A Reading Guide to Intuitive Biostatistics

Nathan Brouwer

2018-08-14

Contents

Preface

This is a reading guide to Harvey Motulsky's Intuitive Biostatistics: A Nonmathematical Guide to Statistical Thinking, 4th edition. More information about the book can be found at the book's website, http://www.intuitivebiostatistics.com/, and it can be purchased from Amazon.com. Motulsky is the CEO and Founder of GraphPad, a user-friendly statistical software popular in some branches of the life sciences.

Intutitive Biostatistics is a fabulous book for researchers that need to understand or do basic statistics and either need a concise primer on the key issues and/or are turned off by the equations underlying the statistical methods. Instead of using math to explain statistical methods, Motulsky focuses on written explanations, real-world examples, and novel graphing approaches. An excellent aspect of this book is that it unpacks common misunderstandings that researchers have, such as how to interpret p-values (Chapter 17), and signposts bad practices that must be avoided (like p-hacking). Again, this is done by focusing on intuition, not math. Motulsky also presents best practices in plotting, data presentation, and data reporting, emphasizing the key aspect of adequate and accurate presentation of results.

This reading guide serves several purposes:

- Highlight the parts of the book I focus on in my teaching (and so will be on any tests!)
- Provide additional complementary examples
- Indicate extensions or alternatives
- Provide citations and links to resources for follow-up
- Indicate where others (including myself, though I am not a trained statistician) might disagree with Motulsky

Each part of the reading guide is essentially an outline of each chapter with commentary as needed. In some cases I have written a brief initial commentary to put the chapter in context. I will often indicate the Excel or R functions related to methods or calculations; for a fuller treatment see my other guide *An R Companion to Motulsky's* Intuitive Biostatistics. At the end of each chapter are typically references, a list of R and Excel functions needed to carry out the analyses in the book, and study questions to consider.

My most important notes and comments are generally in **bold** or bulleted. When I've riffed on an idea and its not necessarily key I've usually put in in a block quote, like the one below:

> For example, sometimes I've written about a section, and my text is almost as long as the original section!

This is a work in progress and many sections are not yet annotated; feel free to contact me with suggestions or corrections.

Nathan Brouwer brouwern@gmail.com

Chapter 1

"Statistics & Probablity Are Not Intuitive"

Commentary

In this introductory chapter Motulsky sketches out some major reasons why people struggle with statistics and probability. This chapter assumpes some basic familiarity with statistical ideas. Sometimes this chapter is a bit terse - its meant to highlight key ideas, not fully discuss or demonstate them.

Vocabulary

Motulsky vocab

- sample
- population
- Bayesian
- multiple comparisons
- regression to the mean

Additional vocab

- Bayes theorem
- pre-registration
- exploratory analyses

Key functions

None

Chapter Notes

1.1 We Tend to Jump to Conclusions

Motulsky uses the phrase "**generalize from a sample to a population**" without defining what this means. In general, this means to look at some subset of the world - either something experienced in real life or generated using a scientific study - and conclude that what was seen in the subset occurs elsewhere. In the example he uses, his daughter experienced meeting doctors, and they all were male, so she generalized to the rest of the world that all doctors must be male. While this example is trivial, anytime we generalize from sample to population (or from a part to the whole) we run the risk that our sample is biased. It could be biased becauase we didn't take a good sample, such as relying just on personal experiencce. Or it could be a rigorously collected scientific sample, but still be non-representative. What if he wanted to prove his daughter wrong and so randomly selected 10 doctor's offices for a web search and looked up who the senior physician is. If he happend to find my doctor's office, he'd see that its a women, Dr. Cathy Lamb. However, it is possible that he could look up 10 doctor's and they could all still be male.

1.2 We Tend to Be Overconfident

1.3 We see Patterns in Random Data

1.4 We don't realize that coincidences are common

He doesn't use the specific term, but he is alluding to the concept of **hindsight bias**.

1.5 We don't expect variability to depend on sample size

Motulsky cites a paper by Andrew Gelman here, one of the most thought provoking - though sometimes just provoking – statistics bloggers of the last decade. He blogs regularly at Statistical Modeling, Causal Inference, and Social Science and writes non-technical pieces for a number of outlets, including Slate. He is also prominent Bayesian.

1.6 We Have Incorrect Intuitive Feelings About Probability

1.7 We Find it Hard to Combine Probabilities

1.8 (We Avoid Thinking About Ambiguous Situations)

(This section appear in previous versions; I am not sure where/if it occurs in the 4th edition)

1.9 We Don't Do Bayesian Calculations Intuitively

Motulsky doesn't define **Bayesian** here, though its not central to what he's talking about. In this example, "Bayesian calculations" refers to a particular type of probability calculation using **Bayes Rule**. His example is a classic example of how probability calculations are used for diagnostic testing.

More generally, "Bayesian"" refers to a particular way to use the mathematics of probability to make inference. All mathematicians agree on the basic rules of probability calculations. In contrast, when it comes to using the math of probablity to make inference from a sample to a population - that is, to do statistics - there is a huge rift between **Frequentists** and **Bayesians**.

1.10 We are Fooled By Multiple Comparisons

The study on astroglogical signs here is a great paper intended to "To illustrate how multiple hypotheses testing can produce associations with no clinical plausibility" (Austin et al 2006, Abstract). "Multiple hypotheses testing" means the same thing as "multiple comparisons." As Motulsky indicates, if you test multiple hypotheses or make multiple comparisons between things, sooner or later you'll find a strong association. This is why it important to make specific hypotheses prior to the beginning of a study - ideally even publically **pre-registering** them - and properly indicate which analyses were defined in advance and which are **exploratory analyses**.

Multiple Comparisons is a big topic that Motulsky doesn't go into detail yet. He devotes several excellent chapters to this topic elsewhere. This issue of multiple comparisons is a big and controversial one. For a discussion of multiple comparisons

1.11 We tend to ignore alternative explanations

1.12 We are fooled by regression to the mean

Regression to the mean is a concept that isn't typically taught in intro stats courses, especially for ecology. For its relevance to ecology and evolution see the paper by Kelly and Price (2006) "Correcting for Regression to the Mean in Behavior and Ecology" in *American Naturalist*.

1.13 We let our biases determine how we interpret data

1.14 We crave certainty, but statistics offers probability

1.15 Further reading

1.16 References

Austin, Mamdani, Juurlink and Hux 2006. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. Journal of Clinical Epidemiology 59:964–969 Open Access

1.17 Annotated Bibliography

1.17.1 Multiple comparisons

Bender & Lange 2001. Adjusting for multiple testing—when and how? Journal of Clinical Epidemiology. 54:343–349. Abstract

> **Multiple comparisons** is a thorny issue that Motulsky briefly introduces here in Chapter 1 and discusses in depth elsewhere. Throughout the book Motulsky focuses on the need for multiple comparisons procedures in general, and the most popular ones used; he doesn't go into the broader arguements about their use and the many ways they can be problematic. Bender & Lange (2001) give a taste of the mess made by multiple comparisons issues. They note "…there seems to be a lack of knowledge about statistical procedures for multiple testing. For instance, multiple test adjustments have been equated with the Bonferroni procedure, which is the simplest, but frequently also an inefficient method …" (pg. 343). They discuss the various positions that have been taken for and against multiple comparisons in the biomedical sciences, and advance their particular perspective on the issue. Elsewhere in the book Motulsky discusses the Bonferonni correction under the heading "The Traditional Approach to Correcting For Multiple Comparisons." He then outlines a more contemporary approach, the **False Discovery Rate (FDR)**. Bender & Lange (2001) was written before the FDR became popular and instead briefly disucuss other alternatives, including Holm modificaiton to the Bonferroni procedures and advanced computational methods.

Chapter 2

"The complexities of probability"

Commentary

Probability is central to statistics, but its inherently hard. Most introductory stats books spend at least one chapter to lay out the foundations, which can seem tangential to the main task at hand - analzying data! Advanced stats books typically go back to probability, often in calculations that are unfortunatley not within the comfort zone of most biologists. Motulsky doesn't shirk the responsiblity of reviewing probability, but does so in a conversational style.

2.1 Focal parts of chapter

The entire chapter should be read

Vocabulary

Motulsky vocab

- probablity as long-term frequency
- probability as subjective belief
- model

Aditional vocab

Key functions

None

Chapter Notes

2.2 Basics of probability

2.3 Probability as long-term frequency

2.3.1 Probabilities as predictions from a model

model

2.3.2 Probabilities based on data

2.4 Probabilities As Strength of Belief

2.4.1 Subjective probabilities

The concept of *subjective probabilities* is a big, broad topic that relates to **Bayesian statistics**. Motulsky is pointing towards the process of how oth personal belief and **prior scientific information** can inform our assessment and even formal analysis of a situation. In Bayesian statistics, a **prior** is a formally stated and quantified belief about the topic of interest. In practice its often stated as a **probability distribution.**

2.4.2 "Probabilities" used to quantify ignorance

This is very important point that relates to how we use end up applying probability and statistics. Motulsky uses the example of an unborn child. The child is developing and (except in very rare cases) is either XX or XY for its sex chromosomes. The process of combining the maternal X and paternal chromosomes to put this X-X or X-Y pairing together is done. As Motulsky discusses, we can still talk about the probablity of the child being XX or XY until the birth and we know what pairing occured.

This is similar when we do an experiment and use statistics. Say we're in the early phases of drug development and we don't know whether a drug performs any different than the control (eg a [placebo(https://en.wikipedia.org/wiki/Placebo)]). We can uses statistics to compare patients who recieved the drug and those that didn't. In the early phases of drug developement it often isn't known if a drug works or fully the biological mechanisms by which it works. In reality, the drug does or does not interact biological in humans, and those interactions are typically positive, negative, or neutral. With a lot of work those details can be worked out; a single drug trial on a limited sample of patients only moves us a bit towards that. What it accomplishes, and what the statistics help us do, is get a handle on how ignorant we remain of the details of how the drug works.

2.4.3 Quantitative predictions of one-time events

At times, when there is a one-time event someone will say something like: "the probability is 50%: it either will happen or not." This is a confusion of the fact that the outcomes are binary (yes/no) with the probability that one outcome will happen or not.

The polling around the 2016 elections has provided lots of fodder for commentary on statistics and data analysis. Andrew Gelman has blogged on this on his own site and also for Slate. See ["We Need to Move Beyond Election-Focused Polling"] (http://www.slate.com/articles/technology/future_tense/2017/09/what_

is_the_future_of_polling.html) which has the tagline "Polling didn't fail us in 2016, but what happened made polling's flaws more apparent. Here's how to fix that."

Also see "19 Lessons for Political Scientists From the 2016 Election".

Among political-science orientated statisticians like Gelman the work of FiveThirtyEight.com comes up a lot. I'm not that familiar with it so I checked Wikipedia: "FiveThirtyEight…is a website that focuses on opinion poll analysis, politics, economics, and sports blogging. The website…takes its name from the number of electors in the United States electoral college."

2.5 Calculations with probabilities can be easier if you switch to calcualting with whole numbers

Motulsky presents two version of the same word problem to show how the presentation of probabilities can impact how easily they can be understoon. The first version of the problem is tricky and I didn't see how to get the answer at first. The main hang up I think is the fact that it requires you to think in terms of **conditional probabilities**. The problem states that 0.8% (not the proportion 0.8, which = 80%!) women are diagnosed with breast cancer. The second sentence states "If a woman has breast cancer, the probability is 90 percent that she will have a positive mamogram." This is a condition probability. In words we'd say more formally "If a woman does have breast cancer, the probability that she has a positive mammogram is 0.9." Stating it this way put the temporal sequence out of order, as does the origina sttatement. A better way might be, "Among women with breast cancer, 90% had positive mammograms." THe upshot is this: we need to think in terms of 90% of 0.8%; that is start with the 0;8% that have cancer and then take 90% of though.

I think this is also tricky because we're working accross and order of magnitude. Its much easier to put in real numbers by starting with 1000 women recieving mammograms. 0.8%*1000 = 8 women in that group that actually have cancer. 8*90% = 7ish. The women with cancer and the women with cancer and a positive mammogram are on the same order of magnitude.

The next trick for solving the problem in terms of the first version of the problem is to figure out how many women have **false positive** mammograms. That is, they don't have cancer but their mammogram comes back as positive and the need to go through further screen to determine that everythings actually ok. The first version of the problem on page 17 states "If a women does not have breast cancer, the probability is 7 percent that she will have a positive mammogram." So 7% of women will get a **false positive**.

A common mistake with this problem that I almost did was to calcualte the number of women out of 1000 with false positives as 7%*1000. However, the problem states "If a women does not have breast cancer, the probability is 7 percent." We were already told that 0.8% of women* do* have cancer; we need to subtract them out. So we get 1000-1000*0.8% = 992 women are cancer free. Of those 992 cancer-free women, 7% will end up with false positive mammograms. So 992*7% = 70.

So, summarizing, we have 1000*8% = 8 women with cancer, but only 8*90% = 7 of those women with positive mammograms. Since these women do indeed have cancer and the mammogram also indicates this, they are referred to as **true positives.** We also have 992*7% = 70 women with false-positive mammograms. 7 true positives plus 70 false positives equals 77 total positive mammmograms.

To get the probablity that a women with a positive mammogram actually has cancer we take the total number of positive mammograms (77) the number of those with cancenr (7): 7/77 = 0.09 = 9%.

Because this problem is difficult it comes up frequently when discussing statistics and medicine.

As an aside, I'll show how these calculations could be written out in R. I'll use "*" as in Excel for multiplication and "/" for division. I'll store values for future use using the assignment operator "<-". I'll also use the round() function to round things off as is done in the original problem.

```
#number of cancer cases out of 1000
## 0.8% = 0.008
true.incidence <- 1000*0.008

#number of "true positives""
## (those with cancer)*(probablity of positive mammogram)
true.positive <- true.incidence*0.9

#exact value is 7.2; round it off to 7 as in the example
true.positive
```

```
## [1] 7.2
```

```
true.positive <- round(true.positive)

#cancer free individuals
cancer.free <- 1000-true.incidence

# false positives
# 7% = 0.07
false.positive <- cancer.free*0.07

#to make the math easy they round up
false.positive
```

```
## [1] 69.44
```

```
false.positive <- 70

#total number of positive mammograms
total.positive <- true.positive+false.positive

#probability that a positive mammogram is a true indication of cancer
true.positive/total.positive
```

```
## [1] 0.09090909
```

2.6 Common Mistakes: Probability

2.6.1 Mistake: Ignoring assumptions

2.6.2 Mistake: Trying to understand probability without clearly defining both the numerator & the denominator

2.6.3 Mistake: Reversing probability statements

2.6.4 Mistake: Believing the probability has a memory

gambler's fallacy

2.7 Lingo

2.7.1 Probability vs. odds

2.7.2 Probability vs. statistics

This is a key idea that I don't think I've though about a lot: * probablity: general principals -> specific situation * statistics: general population <- specific dataset

To relate to his earlier example, if we are interested in the probability of a child being born XY, you can start with a general model (how meiosis works) or data on large population (the CIA database) and make an inference about a specific situation: the birth of a particular child.

2.7.3 Probability vs. likelihood

As Motulsky mentions, **likelihood** has a particular technical meaning in statistics. While this his book doesn't devel into it, you don't have to spend much time doing analyses these days before encountering it. The following topics all involve likelihoods in their current application:

- logistic regression
- analysis of count data with Poisson regression
- generalized linear models (GLMs; of which logistic and Poisson reression are forms)
- mixed models
- generalized linear mixed models (GLMMs)
- Phylogenetic methods (estimating phylogenetic trees; using phylogeneis in statistica analyses)
- Bayesian methods

2.8 Probability In Statistics

Table 2.1 is a good summary. A great question on a test would be to blank out some of the words and ask students to fill them in.

2.9 Further reading

2.9.1 References

2.9.2 Annotated Bibliography

2.9.2.1 Multiple comparisons

Chapter 3

"Confidence Interval of a Proportion"

Preamble

On proportions, frequencies, and percentages

Like many books, Motulsky starts of by discussing **proportional data.** Proportional data can also sometimes be called **frequencies**; a useful, mathematically precise term is **binomial proportions**. They occur when you have a certain, discrete number of things happen, such as full-term births, and you count the frequency or calculate the proportion of a specific event, such as a child having brown eyes. Mathematically the "things happening" are often called **"trials"** and the outcome of interest are often called **"successes"**, though "events" or "outcomes" makes more sense to me. In stats books you will often see the term **"Bernouli trial"** used to refer to a single **binomial** trial. Flipping a coin once is a Bernouli trial.

Proportions are often conveyed in terms of **percentages**, such as "20% of child born in Pittsburgh have brown eyes." Percentages are tricky in stats because you have to keep in mind whether the percentage is derived from counting up **discrete events** that can be counted with absolute precision (babies born with blue eyes) or is a continuous quantity (the percentage of a child's face they've covered with splatter from the food they've eaten).

Another term commonly used is **binary data.** Binary indicates that an event can take on one of two values; in practice, the values "0" and "1" are used when doing the underlying math even though "0" might mean "eyes not brown" and 1 means "eyes brown."

Proportional data are very common in biology: number of fruit flies that are virgin, number of flowers eaten by deer, number of tadpoles surviving a exposure to a toxin. Proportion data are also easy to work with and the math for calculating things like **confidence intervals** intervals and especially **p-values** is much easier than for **continuous varibles** such as the length of a fruit fly wing or the height of a plant.

On counting "events" versus counting things

Imagine you work for the Demography Department at Magee Women's Hospital in Pittsburgh. Your job every day is to visit everyone room in the hospital and count two things: 1) the number of visitors in each room and 2)ikm 5444155203/the number of brown-eyed babies out of the total number of babies. The first task is so that the hopital can determine how many people are visiting and the second task is to determine the percentage of brown-eyed babies born to the hosptial.

For both tasks you are counting, but there is a very important but subtle difference between these two tasks. For the first task, you are counting up the number of people in each room, which could vary from zero

(un-occupied) to a potentially large number if many people are visiting a newborn. Each data point is a number, either zero or something larger.

For the second task you can think of it as counting up the number of brown-eyed babies, and this count could take on many different values depending on how many babies are born. However, this count is **bounded** by the total number of babies born. Similarly, key to what you want to know is the number of brown-eyed babies out of the total number of babies; you are therefore setting up a proportion. The former example doesn't invovle a proportion and is simply an **un-bounded count**.

I'm belaboring this because this because I have frequently seen where misunderstanding the different statistical uses of the term "count" has resulted in biologists (to be fair, only ecologists so far) selecting an incorrect statistical procedure.

On confidence intervals versus p-values

Most books start with **p-values** then move on to **confidence intervals**; while the two things are intimately linked and derived from the same calculations, confidence intervals convey much more information. Motulsky starts off with confidence intervals in this chapter.

Vocabulary

Motulsky vocab

- Bias
- Binomial variable
- Binomial distribution
- Confidence interval (CI)
- Confidence level
- Credible interval
- Point estimate
- Random sampling
- Sampling error
- Simulation
- Uncertainty interval

Additional vocab

- binary data
- binomial proportion
- frequency
- proportion

Chapter Notes

3.1 "Data Expressed as Proportions"

3.2 "The Binomial Distribution: From Population to Sample"

3.3 "Example: Free Throws in Basketball"

3.4 ("Example: Deaths of Premature Babies")

Alive vs. dead is one of the most basic binary conditions in biology.

3.5 "Example: Polling Voters"

3.6 "Assumptions: Confidence Interval of a Proportion"

3.6.1 "Assumption: Random (or representative) sample"

3.6.2 "Assumption: Independent observations"

Proportional data can be tricky because the key to the statistics used to analyze them in most cases is that all of the events are **independent**. For example, each non-twin child born in a hosptial on the first day of month is essentially an **independent trial**. Each child has different parents, a different gestational environment, and most relevant if you are counting up the number with brown hair, different genetics. So, the hair color of one child born on the first day of the month has no impact on the hair color of another child; they are unlinked and unrelated.

In contrast, there its possible that the fates of mothers while giving birth are not independnet. For example, what if we want to know the number of women who originally intended to give birth vaginally but ended up having a cecarian (c-section)? Each women is different, but they are likely to be attended to by the same attending physician, who can vary in their approach to delivery and when they recommend a c-section. So if 20 women give birth on the first day of the month, they hair color of their babies are all independent data points, but whether these women had c-sections or not is potentially not independent.

3.7 "Assumption: Accurate Data"

3.8 "What Does 95% Confidence Really Mean"

3.8.1 "A simulation"

[ ] Figure 4.2: What would happen if you collected many samples and computed a 95% CI for each

This figure is **very important.** The thought experiment where you hypothetically re-run your study or experiment many times is central to the concept of what **confidence intervals** and **p-values** are.

3.8.2 "95% Change of What?""

3.8.3 "What is Special About 95%"

[ ] Nothing. Absolutely nothing. This cannot be repeated enough. There is nothing sacred scientifically or mathematically about 95% or a p-value that is less than 0.05.

[ ] This is so important I will repeate it again. **There is nothing sacred scientifically or mathematically about 95% or a p-value that is less than 0.05.**

[ ] There has even recently been a call to try to get people to not call something **"significant"** unless is **<0.005** (equivalent to using a 99.5 % CI). This has resulted in a lot of discussion in journals, blogs, and twitter, with **frequentists** arguing with frequentists, some **Bayesians** offering their alternatives to signficance tests (eg Wagenmakers) and other Bayesian saying we need to get rid of hypothesis testing entirely (Gelman). There are many interesting blog posts and published opinion pieces on this now.

Like most stats books Motulsky mentions the possiblity of calculating 90% CIs that are more lax, or 99% CIs that are more stringent. Most books have you do exercises where you calcualte different CIs. In the biological sciences I have never seen anything but a 95% CI. I think in manufacturing applications of statistics and other fields perhaps this is more common.

[ ] There has been some discussion that is should be more common to think about what level of "confidence" you want or need to make a descision and adjucting your CI accordingly. This is discussed in print and via the blogs by the pyschologist Daniel Lakens, and I believe Richard Morey.

3.8.4 ("What If The Assumption Are Violated")

THis section appeared in previous editions of the book

3.9 "Are You Quantifying the Event You Really Care About?"

3.10 Lingo

3.10.1 CI versuse confidence limits

"confidence limit" isn't used too much in practice.

3.10.2 Estimate

3.10.3 Confidence level.

[ ] Again, there is nothing special about 95%.

3.10.4 Uncertainty Interval

Uncertainty interval is a proposed replacement term for confidence interval. I have never seen it used, except by those who have proposed it.

3.11 "Calculating The CI of a Proportion""

3.11.1 "Several methods are commonly used""

There are many ways to deal with binomial data in general, and in R. The basic ones usually show up in an intro stats course are

- Binomial test: binom.test()
- Test for equal proportions: prop.test()
- Chi^2 test: chisq.test()

All of these produce similar result and are probably mathematically related if you start to dig into them, which I haven't done lately.

This profusion of different tests is one annoying feature of the traditional way statistics is typically taught and the way most intro-level stats books are written. In contemporary applied statistics, binomial data like this are likely to be analyzed using something called **"logistic regression"** or a **"binomial general linear model"**. A general linear model is often called a **GLM** for short. Motulsky doesn't go all the way into developing GLMs but he is generally oriented in that direction, which is good.

To be more precise, there are both

- Multiple ways to analyze these data to get a p-value
- Multiple ways to calcualte a confidence interval

The confidence interval issue is what Motulsky focuses on here. I will work through these calculations in R by hand and also how to use common functions to get them, which also yield p-values.

3.11.2 "How to compute the modified Wald method by Hand"

This is a good computational exercise and so we'll work through the details. For published papers use a computer to do the work!

We'll work with the following quantities

- S = the observed number of "successes"
- n = number of binomial "trials"

Is this the same as gets repeated below? What was I doing ... :(

```
# Step 0: the data
S <- 31   #S = successes = num. infants surviving to 6 months
n <- 39   #n = number of trials = total num. infants in study
z <- 1.96 #z =

#calculate z^2
z2 <- 1.96^2

#round it off to 3 decimal places
z2 <- round(z2, 3)

# Step 1: calculate the Wald-corrected proportion
## observed proportion
P.obs  <- S/n
P.obs  <-  round(P.obs, 3)

## "corrected" proportion
```

```
### Set up whole formula
P.corr <-  (S+z)/(n+z^2)

### might be easier to see if done in steps
numerator <- S+z
denominator <- n+z^2
P.corr <- numerator/denominator

### round off
P.corr <- round(P.corr, 3)

# Compute half-width of the CI
## In one step
W <- z*sqrt(P.corr*(1-P.corr)/(n+z^2))

## in parts
### calculate numerator and denominator
W.numerator   <- P.corr*(1-P.corr)
W.denominator <- n+z^2

### calcualte W
W             <- z*sqrt(W.numerator/W.denominator)

## round W off
W <- round(W, 3)

# Calculate CI bounds
lower.CI <- P.corr - W
upper.CI <- P.corr + W

P.obs.calc  <- "=31/39"
P.corr.calc <- "=(31+1.96)/(39+1.96^2)"
```

The calculations can be laid out in a table like this

A

B

C

D

E

Data

Calcualted.val

survived

31

P.observed =

0.795

deceased

8

P.corrected =

0.769

n.total

39

W.numerator =

32.96

z

1.96

W.denominator=

42.8416

z.squared

3.842

W =

0.126

CI.lower =

0.643

CI.upper =

0.895

We can code Motulsky's analysis using the **Wald method** like this:

```
# Step 0: the data
S <- 31   #S = successes = num. infants surviving to 6 months
z <- 1.96 #z =
n <- 39   #n = number of trials = total num. infants in study

# Step 1: calculate the Wald-corrected proportion
P.obs  <-  S/n           #observe proportion
P.corr <-  (S+z)/(n+z^2) #corrected proportion

#or, approximately, since 1.96^2 requires a calcualte
## here what he is doing is just rounding 1.96 up to 2.
## this if any makes the confidence a bit wider and
## therefore a bit more conservative
## of course, 33/43 typically will require a calculator
P.corr.alt <-  (S+2)/(n+4)

#If you get confused by the parentheses you can always break things up

numerator <- S+z
denominator <- n+z^2

P.corr <- numerator/denominator

# Compute half-width of the CI
W <- z*sqrt(P.corr*(1-P.corr)/(n+z^2))
```

```
lower.CI <- P.corr - W
upper.CI <- P.corr + W
```

Here is the output of the three statistical "tests" can can be applied to these data.

```
#packages to clean up the output
library(tidyr)
library(broom)
library(pander)
```

estimate	p.value	conf.low	conf.high	method
0.7949	0.0002941	0.6354	0.907	binom.test
0.7949	0.000427	0.6306	0.9013	prop.test
NA	0.0002306	NA	NA	chi^2

Here is the output using a binomial GLM (aka logistic regression)

estimate	std.error	p.value
0.775	0.3786	0.00109

3.11.3 Shortcut for proportion near 50% (OPTIONAL)

3.11.4 Shortcut for proportion far from 50% (OPTIONAL)

3.11.5 Shortcut when the numerator is zero: The rule of three (OPTIONAL)

3.12 "Ambiguity if The Proportion Is 0% or 100%"

3.13 "An Alternative Approach: Bayesian Credible Intervals"

3.14 "Common Mistakes: CI of A Proportion"

3.14.1 "Mistake: Using 100 as the denominator when the value is a percentage"

3.14.2 "Mistake: Computing binomial ICs from percentage change in a continous variable"

3.14.3 "Mistake: Computating a CI from data that look like a proportion but really is not"

3.14.4 "Mistake: Interpretting a Bayesin credible interval wihtout knowing what prior probabilities (or probabilitiey distribuiton) were assumed for the analysis"

3.15 Q & A

All of these are good points

[ ] **Figure 4.3. Effect of sample size on the width of a CI** This is a very important idea. Sample size is key to increasing the confidence we have in a result

[ ] **Figure 4.4. Asymmetrical CI** Proportions, percentages, etc are all bounded between 0 and 1, or 0% and 100%. Most methods of calculating confidence intervals for this type of data (but not all!) will produce assymmetric CI. If you see a CI for this type of data that crosses 0% or 100%, there's a good chance the authors did not use an appropriate method for calculating the confidence intervals. I see this most often when data are percentages, like mean percentage of the ground covered by an invasive species.

Chapter 4

"Graphing Continous Data"

Commentary

Vocabulary

Motulsky vocab

- Arithmetic mean
- Bias
- Box-and-wisker plot (boxplot)
- Continous data
- Dotplot (or scatter plot) (or beeswarm)
- Error
- Frequency distribution
- Histogram
- Interquartile range
- Mean
- Median
- Mode
- Outlier
- Percentile
- Quartile
- Precision
- Smoothed data
- Trimmed mean
- Violin plot

Additional vocab

Key functions

- boxplot
- beeswarm::beeswarm
- stripplot

Chapter Notes

4.1 "Continuous Data"

Ecological examples of contiunous data include: the mass of a lizard, the volume of bird feces, the length of a spider appendage, the height of a tree.

Lab biology example of continuous data include: the concentration of protein in solution, the intensity of a band on a gel, the molecular weight of different salt ions.

4.2 "The Mean and Median"

- Median: Medians are useful because the provide a better idea of the center of a distribution even if there are **outliers** or **skew**. Medians are very useful to think about and plot in your graphs, but suprisingly rarely comes into play for actual statistical calculations. This is because the math related to medians causes problems; means are much easier to work with. The field of (**robust statistics**)[https://en.wikipedia.org/wiki/Robust_statistics] frequently works with medians. **Quantile regression** is one method that works particularly with medians.
- The geometric mean: good to know about but results are rarely presented in terms of geometric means. (An exception in ecology is stochastic demography.)
- Harmonic mean: like the geometric mean, results are rarely presented this way.
- Trimmed mean: not currently use much in biology, but potentially used. Discussion of robust statistics sometimes include trimmed means.
- Mode: Like the median, useful to think about but rarely used in statistic computation.

4.3 "Lingo: Terms used to Explain Variability"

4.3.1 "Biological variability"

4.3.2 "Precision"

[?? do I agree with this way of framing things]

4.3.3 "Bias"

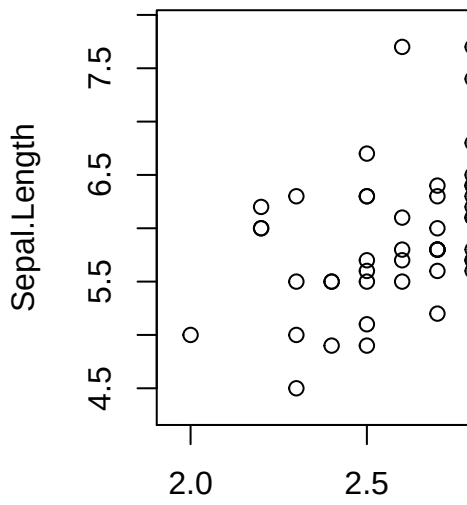
4.3.4 "Accuracy"

4.3.5 "Error"

4.4 "Percentiles"

4.5 "Graphing Data to Show Variation"

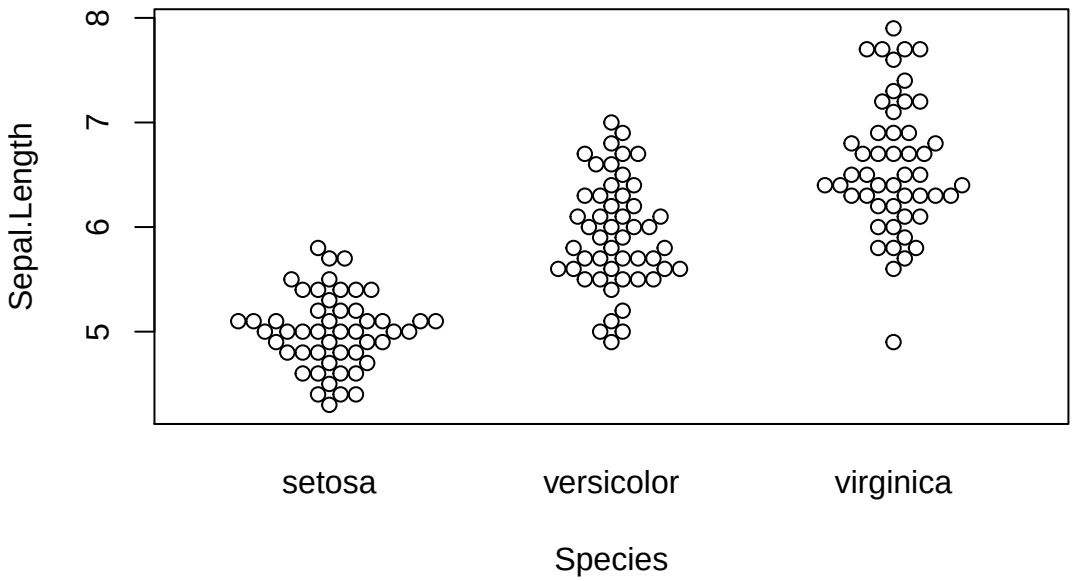
4.5.1 "Scatter plots"



Scatter plots often refer to plots used when you have two numeric variables, like this

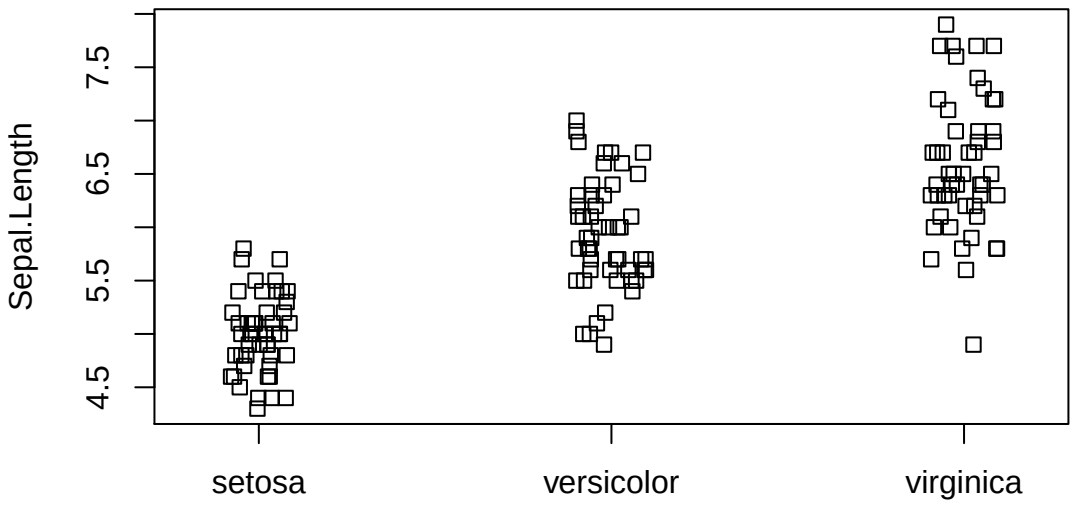
What Motulsky shows in Figure 7.1 is sometimes now called a **beeswarm plot**, a name I like. They can be made in R using the beeswarm package.

```
beeswarm(Sepal.Length ~ Species,
         data = iris)
```



A similar type of plot is a stripchart. These work best when they are set up to not have their points overlapping, which is called **jittering**.

```
stripchart(Sepal.Length ~ Species,
           data = iris,
           vertical = TRUE,
           method="jitter")
```



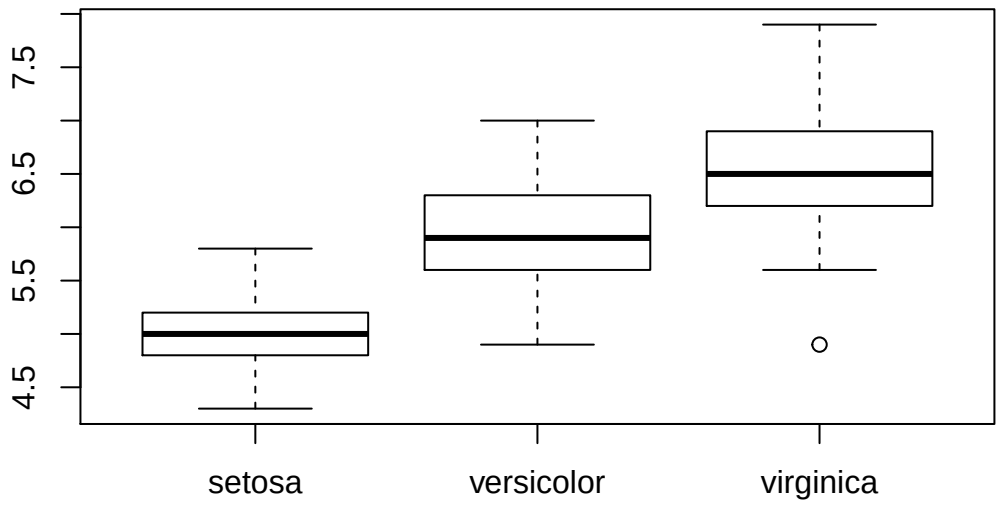
A beeswarm plot is basically a jitter stripchart that has been well organized. And has a cooler name.

4.5.2 "Box-and-whiskers plots"

Usually just called a **boxplot**.

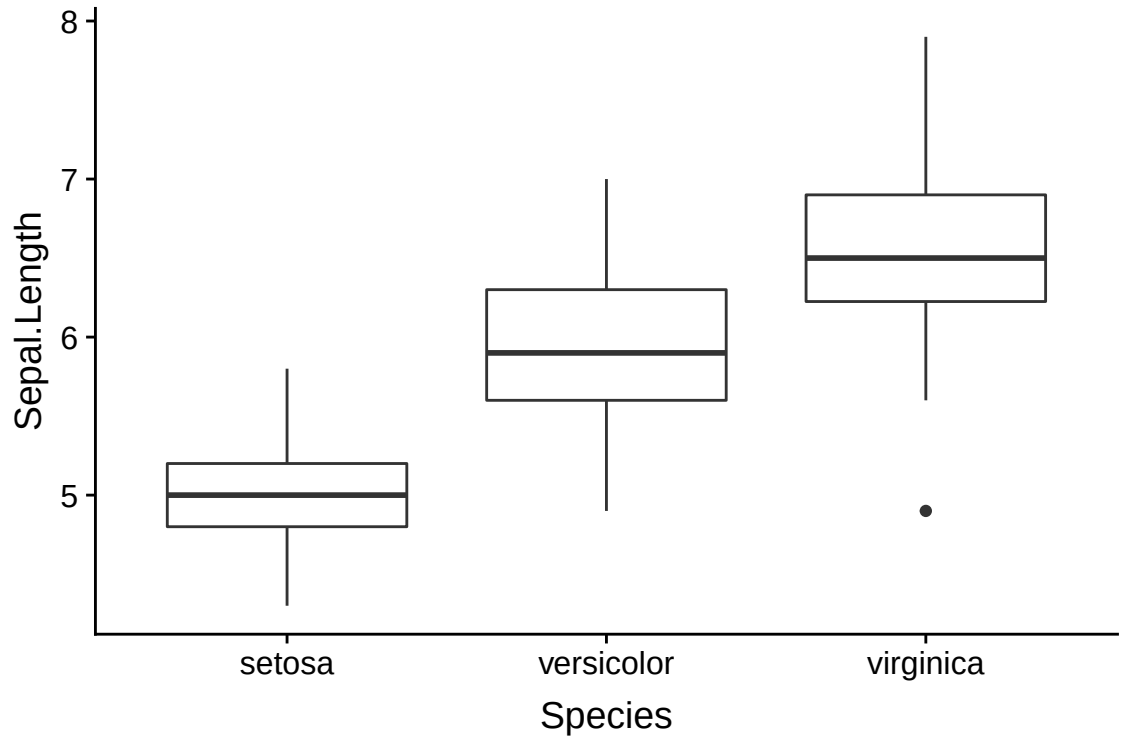
In base R they are made with the boxplot() command.

```
boxplot(Sepal.Length ~ Species,
          data = iris,
          vertical = TRUE,
          method="jitter")
```



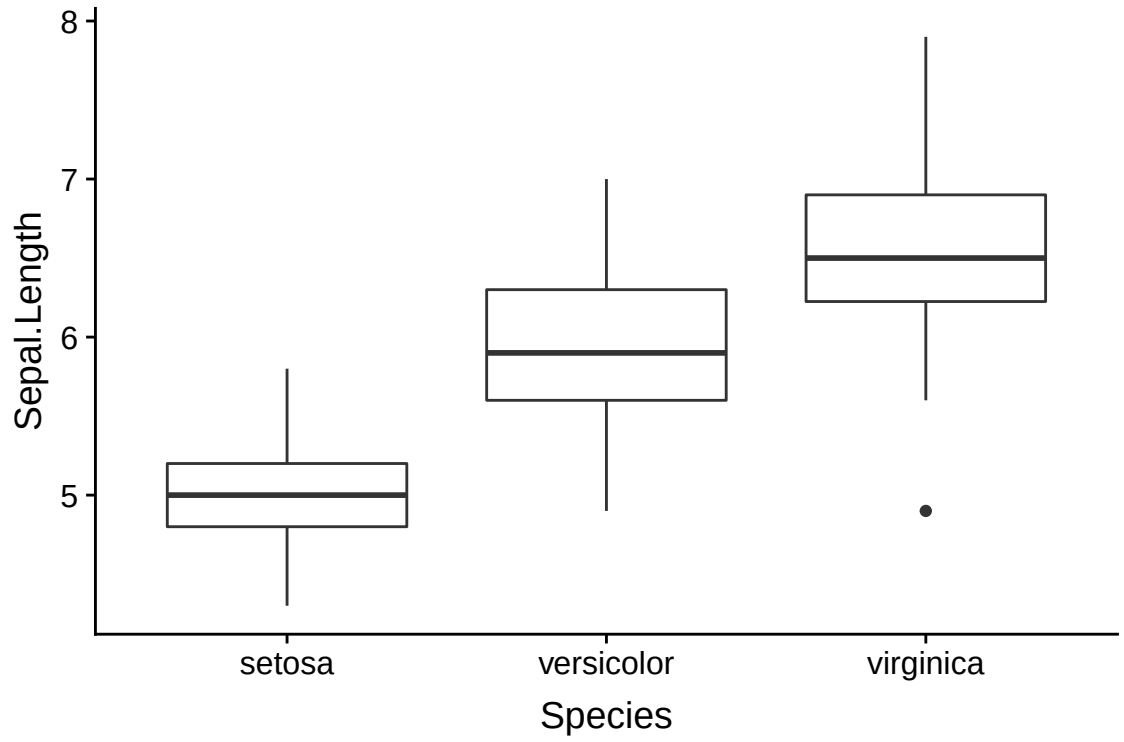
In ggplot they can be made like this with the qplot() function

```
qplot(data = iris,
      y = Sepal.Length,
      x = Species,
      geom = "boxplot")
```



Or directly with ggplot() using geom_boxplot()

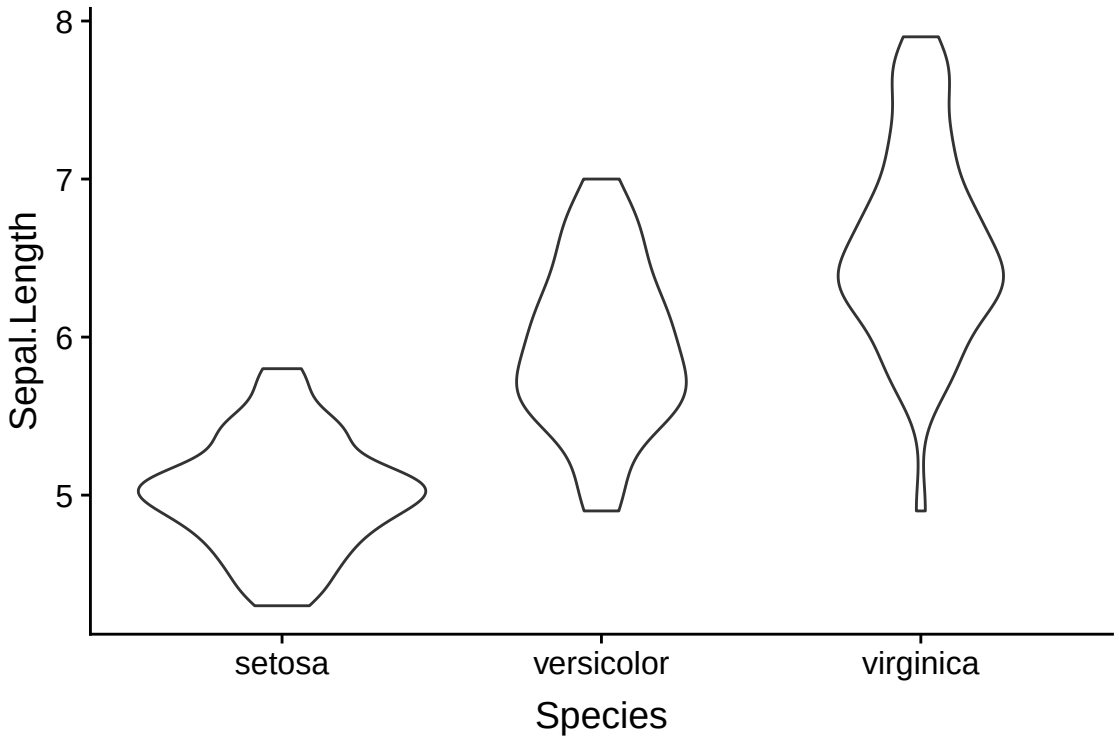
```
ggplot(data = iris,
       aes(y = Sepal.Length,
           x = Species)) +
  geom_boxplot()
```



4.5.3 "Violin plots"

Violin plot can be useful when you want more information than given by a boxplot but have too much data for a beeswarm. There's a package in R which implements violin plots for basic R graphics. In ggplot you use geom_violin().

```
ggplot(data = iris,
       aes(y = Sepal.Length,
           x = Species)) +
  geom_violin()
```



4.6 "Graphing Distributions"

4.6.1 "Frequency distributions"

4.6.2 "Cumulative frequency distribution" (OPTIONAL)

Good to know about but not applicable to most entry-level stats.

4.7 "Beware of Data Massage"

4.7.1 "Beware of filtering out impossible values"

4.7.2 "Beware of adjusting data"

4.7.3 "Beware of smoothing"

4.7.4 "Beware of variable that are the ratio of two measurements"

4.7.5 "Beware of normalized data"

Beware of ratios of ratios

Certo, et al. 2018. Divided We Fall: How Ratios Undermine Research in Strategic Management http://journals.sagepub.com/doi/abs/10.1177/1094428118773455

Curran-Everett, D. 2013. Explorations in statistics: the analysis of ratios and normalized data. Advances in Physiology Education.

Motulsky doesn't mention this. It can be a problem, though. Some papers related to this topic:

Karp et al. 2012. The fallacy of ratio correction to address confounding factors. Laboratory Animals 46: 245–252.

Koch et al. 2015. Overcoming problems with the use of ratios as continuous characters for phylogenetic analyses. Zoologica Scripta.

4.8 Q & A

Further reading

References

Annotated Bibliography

Chapter 5

Chapter "Types of Variables"

Commentary

Vocabulary

- Binomial variable
- Continous variable
- Discrete varible
- Interval variable
- Nominal variable
- ordinal variable
- Ratio variable

5.0.1 Motulsky vocab

5.0.2 Aditional vocab

5.0.3 Key functions

None

Chapter Notes

5.1 "Continous Variables"

5.1.1 "Interval variables"

5.1.2 "Ratio variables"

5.2 "Discrete Variables"

5.2.1 "Ordinal variables"

5.2.2 "Nominal and binomial variables"

5.3 "Why It Matters"

5.4 "Not Quite As Distinct As They Seem"

5.5 Q&A

Further reading

References

Annotated Bibliography

Chapter 6

Chapter "Quantifying Scatter"

Commentary

Vocabulary

6.0.1 Motulsky vocab

6.0.2 Aditional vocab

6.0.3 Key functions

None

Chapter Notes

6.1 "Interpretting A Standard Deviation"

6.2 How It Works: Calculating SD"

6.3 "Why n-1?"

6.3.1 "When it somes makes sense to use n in the denominator"

6.3.2 "Why it usually makes sense to use n-1 in the denominator"

6.3.3 "The fine print"

6.4 "Situations in Which n Can Seem Ambiguous"

6.4.1 "Replicate measurements within repae experiments"

6.4.2 "Eyes, ears and elbos"

6.4.3 "Representaitve experiments"

6.4.4 "Trials with one subject"

6.5 "SD and Sample Size"

6.6 "Other Ways to Quantify & Display Variability"

6.6.1 "Coefficient of variation"

6.6.2 "Variance"

6.6.3 "Interquartile range"

6.6.4 "Five-number smmary"

6.6.5 "Median absolute deviation"

6.7 Q&A

Further reading

References

Annotated Bibliography