

The **median** is the midpoint of a distribution, with half of the observations smaller than it and half of the observations are larger than it. The **lower quartile Q1** is the median of the observations between the minimum observation and the overall median, whereas the **upper quartile Q3** is the median for the observations from the overall median to the maximum observation. Therefore Q1, the overall median and Q3 divide the observations into quarters, hence the term quartile. Using the median, Q1, Q3, the minimum observation and the maximum observation, we can obtain the **five-number summary** of a set of observations.

In symbols the five-number summary is: Minimum Q1 Median Q3 Maximum

A **boxplot** is a graph of the five-number summary.

- A central box spans the quartiles Q1 and Q3
- A line in the box marks the median M
- Lines extend from the box out to the smallest and largest observations

The **interquartile range** is the distance from the first and third quartiles, $IQR=Q3-Q1$. An observation is an **outlier** if it falls more than $1.5 \cdot IQR$ above Q3 or below Q1.

Example #1

Here are the passing yards for each of the 13 games the Gators played in the 2003 season.

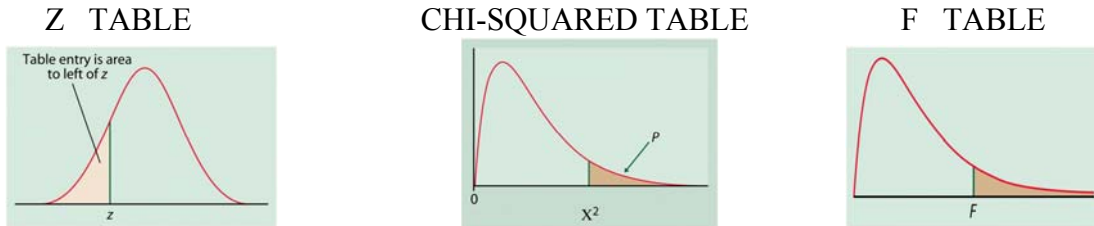
357 219 227 281 268 234 229 269 235 179 121 303 268

a) Place the numbers in increasing order and find M, Q1 and Q3.

b) Draw the boxplot for the passing yards.

DISTRIBUTIONS AND DENSITIES

In Stat I, we focused on a unimodal or mound-shaped distribution called the **normal** distribution. In Stat II, we will focus on two skewed distributions, the **chi-squared** and **F** distributions.



Instead of using the previously mentioned graphs (boxplots, etc...), when describing the overall pattern of a distribution we use a **density curve**, which always has an **area exactly equal to 1 underneath it**. Two important measures for analyzing data and density curves are the **mean μ** and **standard deviation σ** .

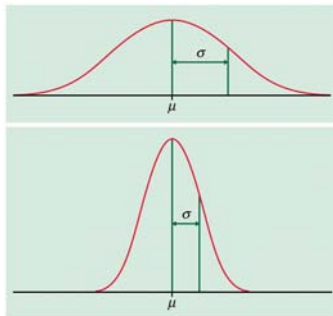


Figure 1.24 page 69

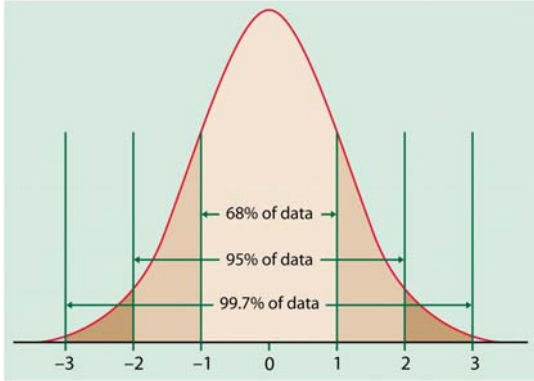
The mean μ along with the median describes the **center** of the distribution.

The standard deviation σ describes its **spread**.

THE NORMAL DISTRIBUTION

The normal distribution is a **symmetric, bell-shaped** distribution with its mean equal to its median.

The **68-95-99.7% Rule** allows us to test whether a group of data is in fact normally distributed.



If data are normally distributed with mean μ and standard deviation σ , then

- Approximately 68% of the observations fall within σ of the mean μ ($\mu \pm \sigma$)
- Approximately 95% of the observations fall within 2σ of the mean μ ($\mu \pm 2\sigma$)
- Approximately 99.7% of the observations fall within 3σ of the mean μ ($\mu \pm 3\sigma$)

When we are dealing with samples and do not have the mean μ or standard deviation σ of our population, we can use the **mean of our sample**, \bar{x} , and the **standard deviation of our sample**, s , instead.

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$$

Example #2

The following is a sample of 25 measurements:

7 6 6 11 8 9 11 9 10 8 7 7 5 9 10 7 7 7 7 9 12 10 10 8 6

a) Compute \bar{X} and s for this sample

b) Count the number of measurements in the intervals $\bar{x} \pm s, \bar{x} \pm 2s, \bar{x} \pm 3s$ and express each count as a percentage

c) Compare the percentages found in part (b) to the percentages given by the 68-95-99.7% Rule. Is the data normally distributed?

STANDARD NORMAL DISTRIBUTION

In order to be able to **compare** values that are measured on **different scales**, like a math score on the ACT test which is out of 36, vs. a math score on the SAT which is out of 800, we use the **standardized value of an observation x** . If x is from a distribution with mean μ and standard deviation σ , the standardized value of x , also known as its **z-score**,

$$\text{is } z = \frac{x - \mu}{\sigma}.$$

When we standardize an observation x which is from a population that is normally distributed with mean μ and standard deviation σ , symbolized by $\mathbf{X} \sim \mathbf{N}(\mu, \sigma)$, then $Z = \frac{X - \mu}{\sigma}$ (where **z is just the standardized value of x**) will be normally distributed with mean 0 and standard deviation 1, $\mathbf{Z} \sim \mathbf{N}(0, 1)$. When an observation is said to be normally distributed with mean 0 and standard deviation 1, it is said to be from the **standard normal distribution**.

We can use the area under a density to talk about proportions or probabilities. The **Standard Normal table** allows us to **calculate proportions** pertaining to our population.

Example #3 (Exercise 1.86)

Eleanor scores 680 on the math part of the SAT exam. The distribution of SAT scores in a reference population is normal with mean 500 and standard deviation 100. Gerald takes the ACT math test and scores 27. ACT scores are normally distributed with mean 18 and standard deviation 6. Find the z-scores for both students. Assuming that both tests measure the same kind of ability, who has the higher score?

Example #4 (Exercise 1.92)

In 2000, the scores of men on the math part of the SAT approximately followed a normal distribution with mean 533 and standard deviation 115.

a) What proportion of men scored above 500?

b) What proportion scored between 400 and 600?

Example #5

High school boys are timed in their ability to run a mile as part of a physical fitness test. The times required to complete the run are approximately normal with mean 450 seconds and standard deviation 40 seconds. You are told that your run is at the upper quartile of run times. What is your run time?

MAKING DECISIONS ABOUT POPULATIONS

Suppose we want to know an average (μ), or proportion (p) of a population, like the average number of alcoholic beverages UF students drink on a Saturday night or the proportion of Americans who wear glasses or contact lenses. Instead of asking everyone in our population our question, which would be extremely time consuming and expensive, we pick a sample and use the information from our sample to estimate what our answer would be for our population. The best way to pick a sample is to choose the people in our sample randomly (hence picking a simple random sample – **SRS**) from the population so as to avoid any bias. The **population parameter** is what we **want to know** about our population, like the population mean or proportion. The **sample statistic** is the value that we calculate based on the data in our sample, like the sample mean or proportion, which we use to **estimate our unknown population parameter**.

SAMPLING DISTRIBUTIONS

The distribution of a sample statistic is its **sampling distribution**. It is a density curve which shows how a statistic would vary in repeated sampling.

When we want to ask our population a question which deals with **yes or no answers**, like whether the members of the population wear glasses or contact lenses, we are dealing with **counts and proportions**.

The random variable X is a **count** of the number of successes in our sample, for example it would be the number of people who wear glasses or contact lenses. For a sample with n observations, the **sample proportion** is $\hat{p} = \frac{X}{n}$. It is the ratio of the number of successes in our sample to the total number of people in our sample.

When our population is at least 10 times the size of our sample, the count X of successes in a SRS of size n has approximately a **Binomial distribution** with size n and probability of success p .

When n is large, it is too difficult (without a computer software) to calculate binomial probabilities. Therefore, we use the **normal distribution as an approximation** for the distribution of a binomial with large n .

When n is large:

- There are at least 10 successes, symbolized by $np \geq 10$
- There are at least 10 failures, symbolized by $n(1-p) \geq 10$
- The number of successes, X , is normally distributed with mean np and standard deviation $\sqrt{np(1-p)}$, symbolized by $X \sim N(np, \sqrt{np(1-p)})$
- The sample proportion, \hat{p} , is normally distributed with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$, symbolized by $\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$

Example #6 (Exercise 5.25)

Here is a simple probability model for multiple-choice tests. Suppose that each student has probability p of correctly answering a question chosen at random from a universe of possible questions. (A strong student has a higher p than a weak student. The correctness of answers to different questions are independent. Jodi is a good student for whom $p = 0.75$).

a) Use the normal approximation to find the probability that Jodi scores 70% or lower on a 100-question test.

b) If the test contains 250 questions, what is the probability that Jodi will score 70% or lower?

When we want to ask our population a question that deals with an **average**, like the average number of alcoholic beverages UF students drink on a Saturday night, we are dealing with **means**.

The random variable X is a measure on each individual in the sample, like the number of drinks or the score on an exam. In a sample of size n , there will be n values of X_i , where X_i is the measurement of X from person i in our sample and so i can be any number from 1 to n . Therefore, the **sample mean** of an SRS of size n is $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$.

When \bar{X} is the mean of an SRS of size n from a population that is normally distributed with mean μ and standard deviation σ , then $\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

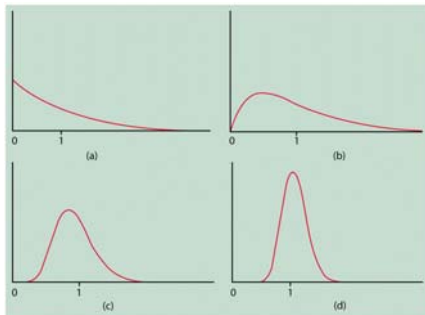


Figure 5.10

Furthermore, the **Central Limit Theorem** states that no matter what the distribution of our original population, if it has mean μ and standard deviation σ , then when n is large \bar{X} will always have a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$. Figure 5.10 shows how as our sample size increases, the distribution of \bar{X} becomes more and more normal. The distribution of \bar{X} usually reaches normality when n is 30.

Example #7 (Exercise 5.32)

The scores of students on the ACT college entrance exam in 2001 had mean $\mu=21.0$ and standard deviation $\sigma=4.7$. The distribution of scores is only roughly normal.

a) What is the approximate probability that a single student randomly chosen from all those taking the test scores 23 or higher?

b) Now take an SRS of 50 students who took the test. What are the mean and standard deviations of the sample mean score \bar{X} of these 50 students?

c) What is the approximate probability that the mean score \bar{X} of these students is 23 or higher?

d) Which of your two normal probability calculations in (a) and (c) is more accurate? Why?

CONFIDENCE INTERVALS

As previously mentioned, we use our sample statistic to estimate our population parameter. Therefore, if we want to find out a population mean but we don't want to find our measure X (for example an exam score) for every individual in our population, we pick an SRS and find our measure X for every individual in our sample. Then we calculate our sample mean and use it to estimate our population mean. The same method holds true for estimating a population proportion.

However, if we want to estimate our population parameter with more confidence, we use a confidence interval. Our confidence interval allows us to be more confident in our estimate of the true population parameter since it not only uses the sample statistic as the estimate, but the confidence interval also allows for some error. The interval is therefore

made up of our estimate and a margin of error, which in turn shows us how accurate we believe the estimate is.

Confidence intervals: estimate \pm margin or error

A **level C confidence interval** for a parameter is an **interval** computed from sample data by a method **that has probability C of producing an interval containing the true value of the parameter**. So for example, a 95% CI for μ would be an interval that includes μ with a probability of 0.95.

Confidence Interval for a population mean, μ : $\bar{X} \pm z^* \frac{\sigma}{\sqrt{n}}$

Here z^* is the value on the standard normal curve with area C (confidence level) between $-z^*$ and z^* , when the data are normally distributed and $n \geq 30$.

Confidence Interval for a population proportion, p : $\tilde{p} \pm z^* \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}$

Here \tilde{p} is Wilson's estimate of the population proportion, $\tilde{p} = \frac{X+2}{n+4}$, where $n \geq 5$.

Example #8 (Exercise 8.1)

To profitably produce a planned upgrade of a software product you make, you must charge customers \$100. Are your customers willing to pay this much? You contact a random sample of 40 customers and find that 11 would pay \$100 for the upgrade. Find a 95% CI for the proportion of all of your customers (the population) who would be willing to buy the upgrade for \$100.

TESTS OF SIGNIFICANCE

What if we think that our population parameter might be a specific value? We can perform a significance test to assess whether our hypothesized value of our parameter can be true.

The setup for a significance test is to state our **null hypothesis** (H_0 , a statement with your claim of the true value of your parameter) and our **alternative hypothesis** (H_a , a statement that we might think is true instead of H_0) and to calculate our **test statistic** and

p-value. The test statistic calculates the likelihood that the null hypothesis is true given our set of data.

In general, our **null hypothesis** will specify the value that we suspect our true population parameter to be.

There are three possible **alternative hypotheses**:

- i) The true population parameter could be **bigger** than the value we suspect

- ii) The true population parameter could be **smaller** than the value we suspect

- iii) The true population parameter is **not equal** to the value that we suspect

Remember that when writing a hypothesis, you must always specify the population parameter you are hypothesizing about.

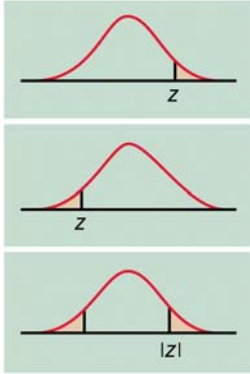
Our **test statistic** allows us to **standardize our data** so that we can use the standard normal table to find the probability of H_0 being true given our data. Our test statistic will be in the form of:

$$\frac{\text{estimate-parameter}}{\text{standard deviation of estimate}}$$

For hypotheses pertaining to μ , with $H_0 : \mu = \mu_0$, our **test statistic** is $z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$, for normally distributed data with a sample size of $n \geq 30$.

For hypotheses pertaining to p , with $H_0 : p = p_0$ our **test statistic** is $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, when there are at least 10 successes and 10 failures.

A **p-value** is the probability, assuming H_0 is true, that the test statistic would take a value as or more extreme than that actually observed. If the p-value is low, usually less than 0.05, we reject H_0 in favor of H_a .



The shaded area is equal to the p-value for a test of H_0 against

- i) $H_a: \mu > \mu_0$ or $H_a: p > p_0$
Therefore, the p-value is $P(Z \geq z)$
- ii) $H_a: \mu < \mu_0$ or $H_a: p < p_0$
Therefore, the p-value is $P(Z \leq z)$
- iii) $H_a: \mu \neq \mu_0$ or $H_a: p \neq p_0$
Therefore, the p-value is $P(Z \geq |z|)$ or $2 * P(Z \geq z)$
(since the normal distribution is symmetric)

Example #9

Here are the number of visits to a bar in a year for a sample of 44 college students.

40 26 39 14 42 18 25 43 46 27 19 47 19 26 35 34 15 44
 40 38 31 46 52 25 35 35 33 29 34 41 49 28 52 47 35 48
 22 33 41 51 27 14 54 45

These students can be considered to be an SRS of the college students in a town. Bar visits are approximately normal. Suppose that the standard deviation of bar visits is known to be $\sigma=11$. The researcher believes that the mean number of visits to a bar in a year of all college students in this town is higher than the national mean μ , which is 32.

a) State the appropriate H_0 and H_a to test this suspicion.

b) Carry out the test. Give the p-value, and then interpret the result in plain language.

TWO-SIDED SIGNIFICANCE TESTS AND CONFIDENCE INTERVALS

We can make a decision to reject or fail to reject H_0 based on a given confidence interval. A level C two-sided significance test rejects a hypothesis $H_0: \mu = \mu_0$ (or equivalently $H_0: p = p_0$), exactly when the value μ_0 (or p_0) falls outside a level $1-C$ confidence interval for μ (or p).

Remember, a CI is an interval with level C of all the possible values of our true population parameter. Therefore, if our hypothesized value of our parameter is included in our CI, then it is a possible value of our true population parameter, and therefore we cannot reject H_0 .

Example #10 (Exercise 6.6 and 6.64)

You measure the weights of 24 male runners and find the sample mean $\bar{X} = 61.79$ kg and population standard deviation $\sigma = 4.5$ kg.

a) Give the 95% CI for the mean weight of the population of all such runners.

b) Based on this CI, does a test of $H_0: \mu = 61.3$ kg vs. $H_a: \mu \neq 61.3$ kg reject H_0 at the 5% significance level?

c) Would $H_0: \mu = 63$ kg be rejected at the 5% level if tested against a two-sided alternative?

THE t DISTRIBUTION

When the population standard deviation σ is unknown, we estimate it by using the **sample standard deviation, s**. When this occurs, we use the t-statistic instead of the z-statistic in our calculations. The t-statistic is $t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$ and it has the **t distribution**

with n-1 degrees of freedom.

When the size of our SRS is small, **n < 30**, we also use the t-statistic.

Confidence intervals and significance tests are performed in the same way as before, but we now use the t-statistic instead of the z-statistic.

Example #11 (Exercise 7.2)

You want to rent an unfurnished one-bedroom apartment for next semester. You take a random sample of 10 apartments advertised in the local newspaper and record the rental rates. Here are the rents (in dollars per month):

500, 650, 600, 505, 450, 550, 515, 495, 650, 395

Find a 95% confidence interval for the mean monthly rent for unfurnished one-bedroom apartments available for rent in this community.

COMPARING TWO MEANS OR TWO PROPORTIONS

When comparing two parameters from two independent populations (the two groups are **not** related), we are looking to see if there is a **difference between the two groups**. Therefore, we want to know if the difference between the two parameters is **zero, positive or negative**.

Let A represented either the mean or proportion of the 1st population and B represented the corresponding mean or proportion of the 2nd population.

- A difference equal to zero implies that the two population parameters are equal.
A-B=0 \rightarrow A=B

- A positive difference implies that A is larger than B, $A - B > 0 \rightarrow A > B$.

- A negative difference implies that A is smaller than B, $A - B < 0 \rightarrow A < B$.

Two-sample Test Statistics:
$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \text{ with d.f. the smaller of } n_1 - 1 \text{ and } n_2 - 1$$

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Two-sample Confidence Intervals:
$$(\bar{X}_1 - \bar{X}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(\bar{X}_1 - \bar{X}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \text{ with d.f. the smaller of } n_1 - 1 \text{ and } n_2 - 1$$

$$(\tilde{p}_1 - \tilde{p}_2) \pm z^* \sqrt{\frac{\tilde{p}_1(1 - \tilde{p}_1)}{n_1 + 2} + \frac{\tilde{p}_2(1 - \tilde{p}_2)}{n_2 + 2}}$$

Example #12 (Exercise 7.58 & 7.59)

Pat wants to compare the cost of one- and two-bedroom apartments in the area of your campus. She collects data for a random sample of 10 advertisements of each type.

Here are the rents for the two-bedroom apartments (in dollars per month):
595, 500, 580, 650, 675, 675, 750, 500, 495, 670

Here are the rents for the one-bedroom apartments:
500, 650, 600, 505, 450, 550, 515, 495, 650, 395

a) Find a 95% confidence interval for the additional cost of a second bedroom.

Pat wonders if two-bedroom apartments rent for significantly more than one-bedroom apartments.

b) State appropriate null and alternative hypotheses.

c) Report the test statistic, its degrees of freedom, and the p-value. What do you conclude?

d) Can you conclude that every one-bedroom apartment costs less than every two-bedroom apartment?

e) Which is more useful to someone planning to rent an apartment, the confidence interval or the significance test? Why?

Example #13

A 95% CI for the difference in the proportion of binge drinkers between women (p_1) and men (p_2) is given to be (-0.069, -0.045).

a) The p-value of $H_0: p_1 - p_2 = 0$ would be ...

- i) equal to 0.05 ii) >0.05 iii) <0.05 iv) unknown

- b) The p-value for a test of $H_0: p_1 - p_2 = 0$ vs. $H_a: p_1 - p_2 > 0$ would be ...
- i) equal to 0.05 ii) > 0.05 iii) < 0.05 iv) unknown

RELATIONSHIPS AMONG DATA

CORRELATION

A **response variable** measures an **outcome** of a study. An **explanatory variable** **explains or causes changes in the response** variable. The response variable is usually denoted by y and plotted on the y -axis. The explanatory variable is denoted by x , and is plotted on the x -axis. For example, a person's height (explanatory) can explain their weight (response), the number of beers you drink (explanatory) can explain your score on a test (response).

Correlation, denoted by r , measures the **direction** and **strength** of the **linear relationship** between two quantitative variables.

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{X}}{\sigma_x} \right) \left(\frac{y_i - \bar{Y}}{\sigma_y} \right)$$

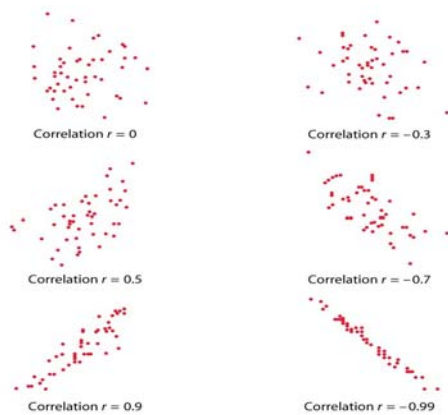


Figure 2.10

Correlation is a number between -1 and 1, where a **negative** values of r means that the two quantitative variables have a **negative** relationship (one variable increases as the other decreases), $r=0$ means that the two variables are not related and a positive value of r means that they have a **positive** relationship (they are either simultaneously increasing or decreasing). Values of r close to 0 imply that the relationship is **weak**, whereas values of r close to either -1 or 1 indicate a **strong** relationship.

LEAST SQUARES REGRESSION

A regression line is a **straight line** that describes the **linear relationship** between a response variable and an explanatory variable. In other words, it shows the dependence of the response variable y on the explanatory variable x .

Based on a set of data, we can fit a straight line to the data which describes their linear pattern. Since our regression line gives us the linear pattern of a group of data, we can use this line to **predict** what value y might be given a value of x . The **least squares regression line** is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible. In other words, we choose the line that comes as close as possible to every data point.

A straight line that relates y to x has an equation of the form $y = \mathbf{a} + \mathbf{bx}$, where **a is the intercept** (the value of y when $x=0$) and **b is the slope** (the amount by which y changes when x increases by one unit).

The equation of the **least squares regression line** of y on x is $\hat{y} = \mathbf{a} + \mathbf{bx}$, where the slope b is equal to $b = \frac{S_{XY}}{S_{XX}} = r\sqrt{\frac{S_{YY}}{S_{XX}}}$, and the intercept a is equal to $a = \bar{Y} - b\bar{X}$.

$$S_{xx} = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{1}{n} \left(\sum_{i=1}^n X_i \right)^2 \quad \sqrt{S_{xx}} = \sigma_x$$

$$S_{YY} = \sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n Y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2 \quad \sqrt{S_{YY}} = \sigma_y$$

$$S_{XY} = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^n X_i Y_i - \frac{1}{n} \sum_{i=1}^n X_i \sum_{i=1}^n Y_i$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

Example #14 (Exercise 2.52)

In Professor Friedman's economics course the correlation between the students' total scores before the final exam and their final exam scores is $r=0.6$. The pre-

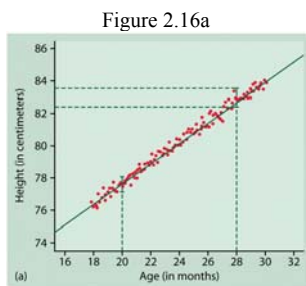
exam totals for all students in the course have mean 280 and standard deviation 30. The final exam scores have mean 75 and standard deviation 8. Professor Friedman has lost Julie's final exam, but knows that her total before the exam was 300. He decides to predict her final exam score from her pre-exam total.

a) What is the slope and intercept of the least squares regression line of final exam scores on pre-exam total scores in this course?

b) Use the regression line to predict Julie's final exam score.

COEFFICIENT OF DETERMINATION

The coefficient of determination is the square of the correlation, R^2 . It is the fraction of the variability in the response variable y that can be explained by the explanatory variable x .



R^2 is a value between 0 and 100%. The closer R^2 is to 100%, the better our explanatory variable is at explaining the variation in the response variable. That is, the better the least squares regression line is at describing the linear relationship between x and y . Our observations in figure 2.16a fall much more tightly around our LSR line than those in figure 2.16b. Therefore, the LSR line in figure 2.16a is better at describing the relationship between x and y and so it has a larger r^2 value than that for figure 2.16b.



Example #14 continued

c) Julie doesn't think this method accurately predicts how well she did on the final exam. Calculate r^2 and use the value you get to argue that her actual score could have been much higher or much lower than the predicted value.

Example #15

One vodka manufacturing company seeks to predict a sweetness index (y) in their new cranberry vodka drink based on the amount of sugar (x) in the juice. Based on a sample of n=8 manufactured drink batches, pairs of these quantities were recorded. We are told that:

$$\sum_{i=1}^n X_i = 45.8 \quad \sum_{i=1}^n Y_i = 18.6 \quad \sum_{i=1}^n X_i Y_i = 106.5 \quad \sum_{i=1}^n X_i^2 = 262.74 \quad \sum_{i=1}^n Y_i^2 = 43.56$$

a) Find the slope and intercept of this best line.

b) Predict the sweetness index y when the sugar is measured to be 5.5

c) What is the correlation between x and y based on these data pairs?

d) Find the value of R^2 and interpret it.

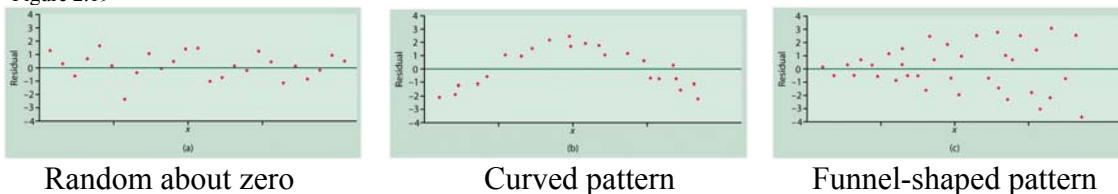
RESIDUALS

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line.

$$\text{Residual} = \text{Observed } y - \text{Predicted } y = y - \hat{y}$$

We can assess the fit of a regression line (how well our LSR line describes the true pattern between x and y) by looking at **scatterplots** of our residuals. If our LSR line is a good fit, then we expect to see **NO pattern** in our scatterplot, the residuals should be scattered at random around zero. Any residual plot which has a pattern, like a curve or funnel shape, signifies that our LSR line is not an appropriate description of the relationship between our response and explanatory variables.

Figure 2.19



An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the y direction of a scatterplot have large regression residuals. An observation is **influential** for a statistical calculation if removing it from our data set would have a big impact on the result of the calculation. Points that are outliers in the x direction of a scatterplot are often influential for the LSR line.