Statistics, R, and Music

David John Baker

2017-09-07

Contents

1 Introduction to Statistics **5**

2 Sampling Distributions and T-Tests **7**

3 Confidence Intervals and Power **9**
3.1 Confidence Interval Estimation 9
3.2 Example 10
3.3 Computation of Confidece Interval 10

4 Methods **13**

5 Applications **15**
5.1 Example one 15
5.2 Example two 15

6 Final Words **17**

Chapter 1

Introduction to Statistics

What this book will assume

- Basic knowledge of math
- Basic computer literacy
- Include chapter on levels of measuremet? Aka Chapter 1-4 Cohen?
- Ability to look for answers self, working with R you need to be able to troubleshoot

What book plans to accomplish

- Grounding in basic NHST statistics
- Teach basic calculations by hand
- Working knowldge of R for all tests
- Manipulating data in R
- Basic visualization of data in R
- Guide to finding similar values from SPSS

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

For now, you have to install the development versions of **bookdown** from Github:

```
devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need to install XeLaTeX.

Chapter 2

Sampling Distributions and T-Tests

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter 4.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2017) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

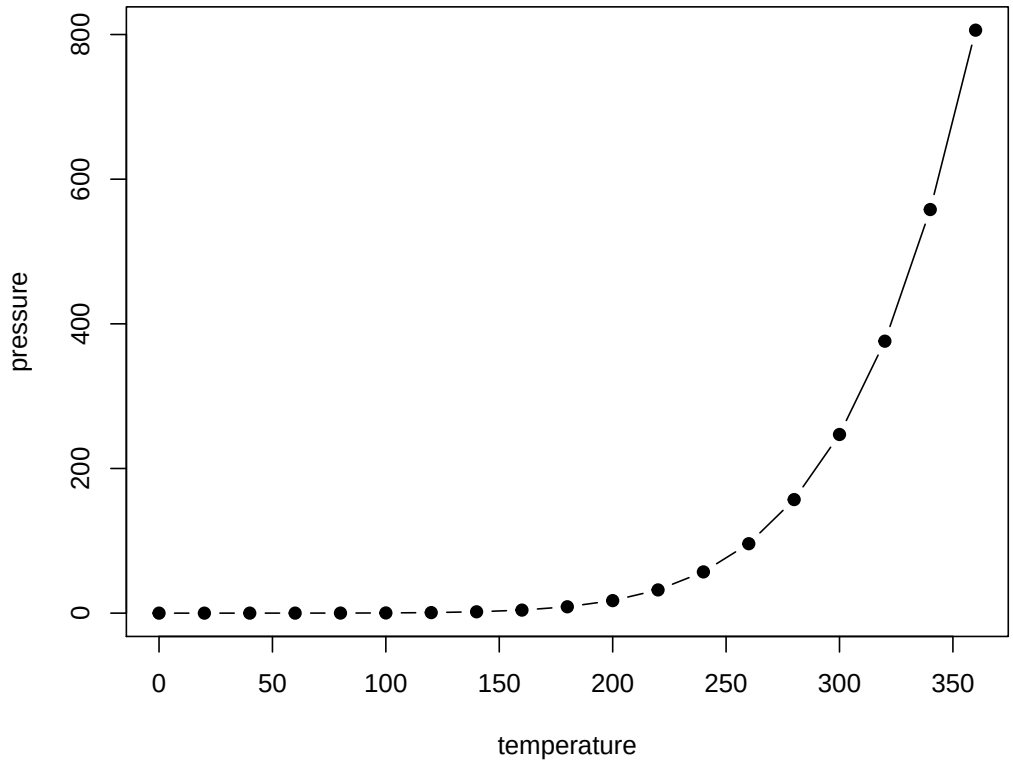

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Confidence Intervals and Power

Class notes from September 5th, 2017 - UNEDITED

3.1 Confidence Interval Estimation

Confidence intervals are a way to get at a range of likely values for an estimate. It's an inferrential way of thinking about values from your sample. It differs from things like the mean, median, and mode in that those are point esitmates and a CI is a range.

We know that over repeated sampling, the CI represents the percentage of sugt intervals that contain the population mean. Just like a p value, you can set the CI to any percentage you want. So just like you can think of a distribution of sample means as a way of describing your parameters, you can also do this with CIs.

A CI gives you an educated guess of identifying the population mean than just knowing one number, which would be the mean itself. Keep in mind for all of this we are operating on the sampling distribution informaiton. Just like the set of t distributions are a set of sampling distributions, CIs are computed from the sampling distributions.

There is a relationship between CIs and p values because at their core is the sampling distribution. They are not the same thing, but they represent information coming from the sampling distribution.

Take a look at the image below. We see here in an image taken from THIS that we can see what a confidence interval is. Image you take a sample from known populaiton paramters. We tell it we want a certain N, mu, and standard deviation. What it then does is construct a CI around that mean. We have a certain amount of information that comes from our sample. Our CI is a symmetrical range that gives you a sense of the possibilites of where the populatin mean might be.

IMAGE WITH ONE

It's not a probability of the mean being in that CI! A CI represents the percentage of intervals that capture the population mean. This only makes sense if you think about repeated sampling over many experiments. You then calculate the number of intervals that contain the actual population mean.

IMAGE BLACK RED

IMAGE WITH ALL OF THEM.

So now below you can see 25 samples. You can start to see teh sampling distribution get created at the bottom They start to form a normal curve. Notice that every sample mean has a symmetrical intervals. Note that there are 2 sample means and their CI, the entire interval does not have the population mean. But for all of the other ones, they do have the population mean. So 23/25 is 92%, but as we add more in

we would eventually add more and approach 95%. Any one of these intervals gives you an estimate of the populatino mean. That does not mean its 95% likely to be right. It's that in repeated samples 95% of the samples will fall in that range. What we have now is a range of values that might hold the mean. 95% of these will have mean in the long run.

Note that with the sampling distribution at the bottom, we are still using information. We still use mean, the standard error. We still know the degrees of freedom.

Note that this is just a visualzation. Note some of the intervals are wider, some are shoter. This is because your samples SD is goign to hop around. The sample's CI takes into account the standard error.

We can see now see the difference in means between two sample means. We now plot the difference from TTESTFROM EARLIER. We subtract the two means and can see in the long run, we had a population mean of 10. But if you look at it, note that the populations hop around. Not only that, but they all have CIs.

With this image, if the Null was true, we would get means around 0. Note that a lot of these random samples from a known population that is not the null, over HALF of them make it look like the null could be a reasonable guess. Just like a NHT, sometimes you are going to get a high P value even though we know the Null is false. Why might this happen? Well it might just be that we need to figure out how to set up.

3.2 Example

Let's return to the example that we are looking at SAT scores with a population mean of 455. We then calculate the 95% confidence interval. What it's trying to do is capture the inside part. Note your CI is not exactly the mean and the cut off boundary, your mean changes. This is just a reminder we use the sampling distribution interval.

SAMPLING DISTRIBUTION SAT IMAGE

3.3 Computation of Confidece Interval

3.3.1 On a single sample

Confidence interval for a single sample mean.

$$\bar{X} \acute{s} (t_{cv} s_{\bar{X}})$$

Note all we are doing is rearranging the t formulat. We multiply both times by the standard error. It's not your actual t value, it's the critical value of t. When you have two samples, you just rearrange the values.

$$(\bar{X_1} - \bar{X_1}) \acute{s} (t_{cv})(s_{\bar{X_1} - \bar{X_1}})$$

Remember we are assuming we are constructing this from the sample means. Let's return to our example from last chapter. We know the specific numbers from each of our sample means.

PUT IN NUMBERS HERE IN TABLE.

Let's find out what happens when we look at each on interms of its own mean and a difference. We don't know the population mean, so we go with the sample mean.

$t_{cv,df=149,\alpha=.05,twotailed} = \acute{s} 1.97$

CALCULATION HERE FOR SPRING AND FALL CLASS! *NOte they are almost identical*

SEE IF YOU CAN COLOR CODE THIS!!!

Note that at this point we are not comparing them against each other. This says beyond just knowing the mean, says interval. This does not mean the population mean is 95% likely to be in that interval. It is if I did this over and over, that 95% of those intervals will hold the sample mean.

The reasons we know this is because it's better than using the mean alone. Report the CI like it is a descritive statistc. Though note that this comes from the sampling distribution, which comes from an infferential.

3.3.2 Differnence

See slides (updated)

Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2017). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.4.9.