Statistics Overview [1]

Sample book subtitle [2]

First-name Last-name[3]

May 27, 2016

[1]Things I want to remember
[2]This is yet another footnote.
[3]`www.example.com`

Dedicated to Calvin and Hobbes.

Contents

List of Figures

List of Tables

Preface

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis risus ante, auctor et pulvinar non, posuere ac lacus. Praesent egestas nisi id metus rhoncus ac lobortis sem hendrerit. Etiam et sapien eget lectus interdum posuere sit amet ac urna.

Un-numbered sample section

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis risus ante, auctor et pulvinar non, posuere ac lacus. Praesent egestas nisi id metus rhoncus ac lobortis sem hendrerit. Etiam et sapien eget lectus interdum posuere sit amet ac urna. Aliquam pellentesque imperdiet erat, eget consectetur felis malesuada quis. Pellentesque sollicitudin, odio sed dapibus eleifend, magna sem luctus turpis.

Another sample section

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Duis risus ante, auctor et pulvinar non, posuere ac lacus. Praesent egestas nisi id metus rhoncus ac lobortis sem hendrerit. Etiam et sapien eget lectus interdum posuere sit amet ac urna. Aliquam pellentesque imperdiet erat, eget consectetur felis malesuada quis. Pellentesque sollicitudin, odio sed dapibus eleifend, magna sem luctus turpis, id aliquam felis dolor eu diam. Etiam ullamcorper, nunc a accumsan adipiscing, turpis odio bibendum erat, id convallis magna eros nec metus.

Structure of book

Each unit will focus on <SOMETHING>.

About the companion website

The website[1] for this file contains:

- A link to (freely downlodable) latest version of this document.
- Link to download LaTeX source for this document.
- Miscellaneous material (e.g. suggested readings etc).

Acknowledgements

- A special word of thanks goes to Professor Don Knuth[2] (for TeX) and Leslie Lamport[3] (for LaTeX).
- I'll also like to thank Gummi[4] developers and LaTeXila[5] development team for their awesome LaTeX editors.
- I'm deeply indebted my parents, colleagues and friends for their support and encouragement.

Amber Jain
`http://amberj.devio.us/`

---

[1] `https://github.com/amberj/latex-book-template`
[2] `http://www-cs-faculty.stanford.edu/~uno/`
[3] `http://www.lamport.org/`
[4] `http://gummi.midnightcoding.org/`
[5] `http://projects.gnome.org/latexila/`

1

Coding Tidbits to remember

I was writing C++ code using Rcpp and wanting to debug the code. This was possible on my Mac in the terminal by doing the following:

1. In the terminal type `R -debugger=lldb`
2. Type `run`
3. This starts an R session. Then load the necessary libraries.
4. Type `c`. This then goes back to the lldb process
5. Type `b foo_function`. This sets the breakpoint at the beginning of the function.
6. Typing `n` will then go through line by line. This video was a pretty good tutorial - `https://vimeo.com/11937905`

2

Probability Theory

> *"This is a quote and I don't know who said this."*
>
> – Author's name, *Source of this quote*

2.1 Exponential Families

A pdf is a member of an exponential family if the pdf can be written as:

$$f(x|\theta) = h(x)c(\theta)exp(\sum_{i=1}^{k} w_i(\theta)t_i(x)) \tag{2.1}$$

So for example, is the normal distribution part of the exponential family? The normal pdf is written as:

$$f(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \tag{2.2}$$

If μ is known, then let $h(x) = \frac{1}{\sqrt{2\pi}} I_{[-\infty,\infty](x)}$, $c(\sigma) = \frac{1}{\sigma}$, $t_1(x) = -(x - \mu)^2$, and $w_1(\sigma) = \frac{1}{2\sigma^2}$. Since we can write this way, the normal with μ known (and it can be shown when σ is known), is part of the exponential family.

To find the natural parameter space we look at whatever is in front of x in the pdf, in the exponent. So in the case of the normal if we foil out $-(x-\mu)^2$ in the exponent we get $\eta_1 = \frac{-1}{2\sigma^2}$ and $\eta_2 = \frac{\mu}{\sigma^2}$. When we look at the values that both μ and σ^2 can take on, we get the natural parameter space $\{(\eta_1, \eta_2) : \eta < 0, -\infty < \eta_2 < \infty\}$.

2.2 Likelihood

Maximum likelihood is actually a pretty simple concept. In statistics we collect data on a population. We want to know what the population looks like, and we assume that it follows some probability distribution. We might have some idea of what the general shape of the distribution looks like and we represent it as $f(x|\theta)$.

The maximum likelihood method wants to use the data to find the optimal parameters that describe the data. For example, say we use maximum likelihood on data that came from a normal distribution. We don't know what the true mean and true standard deviation really are, but we do have data that come from the true distribution with those true parameters. The maximum likelihood technique would then find the best estimate of the true mean and the true standard deviation, based on the given data, the parameters of the distribution the data is most likely to come from.

So given a bunch of observation $x_1, x_2, ..., x_n$, we have for each datapoint the distribution $f(x_1, x_2, ..., x_n|\theta) = (f(x_1|\theta) * f(x_2|\theta)*, ..., f(x_n|\theta)$ if $vecx$ is iid.

3

Generalized Linear Model

The basic generalized linear model is given as:

$$y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.1}$$

where y is Nx1, X is Nxp, β is p x1, and ϵ is Nx1.

3.0.1 Theorems

3.0.2 Hypothesis Testing

3.0.3 Constrained Estimation

3.0.4 Missing Cells Problem

4

Linear Algebra

Three types of functions. Those that go from $\mathbb{R}$ to $\mathbb{R}$, $\mathbb{R}^n$ to $\mathbb{R}$, and $\mathbb{R}^n$ to $\mathbb{R}^m$. Jacobian -

idempotent - a matrix that multiplied with itself, gives itself.

5

Analysis

Injective - Surjective (Onto) - If every element in$y$ in Y has as corresponding element x in X such that $f(x) = y$ Bijective - means

6

Stat 121

R^2 is the percent of variability in y explained by the model.

7

Stat 624

- Monte Carlo Integration/Simulation Studies
  - Look over Monte Carlo Integration paper
    - Monte Carlo Integration has to do when we want to find the expected value (which is an integral) of some function of a random variable
    - We use the LLN to justify just taking the empirical mean of the function evaluated at samples of the distribution
  - Assess Monte Carlo error, for example confidence intervals using the CLT
  - Practice problems - birthday.R, board game simulation, hw 3 with cereal boxes, urn simulation, m+m sim, tack problem
- Welch's stuff?
- Permutation tests
  - The best way I can think of it is if we take two groups (example off of Wiki). We take the means of those two groups. If our statistic is determining the difference between the two means, then we take the difference of those two means. Then we pool all of the observations from both groups into one big group and find all possible groups of the same size as the original groups. We find the difference between the means of these groups and create a distribution of these statistics. We find the p-value by calculating the number of differences that are greater than our original test statistic. If this p-value is small, then this means that the chances of getting a difference that big is really rare if we consider the two groups from the same distribution, so we will reject the null that they come from the same distribution.
- Drawing from a random distribution

  - One method we can use is the probability integral transform. If we want a random observation from some distribution with cdf $F(x)$, then all we need is a random draw from a Uniform and then solve $F(x) = u$.

- Box-Muller transform
- Markov chain Monte Carlo
-
- The idea behind MCMC is to get random draws from a distribution, and therefore we can approximate the distribution
- We start at an initial state, and then have a proposal density that determines what the next state is. If we are doing Metropolis then our proposal density must be symmetric, if we are doing Metropolis-Hastings, it doesn't have to be. Once we have generated a proposal, we accept that proposal with Metropolis $min(1, \frac{p(x^*)}{p(x)})$ or Metropolis-Hastings $min(1, \frac{p(x^*)}{p(x)} \frac{g(x|x^*)}{g(x^*|x)})$. If it isn't accepted then we just stay in the same state