Statistics for Scared People

C.M. Curry

2024-02-20

Contents

Who is this book for?

Part I

Questions about your goals

Chapter 1

What is your goal?

1.1 Exploratory or hypothesis generation

1.2 Inferential or hypothesis testing "Are things different"

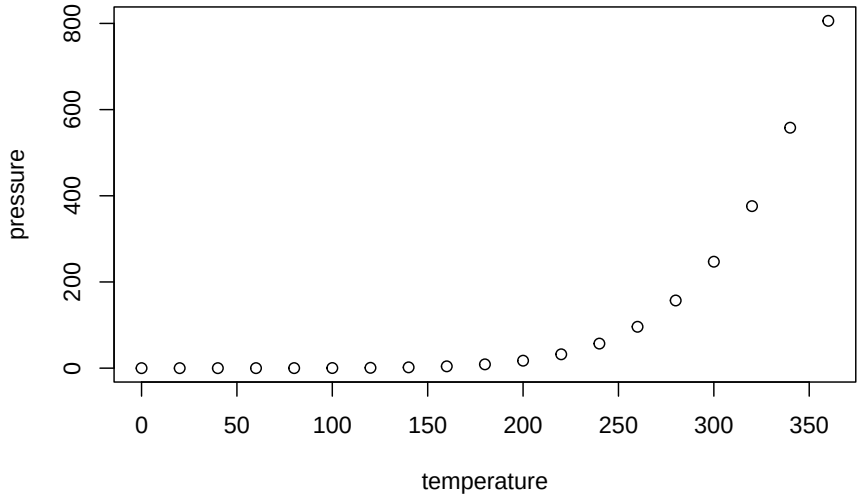
This is a hypothesis, not a description. Description can highlight, but doesn't test what's different. Descriptions can still have a bias (mean vs median vs range all show different things descriptively, PCA problems). Doesn't mean it's an experiment.

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

1.3 Physical or mechanistic predictions - you can only statistics them away sometimes

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Chapter 2

Types of resources

Peer-reviewed vs not: what can you cite? What helps?

Chapter 3

Distributions

3.1 Bounded

3.2 Heteroscedascitity vs homoscedasicity

3.3 Theoretical, existing, known

3.4 Simulated, randomized, computational

3.5 When to use either?

It seems like objections to bootstrapping linear models (and presumably other complex models) fall into two categories: 1. Sampling design isn't accounted for by complete randomization (ignoring stratification of categories or other sampling vagaries) 2. It's less elegant (???).

Venables and Ripley 2002, pg 164, say "we see bootstrapping as having little place in least-squares regression. If the errors are close to normal, the standard theory suffices. If not, there are better methods of fitting than least squares, or perhaps the data should be transformed [...]" Hastie et al. 2008 (Elements of Statistical Learning) seem in favor of bootstrapping

Johnston and Faulkner (2021) are enthusiastically in favor of the bootstrap at least for their relatively simple design to replace a t-test. - https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8613103/#CR47 – says do all the other stuff like deal with random effects and autocorrelation first. We have done this already.

Example: should we run a Redundancy Analysis (RDA) a la https://r.qcbs.ca/workshop10/book-en/redundancy-analysis.html , which I understand has multivariate normality assumptions. He has a small sample size (around 35 I believe) and residuals are not coming out normal in smaller linear models. - https://journals.sagepub.com/doi/10.1177/0049124189018002003 - https://statisticsbyjim.com/hypothesis-testing/bootstrapping/ - https://online.stat.psu.edu/stat555/node/119/ - https://www.sagepub.com/sites/default/files/upm-binaries/21122_Chapter_21.pdf - https://link.springer.com/referenceworkentry/10.1007/978-1-4419-1153-7_84

Part II

Specific tests

How to use this section

Each section will contain some potential descriptions if needed OR direct citations and links to relevant literature if those explanations are clearest.

Principal components analysis

3.6 Explanation.

Cite Allison Horst's whale figure here.

3.6.1 Questions and data types

3.6.2 Key assumptions

3.6.3 Key distinctions among methods within PCA

3.7 email text

3.7.1 CART/ctree explanations

- Start with this one, CART section mainly: [http://www.jstor.org/stable/10.1086/587826](http://w

- [https://stats.stackexchange.com/questions/12140/conditional-inference-trees-vs-traditional-dec

- [https://stats.stackexchange.com/questions/255150/how-to-interpret-this-decision-tree](https://

3.7.2 Examples of PCA in the wild:

- [https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/1051-0761%282006%29016%5B0687%3AH

- Uses R's ctree: [https://link.springer.com/article/10.1007/s11252-019-00896-0](https://link.spr

3.7.3 Once you have decided to use it, check implementation

Supervised learning

3.8 Decision trees/CART/classification tree/regression tree/ctree email text

3.8.1 CART/ctree explanations

- Start with this one, CART section mainly: [http://www.jstor.org/stable/10.1086/587826](http://w

- [https://stats.stackexchange.com/questions/12140/conditional-inference-trees-vs-traditional-dec

- [https://stats.stackexchange.com/questions/255150/how-to-interpret-this-decision-tree](https://

3.8.2 Examples of CART in the wild:

- [https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/1051-0761%282006%29016%5B0687%3AH

- Uses R's ctree: [https://link.springer.com/article/10.1007/s11252-019-00896-0](https://link.spr

What each section has

3.9 Explanation

3.9.1 The basics

A simple explanation and hopefully figure of what the test does or gets at.

3.9.2 More technical

3.9.2.1 Questions and data types

Example problem structures and types of data you need.

3.9.2.2 Key assumptions

This is how to know if you can use the method.

3.9.2.3 Key distinctions among related methods

Within and among methods - related?

3.9.2.4 Implementations and controversies

3.9.3 Most technical

The key citations.

3.10 Examples “in the wild”

Citations and what is useful in the paper.

Bibliography

Johnston, M. G. and Faulkner, C. (2021). A bootstrap approach is a superior statistical method for the comparison of non-normal data with differing variances. *New Phytologist*, 230(1):23–26. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/nph.17159.