Statistical_Inference_Notes

Coursera Course by John Hopkins University

INSTRUCTORS: Dr. Brian Caffo, Dr. Roger D. Peng, Dr. Jeff Leek

Contents

Intro

Instructor's Note:
"*Statistical inference is the process of drawing conclusions about populations or scientific truths from data. There are many modes of performing inference including statistical modeling, data oriented strategies and explicit use of designs and randomization in analyses. Furthermore, there are broad theories (frequentists, Bayesian, likelihood, design based, . . .) & numerous complexities (missing data, observed and unobserved confounding, biases) for performing inference. A practitioner can often be left in a debilitating maze of techniques, philosophies and nuance. This course presents the fundamentals of inference in a practical approach for getting things done. After taking this course, students will understand the broad directions of statistical inference and use this information for making informed choices in analyzing data.*

All the best,

Brian Caffo"

Statistical inference help us extend beyond a small subset of data to give answers about a population.

Course Description:
"*In this class students will learn the fundamentals of statistical inference. Students will receive a broad overview of the goals, assumptions and modes of performing statistical inference. Students will be able to perform inferential tasks in highly targeted settings and will be able to use the skills developed as a roadmap for more complex inferential challenges.*"

GitHub Link for Lectures

Statistical Inference Lectures on GitHub

Course Book

The book for this course is located on LeanPub

Data Science Specialization Community Site

The site is created using GitHub Pages

Homework Problems

The homework problems are optional, they are a good opportunity to practice the skills covered in the course. There are also worked out solutions on youtube (linked to from the book)

Here's all four homeworks as interactive web pages (it's probably better to just keep up with them from the book):
* **HW 1**
* **HW 2**
* **HW 3**
* **HW 4**

Probability & Expected Values

Introduction to Probability

Intro

Probability assigns a number between 0 and 1 to events to give a sense of the "chance" of the event. These sections will look at the basics of probability calculus.

An additional resource is the class Mathematical Biostatistics Boot Camp 1
Probability

Given a random experiment (i.e. rolling a die) a probability measure is a population quantity that summarizes the randomness

Specifically, probability takes a possible outcome from the experiment and:
* assigns it a number between 0 and 1
* so that the probability that something occurs is 1 (the die must be rolled)
* so that the probability of the union of any two sets of outsomes that are mutually exclusive is the sum of their repective probabilities (`P(E|F) = P(E) + P(F)`)

Rules probability must follow

- The probability that nothing occurs is 0
- The probability that something occurs is 1
- The probability of something is 1 minus the probability that the opposite occurs, the conjugate
- If an event **A** implies the occurrence of event **B**, then the probability of **A** occuring is less than or equal to the probability that **B** occurs
- For any two events the probability that at least one occurs is the sum of thier probabilites minus their intersection (`P(E|F) = P(E)+P(F)-P(E&F)`)

Probability mass functions

- Probability calculus is useful for understanding the rules that probabilities must follow.
- We need ways to model and think about probabilites for numeric outcomes of experiments
  + Densities and mass functions for random variables are the best starting point for this
- The goal is to use the data to estimate properties of the population
- A **random variable** is a numeric outsome of an experiment and can be **discrete** or **continuous**
  + For **discrete** a probability can be assigned for every value it can take
  + for **continous** a probability can be assigned for the ranges of values it can take

Some examples of variables that can be seen as random variables
* The outcome of the flip of a coin (discrete)
* The outcome from the roll of a die (discrete)
* The web site traffic on a given day
+ Since this discrete variable has no upper bound we'd view it as a poisoon distribution
* The BMI of a subject four years after a baseline measuremnt (continous)
* The hypertension status of a subject randomly drawn from a population (binomial discrete variable)
* The number of people who click on an ad (discrete; poisson)
* Intelligence quotients for a sample of children (continous)

PMF

- A **Probability mass function** evaluated at a value corresponds ot the probability that a random variable takes that value. To be a valid pmf a function, p, must satisfy:
  1) It must always be larger than or equal to 0
  2) The sum of the possible values that the random variable can take has to add up to one

Example: Coin Flip
$X = 0$ represents tails and $X = 1$ represents heads
$p(x) = (1/2)^x * (1/2)^{1-x} for x = 0, 1$
And for a loaded coin this could be generalized as
$p(x) = \theta^x * (1-\theta)^{1-x} for x = 0, 1$ where θ represents propabilty of heads
When evaluating this we get..
Probability of heads is $p(1) = \theta^1 * (1-\theta)^{1-1} = \theta$ and
Probability of tails is $p(0) = \theta^0 * (1-\theta)^{1-0} = 1-\theta$

Probability densisty functions

- A **probability density function** (pdf), is a function associated with a continouous random variable

- To be a valid pdf, a function must:

1) Be larger than or equal to zero everywhere

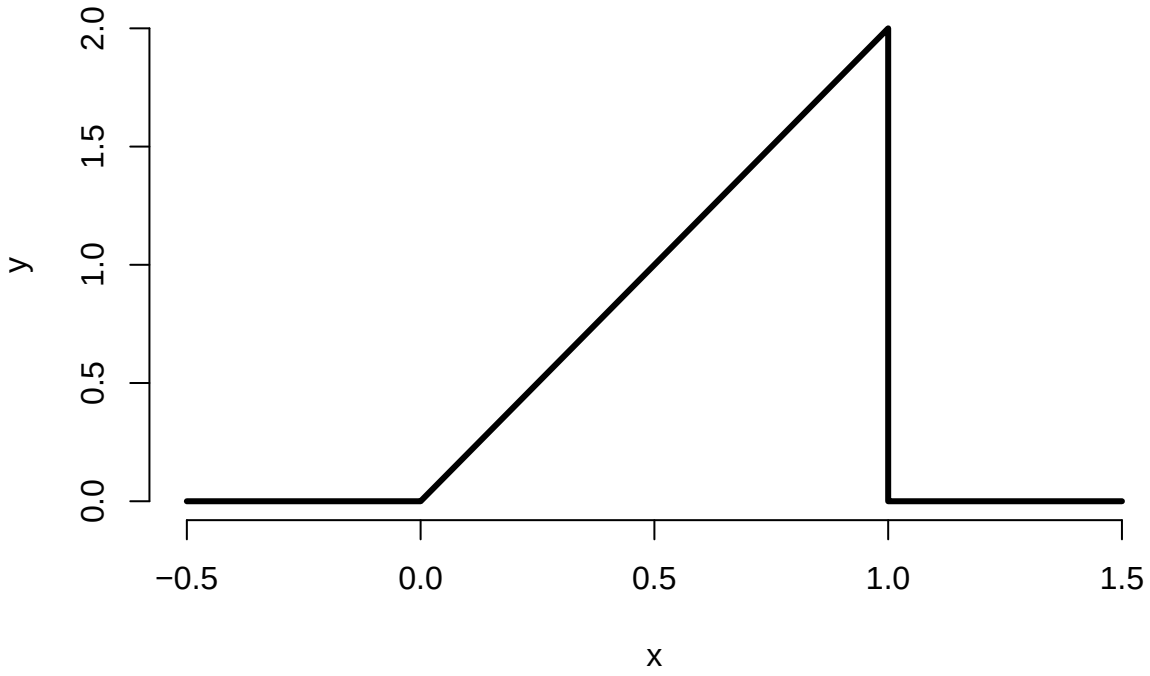
2) The total area under it must be one

- Areas under pdfs correspond to probabilities for that random variable

Example with beta density (triangle & piecewise fn)

$$f(x) = \begin{cases} 2x & for \quad 0 < x < 1 \\ 0 & otherwise \end{cases}$$

```
x <- c(-0.5, 0, 1, 1, 1.5)
y <- c(   0, 0, 2, 0,   0)
plot(x, y, lwd = 3, frame = FALSE, type = "l")
```



To check if this is a valid PDF we can calculate the total area, using geometry we have a triangle with a *base of 1* and a *height of 2* $A = 1/2bh = h/2 * b = 2/2 * 1 = 1$, therefore it statisfies the rules for a PDF.

$\int_0^1 2\text{xd}x = [x^2]|_0^1 = 1^2 - 0^2 = 1$

Assume this pdf is for the porportion of health calls that get addressed in a given day. What's the probability that 75% or fewer of calls get addressed?
$\int_0^{0.75} 2\text{xd}x = [x^2]|_0^{0.75} = 0.75^2 - 0^2 = 0.5625$
The beta function is also a function in R:

```
#first param is for the quantile
#second param is the height of the distrbution
#third param is the width of the dist.
pbeta(0.75, 2, 1)
```

```
## [1] 0.5625
```

Cumulative Distribution Function (CDF) and Survival Function

The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random varaible is less than or equal to the value of x
$F(x) = P(X \leq x)$

The **survival function** of a random variable X is defined as the probability that the random variable is greater than the value x
$S(x) = P(X > x)$
Notice that $S(x) = 1 - F(x)$

We can evaluate multiple quantiles at once with `pbeta`

```
pbeta(c(0.4, 0.5, 0.6), 2, 1)
```

```
## [1] 0.16 0.25 0.36
```

Quantile

The α^{th} **quantile** of a distribution with distribution function F is the point x_α such that $F(x_\alpha) = \alpha$
* A **percentile** is simply a quantile with α expressed as a percent
* The **median** is the 50^{th} percentile
* The `qbeta` function will take a quantile and return the value of x_α

```
sqrt(.5) #Solving x^2 = 0.5 manually
```

```
## [1] 0.7071068
```

```
qbeta(0.5, 2, 1)
```

```
## [1] 0.7071068
```

Conditional Probability

- Conditional Probability is a probability of an event (E) given some condition (F) represented as $P(E|F)$ (Probability of E given F)
- The probability of a die landing on 1 is 1/6, however the probability of a die landing on a 1 *given* the die landed on an odd number is 1/3
- The general formula for conditional probability is $P(A|B) = \frac{P(A \bigcap B)}{P(B)}$

Bayes' rule

- Named after **Thomas Bayes**
- Used to find $P(B|A)$ when one knows $P(A|B)$, however one also has to know $P(B)$ and $P(B^c)$
$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^c)P(B^c)}$

Diagnostic Tests

Bayes' rule is useful in diagnostic tests
* Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative repectively
* Let D and D^c be the event that the subject of the test has or does not have the disease repectively
* Then the **sensitivity** of the test can be evaluated as $P(+|D)$, the probability the test is positive given the subject has the disease
* Likewise, the **specificity** of the test is evaluated as $P(-|D^c)$, the probability the test is negative given the subject does not have the disease
+ A good test has high specificity
* **Positive predictive value** - probability of having a disease given a positive test, $P(D|+)$
* **Negative predictive value** - probability of not having the disease given a negative test, $P(D^c|-)$
* **Prevalence of the disease** - just the probability of having the disease, $P(D)$

Example:
Say there is a test for HIV such that it has...
* a sensitivity of 99.7%; $P(+|D) = 0.997$
* a specificity of 98.5%; $P(-|D^c) = 0.985$
* population has a prevalence of HIV of 0.1%; $P(D) = 0.001$
What is the associated positive predictive value, $P(D|+)$?
$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D)+P(+|D^c)P(D^c)}$
$P(D|+) = \frac{0.997*0.001}{0.997*0.001+(1-P(-|D^c))*(1-P(D))}$
0.997 * 0.001 = 9.97×10^{-4}
$P(D|+) = \frac{0.000997}{0.000997+(1-0.985)*(1-0.001)}$
1-0.985 = 0.015
1-0.001 = 0.999
0.015 * 0.999 = 0.014985
$P(D|+) = \frac{0.000997}{0.000997+0.014985}$
$P(D|+) = \frac{0.000997}{0.000997+0.014985}$
$P(D|+) = \frac{0.000997}{0.015982}$
$P(D|+) = 0.06238...$
So the positive predictive value is about 6.2%

The low prevalence in the population is the reason for the low positive predictive value.

We can also look at the probability of not having the disease given a positive test,
$P(D^c|+) = \frac{P(+|D^c)P(D^c)}{P(+|D^c)P(D^c)+P(+|D)P(D)}$
It can be seen that the denominator is equivalent to the denominator is $P(D|+)$ through the communitative property
$P(D^c|+) = \frac{P(+|D^c)P(D^c)}{P(+|D)P(D)+P(+|D^c)P(D^c)}$ as such we can evaluate this as so:
$P(D^c|+) = \frac{(1-P(-|D^c))(1-P(D))}{0.015982}$
$P(D^c|+) = \frac{(1-0.985)(1-0.001)}{0.015982}$
1-0.985 = 0.015 ; 1-0.001 = 0.999 0.015 * 0.999 = 0.014985
$P(D^c|+) = \frac{0.014985}{0.015982}$
$P(D^c|+) = 0.9376...$

Likelihood Ratio

But we are more interested in the **likelihood ratio**, that is given a positive test how does this increase your "*odds*" of having the disease. * **odds** are the ratio of $\frac{P(E)}{P(E^c)}$

In the case of the example we'd want to find the Diagnostic likelihood ratio for a positive test result; how does getting a positive result affect your odds. This would be expressed as such:
$\frac{P(D|+)}{P(D^c|+)}$
Since the denominators of these two are equivalent this would simplify to:
$\frac{P(+|D)}{P(+|D^c)} * \frac{P(D)}{P(D^c)}$
The first expression is your Diagnostic Likelihood Ratio, the value your pre-test odds, $\frac{P(D)}{P(D^c)}$,are multiplied by for recieving a positive test result.

$DLR_+ = \frac{0.997}{(1-0.985)} \approx 66$
Meaning that the odds of you having the disease after a positive test result is 66 times more than before the test.

Independence

Event A is independent of event B if $P(A|B) = P(A)$ where $P(B) > 0$
As well as if $P(A \cap B) = P(A)P(B)$
As such $P(A_1 \cap A_2)$ if & only if A_1 and A_2 are independent events
* **Independent Identically Distributed (iid)** - random variables that are independent and identically distribtued
+ Independent - statistically unrelated from one and another
+ Identically distributed - all having been drawn from the same population distribution
* iid random variables are the default model for random samples

Reminder to commit (03), delete this line *AFTER* committing

Expected values

Expected values, simple examples

Expected values for PDFs

Reminder to commit (04), delete this line *AFTER* committing

Lessons with `swirl()`

Quiz 1

Reminder to commit (S1), delete this line *AFTER* committing

Variability, Distribution, & Asymptotics

Introduction to Variability

Variance Simulation Examples

Standard Error of the Mean

Variance Data Example

Reminder to commit (05), delete this line *AFTER* committing

Binomial Distrubtion

Normal Distribution

Poisson

Reminder to commit (06), delete this line *AFTER* committing

Asymptotics and the Law of Large Numbers (LLN)

Asymptotics and the Central Limit Theorem (CLT)

Asymptotics and Confidence Intervals

Reminder to commit (07), delete this line *AFTER* committing

Lessons with `swirl()`

Quiz 2

Reminder to commit (S2), delete this line *AFTER* committing

Intervals, Testing, & P-values

T Confidence Intervals

T Confidence Intervals Example

Independent Group T Intervals

A Note on Unequal Variance

Reminder to commit (08), delete this line *AFTER* committing

Hypothesis Testing

Example of Choosing a Rejection Region

T Tests

Two Group Testing

Reminder to commit (09), delete this line *AFTER* committing

P-Values

P-Value Further Examples

Reminder to commit (10), delete this line *AFTER* committing

Lessons with `swirl()`

Quiz 3

Reminder to commit (S3), delete this line *AFTER* committing

Power, Bootstrapping, & Permutation Tests

Power

Calculating Power

Notes on Power

T Test Power

Reminder to commit (11), delete this line *AFTER* committing

Multiple Comparisons

Reminder to commit (12), delete this line *AFTER* committing

Bootstrapping

Bootstrapping Example

Notes on the Bootstrap

Permutation Tests

Reminder to commit (13), delete this line *AFTER* committing

Lessons with `swirl()`

Quiz 4

Reminder to commit (S4), delete this line *BEFORE* committing