confidence Interval above), and the random error for the new case (with standard deviation σ) must be taken into account. The point prediction is the same as for the mean. The prediction, standard error of prediction, and $(1-\alpha)100\%$ Prediction Interval are given below.

$$\hat{Y}^*_{\text{New}} = \hat{\beta}_0 + \hat{\beta}_1 X^* \qquad \hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{MSE\left[1 + \frac{1}{n} + \frac{\left(X^* - \overline{X}\right)^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right]}$$

$$(1-\alpha)100\%\text{ PI }: \hat{Y}^*_{\text{New}} \pm t_{\alpha/2,n-2}\hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\}$$

Note that the Prediction Interval will tend to be much wider than the Confidence Interval for the mean.

Example 9.3: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, a 95% Confidence Interval for the mean log(Revenue) of all possible films with a Budget of 60 ($X^* = \log(60) = 4.0943$) is obtained. Also a Prediction Interval for a single new movie with a budget of 60 is computed. The predicted value is $\hat{Y}^* = -1.4401 + 1.3195(4.0943) = 3.9623$. A plot of the data, fitted equation, 95% Confidence and Prediction Intervals is given in Figure 9.2.

$$\hat{SE}\left\{\hat{Y}^*\right\} = \sqrt{0.8153\left[\frac{1}{190} + \frac{(4.0943 - 3.5049)^2}{131.043}\right]} = \sqrt{0.8153(0.0079)} = 0.0803$$

$$\hat{SE}\left\{\hat{Y}^*_{\text{New}}\right\} = \sqrt{0.8153(1.0079)} = 0.9065$$

$$95\%\text{ CI for Mean: } 3.9623 \pm 1.973(0.0803) \equiv 3.9623 \pm 0.1585 \equiv (3.8038, 4.1208)$$

$$95\%\text{ PI for Individual: } 3.9623 \pm 1.973(0.9065) \equiv 3.9623 \pm 1.7885 \equiv (2.1738, 5.7508)$$

To convert back to the original units, the bounds of the Confidence and Prediction Intervals are exponentiated. The predicted revenue is $e^{3.9623} = 52.58$ and the 95% Confidence Interval and Prediction Interval are given below.

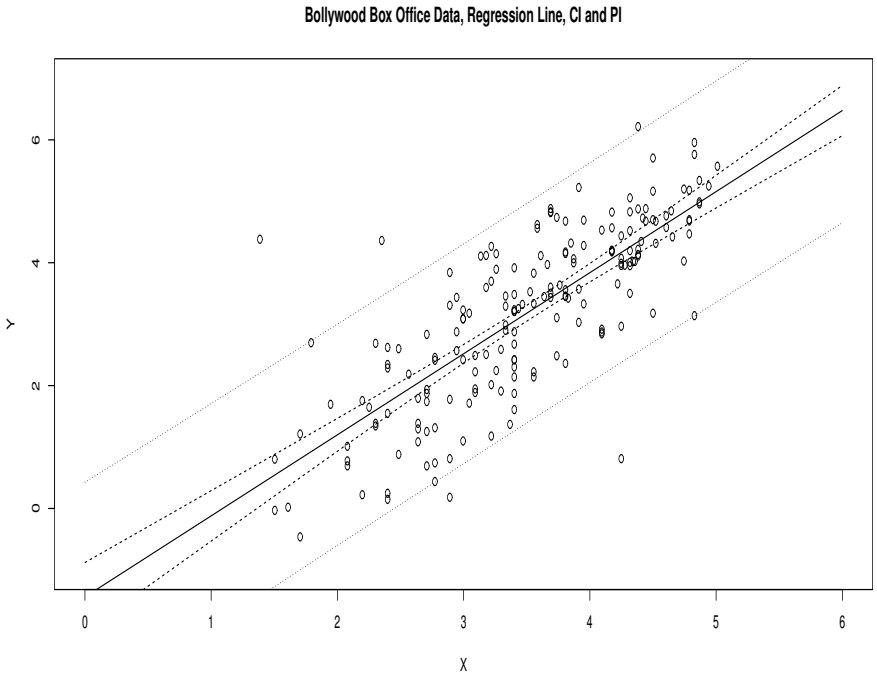

Figure 9.2: Bollywood Data, Fitted Equation 95% Confidence Interval for the mean and Prediction Interval for individual films

95% CI for Mean: $\left(e^{3.8038} = 44.87, e^{4.1208} = 61.61\right)$ 95% PI for Individual: $\left(e^{2.1738} = 8.79, e^{5.7508} = 314.44\right)$

R Commands and Output

```
## Commands
## Using predict function based on bolly.mod1 object with X*=log(60)
# CI for mean
ci.log60 <- predict(bolly.mod1, list(X=log(60)), interval="c")
# PI for individual movie
pi.log60 <- predict(bolly.mod1, list(X=log(60)), interval="p")

cipi.out1 <- rbind(ci.log60, pi.log60, exp(ci.log60), exp(pi.log60))
colnames(cipi.out1) <- c("Estimate", "Lower Bound", "Upper Bound")
rownames(cipi.out1) <- c("CI(log scale)", "PI(log scale)",
                         "CI(original scale)", "PI(original scale)")
round(cipi.out1, 4)

## Output

> round(cipi.out,4)
                   X*  Y-hat* CI Lower CI Upper PI Lower PI Upper
Log Scale      4.0943  3.9624   3.8040   4.1209   2.1743   5.7506
Original Scale 60.0000 52.5856  44.8797  61.6147   8.7959 314.3788

> round(cipi.out1, 4)
                  Estimate Lower Bound Upper Bound
CI(log scale)       3.9624      3.8040      4.1209
PI(log scale)       3.9624      2.1743      5.7506
CI(original scale)  52.5856     44.8797     61.6147
PI(original scale)  52.5856      8.7959    314.3788
```

∇

9.1.4 Analysis of Variance

When there is no association between Y and X ($\beta_1 = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total variation can be denoted as $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **Total Sum of Squares**.

When there is an association between Y and X ($\beta_1 \neq 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ (in terms of minimizing sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on X (as opposed to having ignored X). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**.

$$TSS = SSE + SSR \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

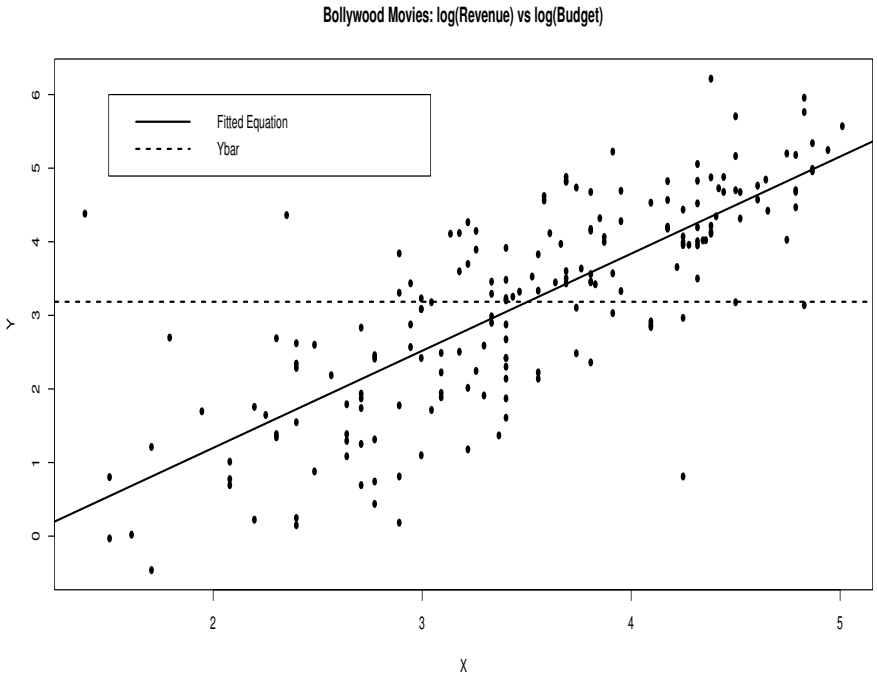

Figure 9.3: Plot of Data (points), Fitted Equation and Mean of Y - Bollywood movie regression with Y=log(Revenue) and X=log(Budget)

A plot including the data (Y), the horizontal line at the mean response ($\overline{Y}$) and the fitted equation is given in Figure 9.3. The sum of the squared vertical distances from the data Y_i to $\overline{Y}$ is the Total Sum of Squares TSS. The sum of the squared vertical distances from Y_i to their fitted values $\hat{Y}_i$ is the Error Sum of Squares SSE. The sum of the squared vertical distances from $\hat{Y}_i$ to $\overline{Y}$ is the Regression Sum of Squares SSR.

Each sum of squares has a **degrees of freedom** associated with it. The **Total Degrees of Freedom** is $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - 2$ (for simple regression). The **Regression Degrees of Freedom** is $df_{\text{Regression}} = 1$ (for simple regression).

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - 2 + 1$$

The Error and Regression sums of squares have **Mean Squares**, which are the sum of squares divided by their corresponding degrees of freedom: $MSE = SSE/(n-2)$ and $MSR = SSR/1$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} = \sigma^2 + \beta_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2$$

Source	*df*	*SS*	*MS*	F_{obs}	*P*-value
Regression (Model)	1	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{1}$	$F_{obs} = \frac{MSR}{MSE}$	$P\left(F_{1,n-2} \geq F_{obs}\right)$
Error (Residual)	$n-2$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-2}$		
Total (Corrected)	$n-1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 9.1: Analysis of Variance Table for Simple Linear Regression

Source	*df*	*SS*	*MS*	F_{obs}	*P*-value
Regression (Model)	1	228.1725	228.1725	279.8667	≈ 0
Error (Residual)	188	153.2676	0.8152		
Total (Corrected)	189	381.4360			

Table 9.2: Analysis of Variance Table for Bollywood Box Office Data

Note that when $\beta_1 = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A second way of testing whether $\beta_1 = 0$ is by the following F-test.

$$H_0 : \beta_1 = 0 \qquad H_A : \beta_1 \neq 0 \quad TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,1,n-2} \qquad P = P\left(F_{1,n-2} \geq F_{obs}\right)$$

The Analysis of Variance is typically set up in a table as in Table 9.1.

A measure often reported from a regression analysis is the **Coefficient of Determination** or r^2. This represents the variation in Y "explained" by X, divided by the total variation in Y.

$$r^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq r^2 \leq 1$$

The interpretation of r^2 is the proportion of variation in Y that is "explained" by X, and is often reported as a percentage ($100r^2$).

Example 9.4: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, the Analysis of Variance and F-test are given Table 9.2. Note that the Total Sum of Squares and Error Sum of Squares were computed in Example 9.1. The Regression Sum of Squares is the difference $SSR = TSS - SSE = 381.4360 - 153.2636 = 228.1725$.

The coefficient of determination, r^2, is 228.1725/381.4360=0.5982. Approximately 60% of the variation in log Revenue is "explained" by log Budget.

R Commands and Output

```
## Commands
```

```
bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
anova(bolly.mod1)

## Output

> round(aov.out,4)
         TSS      SSE      SSR    MSE    F_obs F(.05) P-value    R^2
[1,] 381.436 153.2636 228.1725 0.8152 279.8867 3.8914       0 0.5982

> summary(bolly.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.44023    0.28410  -5.069 9.51e-07 ***
X            1.31955    0.07887  16.730  < 2e-16 ***

Residual standard error: 0.9029 on 188 degrees of freedom
Multiple R-squared:  0.5982,    Adjusted R-squared:  0.5961
F-statistic: 279.9 on 1 and 188 DF,  p-value: < 2.2e-16

> anova(bolly.mod1)
Analysis of Variance Table
Response: Y
           Df Sum Sq Mean Sq F value    Pr(>F)
X           1 228.17 228.172  279.89 < 2.2e-16 ***
Residuals 188 153.26   0.815
```

∇

9.1.5 Correlation

The regression coefficient β_1 depends on the units of Y and X. It also depends on which variable is the dependent variable and which is the independent variable. A second widely reported measure is the **Pearson Product Moment Coefficient of Correlation**. It is invariant to linear transformations of Y and X, and does not distinguish which is the dependent and which is the independent variable. This makes it a widely reported measure when researchers are interested in how two random variables vary together in a population. The population correlation coefficient is labeled ρ, and the sample correlation is labeled r, and its formula is given below.

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2}} = \left(\frac{s_X}{s_Y}\right)\hat{\beta}_1$$

where s_X and s_Y are the standard deviations of X and Y, respectively. While $\hat{\beta}_1$ can take on any value, r lies between -1 and $+1$, taking on the extreme values if all of the points fall on a straight line. The test of whether $\rho = 0$ is mathematically equivalent to the t-test for testing whether $\beta_1 = 0$. The 2-sided test is given below.

$$H_0 : \rho = 0 \qquad H_A : \rho \neq 0 \quad TS : t_{obs} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \qquad RR : |t_{obs}| \geq t_{\alpha/2,n-2} \qquad P = 2P(t_{n-2} \geq |t_{obs}|)$$

To construct a large-sample confidence interval, **Fisher's z transform** is used to make the transformed

r to have a sampling distribution that is approximately normal. A confidence interval is obtained on the transformed correlation, then "back transformed" to the end points in terms of ρ.

$$z' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \qquad (1-\alpha)100\% \text{ CI for } \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right): \quad z' \pm z_{\alpha/2}\sqrt{\frac{1}{n-3}}$$

Labeling the endpoints of the Confidence Interval as (a, b), the Confidence Interval for ρ is computed as follows.

$$(1-\alpha)100\% \text{ Confidence Interval for } \rho: \left(\frac{e^{2a}-1}{e^{2a}+1}, \frac{e^{2b}-1}{e^{2b}+1}\right)$$

Example 9.5: Bollywood Films' Revenues and Budgets 2013-2017

Continuing with the Bollywood data with both Revenues and Budget on logarithmic scales, the sample correlation, a test of whether $\rho = 0$, and a 95% Confidence Interval for ρ are computed below.

$$r = \frac{172.9174}{\sqrt{131.0430(381.4360)}} = 0.7734 \qquad t_{obs} = \frac{0.7734}{\sqrt{\frac{1-0.7734^2}{190-2}}} = 16.73$$

$$z' = \frac{1}{2}\ln\left(\frac{1+0.7734}{1-0.7734}\right) = 1.0287 \qquad 1.0287 \pm 1.96\sqrt{\frac{1}{190-3}} \equiv 1.0287 \pm 0.1433 \equiv (0.8854, 1.1720)$$

$$\Rightarrow \quad (1-\alpha)100\% \text{ CI for } \rho: \left(\frac{e^{2(0.8854)}-1}{e^{2(0.8854)}+1}, \frac{e^{2(1.1720)}-1}{e^{2(1.1720)}+1}\right) \equiv \left(\frac{4.8756}{6.8756}, \frac{9.4228}{11.4228}\right) \equiv (.7091, .8249)$$

R Commands and Output

```
## Commands
cor.test(X,Y)
## Output
> cor.test(X,Y)
        Pearson's product-moment correlation
data:  X and Y
t = 16.73, df = 188, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.7091543 0.8249551
sample estimates:
      cor
0.7734296
```

9.1.6 Checking Linearity

A plot of the residuals versus X should be a random cloud of points centered at 0 (they sum to 0). A "U-shaped" or "inverted U-shaped" pattern is inconsistent with linearity.

A test for linearity can be conducted when there are repeat observations at certain X-levels (methods have also been developed to "group" X values). Suppose there are c distinct X-levels, with n_j observations at the j^{th} level. The data need to be re-labeled as Y_{ij} where j represents the X group, and i represents the individual case within the group ($i = 1, \ldots, n_j$). The following quantities are computed.

$$\overline{Y}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j} \qquad \hat{Y}_j = \hat{\beta}_0 + \hat{\beta}_1 X_j$$

Then decompose the Error Sum of Squares into **Pure Error** and **Lack of Fit**.

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 + \sum_{j=1}^{c} n_j\left(\overline{Y}_j - \hat{Y}_j\right)^2 \qquad SSE = SSPE + SSLF$$

Partition the error degrees of freedom $(n-2)$ into Pure Error $(n-c)$ and Lack of Fit $(c-2)$. This leads to an F-test for testing H_0: Relation is Linear versus H_A: Relation is not Linear.

$$TS: F_{obs} = \frac{[SSLF/(c-2)]}{[SSPE/(n-c)]} = \frac{MSLF}{MSPE} \qquad RR: F_{obs} \geq F_{\alpha,c-2,n-c} \qquad P = P\left(F_{c-2,n-c} \geq F_{obs}\right)$$

Note that the Pure Error sum of squares is the Error sum of squares for the 1-Way ANOVA model (treating the distinct X levels as nominal categories).

$$SSPE = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(Y_{ij} - \overline{Y}_j\right)^2 = \sum_{j=1}^{c}\left(n_j - 1\right)s_j^2$$

If the relationship is not linear, polynomial terms can be added to the model to allow for "bends" in the relationship between Y and X using multiple regression.

Example 9.6: Chewiness of Berries of Various Sugar Contents

A study of physical and mechanical properties of berries of $c = 6$ sugar contents (176.5, 192.6, 209.3, 225.0, 242.1, and 258.5) included the response chewiness (Zouid, et al, 2013, [54]). There were $n_j = 15$ replicates at each sugar content level. Figure 9.4 gives data that have been simulated to match the mean and standard deviation at each of the six sugar contents.

Computations for the Lack of Fit F-test are given in Table 9.3. The fitted simple linear regression equation and error sum of squares are given below.

$$\hat{Y}_j = 7.6629 - 0.0228X_j \qquad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 74.126 \qquad df_E = 90 - 2 = 88$$

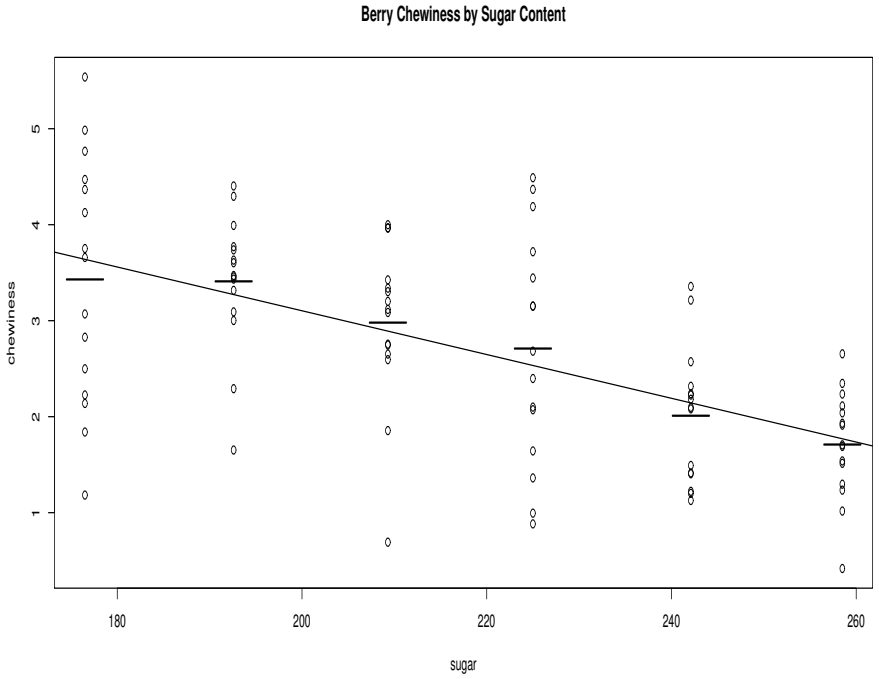

Figure 9.4: Berry Chewiness and Sugar Content with fitted regression line and group means

j	X_j	n_j	$\overline{Y}_j$	s_j	$\hat{Y}_j$	$\overline{Y}_j - \hat{Y}_j$
1	176.5	15	3.43	1.29	3.639	-0.209
2	192.6	15	3.41	0.71	3.272	0.138
3	209.3	15	2.98	0.86	2.891	0.089
4	225.0	15	2.71	1.20	2.534	0.176
5	248.1	15	2.01	0.70	2.144	-0.134
6	258.5	15	1.71	0.57	1.770	-0.060

Table 9.3: Lack-of-Fit summary statistics for Berry Chewiness experiment

The F-test for lack-of-fit is conducted as follows. There is no evidence that the true relationship between chewiness and sugar content is not linear.

$$H_0 : E\{Y_{ij}\} = \beta_0 + \beta_1 X_j \qquad H_A : E\{Y_{ij}\} = \mu_j \neq \beta_0 + \beta_1 X_j$$

$$SSLF = 15\left[(-0.209)^2 + \cdots + (-0.060)^2\right] = 1.848 \quad df_{LF} = 6 - 2 = 4 \quad MSLF = \frac{1.848}{4} = 0.4620$$

$$SSPE = (15-1)\left[(1.29)^2 + \cdots + (0.57)^2\right] = 72.278 \quad df_{PE} = 90 - 6 = 84 \quad MSPE = \frac{72.278}{84} = 0.8605$$

$$TS : F_{LF} = \frac{0.4620}{0.8605} = 0.5369 \quad RR : F_{LF} \geq F_{.05,4,84} = 2.480 \quad P = P\left(F_{4,85} \geq 0.5369\right) = .7090$$

R Commands and Output

```
### Commands

berry1 <- read.csv("http://www.stat.ufl.edu/~winner/data/berry_sugar_chewy.csv")
attach(berry1); names(berry1)

chewy1 <- lm(chewiness ~ sugar)      ### Fit Linear Regression
anova(chewy1)
chewy2 <- lm(chewiness ~ factor(sugar)) ### Fit 1-Way ANOVA
anova(chewy2)
anova(chewy1, chewy2)       ### Compare Linear Reg w/ 1-Way ANOVA

### Output

> round(LF.out,4)
     df(LF)  SSLF  MSLF df(PE)    SSPE   MSPE  F_LOF F(.05) P(F>=F_LOF)
[1,]      4 1.848 0.462     84 72.2784 0.8605 0.5369 2.4803       0.709

> anova(chewy1)
Analysis of Variance Table
Response: chewiness
          Df Sum Sq Mean Sq F value    Pr(>F)
sugar      1 36.721  36.721  43.594 2.951e-09 ***
Residuals 88 74.126   0.842

> anova(chewy2)
Analysis of Variance Table
Response: chewiness
              Df Sum Sq Mean Sq F value    Pr(>F)
factor(sugar)  5 38.569  7.7138  8.9647 7.423e-07 ***
Residuals     84 72.278  0.8605

> anova(chewy1, chewy2)       ### Compare Linear Reg w/ 1-Way ANOVA
Analysis of Variance Table
Model 1: chewiness ~ sugar
Model 2: chewiness ~ factor(sugar)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     88 74.126
2     84 72.278  4     1.848 0.5369  0.709
```

∇

9.2 Multiple Linear Regression

When there is more than one predictor variable, the model generalizes to multiple linear regression. The calculations become more complex, but conceptually, the ideas remain the same. We will use the notation of p as the number of predictors, and $p' = p+1$ as the number of regression coefficients in the model (including the intercept). The model can be written as follows with the same assumptions about the errors as in simple regression.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \qquad \epsilon \sim N(0, \sigma^2) \text{ independent}$$

Least squares (and maximum likelihood) estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ minimize the error sum of squares. The fitted values, residuals, and error sum of squares are given below.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots \hat{\beta}_p X_{ip} \qquad e_i = Y_i - \hat{Y}_i \qquad SSE = \sum_{i=1}^{n} e_i^2$$

The degrees of freedom for error are now $n - p' = n - (p + 1)$, as the model estimates $p' = p + 1$ parameters. The degrees of freedom for regression is p.

In the multiple linear regression model, β_j represents the change in $E\{Y\}$ when X_j increases by 1 unit, with all other predictor variables being held constant. It is referred to as the **partial regression coefficient**.

9.2.1 Testing and Estimation for Partial Regression Coefficients

Once the model is fit, for each predictor variable, the estimated regression coefficient, its estimated standard error, t-statistic and confidence interval are obtained. Technically, the estimated variance-covariance matrix for the vector of regression coefficients is computed, with the standard errors being the square root of the variances of the individual coefficients.

To test whether Y is associated with X_j, after controlling for the remaining $p-1$ predictors, the test is whether $\beta_j = 0$. This is equivalent to the t-test from simple regression (in general, the test can be whether a regression coefficient is any specific number, although software packages are testing whether it is 0).

$$H_0 : \beta_j = \beta_{j0} \qquad H_A : \beta_j \neq \beta_{j0} \quad TS : t_{obs} = \frac{\hat{\beta}_j - \beta_{j0}}{\hat{SE}\{\hat{\beta}_j\}} \qquad RR : |t_{obs}| \geq t_{\alpha/2, n-p'} \qquad P = 2P(t_{n-p'} \geq |t_{obs}|)$$

One-sided tests make the same adjustments as in simple linear regression.

$$H_A^+ : \beta_j > \beta_{j0} \qquad RR : t_{obs} \geq t_{\alpha, n-p'} \qquad P = P(t_{n-p'} \geq t_{obs})$$

$$H_A^- : \beta_j < \beta_{j0} \qquad RR : t_{obs} \leq -t_{\alpha,n-p'} \qquad P = P(t_{n-p'} \leq t_{obs})$$

A $(1-\alpha)100\%$ confidence interval for β_j is obtained as:

$$\hat{\beta}_j \pm t_{\alpha/2,n-p'} \hat{SE}\{\hat{\beta}_j\}$$

Note that the confidence interval represents the values of β_{j0} for which the two-sided test: $H_0 : \beta_j = \beta_{j0} \quad H_A : \beta_j \neq \beta_{j0}$ fails to reject the null hypothesis.

Example 9.7: How Stature (Height) Relates to Hand and Foot Length among Females

A regression model was fit, relating stature (Y, height, in mm) to hand length (X_1, mm) and foot length (X_2, mm) for a sample of $n = 75$ female adult Turks (Sanli, Kizilkanat, Boyan, et al., 2005, [45]). The data have been simulated to match means, standard deviations, and bivariate correlations. A matrix plot of the variables is given in Figure 9.5. The model, fitted equation, Error sum of squares and mean square are given below ($n = 75, p' = 2 + 1 = 3$).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \quad \hat{Y}_i = 743.970 + 2.375X_1 + 1.727X_2 \quad SSE = 68924.42 \quad MSE = 957.284$$

The estimated standard errors are 0.486 for $\hat{\beta}_1$ and 0.375 for $\hat{\beta}_2$, respectively. The t-tests and 95% Confidence Intervals for β_1 and β_2 are given below.

Hand: $H_0 : \beta_1 = 0 \quad H_A : \beta_1 \neq 0 \quad TS : t_{obs} = \dfrac{2.375}{0.486} = 4.89 \quad RR : |t_{obs}| \geq t_{.025,72} = 1.993 \quad P = P\left(t_{72} \geq 5.63\right) \approx 0$

Foot: $H_0 : \beta_2 = 0 \quad H_A : \beta_2 \neq 0 \quad TS : t_{obs} = \dfrac{1.727}{0.375} = 4.61 \quad RR : |t_{obs}| \geq t_{.025,72} = 1.993 \quad P = P\left(t_{72} \geq 4.61\right) \approx 0$

95% CI for β_1: $2.375 \pm 1.993(0.486) \equiv 2.375 \pm 0.969 \equiv (1.406, 3.344)$

95% CI for β_2: $1.727 \pm 1.993(0.375) \equiv 1.727 \pm 0.747 \equiv (0.980, 2.474)$

R Commands and Output

```
### Commands
shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
    header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)

f.height <- height[gender == 2]     ### Female Heights
f.hand <- hand[gender == 2]         ### Female Hand Lengths
f.foot <- foot[gender == 2]         ### Female Foot Lengths

f.stature <- data.frame(f.height, f.hand, f.foot)
plot(f.stature)

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
summary(shf.mod1)
confint(shf.mod1)

#### Output
```

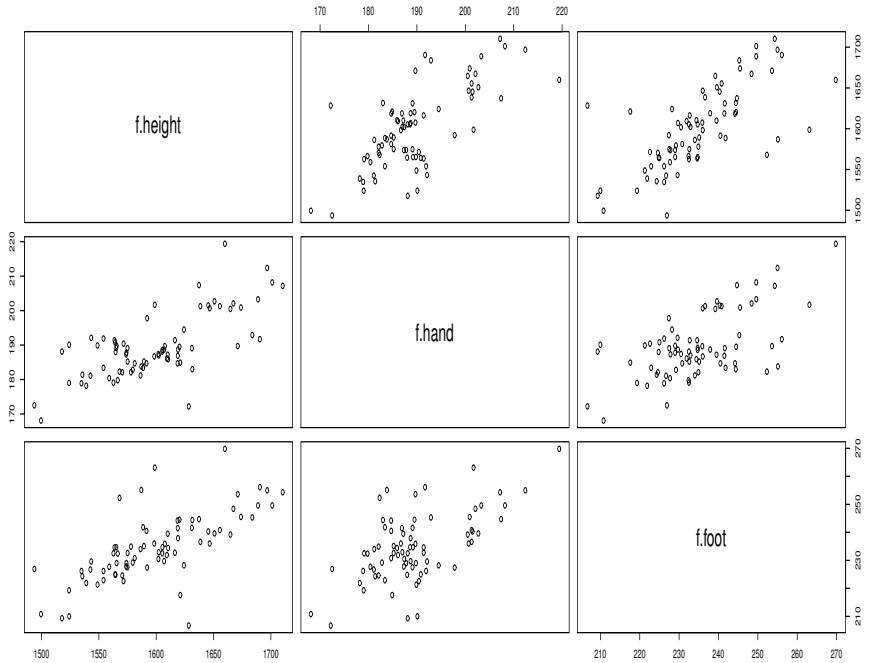

Figure 9.5: Heights, Hand Lengths and Foot Lengths among a Sample of 75 Adult Female Turks

```
> summary(shf.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 743.9696    79.7499   9.329 5.12e-14 ***
f.hand        2.3748     0.4858   4.888 5.99e-06 ***
f.foot        1.7271     0.3745   4.611 1.69e-05 ***

Residual standard error: 30.94 on 72 degrees of freedom
Multiple R-squared:  0.6159,    Adjusted R-squared:  0.6053
F-statistic: 57.73 on 2 and 72 DF,  p-value: 1.093e-15

> confint(shf.mod1)
                 2.5 %     97.5 %
(Intercept) 584.9911070 902.948034
f.hand        1.4062645   3.343310
f.foot        0.9804939   2.473711
```

∇

9.2.2 Analysis of Variance

When there is no association between Y and $X_1, \ldots, X_p$ ($\beta_1 = \cdots = \beta_p = 0$), the best predictor of each observation is $\overline{Y} = \hat{\beta}_0$ (in terms of minimizing sum of squares of prediction errors). In this case, the total

variation can be denoted as $TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, the **Total Sum of Squares**, just as with simple regression.

When there is an association between Y and at least one of $X_1, \ldots, X_p$ (not all $\beta_i = 0$), the best predictor of each observation is $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$ (in terms of minimizing the sum of squares of prediction errors). In this case, the error variation can be denoted as $SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$, the **Error Sum of Squares**.

The difference between TSS and SSE is the variation "explained" by the regression of Y on $X_1, \ldots, X_p$ (as opposed to having ignored $X_1, \ldots, X_p$). It represents the difference between the fitted values and the mean: $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$ the **Regression Sum of Squares**. Note that when there is $p > 1$ predictor, the fitted equation is no longer a straight line in 2-dimensions. This makes visualization more difficult, but the concept of distance from observed to predicted value is the same. For the stature example, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$ represents a 2-dimensional plane in 3-dimensional space.

$$TSS = SSE + SSR \qquad \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

The **Total Degrees of Freedom** remains $df_{\text{Total}} = n - 1$. The **Error Degrees of Freedom** is $df_{\text{Error}} = n - p'$. The **Regression Degrees of Freedom** is $df_{\text{Regression}} = p$. Note that when there is $p = 1$ predictor, this generalizes to simple regression.

$$df_{\text{Total}} = df_{\text{Error}} + df_{\text{Regression}} \qquad n - 1 = n - p' + p$$

The Mean Squares for Error and Regression are: $MSE = SSE/(n - p')$ and $MSR = SSR/p$. It can be shown that these mean squares have the following **Expected Values**, average values in repeated sampling at the same observed X levels.

$$E\{MSE\} = \sigma^2 \qquad E\{MSR\} \geq \sigma^2$$

Note that when $\beta_1 = \cdots \beta_p = 0$, then $E\{MSR\} = E\{MSE\}$, otherwise $E\{MSR\} > E\{MSE\}$. A way of testing whether $\beta_1 = \cdots \beta_p = 0$ is by the F-test.

$$H_0 : \beta_1 = \cdots \beta_p = 0 \qquad H_A : \text{ Not all } \beta_j = 0$$

$$TS : F_{obs} = \frac{MSR}{MSE} \qquad RR : F_{obs} \geq F_{\alpha,p,n-p'} \qquad P = P\left(F_{p,n-p'} \geq F_{obs}\right)$$

The Analysis of Variance is typically set up in a table as in Table 9.4.

The **Coefficient of Determination** is labeled R^2 for the multiple regression model. This represents the variation in Y "explained" by $X_1, \ldots, X_p$, divided by the total variation in Y. Note that the **summary** function in R reports "Multiple R-squared" even when there is only a single predictor.

Source	df	SS	MS	F_{obs}	$P(> F)$
Regression (Model)	p	$SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$	$MSR = \frac{SSR}{p}$	$F_{obs} = \frac{MSR}{MSE}$	$P(F_{p,n-p'} \geq F_{obs})$
Error (Residual)	$n - p'$	$SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$	$MSE = \frac{SSE}{n-p'}$		
Total (Corrected)	$n - 1$	$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$			

Table 9.4: Analysis of Variance Table for Multiple Linear Regression

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} = \frac{SSR}{TSS} = 1 - \frac{SSE}{TSS} \qquad 0 \leq R^2 \leq 1$$

Example 9.8: Stature (Height) as Function of Hand and Foot Length among Females

In a continuation of the Turkish adult females' model relating stature to hand and foot lengths, the following sums of squares and F-test are computed.

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 179409 \quad SSE = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 68924 \quad SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 110504$$

$$MSE = \frac{68924}{75-3} = 957.3 \quad MSR = \frac{110504}{2} = 55252$$

$$H_0 : \beta_1 = \beta_2 = 0 \quad TS : F_{obs} = \frac{55252}{957.3} = 57.72 \quad RR : F_{obs} \geq F_{.05,2,72} = 3.124 \quad P\left(F_{2,72} \geq 57.72\right) \approx 0$$

The Coefficient of Determination is $R^2 = 110504/179409 = .616$, approximately 62% of the variation in height is "explained" by hand and foot length.

R Commands and Output

```
### Commands

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
summary(shf.mod1)
anova(shf.mod1)
drop1(shf.mod1, test="F")

### Output

> summary(shf.mod1)
Residual standard error: 30.94 on 72 degrees of freedom
Multiple R-squared:  0.6159,    Adjusted R-squared:  0.6053
F-statistic: 57.73 on 2 and 72 DF,  p-value: 1.093e-15

> anova(shf.mod1)
Analysis of Variance Table

Response: f.height
          Df Sum Sq Mean Sq F value    Pr(>F)
```

```
f.hand     1  90153   90153  94.203 1.027e-14 ***
f.foot     1  20351   20351  21.265 1.694e-05 ***
Residuals 72  68905     957
```

Note that $SSR = SSR(X_1) + SSR(X_2|X_1) = 90153 + 20351 = 110504$. The sums of squares for the **anova** function are the **Sequential Sums of Squares** and sum up to the Regression Sum of Squares.

∇

9.2.3 Testing a Subset of $\beta^s = 0$

The F-test from the Analysis of Variance and the t-tests represent extremes of model testing (all variables simultaneously versus one-at-a-time). Often interest in testing whether a group of predictors do not improve prediction, after controlling for the remaining predictors.

Suppose that after controlling for g predictors, we wish to test whether the remaining $p - g$ predictors are associated with Y. That is, we wish to test the following hypotheses.

$$H_0 : \beta_{g+1} = \cdots \beta_p = 0 \qquad H_A : \text{ Not all of } \beta_{g+1}, \ldots, \beta_p = 0$$

Note that, the t-tests control for all other predictors, while here, we want to control for only $X_1, \ldots, X_g$. To do this, fit two models: the **Complete** or **Full Model** with all p predictors, and the **Reduced Model** with only the g "control" variables. For each model, obtain the Regression and Error sums of squares, as well as R^2. Let (F) represent the Full model and (R) represent the Reduced model. This leads to the following test statistic and rejection region.

$$TS : F_{obs} = \frac{\left[\frac{SSE(R)-SSE(F)}{(n-g')-(n-p')}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{SSR(F)-SSR(R)}{p-g}\right]}{\left[\frac{SSE(F)}{n-p'}\right]} = \frac{\left[\frac{R_F^2-R_R^2}{p-g}\right]}{\left[\frac{1-R_F^2}{n-p'}\right]}$$

$$RR : F_{obs} \geq F_{\alpha,p-g,n-p'} \qquad P = P\left(F_{p-g,n-p'} \geq F_{obs}\right)$$

Example 9.9: Energy Consumption of Luxury Hotels

A study considered factors relating to Energy Consumption (Y, in millions of kilowatt-hours) for a sample of $n = 19$ luxury hotels in Hainan Province, China (Xin, Lu, Xu, and Wu, 2012, [53]). The model had 3 predictors: Area (X_1, in 1000s of square meters), Age (X_2, in years), and Effective number of guest rooms (X_3, # rooms times occupancy rate).

Consider two models: Model 1 with X_1, X_2, X_3 as predictors and Model 2 with only X_1 as a predictor. The goal is to determine whether age and/or effective guest rooms is associated with energy consumption, after controlling for the hotel's size (Area). The data, fitted values and residuals for Models 1 and 2 are given in Table 9.5. The fitted equations and Error Sums of Squares are given below ($n = 19, p = 3, p' = 4, g = 1$).

Hotel	Y	X_1	X_2	X_3	$\hat{Y}_1$	e_1	$\hat{Y}_2$	e_2
1	1.95	43.00	6.00	44.64	5.61	-3.66	6.31	-4.36
2	1.05	19.98	16.00	85.33	3.31	-2.26	2.64	-1.60
3	4.25	46.53	7.00	115.52	6.48	-2.24	6.87	-2.63
4	2.13	20.96	6.00	110.34	2.48	-0.32	2.80	-0.67
5	2.79	24.21	5.00	230.27	3.82	-1.04	3.32	-0.53
6	13.83	112.20	4.00	188.73	17.11	-3.28	17.33	-3.50
7	5.56	45.00	3.00	78.03	5.70	-0.14	6.63	-1.07
8	4.00	28.55	6.00	54.37	3.27	0.73	4.01	-0.01
9	4.67	32.87	8.00	89.75	4.58	0.09	4.70	-0.03
10	8.92	59.41	5.00	167.23	8.82	0.10	8.92	0.00
11	6.87	45.00	10.00	368.20	7.83	-0.96	6.63	0.24
12	6.01	37.44	13.00	197.29	6.44	-0.43	5.42	0.59
13	8.19	50.83	4.00	83.31	6.74	1.45	7.56	0.63
14	11.74	68.00	13.00	187.53	11.02	0.72	10.29	1.45
15	14.84	78.87	8.00	206.12	12.25	2.58	12.02	2.82
16	5.37	28.45	13.00	128.30	4.42	0.95	3.99	1.37
17	13.52	70.00	4.00	228.74	10.56	2.95	10.61	2.91
18	3.88	20.00	5.00	85.81	2.04	1.85	2.65	1.24
19	10.57	50.00	12.00	120.28	7.67	2.90	7.42	3.15

Table 9.5: Hotel Energy Consumption Data, Fitted Values, and Residuals for Model 1 and Model 2

Model 1: Full: $\hat{Y}_F = -2.1320 + 0.1540X_1 + 0.0959X_2 + 0.0075X_3 \quad SSE(F) = 67.846 \quad df_E(F) = n - p' = 19 - 4 = 15$

Model 2: Reduced: $\hat{Y}_R = -0.5380 + 0.1593X_1 \quad SSE(R) = 75.129 \quad df_E(R) = n - g' = 19 - 2 = 17 \quad p - g = 3 - 1 = 2$

The test of $H_0 : \beta_2 = \beta_3 = 0$ versus $H_A : \beta_2$ and/or $\beta_3 = 0$ is given below.

$$TS : F_{obs} = \frac{\left[\frac{75.129 - 67.846}{17 - 15}\right]}{\left[\frac{67.846}{15}\right]} = \frac{3.642}{4.523} = 0.805 \quad RR : F_{obs} \geq F_{.05,2,15} = 3.682 \quad P(F_{2,15} \geq 0.805) = .4634$$

After controlling for Area, neither Age or Effective guest rooms are associated with Energy Consumption.

R Commands and Output

```
### Commands
hotel_ec <- read.csv("http://www.stat.ufl.edu/~winner/data/hotel_energy.csv")
attach(hotel_ec); names(hotel_ec)

enrgcons <- enrgcons/1000000
area <- area/1000

## Full Model
hec.mod1 <- lm (enrgcons ~ area + age + effrooms)
summary(hec.mod1)
```

```
anova(hec.mod1)

## Reduced Model
hec.mod2 <- lm (enrgcons ~ area)
summary(hec.mod2)
anova(hec.mod2)

## Full versus Reduced F-test
anova(hec.mod2, hec.mod1)

### Output

> summary(hec.mod1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.252767   1.781202  -1.265 0.225260
area         0.148709   0.029066   5.116 0.000127 ***
age          0.113045   0.134527   0.840 0.413924
effrooms     0.005777   0.007096   0.814 0.428315

Residual standard error: 2.127 on 15 degrees of freedom
Multiple R-squared:  0.7946,    Adjusted R-squared:  0.7535
F-statistic: 19.35 on 3 and 15 DF,  p-value: 2.049e-05

> anova(hec.mod1)
Response: enrgcons
          Df  Sum Sq Mean Sq F value    Pr(>F)
area       1 255.218 255.218 56.4258 1.854e-06 ***
age        1   4.286   4.286  0.9475    0.3458
effrooms   1   2.998   2.998  0.6628    0.4283
Residuals 15  67.846   4.523

> summary(hec.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.53804    1.08509  -0.496    0.626
area         0.15925    0.02096   7.599 7.29e-07 ***

Residual standard error: 2.102 on 17 degrees of freedom
Multiple R-squared:  0.7726,    Adjusted R-squared:  0.7592
F-statistic: 57.75 on 1 and 17 DF,  p-value: 7.294e-07

> anova(hec.mod2)
Response: enrgcons
          Df  Sum Sq Mean Sq F value    Pr(>F)
area       1 255.218 255.218   57.75 7.294e-07 ***
Residuals 17  75.129   4.419

> anova(hec.mod2, hec.mod1)
Analysis of Variance Table
Model 1: enrgcons ~ area
Model 2: enrgcons ~ area + age + effrooms
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     17 75.129
2     15 67.846  2    7.2834 0.8051 0.4654
```

9.2.4 Models With Categorical (Qualitative) Predictors

Often, one or more categorical variables are included in a model. If a categorical variable has m levels, there will need to be $m-1$ **dummy** or **indicator variables** to reflect the effects of the variable's levels. The variable will take on 1 if the i^{th} observation is in that level of the variable, 0 otherwise. Note that one level of the variable will have 0^s for all $m-1$ dummy variables, making it the reference category. The β^s for the other groups (levels of the qualitative variable) reflect the difference in the mean for that group with the reference group, controlling for all other predictors.

Note that if the qualitative variable has 2 levels, there will be a single dummy variable, and we can test for differences in the effects of the 2 levels with a t-test, controlling for all other predictors. If there are $m-1>2$ dummy variables, the F-test can be used to test whether all $m-1$ β^s are 0, controlling for all other predictors. An example is given below.

9.2.5 Models With Interaction Terms

When the effect of one predictor depends on the level of another predictor (and vice versa), the predictors are said to **interact**. The way to model interaction(s) is to create a new variable that is the product of the 2 predictors. Suppose the model has Y, and 2 numeric predictors: X_1 and X_2. Create a new predictor $X_3 = X_1X_2$. Now, consider the following model.

$$E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_1X_2 = \beta_0 + \beta_2X_2 + (\beta_1 + \beta_3X_2)\,X_1$$

The slope with respect to X_1 depends on the level of X_2, unless $\beta_3 = 0$, which can be tested with a t-test of $H_0 : \beta_3 = 0$. This logic extends to qualitative variables as well. Create cross-product terms between numeric (or other categorical) predictors with the $m-1$ dummy variables representing the qualitative predictor. Then the t-test ($m-1=1$) or F-test ($m-1\geq 2$) can be conducted to test for interactions among predictors. This is demonstrated by adding males to the stature data below.

Example 9.10: Heights, Hand and Foot Lengths in Males and Females

In the stature study (Sanli, Kizilkanat, Boyan, et al., 2005, [45]), there were also 80 males, for a total of $n = 75 + 80 = 155$ adults. For these models, Y is height, X_1 is hand length, and X_2 is foot length. Create the dummy (indicator) variable $X_3 = 1$ if male, $X_3 = 0$ if female. Then consider three models: Common slopes and intercept by gender (Model 1), Common slopes but different intercepts by gender (Model 2), and Different slopes and intercepts by gender (Model 3). The models are given below.

$$\text{Model 1: } E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2$$

$$\text{Model 2: } E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3$$

$$\text{Females: } E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 \quad \text{Males: } E\{Y\} = (\beta_0 + \beta_3) + \beta_1X_1 + \beta_2X_2$$

$$\text{Model 3: } E\{Y\} = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_1X_3 + \beta_5X_2X_3$$

$$\text{Males: } E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)\,X_1 + (\beta_2 + \beta_5)\,X_2$$

The fitted equations and their Error Sums of Squares are given below (the regression coefficients are taken from the R output given below).

$$\text{Model 1: } \hat{Y}_F = \hat{Y}_M = 372.64 + 3.32X_1 + 2.58X_2 \quad SSE_1 = 189029 \quad df_{E1} = 155 - 3 = 152$$

$$\text{Model 2: } \hat{Y}_F = 581.99{+}2.81X_1{+}2.06X_2 \quad \hat{Y}_M = 621.55{+}2.81X_1{+}2.06X_2 \quad SSE_2 = 165341 \quad df_{E2} = 155{-}4 = 151$$

$$\text{Model 3: } \hat{Y}_F = 743.97{+}2.38X_1{+}1.73X_2 \quad \hat{Y}_M = 439.27{+}3.29X_1{+}2.38X_2 \quad SSE_3 = 157360 \quad df_{E3} = 155{-}6 = 149$$

Tests comparing the different models include Model 2 versus Model 1, where the null hypothesis is common slopes and intercepts (Model 1) and the alternative is common slopes and different intercepts (Model 2). The null hypothesis is $H_0 : \beta_3 = 0$.

$$TS : F_{12} = \frac{\left[\frac{189029-165341}{152-151}\right]}{\left[\frac{165341}{151}\right]} = \frac{23688}{1095} = 21.63 \quad RR : F_{12} \geq F_{.05,1,151} = 3.904$$

A second test comparing the different models include Model 3 versus Model 2, where the null hypothesis is common slopes and different intercepts (Model 2) and the alternative is different slopes and intercepts (Model 3). The null hypothesis is $H_0 : \beta_4 = \beta_5 = 0$.

$$TS : F_{23} = \frac{\left[\frac{165341-157360}{151-149}\right]}{\left[\frac{157360}{149}\right]} = \frac{3990.5}{1056} = 3.78 \quad RR : F_{23} \geq F_{.05,2,149} = 3.057 \quad P = .0251$$

The "full model" allowing for different slopes and intercepts for males and females gives the best fit.

R Commands and Output

```
### Commands
shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)

male <- 2-gender ### male = 1 if male, 0 if female

## Model 1: Common slope/intercept
shf.mod1 <- lm(height ~ hand + foot)
summary(shf.mod1)
anova(shf.mod1)

## Model 2: Common slope/Different intercept
shf.mod2 <- lm(height ~ hand + foot + male)
summary(shf.mod2)
anova(shf.mod2)

## Model 3: Different slope/intercept
shf.mod3 <- lm(height ~ hand + foot + male + I(hand*male) + I(foot*male))
summary(shf.mod3)
anova(shf.mod3)
```

```
anova(shf.mod1,shf.mod2) ### Compare Models 1 and 2
anova(shf.mod2,shf.mod3) ### Compare Models 2 and 3



### Output
> ## Model 1: Common slope/intercept
> shf.mod1 <- lm(height ~ hand + foot)
> summary(shf.mod1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 372.6378    43.2581   8.614 8.41e-15 ***
hand          3.3175     0.3461   9.586  < 2e-16 ***
foot          2.5816     0.2490  10.370  < 2e-16 ***

Residual standard error: 35.26 on 152 degrees of freedom
Multiple R-squared:  0.8608,    Adjusted R-squared:  0.859
F-statistic: 470.1 on 2 and 152 DF,  p-value: < 2.2e-16

> anova(shf.mod1)
Analysis of Variance Table
Response: height
           Df  Sum Sq Mean Sq F value    Pr(>F)
hand        1 1035412 1035412  832.59 < 2.2e-16 ***
foot        1  133728  133728  107.53 < 2.2e-16 ***
Residuals 152  189029    1244

> ## Model 2: Common slope/Different intercept
> shf.mod2 <- lm(height ~ hand + foot + male)
> summary(shf.mod2)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 581.9858    60.6099   9.602  < 2e-16 ***
hand          2.8116     0.3425   8.210 9.11e-14 ***
foot          2.0643     0.2587   7.979 3.43e-13 ***
male         39.5640     8.5064   4.651 7.16e-06 ***

Residual standard error: 33.09 on 151 degrees of freedom
Multiple R-squared:  0.8783,    Adjusted R-squared:  0.8758
F-statistic: 363.1 on 3 and 151 DF,  p-value: < 2.2e-16

> anova(shf.mod2)
Analysis of Variance Table
Response: height
           Df  Sum Sq Mean Sq F value    Pr(>F)
hand        1 1035412 1035412 945.602 < 2.2e-16 ***
foot        1  133728  133728 122.128 < 2.2e-16 ***
male        1   23687   23687  21.633 7.157e-06 ***
Residuals 151  165341    1095

> ## Model 3: Different slope/intercept
> shf.mod3 <- lm(height ~ hand + foot + male + I(hand*male) + I(foot*male))
> summary(shf.mod3)
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    743.9696    83.7772   8.880 1.98e-15 ***
hand             2.3748     0.5104   4.653 7.17e-06 ***
foot             1.7271     0.3934   4.390 2.14e-05 ***
male          -304.7039   125.5987  -2.426   0.0165 *
I(hand * male)    0.9120     0.6809   1.340   0.1824
I(foot * male)    0.6537     0.5162   1.266   0.2074

Residual standard error: 32.5 on 149 degrees of freedom
```

```
Multiple R-squared:  0.8841,    Adjusted R-squared:  0.8803
F-statistic: 227.4 on 5 and 149 DF,  p-value: < 2.2e-16

> anova(shf.mod3)
Analysis of Variance Table
Response: height
               Df  Sum Sq Mean Sq  F value    Pr(>F)
hand            1 1035412 1035412 980.4040 < 2.2e-16 ***
foot            1  133728  133728 126.6232 < 2.2e-16 ***
male            1   23687   23687  22.4289 5.035e-06 ***
I(hand * male)  1    6288    6288   5.9538   0.01586 *
I(foot * male)  1    1694    1694   1.6036   0.20737
Residuals     149  157360    1056
>
> anova(shf.mod1,shf.mod2) ### Compare Models 1 and 2
Analysis of Variance Table

Model 1: height ~ hand + foot
Model 2: height ~ hand + foot + male
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    152 189029
2    151 165341  1     23687 21.633 7.157e-06 ***

> anova(shf.mod2,shf.mod3) ### Compare Models 2 and 3
Analysis of Variance Table

Model 1: height ~ hand + foot + male
Model 2: height ~ hand + foot + male + I(hand * male) + I(foot * male)
  Res.Df    RSS Df Sum of Sq      F  Pr(>F)
1    151 165341
2    149 157360  2    7981.4 3.7787 0.02507 *
```

∇

9.3 R Code for Chapter 9

```
### Chapter 9

### Examples 9.1-9.5  Bollywood Revenue/Budget Analysis

### Example 9.1

bolly <- read.csv("http://www.stat.ufl.edu/~winner/data/bollywood_boxoffice.csv")
attach(bolly); names(bolly)

### Figure 9.1
par(mfrow=c(2,2))
plot(Gross ~ Budget, main="Y=Gross, X=Budget")
abline(lm(Gross~Budget))
plot(log(Gross) ~ Budget, main="Y=log(Gross), X=Budget")
abline(lm(log(Gross) ~ Budget))
plot(Gross ~ log(Budget), main="Y=Gross, X=log(Budget)")
abline(lm(Gross ~ log(Budget)))
plot(log(Gross) ~ log(Budget), main="Y=log(Gross), X=log(Budget)")
abline(lm(log(Gross) ~ log(Budget)))
### End Figure 9.1
```

```
Y <- log(Gross)
X <- log(Budget)
n <- length(Y)

SSYY <- sum((Y-mean(Y))^2)
SSXX <- sum((X-mean(X))^2)
SSXY <- sum((X-mean(X))*(Y-mean(Y)))

b1 <- SSXY/SSXX
b0 <- mean(Y) - b1*mean(X)
SSE <- SSYY - b1^2*SSXX
MSE <- SSE/(n-2)
se.b1 <- sqrt(MSE/SSXX)
se.b0 <- sqrt(MSE*(1/n + mean(X)^2/SSXX))

ss.out <- cbind(SSYY, SSXX, SSXY, SSE, MSE, b1, b0, se.b1, se.b0)
colnames(ss.out) <- c("SSYY", "SSXX", "SSXY", "SSE", "MSE",
          "beta1-hat", "b0-hat", "SE{b1}", "SE{b0}")
round(ss.out, 4)

### Example 9.2

t.b1 <- b1/se.b1
p.b1 <- 2*(1-pt(abs(t.b1),n-2))
beta1.LB <- b1 - qt(.975,n-2)*se.b1
beta1.UB <- b1 + qt(.975,n-2)*se.b1
b1.out <- cbind(b1, se.b1, t.b1, p.b1, beta1.LB, beta1.UB)
t.b0 <- b0/se.b0
p.b0 <- 2*(1-pt(abs(t.b0),n-2))
beta0.LB <- b0 - qt(.975,n-2)*se.b0
beta0.UB <- b0 + qt(.975,n-2)*se.b0
b0.out <- cbind(b0, se.b0, t.b0, p.b0, beta0.LB, beta0.UB)

b.out <- rbind(b0.out, b1.out)
rownames(b.out) <- c("Intercept", "log(Budget)")
colnames(b.out) <- c("Estimate", "Std. Error", "t", "P-Value", "Lower Bound",
"Upper Bound")
round(b.out, 4)

## Analysis using lm (linear model) function in R

bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
confint(bolly.mod1)


### Example 9.3

x.star <- log(60)
yhat.star <- b0 + b1*x.star
CIstar.LB <- yhat.star - qt(.975,n-2)*sqrt(MSE*(1/n+(x.star-mean(X))^2/SSXX))
CIstar.UB <- yhat.star + qt(.975,n-2)*sqrt(MSE*(1/n+(x.star-mean(X))^2/SSXX))
PIstar.LB <- yhat.star - qt(.975,n-2)*sqrt(MSE*(1+1/n+(x.star-mean(X))^2/SSXX))
PIstar.UB <- yhat.star + qt(.975,n-2)*sqrt(MSE*(1+1/n+(x.star-mean(X))^2/SSXX))

cipi_log.out <- cbind(x.star,yhat.star,CIstar.LB,CIstar.UB,PIstar.LB,PIstar.UB)
cipi_orig.out <- cbind(exp(x.star),exp(yhat.star),exp(CIstar.LB),
               exp(CIstar.UB),exp(PIstar.LB),exp(PIstar.UB))
cipi.out <- rbind(cipi_log.out, cipi_orig.out)
rownames(cipi.out) <- c("Log Scale", "Original Scale")
colnames(cipi.out) <- c("X*","Y-hat*","CI Lower","CI Upper","PI Lower","PI Upper")
round(cipi.out,4)
```

```
X.grid <- seq(0,6,0.01)
yhat <- b0 + b1*X.grid

CI.LO <- yhat - qt(.975,n-2)*sqrt(MSE*(1/n + (X.grid-mean(X))^2/SSXX))
CI.HI <- yhat + qt(.975,n-2)*sqrt(MSE*(1/n + (X.grid-mean(X))^2/SSXX))
PI.LO <- yhat - qt(.975,n-2)*sqrt(MSE*(1 + 1/n + (X.grid-mean(X))^2/SSXX))
PI.HI <- yhat + qt(.975,n-2)*sqrt(MSE*(1 + 1/n + (X.grid-mean(X))^2/SSXX))

## Using predict function based on bolly.mod1 object with X*=log(60)
# CI for mean
ci.log60 <- predict(bolly.mod1, list(X=log(60)), interval="c")
# PI for individual movie
pi.log60 <- predict(bolly.mod1, list(X=log(60)), interval="p")

cipi.out1 <- rbind(ci.log60, pi.log60, exp(ci.log60), exp(pi.log60))
colnames(cipi.out1) <- c("Estimate", "Lower Bound", "Upper Bound")
rownames(cipi.out1) <- c("CI(log scale)", "PI(log scale)",
                         "CI(original scale)", "PI(original scale)")
round(cipi.out1, 4)

## Figure 9.2

par(mfrow=c(1,1))
plot(X,Y,xlim=c(0,6), ylim=c(-1,7),
main="Bollywood Box Office Data, Regression Line, CI and PI")
lines(X.grid,yhat,lty=1)
lines(X.grid,CI.LO,lty=2)
lines(X.grid,CI.HI,lty=2)
lines(X.grid,PI.LO,lty=3)
lines(X.grid,PI.HI,lty=3)
## End Figure 9.2


### Example 9.4

plot(Y ~ X, pch=16,
    main="Bollywood Movies: log(Revenue) vs log(Budget)")
abline(lm(Y ~ X), lwd=2)
abline(h=mean(Y), lty=2, lwd=2)
legend(1.5,6,c("Fitted Equation", "Ybar"), lty=c(1,2), lwd=c(2,2))

SSR <- SSYY-SSE
MSE <- SSE/(n-2)
F_obs <- (SSR/1)/MSE
F_05 <- qf(.95,1,n-2)
p_F <- 1 - pf(F_obs,1,n-2)
aov.out <- cbind(SSYY,SSE,SSR,MSE,F_obs,F_05,p_F,SSR/SSYY)
colnames(aov.out) <- c("TSS","SSE","SSR","MSE","F_obs","F(.05)","P-value","R^2")
round(aov.out,4)

### Using aov function
bolly.mod1 <- lm(Y ~ X)
summary(bolly.mod1)
anova(bolly.mod1)

### Example 9.5

cor.test(X,Y)

rm(list=ls(all=TRUE))


### Example 9.6
```

```
berry1 <- read.csv("http://www.stat.ufl.edu/~winner/data/berry_sugar_chewy.csv")
attach(berry1); names(berry1)
sugar.lev <- unique(sugar)

## Figure 9.3

plot(sugar, chewiness, main="Berry Chewiness by Sugar Content")
abline(lm(chewiness ~ sugar))
for (i in 1:length(sugar.lev)) {
  lines(c(sugar.lev[i]-2,sugar.lev[i]+2),
   c(mean(chewiness[sugar==sugar.lev[i]]),
     mean(chewiness[sugar==sugar.lev[i]])),
    lwd=2)
}
## End Figure 5.3

(ybar_j <- as.vector(tapply(chewiness, sugar, mean)))
(s_j <- as.vector(tapply(chewiness, sugar, sd)))
(n_j <- as.vector(tapply(chewiness, sugar, length)))
c <- length(n_j)
(X_j <- unique(sugar))

chewy1 <- lm(chewiness ~ sugar) ### Fit Linear Regression
summary(chewy1)
anova(chewy1)

(yhat_j <- predict(chewy1,list(sugar=X_j)))
SS_LF <- sum(n_j * (ybar_j - yhat_j)^2)
df_LF <- c-2
MS_LF <- SS_LF/df_LF
SS_PE <- sum((n_j-1)*s_j^2)
df_PE <- sum(n_j)-c
MS_PE <- SS_PE/df_PE
F_LF <- MS_LF / MS_PE
F_05 <- qf(.95,c-2,sum(n_j)-c)
p_F <- 1 - pf(F_LF, c-2, sum(n_j)-c)

LF.out <- cbind(df_LF, SS_LF, MS_LF, df_PE, SS_PE, MS_PE, F_LF, F_05, p_F)
colnames(LF.out) <- c("df(LF)","SSLF","MSLF","df(PE)","SSPE","MSPE",
    "F_LOF", "F(.05)", "P(F>=F_LOF)")
round(LF.out,4)

chewy2 <- lm(chewiness ~ factor(sugar)) ### Fit 1-Way ANOVA
anova(chewy2)

anova(chewy1, chewy2) ### Compare Linear Reg w/ 1-Way ANOVA

rm(list=ls(all=TRUE))


### Examples 9.7-9.8


### Example 9.7

shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)

f.height <- height[gender == 2] ### Female Heights
f.hand <- hand[gender == 2] ### Female Hand Lengths
f.foot <- foot[gender == 2] ### Female Foot Lengths

f.stature <- data.frame(f.height, f.hand, f.foot)
```

```
## Figure 9.4
plot(f.stature)
## End Figure 9.4

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
summary(shf.mod1)
confint(shf.mod1)

### Example 9.8

shf.mod1 <- lm(f.height ~ f.hand + f.foot)
summary(shf.mod1)
anova(shf.mod1)
drop1(shf.mod1, test="F")

rm(list=ls(all=TRUE))


### Example 9.9

hotel_ec <- read.csv("http://www.stat.ufl.edu/~winner/data/hotel_energy.csv")
attach(hotel_ec); names(hotel_ec)

enrgcons <- enrgcons/1000000
area <- area/1000

## Full Model
hec.mod1 <- lm (enrgcons ~ area + age + effrooms)
summary(hec.mod1)
anova(hec.mod1)

## Reduced Model
hec.mod2 <- lm (enrgcons ~ area)
summary(hec.mod2)
anova(hec.mod2)

## Full versus Reduced F-test
anova(hec.mod2, hec.mod1)

rm(list=ls(all=TRUE))

### Example 9.10

shf1 <- read.table("http://www.stat.ufl.edu/~winner/data/stature_hand_foot.dat",
header=F, col.names=c("idnum", "gender", "height", "hand", "foot"))
attach(shf1)

male <- 2-gender ### male = 1 if male, 0 if female

## Model 1: Common slope/intercept
shf.mod1 <- lm(height ~ hand + foot)
summary(shf.mod1)
anova(shf.mod1)

## Model 2: Common slope/Different intercept
shf.mod2 <- lm(height ~ hand + foot + male)
summary(shf.mod2)
anova(shf.mod2)

## Model 3: Different slope/intercept
shf.mod3 <- lm(height ~ hand + foot + male + I(hand*male) + I(foot*male))
summary(shf.mod3)
anova(shf.mod3)
```

```
anova(shf.mod1,shf.mod2) ### Compare Models 1 and 2
anova(shf.mod2,shf.mod3) ### Compare Models 2 and 3

rm(list=ls(all=TRUE))
```

Bibliography

[1] Agresti, A. (2002). *Categorical Data Analysis. 2nd Ed.* Wiley, New York.

[2] Agresti, A. and B. Coull (1998). "Approximate is Better than 'Exact' for Interval Estimation of Binomial Proportions," *The American Statistician*, Vol. 52, #2 pp. 119-126.

[3] Agresti, A. and L. Winner (1997). "Evaluating Agreement and Disagreement Among Movie Reviewers," *Chance*, Vol. 10, pp. 10-14.

[4] Ahonen, H., A.J. Stow, R.G. Harcourt, and I. Charrier (2014). "Adult Male Australian Sea Lion Barking Calls Reveal Clear Geographical Variations," *Animal Behaviour*, Vol. 97, pp. 229-239.

[5] Ali, E., W. Guang, and A. Ibrahim (2014). "Empirical Relations Between Compressive Strength and Microfabric Properties of Amphibolites Using Multivariate Regression, Fuzzy Inference, and Neural Networks: A Comparative Study," *Engineering Geology*, Vol. 183, pp. 230-240.

[6] Barnum, D.T. and J.M. Gleason (1994). "The Credibility of Drug Tests: A Multi-Stage Bayesian Analysis," *Industrial and Labor Relations Review*, Vol. 47, #4, pp. 610-621.

[7] Berenson, J.R., A. Lichtenstein, L. Porter, et al. (1996). "Efficacy of Pamidronate in Reducing Skeletal Events in Patients with Advanced Myeloma," *New England Journal of Medicine*, Vol. 334, pp. 488-493.

[8] Bhatnagar, A. and V.K. Mehta (2007). "Efficacy of Deltamethrin and Cyfluthrin Impregnated Cloth Over Uniform Against Mosquito Bites," *Medical Journal Armed Forces India*, Vol. 63, pp. 120-122.

[9] Broders, A.C. (1920). "Squamous-Cell Epithelioma of the Lip," *Journal of the American Medical Association*, Vol. 74, pp. 656-664.

[10] Bruce, A.C., J.E.V. Johnson, and J. Peirson (2012). "Recreational versus Professional Bettors: Performance Differences and Efficiency Implications," *Economic Letters*, Vol. 114, pp. 172-174.

[11] Cameron, A.C. and P.K. Trivedi (2005). *Microeconometrics: Methods and Applications*. Cambridge, Cambridge.

[12] Chambers, G.F. (1889). *Handbook of Astronomy, 4th Ed.* Oxford.

[13] Chang, P.-C., S.-Y. Chou, and K.-K. Shieh (2013). "Reading Performance and Visual Fatigue When Using Electronic Paper Displays in Long-Duration Reading Tasks Under Various Lighting Conditions," *Displays*, Vol. 34, pp:208-214.

[14] Chihara, L. and T. Hesterberg (2011). *Mathematical Statistics with Resampling and R.* Wiley, Hoboken, NJ.

[15] Clarke, R.D. (1946). "An Application of the Poisson Distribution," *Journal of the Institute of Actuaries*, Vol. 72, p. 481.

[16] Cohen, A.M. (1996). "The Hands of Blues Guitarists," *American Music*, Vol. 14, #4, pp. 455-479.

[17] Culp Jr., R.L., and D. Pollage (2002). "The Rhetoric of Strict Products Liability versus Negligence: An Empirical Analysis," *New York University Law Review*, Vol. 77, #4, pp. 874-961.

[18] Dror, I.E., C. Champod, G. Langenburg, D. Charlton, H. Hunt, and R. Rosenthal (2011). "Cognitive Issues in Fingerprint Analysis: Inter- and Intra-Expert Consistency and the Effect of a 'Target' Comparison," *Forensic Science International*, Vol. 208, pp. 10-17.

[19] Durante, C., C. Baschieri, L. Bertacchini, D. Bertelli, M. Cocchi, A. Marchetti, D. Manzini, G. Papotti, S. Sighinolfi (2015). "An Analytical Approach to Sr Isotope Ratio Detemination in *Lambrusco* Wines for Geographic Traceability Purposes," *Food Chemistry*, Vol. 173, pp. 557-563.

[20] Efron, B. and R. Tibshirani (1993). *An Introduction to the Bootstrap.* Chapman & Hall, New York.

[21] Eisenberg, T., S.P. Garvey and M.T. Wells (2001). "Forecasting Life and Death: Juror Race, Religion, and Attitude Toward the Death Penalty," *The Journal of Legal Studies*, Vol. 30, pp. 277-311.

[22] Feller, W. (1968). *An Introduction to Probability Thory and Its Applications, Vol.1, 3rd. Ed.* Wiley, New York.

[23] Gilovich, T. R. Vallone, and A. Tvesky (1985). "The Hot Hand in Basketball: On the Misperception of Random Sequences," *Cognitive Psychology*, Vol. 17, #3, pp. 295-314.

[24] Gueguen, N. and C. Jacob (2013). "Color and Cyber-Attractiveness: Red Enhances Men's Attraction to Women's Personal Ads," *Color Research and Application*, Vol. 38, #4, pp. 309-312.

[25] Hammond, E.C. and D. Horn (1954), "The Relationship Between Human Smoking Habits and Death Rates," *Journal of the American Medical Association*, Vol. 155, pp. 1316-1328.

[26] Holland, T.H. (1902). "The Kanets of Kulu and Lahoul, Punjab: A Study in Contact-Metamorphism," *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 32, pp.96-123.

[27] Hollander, M. and D.A. Wolfe (1999). *Nonparametric Statistical Methods, 2nd. Ed.*, Wiley, New York.

[28] Jeffery, F.R. and J. Stathis (1996). "Function Point Sizing: Structure, Validity, and Applicability," *Empirical Software Engineering*, Vol. 1, #1, pp. 11-30.

[29] Jorgensen, M., U. Indahl, D.Sjoberg (2003). "Software Effort Estimation by Analogy and "Regression to the Mean"," *The Journal of Systems and Software*, Vol. 68, pp. 253-262.

[30] Kahneman, D., P. Slovic, and A. Tversky (1982). *Judgment Under Uncertainty: Heuristics and Biases*, Cambridge University Press, Cambridge, UK.

[31] Kahneman, D. and A. Tversky (1984). "Choices, Values, and Frames," *American Psychologist*, Vol. 35, #4, pp. 341-350.

[32] Kamimura, A., T. Takahishi, and Y. Watanabe (2000). "Investigation of Topical Application of Procyanidin B-2 from Apple to Identify its Potential Use as a Hair Growing Agent," *Phytomedicine*, Vol. 7, #6, pp. 529-536.

[33] Koyama, K., H. Hokunan, M. Hasegawa, S. Kawamura, and S. Koseki (2016). "Do Bacterial Cell Numbers Follow a Theoretical Poisson Distribution? Comparison of Experimentally Obtained Numbers of Single Cells with Random Number Generation via Computer Simulation," *Food Microbiology*, Vol. 60, pp. 49-53.

[34] Liang, D.G., J.R. Dusseldorp , C. van Schalkwyk , S. Hariswamy, S. Wood, V. Rose, P. Moradi (2016). "Running Barbed Suture Quilting Reduces Abdominal Drainage in Perforator-Based Breast Reconstruction," *Journal of Plastic, Reconstructive & Aesthetic Surgery*, Vol. 69, pp. 42-47.

[35] Lim, T.-S. and W.Y. Loh (1996). "A Comparison of Tests of Equality of Variances," *Computational Statistics and Data Analysis*, Vol. 22, pp. 287-301.

[36] Lin, W. and J. Wang (2012). "An Integrated 3D Log Processing Optimization System for Hardwood Sawmills in Central Appalachia, USA," *Computers and Electronics in Agriculture*, Vol. 82, pp. 61-74.

[37] Lister, J. (1870). "Effects of the Antiseptic System of Treatment on the Salubrity of a Surgical Hospital," *The Lancet*, Vol.1, pp. 4-6,40-42.

[38] Mehrgini, B., H. Memarian, M.B. Dusseault, A. Ghavidel, and M. Heydarizadeh (2016). "Geomechanical Characteristics of Common Reservoir Caprock in Iran (Gachsaran Formation), Experimental and Statistical Analysis," *Journal of Natural Gas Science and Engineering*, Vol. 34, pp. 898-907.

[39] Ott, R.L. and M. Longnecker (2016). *Statistical Methods & Data Analysis, 7th Ed.* Cengage Learning, Boston.

[40] Page, L., D.A. Savage, and B. Torgler (2014). "Variation in Risk Seeking Behaviour Following Large Losses: A Natural Experiment," *European Economic Review*, Vol. 71, pp. 121-131.

[41] Pachel, C. and J. Neilson (2010). "Comparison of Feline Water Consumption Between Still and Flowing Water Sources: A Pilot Study," *Journal of Veterinary Behavior*, Vol. 5, pp. 130-133.

[42] Peckmann, T.R., S. Scott, S. Meek, and P. Mahakkanukrauh (2017). "Sex Estimation from the Scapula in a Contemporary Thai Population: Applications for Forensic Anthropology," *Science and Justice*, Vol. 57, pp. 270-275.

[43] Poburka, P.J., R.R. Patel, and D.M. Bless (2017). "Voice-Vibratory Assessment With Laryngeal Imaging (VALI) Form: Reliability of Rating Stroboscopy and High-speed Videoendoscopy," *Journal of Voice*, Vol. 31, No. 4, pp. 513.e1513.e14.

[44] Rusanganwa, J. (2013). "Multimedia as a Means to Enhance Teaching Technical Vocabulary to Physics Undergraduates in Rwanda," *English for Specific Purposes*, Vol. 32, pp. 36-44.

[45] Sanli, G.S., E.D. Kizilkanat, N. Boyan, E.T. Ozsahin, M.G. Bozkir, R. Soames, H. Erol, and O. Oguz (2005). "Stature Estimation Based on Hand Length and Foot Length," *Clinical Anatomy*, Vol. 18, #8, pp. 589-596.

[46] Scheaffer, R.L., W. Mendenhall, and L. Ott (1990). *Elementary Survey Sampling, 4th Ed.* PWS-KENT, Boston.

[47] Sheldrake, R., P. Smart, and L. Avraamides (2015). "Automated Tests for Telephone Telepathy Using Mobile Phones," *Explore*, Vol. 11, #4, pp. 310-319.

[48] Storm, L., P.E. Tressoldi, and L. Di Risio (2010). "Meta-Analysis of Free Response Studies, 1992:2008: Assessing the Noise Reduction Model in Parapsychology," *Psychological Bulletin*, Vol. 136, No. 4, pp. 471-485.

[49] Teerapatsakul, C., C. Bucke, R. Parra, T. Keshavarz, and L. Chitradon (2008). "Dye Decolorisation by Laccase Entrapped in Copper Alginate," *World Journal of Microbiology and Technology*, Vol. 24, pp. 1367-1374.

[50] Thorndike, F. (1926). "Applications of Poisson's Probability Summation," *Bell System Technical Journal*, Vol. 5, pp. 604-624.

[51] Walter, S.R., W.T.M. Dunsmuir, and J.I. Westbrook (2015). "Studying Interruptions and Multitasking in situ: The Untapped Potential of Quantitative Observational Studies," *International Journal of Human-Computer Studies*, Vol. 79, pp. 118-125.

[52] Winner, L. (2006). "NASCAR Winston Cup Race Results for 1975-2003," *Journal of Statistical Education*, Vol. 14, #3.

[53] Xin, Y., S. Lu, N. Zhu, and W. Wu (2012). "Energy Consumption Quota of Four and Five Star Luxury Hotels Buildings in Hainan Province, China," *Energy and Buildings*, Vol. 45, pp. 250-256.

[54] Zouid, I., R. Siret, F. Jourjon, E. Mehinagic, and L. Rolle (2013). "Impact of Grapes Heterogeneity According to Sugar Level on Both Physical and Mechanical Berries Properties and Their Anthocyanins Extractability at Harvest," *Journal of Texture Studies*, Vol. 44, pp. 95-103.