STA 4211 Design and Analysis of Experiments

Deborah Burr

January 9, 2023

What is STA 4211 about?

Design of experiments is an important topic.

We will study:

- Key features common to many designs, and
- Specific, common experimental designs that illustrate these features

Important Questions Can we infer causation from a given design? If yes, how and to what degree can we infer causation?

Let's consider an example.

What is STA 4211 about? (con.)

Note: We will consider situations with continuous, or at least quantitative, response variable Y.

Two major topics:

- Design
- Analysis
  - "By hand" when possible; for balanced designs (as seen in Statistics I)
  - Via the linear model (as in STA 4210 Regression Analysis); with categorical predictors
  - Using R. Why use of code is important.

Additional comments:

- We start with two building block designs and cover them thoroughly.
- Standard inference, plus multiple comparisons.

Ch. 15 Example Effect of Vitamin C on Number of Colds

In the 1970s there was much interest in effect of vitamin supplements on health. In particular, it was believed that taking vitamin C could help prevent colds.

Questions: Does taking vitamin C reduce number of colds, and if so, by how many? Get data to find out.

Method 1: Select 400 individuals from the set of people who take vitamin C every day, and select 400 individuals from set of people who don't. Count the number of colds for each of the 800 people, and compare the numbers of colds in each group.

Method 2: Take a sample of 800 volunteers. Half, selected randomly, are to take vitamin C for the study period, the other half a placebo. At the end of the study, count the number of colds for each of the 800 people, and compare the numbers of colds in each group.

Example. Experiment on Vitamin C and Colds

Experiment conducted in 1976, with 868 children participating. Half of them were randomly selected for the experimental group; these 434 children received a 1000-mg tablet of vitamin C daily for the test period. The remaining 434 children, the control group, received an identical tablet containing no vitamin C every day.

Results:

Group	Mean	n
Vitamin C	$.38$	434
Placebo	$.37$	434

The difference between the two groups ($.01$ colds per child) was not statistically significant.

Consider the following aspects of design:

- Control
- Randomization
- Blinding (single, double)

Features of experiments

Ex. Vitamin C and Colds

1. This is a *controlled experiment* because the experimenter decides who gets the treatment, and who gets the placebo.
2. It is also *randomized*, because those who receive Vit C are selected at random from the whole group.
3. It is *blinded*, because a placebo pill is given. Thus the subjects don't know if they received Vitamin C or not. Important in experiments with human subjects where there may be a psychological effect of the treatment; blinding eliminates this bias.
4. The experiment might have been double-blinded. We aren't given enough information to tell for sure. We'd have to know the protocol for diagnosing the colds. If the subjects themselves report how many colds they had during the study period (self-report), then since the subjects are blinded already, this means the study is double-blinded.

We will discuss *statistical significance* later.

Outline of the course

- Some review of Statistics I
- Single-factor studies; completely randomized design (Ch. 16)
- Follow-up analysis of factor level means in single-factor studies (Ch. 17)
- Diagnostics for one-way ANOVA; Section 18.1
- Balanced two-way factorial experiments; two-factor studies; Ch. 19
- Randomized blocks experiment and analysis; Ch. 21
- Analysis of covariance; Ch. 22 (if possible)
- Two-factor studies with unequal sample size (unbalanced); Ch. 23

REVIEW OF STATISTICS I

We will consider some topics from Stats I that are especially important building blocks for regression and Anova.

Three main topics or units of study in Statistics I (2023, 3032, 4321/4322): Data description, Probability, Inference

List of key Statistics I topics

- Features of designs: control, randomization
- Data description: histogram, mean, standard deviation, scatterplot, correlation
- Probability: random variable (r.v.) and its expected value, variance and standard deviation (SD); expected value, variance and SD of a linear combination of r.v.'s; normal distribution; t distribution
- Inference: concepts of confidence interval and hypothesis test; t procedures for two samples; F tests

Let's begin the review of Statistics I by considering the types of studies covered in that course.

Settings for Inference (Statistics I Review)

1. one sample, inference about the mean μ
2. one sample, inference about the proportion p
3. two independent groups, inference about the difference of two means
4. two independent groups, inference about the difference of two proportions
5. inference about the mean of paired differences

Diagram of study features

Role of statistical significance

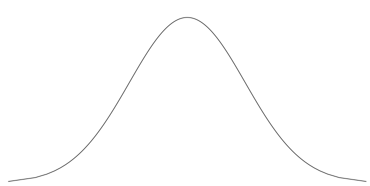
In a controlled, randomized experiment, if the observed difference is found to be statistically significant, then we can conclude that the difference is real, and that the treatment caused the difference.

Reason:

The Normal Distribution

The probability distribution function (more accurately called the *probability density function*, or *pdf*) of the normal distribution is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



We will talk about the *family of normal curves*. There are infinitely many normal curves. You specify one by specifying μ and σ.

To be a pdf, the function $f(x)$ must satisfy two requirements:

- $f(x) > 0$, for all x, and
- The area under the whole curve must be equal to 1.

Therefore (look at the formula), the parameter μ can take on any value on the real line; the parameter σ must be positive.

The normal pdf:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ and σ are parameters of this distribution. Effect on curve if μ increases? Effect on curve if σ decreases?

FACT: The parameter μ is the mean of the distribution; the parameter σ is the standard deviation of the distribution.

Height of curve at $x = \mu$ equals what? This is the maximum height of the curve. The curve is symmetric about μ.

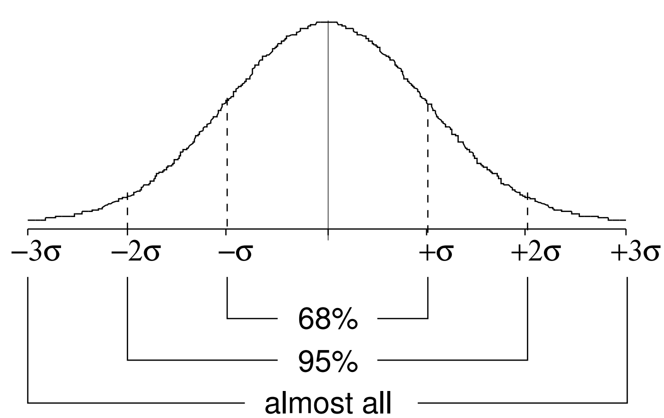
We will write $X \sim \mathcal{N}(\mu, \sigma)$ as shorthand for "the distribution of X is normal, with mean μ and SD σ."

Most important case: $\mu = 0$, $\sigma = 1$.

The curve that is shown is standard normal. Sketch curves for mean 0, sd $.5$; mean 0, sd 2; mean 2, sd 1.

Basic Properties of the Normal Distribution

1. Curve is symmetric, centered at the mean μ.
2. 50% of area lies to right of μ.
3. The SD σ measures the spread: The bigger the value of σ, the more spread out and flatter the curve.
4. Area under the curve is always 1.
5. (a) 68% of area is within σ of μ,
   (b) 95% of area is within 2σ of μ,
   (c) $\sim 99.7\%$ of area is within 3σ of μ.



Probabilities involving the normal distribution can be computed either using familiar tables, or with R or another computer program.

Suppose X is an observation from a population with mean μ and SD σ. The standard score of X, usually denoted Z, is the number of SD's above $(+)$ or below $(-)$ the mean X is. Formula is

$$Z = \frac{X - \mu}{\sigma}.$$

The standard score indicates the relative standing of X in the population.

Calculating the standard score is a way of getting a common scale for different measurements which are approximately normally distributed.

Example. Suppose that a course has two midterms, for which the scores are approximately normally distributed, with means and SD's given below:

	Midterm 1	Midterm 2
Class Avg	55	50
Class SD	14	10
You get	76	67

On which test did you do better relative to the rest of the class?

Expected Value and Variance: Discrete Case

Expected Value: If X is a discrete random variable, then the *expected value of* X, denoted $E(X)$, is the weighted average of all the possible values of X, where the weights are given by the probabilities.

In symbols, if the possible values are $x_1, x_2, x_3, \ldots$, and the probabilities are $p_1, p_2, p_3, \ldots$, then

$$E(X) = \sum x_i p_i.$$

Variance: If X is a discrete random variable then the *variance of* X, denoted $\mathrm{Var}(X)$, is the weighted average of the squared distance between the possible values of X and the expected value of X, where the weights are given by the probabilities.

In symbols, if the possible values are $x_1, x_2, x_3, \ldots$, the probabilities are $p_1, p_2, p_3, \ldots$, and the expected value is μ, then

$$\mathrm{Var}(X) = \sum (x_i - \mu)^2 p_i \qquad \big[\text{in other words, } \mathrm{Var}(X) = E\big((X-\mu)^2\big)\big].$$

Standard Deviation: For a random variable X with variance $\mathrm{Var}(X)$, the SD or standard deviation of X is $\sqrt{\mathrm{Var}(X)}$

Expected Value and Variance: Continuous Case

For continuous r.v. X, $E(X)$ and $\mathrm{Var}(X)$ are defined similarly to the discrete case, but require calculus for the actual formula.

Normal r.v., $X \sim N(\mu, \sigma)$, we have $EX = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

For normal r.v.s, there is an important interpretation of the standard deviation (see p. 14).

Rules for Expected Value and Variance of Linear Combinations

Suppose $X_1, X_2, \ldots, X_n$ are random variables with means $\mu_1, \mu_2, \ldots, \mu_n$, respectively, and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, respectively. Suppose $a_1, a_2, \ldots, a_n$ and c are constants. Then:

Whether or not the X's are independent,

$$\begin{aligned} E(a_1X_1 + \cdots + a_nX_n + c) &= a_1E(X_1) + \cdots + a_nE(X_n) + c \\ &= a_1\mu_1 + \cdots + a_n\mu_n + c \\ &= \Sigma_{i=1}^n a_i\mu_i + c \end{aligned}$$

If the X's are independent,

$$\begin{aligned} \text{Var}(a_1X_1 + \cdots + a_nX_n + c) &= a_1^2\text{Var}(X_1) + \cdots + a_n^2\text{Var}(X_n) \\ &= a_1^2\sigma_1^2 + \cdots + a_n^2\sigma_n^2 \\ &= \Sigma_{i=1}^n a_i^2\sigma_i^2 \end{aligned}$$

Question:

Call the true period of a pendulum μ. Suppose there are two measurement techniques (two different clocks) for measuring the period. Call the measurement for clock 1 X_1, and the mst. for clock 2 X_2. Suppose that both methods are unbiased. That is, $E(X_1) = \mu$, $E(X_2) = \mu$. However, the SDs are different. $SD(X_1) = .2$ and $SD(X_2) = .1$. Find the mean and SD of the weighted average, $.25X_1 + .75X_2$.

Application of Rules to Find $E(\bar{X})$ & Var$(\bar{X})$

Suppose that $X_1, X_2, \ldots, X_n$ are a random sample from a distribution with mean μ and standard deviation σ. Then

$E(\bar{X}) =$

Var$(\bar{X}) =$

Distribution of the sample mean $\bar{X}$

Suppose we draw a random sample of size n from a population with expected value μ and standard deviation σ.

Let $\bar{X}$ be the mean of the *sample*. Then:

1. The mean of the distribution of $\bar{X}$ is equal to μ.
2. The standard deviation of the distribution of $\bar{X}$ is equal to $\sigma/\sqrt{n}$.
3. If n is large, then the distribution of $\bar{X}$ is approximately normal (Central Limit Theorem), and if the distribution of the X's is normal, the distribution of $\bar{X}$ is exactly normal for any sample size n.
   In either case, we use the following as if it were exactly true:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

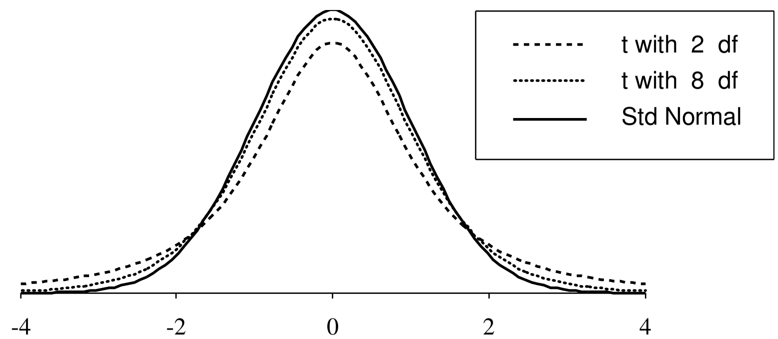
The t distribution

When we substitute the sample standard deviation S for σ, the distribution of $\frac{\bar{Y}-\mu}{S/\sqrt{n}}$ is no longer a normal distribution.

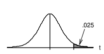
Fact If we have a random sample of size n from a normal distribution with mean μ and standard deviation σ, then

$$\frac{\bar{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1} \quad \text{(“}t\text{ with } n-1 \text{ degrees of freedom”)}$$

The *t* Distribution



df $= 5$ $\quad t_{.025} = 2.571$
df $= 10$ $\quad t_{.025} = 2.228$
df $= 20$ $\quad t_{.025} = 2.086$
df $= 30$ $\quad t_{.025} = 2.042$
df $= \infty$ $\quad t_{.025} = 1.96$



$0 \qquad t_{.025}$

Note: If df $= \infty$, then $t_\alpha = z_\alpha$.

df	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
∞	1.282	1.645	1.960	2.326	2.576

Example: You want to see how much you weigh. Your bathroom has a scale which gives readings which are unbiased (have mean equal to the true weight of the object on the scale) and unknown SD. You weigh yourself three times, getting $149, 151, 150$. What is a 95% CI for your true weight?

Comparison of means from two normal samples (any sample sizes)

Basic Framework

Assume:

1. The Y's are a random sample of size n_Y from a normal population with mean μ_Y and SD σ_Y, both unknown; $\bar{Y}$ is the sample mean and S_Y is the sample SD.
2. The Z's are a random sample of size n_Z from a normal population with mean μ_Z and SD σ_Z, both unknown; $\bar{Z}$ is the sample mean and S_Z is the sample SD.
3. The two samples are independent of one another.
4. The two standard deviations are equal; that is, $\sigma_Y = \sigma_Z$.

Want:

1. A confidence interval for $\mu_Y - \mu_Z$.
2. Hypothesis tests concerning $\mu_Y - \mu_Z$.

We want a **confidence interval** (CI) for $\mu_Y - \mu_Z$. This is an interval estimate of a population parameter. Width of interval is determined by: variance of observations in both groups, sample sizes, confidence level.

True or false: *The higher the confidence level, the narrower the confidence interval.*

Also consider how variance and sample size affect width of the CI.

We also want **hypothesis tests** about $\mu_Y - \mu_Z$. We can test for any value of the parameter $\mu_Y - \mu_Z$, but we almost always test whether it is zero.

Three forms of hypothesis test about $\mu_Y - \mu_Z$:

- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z > 0$
- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z < 0$
- H_0: $\mu_Y - \mu_Z = 0$ vs. H_a: $\mu_Y - \mu_Z \neq 0$

The standard two-sample t procedures: Theory

Basic Facts

1. $E(\bar{Y} - \bar{Z}) = \mu_Y - \mu_Z$
2. $\text{SD}(\bar{Y} - \bar{Z}) = \sigma\sqrt{\dfrac{1}{n_Y} + \dfrac{1}{n_Z}}$
3. $\bar{Y} - \bar{Z}$ has a normal distribution.
4. $\dfrac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\sigma\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim \mathcal{N}(0, 1)$.
5. $\dfrac{(\bar{Y} - \bar{Z}) - (\mu_Y - \mu_Z)}{\hat{\sigma}\sqrt{\frac{1}{n_Y} + \frac{1}{n_Z}}} \sim t_{n_Y + n_Z - 2}$,

where $\hat{\sigma}^2 = \dfrac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}$

Note the general structure of the quantity in Basic Fact 5:

$$\frac{\text{Estimate} \ - \text{Parameter}}{\text{St. error of Estimate}}$$

This is the same structure as for the one-sample t procedures, so the CI for $\mu_Y - \mu_Z$ will have the same structure as before, also:

The standard two-sample t-test:

Form

$$T = \frac{\bar{Y} - \bar{Z}}{\sqrt{\left(\frac{1}{n_Y} + \frac{1}{n_Z}\right)\frac{(n_Y - 1)S_Y^2 + (n_Z - 1)S_Z^2}{(n_Y + n_Z - 2)}}} .$$

If the null hypothesis is true, and the two population variances are equal, then the distribution of this is t with $n_Y + n_Z - 2$ df.

Reminder: The standard two-sample t-test assumes that the two SD's are equal, and goes on to exploit that feature.

Applications of the two sample T procedures

The area of A/B testing (e.g. at Google, LinkedIn, Microsoft, etc) consists mainly of two-sample t tests and confidence intervals.

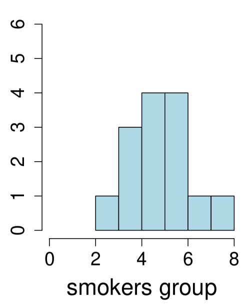
Example: A physician named A.W. Andrews developed way of scoring a photograph of a face for wrinkles. The scores range from 1 to 10, with 1 indicating no wrinkles, 10 indicating a severe case.

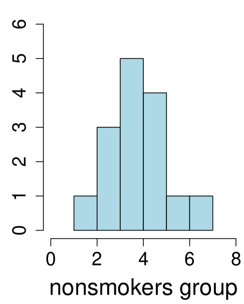
He photographed 29 women aged 45–55, of whom 14 smoked a pack a day and 15 did not smoke. He reported that the 14 who smoked can be considered a random sample from the population of women who smoke a pack a day, and the 15 who didn't smoke can be considered a random sample from the population of women who don't smoke.[1]

Scores:

S: $2, 6, 4, 6, 4, 5, 5, 6, 5, 6, 7, 5, 4, 8$ $\quad n_s = 14 \quad \bar{Y}_s = 5.2 \quad S_s = 1.48$

NS: $1, 5, 3, 5, 3, 4, 4, 5, 4, 5, 6, 4, 3, 7, 4$ $\quad n_{ns} = 15 \quad \bar{Y}_{ns} = 4.2 \quad S_{ns} = 1.42$





```
Welch Modified Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=26.69, p-value=0.0711
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09,  2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20
```

```
Standard Two Sample t-test

data: smokers and nonsmokers
t=1.88, df=27, p-value=0.0706
alternative hyp.: true difference in means is not 0
95 percent confidence interval: (-0.09, 2.12)
sample estimates:  mean of y: 5.21  mean of z: 4.20
```


The Standard Two Sample t-test is the one based on the assumption of equal variances (assumption 4).

The Welch-corrected procedure does not assume equal variances. Current wisdom is to use the Welch-corrected test, to be safe. The standard two sample t-test can be a lot worse if the two variances are not equal.

The two tests are the same in structure; the difference is in how the standard error in the denominator is estimated.

The Standard test is more similar to what we do in regression analysis, and that is why we focus on that one here.

Comments on the Analysis

- Conclusion:

- The CI is $(-0.09, 2.12)$, which means that we are "95% confident" that $-0.09 < \mu_{\text{s}} - \mu_{\text{ns}} < 2.12$. This interval contains 0.

- Need to do an informal check of normality. You do this by looking at the histograms. The assumption of normality is not really critical, and unless there is something glaring, you usually don't worry about it.

- Side note on Welch's method. The number of df is not necessarily an integer. Actually, there are t-distributions with any number of df. They are not tabulated, but this is not a problem, because you will always do this using a software package anyway.