# 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)`)

## Lesson with `swirl()`: Introduction

(No new content)

## 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)

## Lesson with `swirl()`: Probability1

(No new content. Went over some basic probability situations, such as drawing cards from a deck)

### 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 $\theta$ 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

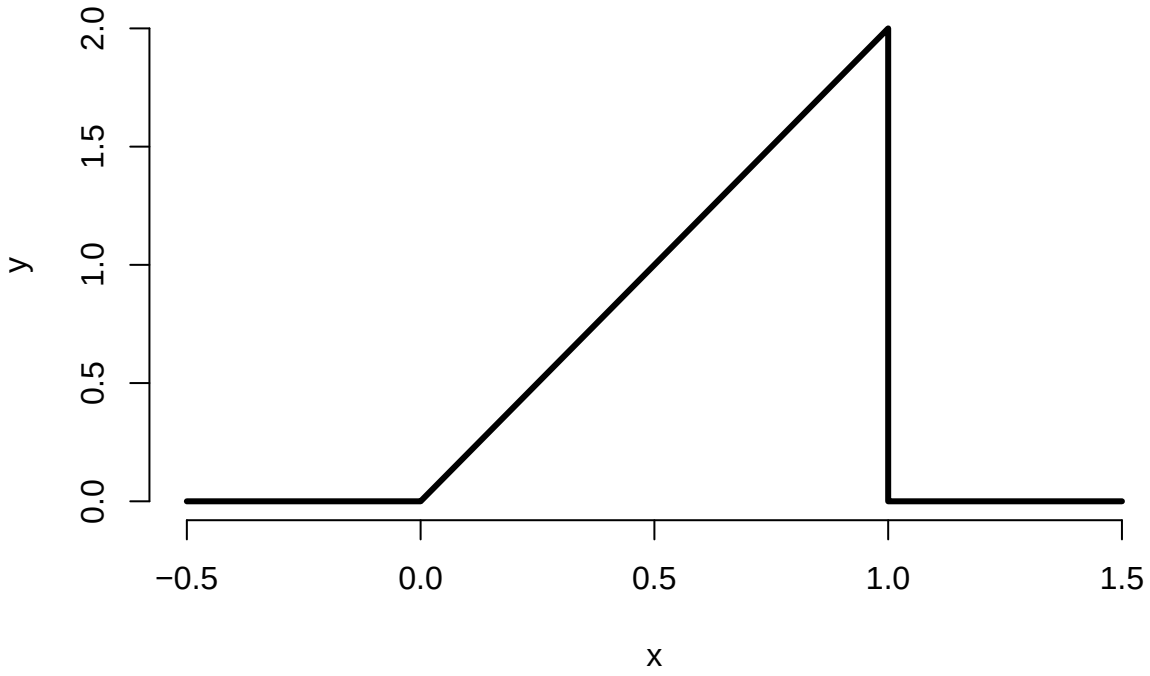
2) The total area under it must be one

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

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

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

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

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

## Lesson with `swirl()`: Probability2

- Continous random variables are usually associated with measurements of time, disance, or some biological process that can take any value
  - Limitations of precision in taking the measurements may imply the values are discrete, but we consider them continuous.

- A sample median is an estimato of a population median (the estimand)

## 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 = $9.97 \times 10^{-4}$
$P(D|+) = \frac{0.000997}{0.000997+(1-0.985)*(1-0.001)}$
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.

## Lesson with `swirl()`: Conditional Probability

- $P(B|A) = \frac{P(B \cap A)}{P(A)} = P(A|B) * \frac{P(B)}{P(A)}$
  - This is a simple form of Bayes' rule

- Suppose we don't know $P(A)$, but only know its conditional probabilities, $P(A|B)$ and $P(A|B^c)$
  - We could deduce $P(A) = P(A|B) * P(B) + P(A|B^c) * P(B^c)$ since this would essentially be the sum of $P(A \cap B) + P(A \cap B^c)$ which with discrete mathematics would be reduced to just $P(A)$ since it's the intersection of A with all of B and all of *not* B, resulting in only A.
  - This is why $P(A|B) * P(B) + P(A|B^c) * P(B^c)$ is the denominator in Bayes' rule

## 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

### Expected values

- Another term for the pop. mean of a random variable
- The sample mean can be thought of as the center of mass of data if each data point is equally likely
- However, since not all points are equally likely the mean is found by getting the expected value of each discrete variable

```
PMtable <- data.frame(value = c(0:4), prob = c(0.1,0.15,0.4,0.25,0.1))
PMtable
```

```
##   value prob
## 1     0 0.10
## 2     1 0.15
## 3     2 0.40
## 4     3 0.25
## 5     4 0.10
```

```
#We can see this is a valid dist. since the probabilites sum to 1
sum(PMtable$prob)
```

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

```
#To get the mean we can get the expected value of each variable...
EVs <- PMtable$value * PMtable$prob
PMtable <- cbind(PMtable, EVs)
PMtable
```

```
##   value prob  EVs
## 1     0 0.10 0.00
## 2     1 0.15 0.15
## 3     2 0.40 0.80
## 4     3 0.25 0.75
## 5     4 0.10 0.40
```

```
#Then take the sum of the Expected values
sum(PMtable$EVs)
```

```
## [1] 2.1
```

```
#This is the same as if we had a sample population with a rel. freq
#that is the same as the probabilities
rFreq <- PMtable$prob*100
sample <- c(rep(0, rFreq[1]),
            rep(1, rFreq[2]),
            rep(2, rFreq[3]),
```

```
            rep(3, rFreq[4]),
            rep(4, rFreq[5]))
mean(sample)
```

```
## [1] 2.1
```

The population mean, denoted $E[X]$, represents the center of mass for a collection of locations, $x$, and weights, $p(x)$
$E[X] = \sum_x xp(x)$

### Expected values, simple examples

- The expected value can be a value that the discrete variables can't take, but represents the center of mass of the areas of each results' probabilities
- In an example for a die the expected value would be:

```
die <- data.frame(value = 1:6, prob = rep(1/6,6))
sum(die$value*die$prob)
```

```
## [1] 3.5
```

### Expected values for PDFs

- Similar to the PMF the pop. mean of the PDF is akin to the center of mass of it's density function
- The sample mean will be centered on the same as the original, population mean.
  + When this occurs, the sample mean is said to be **unbiased** because its distribution is centered at what it's trying to estimate
- I made **a good graph on desmos** that helps visualize this back when I was tutoring. As the sample size increases the probability of the sample mean equaling the population mean tends towards 100%

### Summary

- Expected values are properties of distibutions
- The population mean is the center of mass of population
- The sample mean is the center of mass of the observed data
- The sample mean is an estimate of the population mean

- The sample mean is unbiased
  + The population mean of its distribution is the mean that it's trying to estimate

- The more data that goes into the sample mean, the more concentrated its density / mass function is around the population mean

### Lesson with `swirl()`: Expectations

(No new content)

### Quiz 1

1. Consider influenza epidemics for two parent heterosexual families. Suppose that the probability is 17% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 12% while the probability that both the mother and father have contracted the disease is 6%. What is the probability that the mother has contracted influenza?

```
#P(M or F) = P(M) + P(F) - P(M&F)
#0.17 = P(M) + 0.12 - 0.06
#P(M) = 0.17 - 0.12 + 0.06
0.17 - 0.12 + 0.06
```

```
## [1] 0.11
```

2. A random variable, X is uniform, a box from 0 to 1 of height 1. (So that its density is f(x) = 1 for 0 <= x <= 1. What is its 75th percentile?

- It's a uniform density so it's just at 0.75

3. You are playing a game with a friend where you flip a coin and if it comes up heads you give her X dollars and if it comes up tails she gives you Y dollars. The probability that the coin is heads is p (some number between 0 and 1.) What has to be true about X and Y to make so that both of your expected total earnings is 0. The game would then be called "fair".

- -X * p + (1-p) * Y = 0

4. A density that looks like a normal density (but may or may not be exactly normal) is exactly symmetric about zero. (Symmetric means if you flip it around zero it looks the same.) What is its median?

- if the data is symetric about 0 the median must be 0

5. What is the mean of the following PMF

```
data <- data.frame(X = 1:4, Prob = (1:4)*0.1)
sum(data$X*data$Prob)
```

```
## [1] 3
```

6. A web site (www.medicine.ox.ac.uk/bandolier/band64/b64-7.html) for home pregnancy tests cites the following: "When the subjects using the test were women who collected and tested their own samples, the overall sensitivity was 75%. Specificity was also low, in the range 52% to 75%." Assume the lower value for the specificity. Suppose a subject has a positive test and that 30% of women taking pregnancy tests are actually pregnant. What is the probability of pregnancy given the positive test?

```
sens <- 0.75; spec <- 0.52; preg <- 0.3
pregGivenPos <- (sens * preg) /
                  ((sens * preg) + (1 - spec) * (1 - preg))
pregGivenPos
```

```
## [1] 0.4010695
```

**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**