A Companion to The Analysis of Biological Data

Nathan Brouwer

2018-06-22

Contents

Introduction

This is a reading guide to the excellent introductory biostatistics book *The Analysis of Biological Data, 2nd edition* by Whitlock and Schluter. It consists of annotated outlines of each chapter where I have highlighted the portions of the text I focus on when I teach biostatistics. A primary goal is to indicate to my students the text, equations, and vocabulary that are most closely related to what is covered in the lecture, and which examples are most most complimentary to what we cover in lecture and lab.

Chapter 1

Statistics and Samples

1.1 1.1 What is statistics?

Vocab:

Estimation: The process of inferring an unknown quantity of a population using sample data. (pg 2)

Parameters: a quantity describing a population

Estimate: a quantity calcualted from a sample

1.2 1.2 Sampling populations

The concept of a statistical population and a biological population are different. In statistics, it has a very very specific definition.

1.2.1 Example 1.2: Raining cats

Very important idea. See page 9 for definition of **sample of convenience**. An ecological example would be to use road-killed deer as a sample for estimating the size of deer horns.

1.2.2 Population and samples (pg 4)

Vocab:

Population: All individuals of interest

Sample: Subset of units taken from the population.

1.2.3   Properties of good samples (pg 5)

Vocab:

Sampling error

Precision

Accurate (unbiased)

Bias

Figure 1.2-2 :The classic precision/bias bulls eye figure.

1.2.4   Random sampling

- Can't emphasize this enough!
- You must be able to distinguish a random from a non-random sample
- You need to be able to determine if the study units are independent from each other or not

Vocab

Random sample: "In a random sample, each member of a population has an equal and independent chance of being selected." (pg 7) Random means random. If a random number generator wasn't involved in picking it, then it wasn't random!

Equal chance: Equal means equal. Say I want to randomly sample all the students on campus to determine the mean age of CalU students. I walk through the middle of campus and flip a coin each time I see someone. If I get heads, I ask them their age; tails, I keep walking. This involves randomness, *BUT* if I am interested in all the students at CalU then this sample is biased. All student do not have an *"equal chance"* of being selected - I am only sampling students who are present on campus when I am collecting my data.

1.2.5   How to take a random sample

- Very important section

1.2.6   Figure 1.2-3

- Very important figure!

1.2.7   Sample of convience (pg 9)

1.2.8   Volunteer bias (pg 10)

Vocab: **Sampline of convience Volunteer Bias**

1.2.9   Data in the real world

1.3   1.3 Types of data an variables (pg 11)

- Based on answers on the homework, people seem to have trouble with this topic. Doing practice problems can help.

Vocab: **Variable:** The characteristics being measured: length, height, weight, color, sex, etc. **Data:** ** **The actual measurements** Categorical variable Numeric variable**

1.3.1 Explantory and response variables

- People seem to have a very hard time with this.
- In R we usually express things as "response ~ predictor"
- This reads as "the response depends on the predictor", or, the "response is caused by the predictor."
- We almost always plot things with the response on the y axis and predictor on the x.
- So, when trying to determine response vs. predictor, ask, "how would I plot this?" Even better - draw a little graph and see if it makes sense.

1.4 1.4 Frequency distributions and probablity distributions (pg 13)

Vocab **Frequency Frequency distribution**

[we will talk about the normal distribution after the 1st test]

1.5 1.5 Types of studies

- People seem to struggle with this.
- "Observational study" doesn't mean "something was observed"; something always gets observed in a study or experiment however its done.

- We use the word "observational" to indicate that *only* observations were made - the investigator did not do anything experimental; they did not manipulate, interfere, meddle, mess with, impact or change anything in any way.

- In an observational study, all investigator does is observe the state of nature as it already was before they showed up; if the investigator was not there, things would be (largely) the same (the presence of an observer, however, can impact how things behave).
- In contrast, in an experiment, the investigator is actively and intentionally changing things. In the investigator wasn't present, things would be completely different; in fact, often the investigator sets up everything that is being observed from scratch so without them there would be nothing going on at all.

Vocab

Experimental study A study where the research "assigns different treatment groups or values of an explanatory variable randomly to the individuals units of study" (pg 15). I often call this a "manipulative experiment" b/c the researcher is intentionally manipulating nature in order to answer the question they are interested in.

Observational study A study where the researcher has "no control over which units fall into which categories." Observational studies are sometimes called "natural experiments" when nature has created experiment-like conditions. A study documenting the reproductive rates of herons in an urban wetland would be purely observational. A study comparing reproduction in urban wetlands with high levels of pollutants vs. rural wetlands would be a type of natural experiment.

1.6   1.6 Summary

Read it

1.7   Practice Problems

Do some! (or lots!)

Chapter 2

2 Displaying Data

Displaying data well is crucial in science, but easy to do poorly. An increasing number of people are specializing in the production of "infographics" for displaying complex quantitative information in an effective manner.

2.1   2.1 Guidelines for effective graphs

2.1.1   How to draw a bad graph

Figure 2.1-1 This is a bad graph. 3D plots are popular b/c they are easy to make in Excel but they should be avoided.

2.1.2   How to draw a good graph

The book's rules for good plots

1. "Show the data": this often means showing the raw data, or using a boxplot or histogram to reveal information about the distribution of the data.
2. "Make patterns in the data easy to see": Plot raw data if possible; use color and shapes for different symbols. Use "jittering" so that data points don't overlap.
3. "Represent magnitudes honestly"
4. "Draw graphical elements clearly": make the font large enough to read, clearly label the axes, add a legend; make it color-blind friendly.

My rules are similar, just with some further emphasis on certain things:

1. Show the **raw data** if possible
2. Show **distributional info** if possible
3. ALWAYS Include **error bars** around means
4. ALWAYS Include error bars around means

5. Make patterns in the data easy to see
6. Represent magnitude honestly
7. Draw graphical elements clearly
8. **Include a legend** and label things clearly

By "distributional information" I mean info on the spread, shape, density, variation etc in the data. This is best done with a boxplot or a histogram, or the raw data.

Figure 2.1-2 The graph on the left illustrates the principal of "showing the data", while the one of the right hides the data. If there were more datapoints, a boxplot would be a good choice instead of showing every data point.

Figure 2.1-3 The upper part of the graph is bad for 2 reason. 1st, is unnecessarily uses 3D, which add no information and makes it harder to see exactly how high each bar is. 2nd, the y axis starts at a high number, which accentuates the difference between the far left and far right bars.

2.2   2.2 Showing data for one variable

Note: in 2017 I DID NOT emphasize the vocab terms on page 30

2.2.1   Showing numerical data: frequency table & histogram (pg 33)

I emphasized histograms. I love histograms.

The authors mention that "the shape of a frequency distribution is more obvious in a histogram" than a table with the raw data (page 35). I similarly emphasized that when possible, avoid tables and use graphs.

2.2.2   Describing the shape of a histogram (pg 36)

I did not emphasize this in class but it is very very important. Please be familiar with these terms.

Vocab: **Symmetry** "A frequency distribution is symmetric if the pattern of" data on the "left half of the histogram is the mirror image" more or less "of the pattern on the right half" (pg 36) **Skew**: "asymmetry in the shape of a frequency distribution." **Outliers** Extreme values "well outside the range of values of other observations in the dataset"

[In 2017 I did not emphasize the section "How to draw a good histogram"]

2.3   2.3 Showing association between two variables

[in 2017 I didn't emphasize Showing association between categorical variables, grouped bar plots, or mosaic plots]

2.3.1   Showing association between numerical variables: scatter plot (pg 42)

Scatter plots are very important.

2.3.2 Showing association between a numerical and a categorical variable

These plots are very important, especially boxplots.

Compare **Figure 2.3-4** and **Figure 2.3-5** to see how information can be presented as a boxplot (3-4) or histograms (3-5).

2.4 2.4 Showing trends in time and space

2.5 Line graph

As in **Figure2.4-1**, line graphs are frequently used to show how things change over time, such as populations or rates of infection.

Vocab

2.6 2.5 Howw to make good tables

This is an important section but I did not emphasize it in 2017, except to say that the same principals that apply to plots also apply to tables. Moreover, its better to use plots than tables.

2.7 2.6 Summary

Line graph

Chapter 3

Describing Data

Study Guide: Test 1 - Chapter 3 Analysis of Biological Data 2nd ed.

3.1 3.1 Arithemetic mean and standard deviation

3.1.1 The sample mean

We call the **sample mean** y.bar (In contrast in chapter 4 we talk about "mu", the population mean.)

3.1.2 Variance and standard deviation

I won't ask you to write these equations from memory, but I will expect you to know the difference between the equation for the variance vs. the standard deviation, what the numerator and denominator mean, etc. For example, the square root of the variance is the standard deviation; the numerator of the variance is the "sum of square"; the "n" in the denominator of these equations is the sample size.

Vocab

Variance: A measure of spread / variability of the data; the precursor to the standard deviation. The numerator is the sum of squares. The n in denominator is the sample size. You never plot the SD on a graph.

$$s^2 = \frac{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n - 1}$$

Standard deviation "A common measure of the spread of a distribution. It indicates just how different measurements typically are from the mean." While very useful, we rarely plot the SD on a graph.

$$s = \sqrt{\frac{\sum\limits_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}{n-1}}$$

Sum of square: The numerator in the equations for the variance (s^2; var) and standard deviation (s; SD.

3.2   Rounding means, standard deviatinos, and other quantities

A good little section on how to round numbers

3.2.1   Coefficient of variation

[I did not cover this in 2017]

3.2.2   Calculating mean and SD from a frequency table

[skipped in 2017]

3.2.3   Effect of changing measurement scale

[skip]

3.3   Median and interquartile range (pg 73)

Vocab:

Median

Interquartile range (IQR): "The difference between the 3rd and 1st quartiles of the data. It is the span of the middle 50% of the data.

First quartile:

Third quartile:

Boxplot

Note that the "whiskers" extending above and below the bar plot are NOT error bars! They represent the range.

3.4   How measures of location and spread compare

This is a good section for understanding the difference between the mean and median.

"Median and mean measure different aspects of the location of a distribution. The median is the middle value of the data, whereas the mean is its center of gravity." (pg 80). If the distribution is perfectly **symmetrical** the median = mean. The more **skewed** a distribution is, the further the median is from the mean.

3.4.1 Displaying cumulative relative frequencies

[Did not emphasize in 2017]

3.5 3.5 Proportion

[Did not emphasize in 2017]

3.6 3.6 Summary

Their summaries are always good. An especially important point is the 4th one.

Chapter 4

Chapter 4 Estimating with Uncertainty

4.1   4.1 The sampling distribution of an estimate

Vocab: **Estimation:** "The process of inferring a population parameter [eg "mu"] from sample data" (pg 96)"

Sampling distribution: "the probability distribution of all the values for an estimate that we *might* have obtained when we sampled the population" (pg 96)

4.1.1   Example 4.1 The Length of Human genes

This example is central to the chapter and was the focus of several of my lectures.

Figure 4.1-1 The empirical (real) distribution of lengths all genes in the human genome. This is **NOT** a **sampling distribution.** A sampling distribution is a distribution of possible values of y.bar, the mean from a sample.

4.2   Estimating mean gene length with a random sample

Figure 4.1-2 The distribution of a random sample of 100 genes. This is again **NOT** a sampling distribution, but the distribution of a single random sample taken from the whole population of genes. What they are plotting here is the length of genes. Note the X-axis of the histogram says "gene length." Compare this with Figure 4.1-3.

4.3   The sampling distribution of Y.bar

Vocab: **Sampling distribution:** "The probability distribution of all values for an estimate that we might obtain when we sample a population" (pg 99). That is, a theoretical distribution of many replications of the same study or experiment.

Figure 4.1-3 This **IS** a **sampling distribution.** What they are plotting here is the mean length of genes from many replicated studies. Note what how the x-axis is labeled: "Sample mean length Y.bar"

Figure 4.1-4 These are **sampling distributions.** Note the x-axis. How the shape of the sampling distribution changes as sample size increases is very important. A wide sampling distribution means you could get and estimate (Y.bar) that is not very accurate. A narrow sampling distribution means your Y.bar is likely to be pretty close to the truth.

Key Point: "Increasing sampling size reduces the spread of the sampling distribution of an estimate, increasing precision" (pg 100)

4.4   4.2 Measuring the uncertainty of an estiamte

4.4.1   Standard error (SE)

Vocab **Standard error (SE)** A general definition:"The standard error (SE) of an estimate [Y.bar] is the standard deviation (SD) of the estimate's sampling distribution." (pg 101). Standard errors come in different flavors depending and exactly the kind of data your are working with. This is a generic definition that applies to all SEs.

4.5   The Standard error of Y.bar

4.6   The standard error of Y.bar from data

Like most stats books, they give you 2 equations, one assuming you somehow know the "True" standard deviation of the population (the "population standard deviation of the variable Y.bar", page 101, middle of page) and one equation for the real situation where you have calculate the SD from data. They really should emphasize that "population standard deviation of the variable Y.bar" is never ever known, except in the fanciful world of statisticians brains.

Vocab: **Standard error of the mean** (precise definition for y.bar; this applies to many of the circumstances where you'll want to calculate an SE): "The standard error of the mean is estimate from data as the sample standard deviation (S) divided by the square root of the sample size (n)" (pg 102).

The equation for the standard error of a mean (y.bar) is:

$$SE = \frac{SD}{\sqrt{N}}$$

I won't ask you a question like "write out the equation for the standard error" but I could ask a question like "what is the numerator in this equation" and show you the equation for the SE.

A very important point is that the SE depending on the sample size. The larger the sample size, the smaller the sample size. A large sample size improve precision of the estimate of y.hat and is reflected in a small SE and narrow error bars around the estimate.

4.7 4.3 Confidence Interval (pg 102)

This is a very very important section!

Vocab **Confidence Interval:** "A confidence interval is a range of values surrounding the sample estimate that is likely to contain the population parameter [mu]" (pg 102)

95% confidence Interval They say "The 95% Confidence interval provides a most-plausible range for a parameter. Values lying within the interval or more plausible, whereas those outside are less plausible, based on the data." 95% is a convention that most scientists stick to and relates to the typical p-value of 0.05 (1-0.05 = 0.95.)

4.8 The 2SE Rule of them

Vocab **2SE Rule of Thumb** "A rough approximately of a 95% confidence interval for a mean can be calculated as the sample mean [y.hat] plus and minus 2 standard errors." (page 104) More precisely, 1.96.

4.9 4.4 Error bars

Always put error bars around mean values! IF you report a mean, report the precision of that mean, which requires report the SE or 95% CI and plotting error bars. 95% CI is preferable to SE!

Vocab

Error bars: "Lines on a a graph extending outward from the sample estimate [usually Y.bar] to illustrate uncertainty about the value of the parameter being estimated" (pg 105). Can be an SE or 95%CI. 95%CI are recommended. Very rarely is the SD.

Figure 4.4-2: Great figure demonstrating the difference between the raw data, which has a huge range, the smallness of the SE, the 95%CI, and the SD.

Chapter 5

CHAPTER 6: HYPOTHESIS TESTING

Reading Guide
Chapter 6: Estimating with Uncertainty
Whitlock, MC and D Schluter. 2015. The Analysis of Biological Data, 2nd ed. Roberts & Company, Greenwood Village, Colorado.
Book website: http://whitlockschluter.zoology.ubc.ca/
Publisher: https://goo.gl/mSVUHe
Notes compiled by Nathan Brouwer (brouwern@gmail.com)

A key phrase: "Estimation asks, 'how large is the effect?' Hypothesis testing asks, 'Is there any effect at all.'" (pg 149). This relates to the fact that the alternative hypothesis (Ha) is very general.

Another key prhase: "Hypothesis testing quantifieds how unusual the data are. assuming that the null hypothesis is true" (pg 150). I refered in lecture to the idea of test statistics and p-values as "indices of surprise", which relates to this idea.

Hypothesis testings "compares data to waht we would expect to see if a specific null hypothesis were true. If the data are not too unusal, comapred to what we woudl expect to see if the null hypothesis were true, then the null hypothesis is rejected." (pg 150)

5.1   6.1 Making and using statistical hypothesis

5.1.1   Null hypothesis

aka Ho

Null hypothesis

5.1.2   Alternative hypothesis

aka Ha. When I read this section for the 1st time I finally realized how vauge Ha is...

Alternative hypothesis "inclues all other feasible values for the population parameter [mu] besides the value stated in the null hypothesis"" [usually 0 or a difference of 0; 50% for probabilites] (pg 153)

5.2   6.2 Hypothesis testing: an example

I went over this example in detail during my lecturees

The four steps of hypothesis testing: you should know each one

1. State….
2. Compute….
3. Determine…
4. Draw the appropriate…

5.2.1   Example 6.2 The right hand of toad

I like this example, and hence used it in class

5.2.2   Stating hypotheses

Note that for probabilities and proportions the null hypothesis (Ho) is usually 50%, 0.5, etc. This is in contrast to t-tests and ANOVA, where the null is that there is no difference between groups (mean.1 - mean.2 = 0)

Two-sided test

5.2.3   Test statistic

Test statistic

Sometiems just called "test-stat." For a proportion, as in the frog example, its just the observed number of events (eg, number of Righted handed toads). For t-tests, ANOVA, regression it requires more math. T-tests use the t-stat, ANOVA uses the F-stats.

5.2.4   The null distribution

This is a tricky but key idea. Hypothesis testing is all about comparing the observed test stat to a **sampling distribution** built by assuming that the null hypothesis is true.

Null distribution "the sampling distribution of outcomes for a test statistic under that assumption that the null hypothesis is true" (pg 155). That is, the distribution of test stats you'd get in a world where the null was indeed true.

5.2.5   Table 6.2-1

I replicated this table in my lecture. You should know how it was made.

5.2.6   Quantifying uncertainty: the P-value

P-value "the probability of obtaining the data (or data showing as great or greate difference from the null hypothesis) if the null were true." (pg 157). Is calcualted from the null distribution for a particular study.

You need to know this defintion, and be able to spot incorrect definitions of p-values.

5.2.7 Figure 6.2-2

I replicated this figure in my lecture. You should know how it was made.

5.2.8 Draw the appropriate conclusion

Significance level alpha (pg 159)

5.2.9 Reporting results

Your authors say you need to report

1. test stats
2. sample size
3. p-value

I also say you should report means, effect sizes, and confidence intervals.

5.3 6.3 Errors in hypothesis testings

5.3.1 Type I and Type II errors

5.3.2 Table 6.3-1

I did not present a table like this but its a standard way to discuss type I vs. type II erors

Type I error

Type II error

Power "the power of a test is the probability that a random sample will elad to a rejection of a false null hypothesis." (pg 161). That is, the ability to reject the null when it actually is false.

5.3.3 Other types of errors

I also discussed the following types of errors

Type III errors rejecting a false null for the wrong reason; getting the right answer for the wrong reason.

Type S errors sign errors

Type M errors magnitude errors

5.4 6.4 When the null hypothesis is not rejected

5.4.1 Example 6.4: The genetics of mirror-image flowers

An example of getting a high p value.

5.4.2   Figure 6.4-1

You should understand how this figure is derived

5.4.3   Interpreting a non-significant results.

You should know whether you can or can't "prove" a null hypothesis.

5.5   6.5 One-sided tests

You just need to know that you shouldn't use them.

5.6   Hypothesis testing versus confidence intervals

You should know how they are related, but why CIs provide more information

5.7   6.7 Summary

Chapter 6

Contigency Analysis: Associations between categorical variables

I don't emphasize this chapter because a) contingency tables are usually covered in intro stats courses and b) I want to get to **logistic regression** by the end of the course, which represents a more contemporary approach to the analysis of categorical data. There are several general aspects of the analysis of categorical data using contingency tables and the chi^2 test that I do want students to know. I highlight the relevant sections from the book below. Overall I want you to know what a contginecy table is, how to identify the response and predictors with this type of data, and that chi^2 tests and logistic regression are used to analyze this type of data. I did not discuss any of the calcualtions related to chi^2 tests, or even discuss how to do it in R. Because of this, I don't indicate which sections can be skipped - unless listed below, a section can be skipped.

6.1   Examples of categorical resposne variables

The authors list several types of categorical data on page 235 - these are good examples to understand.

Contingency analysis "estimates and tests for an association between two or more categorical variables" (pg 236).

In R-like code, this would be categorical.response ~ categorical.predictor

6.1.1   Example 9.3: Your litter box and your brain

You should be able to recognize this type of data as being categorical data which can be analyized using a contingency table or chi^2 test (or also logistic regression).

6.1.2   Table 9.3-1

You should be able to recognize this type of data as being categorical data which can be analyzed using a contingency table or chi^2 test (or also logistic regression).

6.2   9.4 The Chi^2 contingency test

Also just called the chi^2 test. You should read this section from page 246-247.

6.2.1   Example 9.4: The gnarly worm gets the bird

You should be able to recognize this type of data as being categorical data which can be analyzed using a contingency table or chi^2 test (or also logistic regression). I discussed this example in lecture.

6.2.2   Hypotheses

I did not phrase the hypothesis like this; I said soemthing more like

- Ho: parasite infection is not related to being eaten
- Ha: paraiste infection is related to being eaten

Instead o "is related to" one could say "predicts" , eg, "parasite infection predicts whether a fish is eaten or not""

6.2.3   Example 9.5: The feeding habits of vampire bats

You should be able to recognize this type of data as being categorical data which can be analyzed using a contingency table or chi^2 test (or also logistic regression).

6.3   Comparing Two Means

Reading Guide
Chapter 12: Comparing Two Means
Whitlock, MC and D Schluter. 2015. The Analysis of Biological Data, 2nd ed. Roberts & Company, Greenwood Village, Colorado.
Book website: http://whitlockschluter.zoology.ubc.ca/
Publisher: https://goo.gl/mSVUHe
Notes compiled by Nathan Brouwer (brouwern@gmail.com)

This chapter introduces paired t-tests and normal t-tests.

6.3.1   12.1 Paried samples versus two independent samples

Need to Know the difference!

6.3.1.1 Figure 12.1-1

Very good figure

Paried design goes with a paired t-test

Two-sample design goes with a 2-sample aka standard t-test.

6.3.2 12.2 Paired comparisons of means

Good examples of usage on pag 329-330.

Key idea: "Paired measurements are converted to a single measurement by taking the difference between them" (pg 330). This is why I say a paired t-test should be called a "1-sample t-test on paired data." A paired t-test is exactly the same as a 1-sample t-test.

If this example was done in R, you should be able to identify and understanding the relevance of the following things: mean of the differences, t and p, and the 95% confidence interval around the mean difference.

6.3.2.1 Estimating mean difference from paired data

6.3.2.2 Example 12.2: So macho it makes you sick?

6.3.2.3 Table 12.2-1

Note how the data is set up and the difference "d" is calculated.

6.3.2.4 Figure 12.2-2

You should know why we'd care about the differences in this situation.

On pages 332-333 Don't worry about teh math, I didn't emaphsize this.

6.3.2.5 Paired t-test

You should know how the hypotheses are formulated (page 333) but don't worry about the math (page 334). eg, You DO NOT need to know how to calculate the standard deviation or standard error for the difference between means. This is provided in R's output

6.3.2.6 Assumptions

Skip this - I will talk about this when I talk about ANOVA assumptions for the final unit of the class

6.3.3 12.3 Two-sample comparison of the means

"2-sample" means two groups, two treatments, two populations, etc.

If this example was done in R, you should be able to identify and know the meaning of the following things from the output of t.test(): t, df, the difference between means, p, the 95% confidence for the differences between means.

6.3.3.1   Exmaple 12.3 Spike or be spiked

6.3.3.2   Confidence interval for the difference between two means

You don't need to know the math, but you should know why we are interested in the difference, and its CI! (So, most of 338 and 339 can be skipped)

6.3.3.3   Hypotheses for Example 12.3

On page 339 they spell out the hypotheses. Check these out.

On page 340 they go back to the math. Don't worry about it

6.3.3.4   Assumptions

We'll cover assumptions along with ANOVA

6.3.3.5   A two-sample t-test when standard devaitions are unequal

We'll cover this along with ANOVA.

6.3.3.6   12.4 Using the correct sampling units

I did not focus on this topic (sampling units) but did use these data in lab as an example for t-tests. The data are also in the *wildlifeR* package.

6.3.3.7   Example 12.4 So long; thanks to all the fish

A good example, used in lab.

6.3.3.8   Figure 12.4-1

You should be able to look a this graph and predict the p-value from the CIs.

6.3.4   12.5 The fallacy of indirect comparisons

6.3.4.1   Example 12.5 Mommy's baby, Daddy's maybe

6.3.4.2   Figure 12.5-1

6.3.5   12.6 Interpreting overlap of confidence intervals

Discussed in lecture as the idea of "inference by eye."

6.3.5.1   Figure 12.6-1

You should be able to predict the outcome of t-test by looking at the CIs.

6.3.6   12.7 Comparing variances -

6.4   Handling Violations of Assumptions

Reading Guide
Chapter 12: Handling Violations of Assumptions
Whitlock, MC and D Schluter. 2015. The Analysis of Biological Data, 2nd ed. Roberts & Company, Greenwood Village, Colorado.
Book website: http://whitlockschluter.zoology.ubc.ca/
Publisher: https://goo.gl/mSVUHe
Notes compiled by Nathan Brouwer (brouwern@gmail.com)

6.5   Introduction: pages 369 - 368

The numbered bullet points 1 to 3 are especially important to know. They were discussed during my lecture on Fri. Nov 4. They are sometimes called "The ANOVA assumptions", "the assumptions of ANOVA", "statistical assumptions", "assumptions of a t-test," etc.

6.6   13.2 When to ignore violations of assumptions

I discussed why you can sometimes ignore violations of these assumptions, and the concepts of "robustness", during my lecture on Fri. Nov 4.

6.6.1   Violations of Normality (pg 375)

Here they mention the mathematical reason why you can get away with ignoring violations of the assumption of normality: the central limit theorem (discussed in chapter 10). You should know that there is indeed a mathematical reason why you can sometimes get away with ignoring the assumption of normaility. Unequal standard deviations (pg 376)

They note that Welch's correction to the t-test largely solves this problem.

6.7   13.3 Data transformations

6.7.1   Log transformation (pg 377)

You should know that this is the most common type of transformation. They note that the natural log (ln) is more common than the base 10 log (log10).

6.7.2   Figure 13.3-1

[Arcsine transformation - SKIP!] This is actually a discredited approach - not recommended.

6.7.3 The square-root transformation (pg 381)

2nd most common type of transformation, with similar effect as log. [Other transformations - SKIP] These are rarely used in practice [Confidence intervals with transformations - SKIP] Important concept, but didn't cover it

[A caveat: Avoid multiple test with transformations - SKIP] Important concept, but didn't cover it

6.8 # 13.4 Nonparameteric alternatives to one-sample and paired t-test

You should know that there are a thing called nonparametric statistical tests that are often used when ANOVA assumptions are not met. These methods typically convert the data into ranks, which allows different analyses to be done but which result in the loss of information. You do not need to know the particular types of tests.

6.8.1 [Sign test - SKIP][Other specific non-parametric methods - skip]

6.9 13.8 Permutation tests

You should know that there are computer-based methods called permutation tests. See the definitions I handed out. The bullet point in the summary section on permutation tests on page 399 is very concise.

6.10 13.9 Summary

This book generally has an excellent summary section. Definitely see the one on permutation tests.

Chapter 7

Experimental Design

Key idea I emphasized: "In an **experimental study**, the researcher assigns treatments to units or subject so the differences in response can be compared. In an **observational study**...nature does the assigning of treatments to subjects. The researcher has no influence over which subjects receive which treatment." (page 423, emphasis mine). As they continue the difference is important b/c only experiments can identify **causal relationships**.

7.1   14.1 Why do experiments?

7.1.1   Confounding variables

Vocab:

Confounding variable: A factor that "masks or distorts the causal relationship between measured variables in a study." (pg 424)

Vocab:

Experimental artifact "a bias in a measurement produced by unintended consequences of experimental procedures" (pg 425)

7.2   14.2 Lessons from clinical trials

Vocab:

Clinical trial: "An experimental study in which 2 or more treatments are applied to human participants." (page 426)

7.2.1   Design components

This section is a good summary of ideas elaborated on subsequently in 14.3 and 14.4

7.3   14.3 How to reduce bias

7.3.1   Simultaneous control group

Vocab:

Control group

Positive control: I don't think the authors use this term explicitly, but they discuss it on page 429 when they they refer to a clinical study that uses the "best existing treatment"

7.3.2   Randomization

Vocab:

Randomization

Completely randomized design

Haphazard

They authors convey a very important point much better than I did in my lecture: "The virtue of randomization is that it **breaks the association** between possible **confounding variables** and the **explanatory variable**, allowing **causal relationships** between **explanatory** and **response variables** to be assessed" (pg 429). This is a hard idea but key. Randomization spreads out the impact of confounding variable evenly in across the treatment so it doesn't contaminate either one. The authors continue "Randomization doesn't eliminate the variation contributed by confounding variables, only their correlation with the treatment. It ensures that variation from confounding variables is spread more evenly between the different treatment groups. If randomization is done properly, any remaining influence of confounding variables occurs by chance alone, which statistical methods can account for." If randomization isn't done, a confounding variable could end up accidentally having more influence on one treatment or the other and bias the results.

7.3.3   Blinding

Blinding "the process of concealing information from participants (sometimes researchers) about which individuals receive which treatment" (pg 431)

Single-blind experiment

double-blind experiment

Triple-blind experiment The authors don't discuss this but I did in class. Triple blinding involves hiding information from the person who does the analysis.

7.3.4   Replication

Vocab:

Replication

Independent

Experimental Unit

Figure 14.4-1 Is a simplistic but good explanation of how to design experiments that have true replication. I used a similar example of plants in one greenhouse vs. several.

The material here is related the "Interleaf" on Pseudoreplication. (Interleaf 2, page 115) which you should read.

[Note: I didn't discuss the version of the standard error equation that appears on the bottom of page 433]

7.3.5 Balance

Vocab:

Balance "In a balanced experimental design, all treatments have equal sample size" (pg 434)

7.3.6 Blocking

Vocab

Blocking "is grouping of experimental units that have similar properties. Within each block, treatments are randomly assigned to experimental units" (pg 435)

Blocks

The term **randomized block design** was not discussed.

Figure 14.4-2 is a good figure of blocking in a plant study.

7.3.7 Example 14.4A

This is an excellent example.

The list of potential types of blocks on page 437 are good.

[I did not discuss the idea of extreme treatments]

7.4 14.5 Experiments with more than on factor

[This section was not discussed

7.5 14.6 What if you can't do experiment

[This section was not discussed]

7.6 14.7 Choosing a sample size

The bottom of page 441 and top of page 442 are some important thoughts on the importance of choosing an appropriate sample size.

7.6.1   Plan for precision

[This section was not discussed]

7.6.2   Plan for power

[This section was not discussed]

7.6.3   Plan for data loss

[This section was not discussed]

7.7   14.8 Summary

As always, the summary section is great.

Chapter 8

Comparing Means of More Than Two Groups

8.1   15.1 The Analysis of Variance

8.1.1   Example 15.1 The knees who say night

You should be able to read an example like this and know that ANOVA needs to be used to analyze the data

8.1.2   Figure 15.1-1

You should be able to look at graph like this and know that ANOVA needs to be used to analyze the data

8.1.3   Hypotheses

You shoudl be able to relate the hypotheses in terms of mu and know what the null hypothesis and alt. hypothesees are.

8.1.4   ANOVA in a nutshell

[2017 - skipped]

8.1.5   ANOVA tables

[2017 - skipped]

8.1.6   Partitioning the sum of square (pg 464)

[2017 - skipped]

8.1.7   Calculating the mean square (pg 466)

[2017 - skipped]

8.1.8   The Variance ratio, F (pg 467)

[2017 - skipped]

8.1.9   Variation explained R2

[2017 - skipped]

8.1.10   ANOVA with two groups

[2017 - skipped]

8.2   15.2 Assumptions and alternatives

Will cover this in test 3

8.2.1   Robustness of ANOVA (pg 470)

Will cover this in test 3

8.2.2   Data transformations (pg 471)

Will cover this in test 3

8.2.3   Nonparametric alternatives to ANOVA (pg 471)

Will cover this in test 3

8.3   15.3 Planned comparisons

We covered this concept in detail

Planned comparison

8.3.1   Planned comparisons between means (pg 472-3)

We skipped the math

8.4   15.4 Unplanned comparisons

The method they focus on, the **Tukey-Kramer test**, is carried out using the TukeyHSD() function in R. They discuss the **Bonferonni correction** in a seperate Interleaf.

8.5   Example 15.4 Wood Wide Web

You should understand how the hypotheses are formulated and the Tukey test interpretted.

8.6   Assumptions (pg 477)

We will cover this on test 3.

8.6.1   [15.5 Fixed and Random effects]

[Skip]

8.6.2   [15.6 ANOVA with randomly chosen groups - SKIP!]

[Skip]

8.7   15.7 Summary

Read appropriate sections.

Chapter 9

Regression

Reading Guide
Chapter 17:Regression
Whitlock, MC and D Schluter. 2015. The Analysis of Biological Data, 2nd ed. Roberts & Company, Greenwood Village, Colorado.
Book website: http://whitlockschluter.zoology.ubc.ca/
Publisher: https://goo.gl/mSVUHe
Notes compiled by Nathan Brouwer (brouwern@gmail.com)

aka "linear regression", "linear model", "line of best fit", "least squares regression"

Done with the function lm() in R, which stands for linear model. lm() is used for regression and for ANOVA, which points towards the underlying similarity of these two methods.

Regression "is a method that predicts values of one numerical variable from values of another numerical variable." (pg 541). That is, the response AND the predictor are both numeric. This is in contrast to t-tests and ANOVA, where the predictors are categorical/factors.

In R-ish notation:

lm(continuous.response ~ continous.predictor.)

9.1   17.1 Linear Regression

9.1.1   Ex 17.1 The lion's nose

This is an interesting example of how to use regression for **prediction.**

9.1.2   The method of least squares (pg 542)

9.1.3   Figures 17.1-2

17.1-2 is a very important figure showing how the residuals change when different lines are fit to the data.

On page 543 they give the formula for the line as: Y = a + bX

Notation can vary between authors. Here, "a" is the intercept and "b" is the slope. It doesn't matter that they put the intercept first; it means the same thign to say Y = a + bX as to say Y = bX + a.

In words, this would be Y = intercept + slope*X

Slope "The slope of a lienar regresion is the rate of change in Y per unit of X" (pg 543). That is, how much Y changes as X changes.

9.1.4 FIgure 17.1-3

17.1-3 shows the difference between positive (+), negative (-), and flat (0) slopes.

9.1.5 Calculating the slope and intercept (pg 544)

SKip the equations on page 544.

On page 545 they present the regression equation for the lion example: Age = 0.88 + 10.65(proportion black)

which could also be written as Age = 0.88 + 10.65*proportion black

and

Age = + 10.65*proportion black + 0.88

Where 10.65 is the slope and 0.88 is the intercept. This equation allows you to predict Age (the response variable) based on the proportion of the nose that is black.

9.1.6 Populations and samples (pg 545)

9.1.7 Predicted values (aka "Y hat", pg 546)

Key idea: "The predicted value of Y [Y hat] from a regression line estiamte the mean value of Y for all individuals having a given value of X" (pg 546). This is if you sample many many lions that had, say, a proption black on their noses of 0.5, their would be variation in their ages. However, the mean of all those many lions with proportion black of 0.05 would be 6.2, because according to the estimated regression equation

- age = 0.88 + 10.65*proportion black
- age = 0.88 + 10.65*0.5 = 6.2

A regression line can be thought of as a continous set of predictions. That is, a continous serioes of Y hats

9.1.8 Residuals (pg 546)

Residuals are the distance from the regression lion to each data pont. That is, the distance from Y hat (the regression line) to a real data point.

[Skipe the equations on pag 547]

9.1.9 Standard error of the slope (pg 547)

We won't worry about the precise definition, but its very important to know that the slope (and intercept) are estimated **with error** and that error is characterized by the standard error.

9.1.10   Confidence interval for the slope (pg 548)

Slopes (and intercepts) also have confidence intervals (CI). As before, 1.96*SE will give you an approximate confidence interval around the parameter.

Note that there is a difference between the SE and CI for a parameter, and the confidence interval the surround an entire line. To put a confidence intervals (or Confidence band) around an entire regression line invovles combining the uncertainty in the slope and in the intercept.

9.2   17.2 Confidence in predictions

There is a subtle differene between a **confidence band** and a **prediction interval**. Prediction intervals will always be bigger than confidence bands. I almost always focus on confidence bands because as the author states I am usually "interested in the overall trend" of the data. When you are interesetd in a particular value, such as a single lion, how black its nose is and what its age might be, things change.

9.2.1   Confidence intervals for predictions (pg 549)

9.2.2   Figure 17.2-1

17.2-1 Shows how CIs and Prediction INtervals difer.

The authors summarize: "**Confidence bands** measure the precision of the prediction *mean* for each value of X. **Prediction intrvals** measure the precision of the prediction single Y-values for each X" (pg 550)

9.2.3   Extrapolation (pg 550)

Extrapolation means making predictions beyond the range of the origianl data.

9.2.4   Figure 17.2-2

17.2-2 shows a dataset where ear length was only measured for people between 30 and 100 years. If you used the regression equation to estimate ear length of infants, you would be extrapolating to ages not in the originaly data. This is generally a bad idea!

Extrapolation (pg 551)

9.3   17.3 Testing hypotheses about a slope

On page 551 they show the equation for the t-statistic associated with the slope of a regression line. You don't need to know the equation, but do need to know that t-statistics do play a roll in regression.

9.3.1   Example 17.3 Prarie Home Campion

9.3.2   Figure 17.3-1

Note that in this example the data are from an experiment, but regression analysis is being used. Also, the y-axis is log transformed.

9.3.3   The t-test of regression slope (pg 552)

Don't worry about the equations; you do need to know that you can do a t-test to test the hypothesis that the slope of the line is different than zero. Note that this is very general hypothesis: the slope or even whether its positive or negative doesn't factor into the hypothesis, just that its *not* exactly 0.0.

9.3.4   The ANOVA approach (pg 554)

The details aren't important. YOu do need to know that regression models can be tested by comparing two models and an F-statistics generated, just like in ANOVA.

9.3.5   Using R2 to measure the fit of the line to data (pg 555)

R^2 tells you what "fraction of variation in Y...is 'explain' by x." That is, how much scatter there is around the regression line. If all the points fall on the line, then R^2 would be 1.0 (100% of variation explained.)

9.4   17.4 Regression toward the mean

[skipped, but very interesting topic]

9.5   17.5 Assumptions of regression

The author's list 4, I tend to focus on the middle 2:

- "At each value of X, the distribution of possible Y-values is normal" (pg 557)
- This relates to the concept discussed below about the "normality of the residuals."

- Figure 17.5-1 Shows what this would look like.
- This can be assessed using resdiual analysis using a qqplot, or a histogram of the residuals. A qqplot is preferred.
- Log transformation and frequently reduce the impact of this violation
- Biologist have frequently focused on this issue, thought its been shown that its not as much of a deal breaker as the next one...
- "The variance of the Y-values is the same at all values of X"
- This is easiest to show in a picture, which the author's don't do. See lecture notes.
- This is often called "unequal variance" or "heteroskedasticity" (dont worry, I won't make you spell it.)

9.5.1   Figure 17.5-1

9.5.2   Outliers

Outliers can have a strong influence on regression. This is because the line is fit by calculating the residuals and squaring them. An outlier will have a large residual because it is far from the line. Square this large residual, and it increases the sum of squared residual (sum of all the residuals after they have been squared.) So regression model fitting will try to reduce how large that residual is, throwing everything off.

9.5.3   Figure 17.5-2

Note the outlier in the lower left hand corner of the plant at around x = 33, y = 36. Imagine how large the residual would be.

9.5.4   Detecting nonlinearity (pg 559)

The discuss **smoothing** here, or what they more specifically call "scatterplot smoothing".

9.5.5   Detecting non-normality and unequal variance (pg 559)

Here they discuss regression diagnostics and residual analysis using a **residual plot**.

Residual plot

9.6   17.6 Transformations

9.6.1   Figure 17.6-1

THey show how curving lines (the 2 left-hand panels) can be straightened using logs. Note that they explicitly using ln for the natural log (but R uses log for the ln and log10 for the log)

9.6.2   Figure 17.6-2

Compare 17.5-3, which plots the raw data, to 17.6-2, where both the x and y axes are transformed.

9.6.3   Figure 17.6-3

These graphs show the impact of transformation on the residuals when there are problems with **unqual variance / heteroskedasticity** (sometiems called non-constant variance)

9.7   [17.7 The effects of measurement error on regression]

important topic, but skipped

9.8   17.8 Nonlinear regression

9.8.1   [A curve with an asymptote]

skip

9.8.2   Quadratic curves (pg 565)

NLB note: quadratic curves are AKA "x^2 terms", "squared terms", "quadratic terms", "squared effect", "quadratic effects"; I will try to be consistent but will probably fail...

9.8.3   Figure 17.8-2

9.8.4   Formula-free curve fitting (pg 566)]

The go into the details of **smoothing** here. You don't need to know the different kinds of smoothers, just what the overall goal is.

9.8.5   Ex 17.8 The incredible shrinking seal

9.8.6   Figure 17.8-3

The line through 17.8-3 is a smoothed curve.

9.9   17.9 Logistic regression: fitting a binary response variable (pg 567)

You should be familiar with the concept that logistic regression is used when you have a categorical response variable, like mortality, and a numeric predictor varialble, like time. You don't need to know the math.

9.9.1   Figure 17.9-1

In this example, the guppies are categorized as alive or dead. This is their categorical "response" to the treatment. The predictor variable is a continous variable: how long they were exposed to a cold treatment.

You don't need to know the math.

9.10   17.10 Summary

Appendix A

Appendix A: Designing Successful Field Studies, Gotelli & Ellison, 1st Edition

A.1 Chapter 6 Designing Successful Field Studies

Read only section in red boxes. Key vocab and the section it occurs in are listed below.

A.1.1 What is the effect of factor X on variable Y? (pg 138)

This is a very common type of study in ecology, conservation, and wildlife. What is the effect of pH on the growth of waterthrush? What is the impact of coyote predation on fawn survival? What is the effect of nitrogen fertilizer on sugar maple growth?

Vocab **Manipulative experiment: Natural experiment:**

A.1.2 Manipulative experiments (pg 139)

Vocab **Manipulative experiment Cause and effect Confounding**

A.1.3 Natural Experiments (pg 141)

Vocab **Natural experiment**

A.1.4 Replication (pg 149)

A.1.4.1 How much replication?

A.1.4.2 How many replicates are Affordable?

A.1.4.3 Large-scale studies and Environmental Impacts

Vocab **BACI design**

I also mention **BA** - before vs after **CI** - control vs. impact (just done after)

A.1.4.4   Ensuring Independence

Vocab **Independence**

Figure 6.6 This figure present a problem where what happens in one plot impacts a nearby plot; I described this as the effect of one plot "bleeding" over into the next plot.

A.2   Avoiding Confounding Factors (153)

Figure 6.6 A confounded experimental design due to an environmental gradient; in this case, temperature varies between plots.

A.3   Replicaiton & randomization (154)

Vocab: **Replication Randomization**

Figure 6.7 Proper randomization in the face of an environmental gradient.

Figure 6.8 An experimental design with lots of replication, but totally confounded. All the controls on one side and all the treatments on the other.

A.3.1   Does the range of Treatments or Census Categories Bracket or Span the Range of Possible Environmental Conditions?

A.3.2   Have Appropriate Controls Been Established to Ensure That Results Relfect Variation only in the Factor of Interest?

A.3.3   Have all replicates been manipulated the same way except for the Intended Treatment Application?

A.3.4   Summary

Haphazard