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 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)

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

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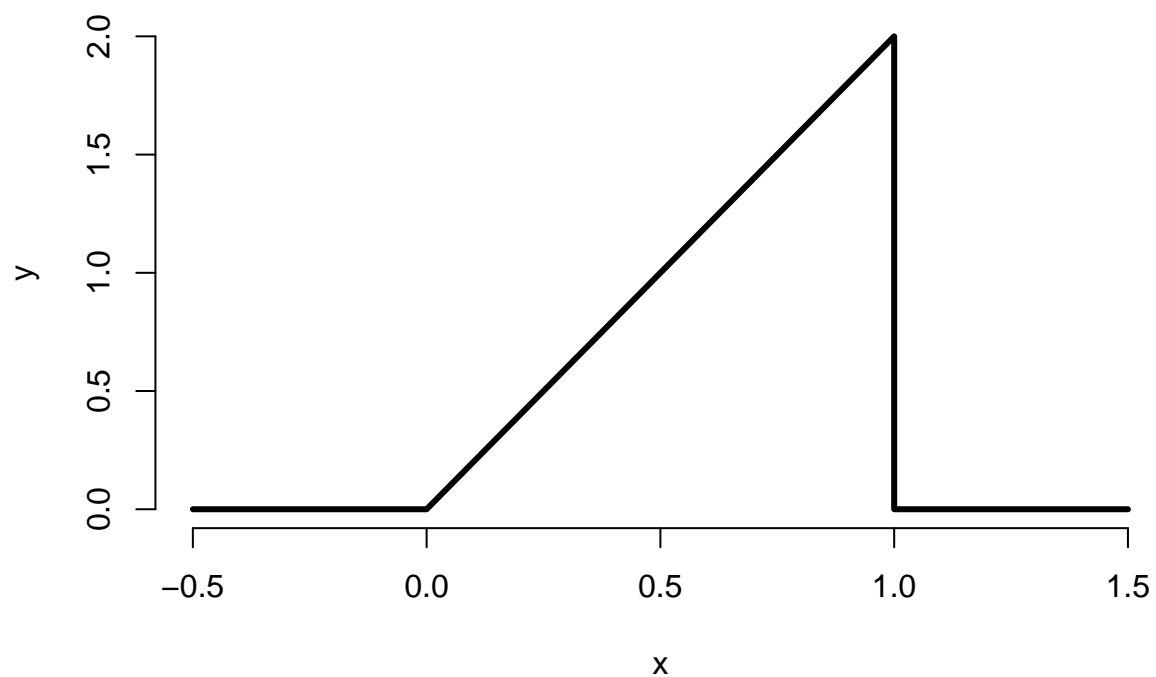
To check if this is a valid PDF we can calculate the total area, using geometry we have a triangle with a *base of 1* and a *height of 2* $A = 1/2bh = h/2 * b = 2/2 * 1 = 1$, therefore it statisfies the rules for a PDF.

Me using integrals for primary school math problem

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

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?

$\int_0^{0.75} 2\mathrm{x}\mathrm{d}x = [x^2]|_0^{0.75} = 0.75^2 - 0^2 = 0.5625$

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
$F(x) = P(X \le x)$

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 αth **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 50th 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
```

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

```r
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
```

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

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

```r
#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
```

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

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

```r
#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
```

Variability, Distribution, & Asymptotics

Introduction to Variability

- Variability is a measuremnt of a sample's variability

- Often we consider the square root of population variability and sample variance, called the standard deviation
    - This is done beecuase the std. dev. has the same units as the population (i.e. if the population is a length in meters, the standard deviation is also in meters (whereas the variance is in meters^2))

- The variance is the sum of the differences of each data point to the mean, squared.
    - $Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2$

Die Example:

$E[X] = 3.5$

$E[X^2] = 1^2 * \frac{1}{6} + 2^2 * \frac{1}{6} + 3^2 \ldots$

$\ldots *\frac{1}{6} + 4^2 * \frac{1}{6} + 5^2 * \frac{1}{6} \ldots$

$\ldots +6^2 * \frac{1}{6} = \ldots$

```
expectedXsquared <- round((1^2+2^2+3^2+4^2+5^2+6^2)*1/6, 2)
expectedX <- sum(1:6*1/6)
```

$= 15.17$

$Var(X) =$ expectedXsquared - expectedX$^2 = 2.92$

- As the variance increases in a normal distribution the curve will flatten as the values are spread more about the mean

- The **sample variance** is the sum of the variation from the mean squared of each data point, then divided by $n - 1$

– $S^2 = \frac{\Sigma_{i-1}(X_i - \bar{X})}{n-1}$

– When taking a sample and measuring the variance we are estimating the population's variance. Since the sample is likely to miss outliers in the population, dividing by $n-1$ increases the value of the variaence, and gets a better estimate of the true population variance. If a sample of a population is taken where the population mean is known, the variancce calculated with the sample mean will always be smaller than the variance calculated with the population mean; since the sample mean will always be generally closer to the sample. Dividing by $n-1$ couteracts this downward bias.

Variance Simulation Examples

```
set.seed(1618033)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.0     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   0.8.5
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
ten <- replicate(1618, sd(sample(1:6, 10, TRUE)))
twenty <- replicate(1618, sd(sample(1:6, 20, TRUE)))
thirty <- replicate(1618, sd(sample(1:6, 30, TRUE)))
data <- rbind(cbind(ten, rep("10", length(ten))),
              cbind(twenty, rep("20", length(twenty))),
              cbind(thirty, rep("30", length(thirty))))
data <- as.tbl(data.frame(data))
data <- rename(data, Standard.Deviation = ten, n = V2) %>%
          mutate(n = as.factor(n),
                 Standard.Deviation =
                   as.numeric(as.character(Standard.Deviation)))
plot <- ggplot(data, aes(Standard.Deviation)) +
  geom_histogram(binwidth = 0.15)  +
  facet_wrap(n~.) +
  geom_vline(xintercept = sqrt(2.92), col = "#FF0000")
plot
```
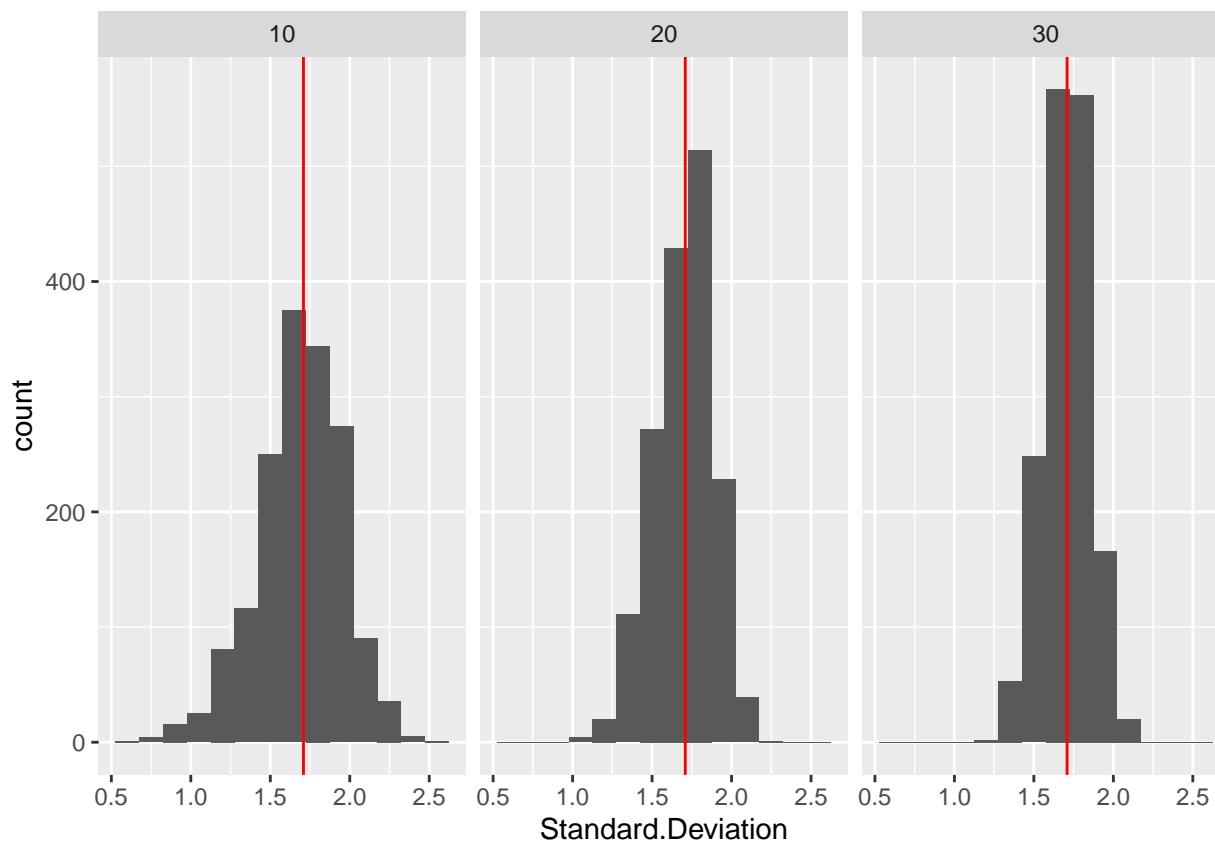
The above code shows the distribution of the standard deviation (sqrt(variance)) for a sample of dice rolls, sample size is 10, 20, and 30. It can be seen that with all samples the mean of the standard deviations tends towards the true population mean that we are estimating with each sample. The sd's get more concentrated around the true pop. sd as the samples are comprised of more dice per sample.

Standard Error of the Mean

- The variance of the sample mean is the population variance divided by the sample size
    - σ^2/n

- As such the standard deviation of the sample mean, **the standard error**, is the population standard deviation divided by the square root of the sample size
    - $\sigma/\sqrt{n}$

- The mean of the sample mean is equivelant to the population mean

Simulations

```
set.seed(1618033)
numSims <- 10000
```

```r
n <- 10
sd(apply(matrix(#v# 1 param indicates the pop sd
  rnorm(numSims * n), numSims), 1, mean))
```

```
## [1] 0.3132377
```

```r
1/sqrt(n)
```

```
## [1] 0.3162278
```

```r
sd(apply(matrix(# runif is a uniform distribution
  runif(numSims * n), numSims), 1, mean))
```

```
## [1] 0.09108488
```

```r
#standard uniform distributions have a variance of 1/12
#sqrt(1/12)/sqrt(n) = 1/sqrt(12*n)
1/sqrt(12 * n)
```

```
## [1] 0.09128709
```

```r
#Poisson(4) have a variance of 4; so a random sample of
#n Poisson(4) have a sd of 2/sqrt(n)
sd(apply(matrix(rpois(numSims * n, 4), numSims), 1, mean))
```

```
## [1] 0.6340419
```

```r
2/sqrt(n)
```

```
## [1] 0.6324555
```

```r
#Fair coin flips have variance of 0.25, so a random sample of
#n coin flips have sd of 1/sqrt(4*n) or 1/(2*sqrt(n))
sd(apply(matrix(sample(0:1, numSims * n, replace = TRUE),
                numSims), 1, mean))
```

```
## [1] 0.1587619
```

```r
1/(2*sqrt(n))
```

```
## [1] 0.1581139
```

Variance Data Example

```r
library(UsingR)
```

```r
data(father.son)
x <- father.son$sheight
n <- length(x)

plot <- ggplot(father.son, aes(sheight)) +
  geom_histogram(aes(y = ..density..), binwidth = 1,
```
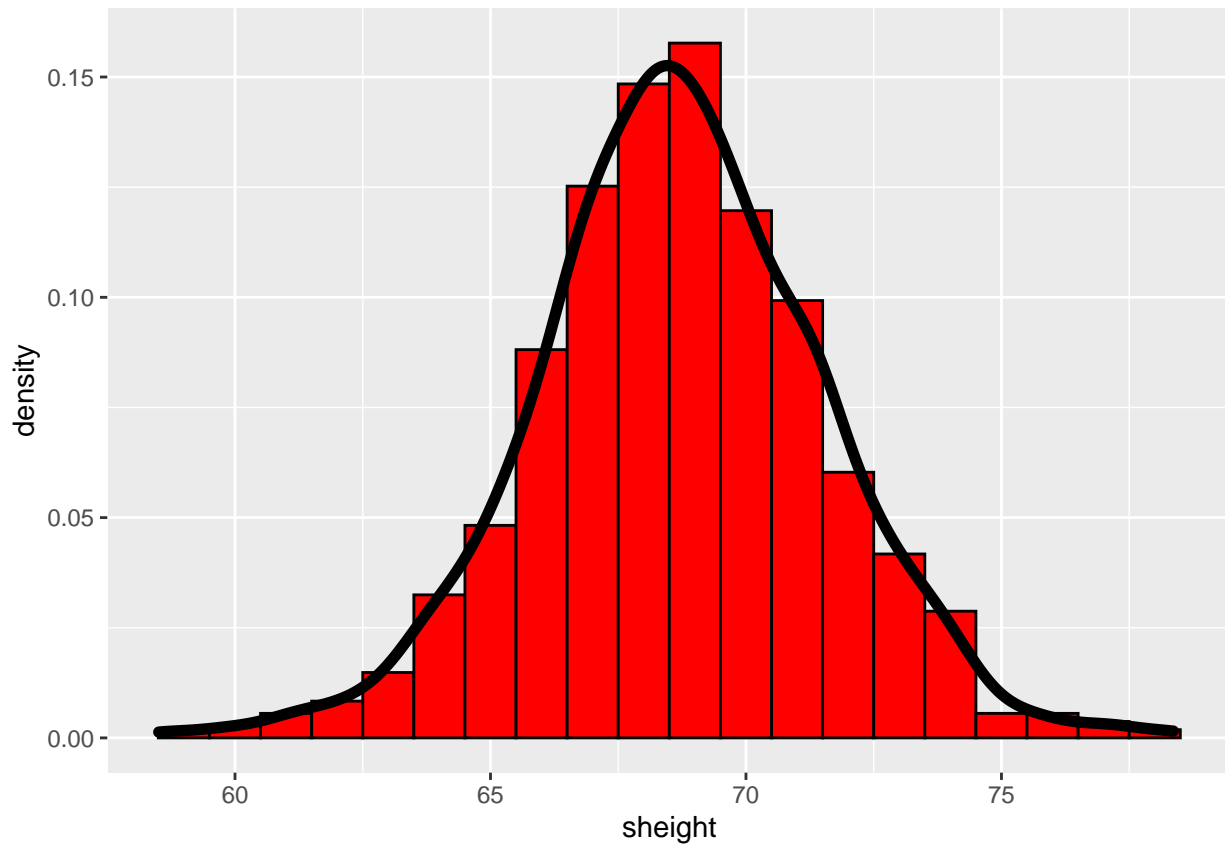
```
                fill = "#FF0000",colour = "#000000")
plot + geom_density(size = 2, colour = "#000000")
```


- The variability of this histogram estimates the variability of son's heights in the population

```
data.frame(Sample.Variance = var(x), Sample.SD = sd(x),
           Variability.Of.Avg = var(x)/n, SD.Of.Avg = sd(x)/sqrt(n))
```

```
##   Sample.Variance Sample.SD Variability.Of.Avg  SD.Of.Avg
## 1        7.922545  2.814702          0.0073493 0.08572806
```

- Both `X.Of.Avg` is an estimate, but it's still the best estimate we can get given the sample data

Summarizing What We Know About Variances

- The sample variance estimates the population variance

- The distribution fo the sample variance is centered at what its estimating
    - It gets more concentrated around the population variance with larger sample sizes

- The variance of the sample mean is the population variance divided by n
    - $Var(\bar{X}) = \sigma^2/n$

Lesson with `swirl()`: Variance

- **Chebyshev's inequality** - Probability that a random variable X is *at least* k standard deviations from it's mean
  - $P(X < (\mu + k * \sigma)) < \frac{1}{k^2}$

Binomial Distrubtion

- Formally called **The Bernoulli Distribution**, named after **Jacob Bernoulli** (whom sided with Leibniz).

- The distribution arrises from a coin flip, that may be biased

- Bernoulli random variables take (only) the values 1 and 0 with probabilities of p and $1 - p$ repectively
  The probability mass function:

$P(X = x) = p^x(1 - p)^{1-x}$

- The mean is p and the variance is $p(1 - p)$

- A 1 is considered a "success" regardless if the case of a 1 is successful or not.

Binomial Trials

A binomial random variable is the sum of successes in a trial
* In specific, let $X_1, ..., X_n$ be independent and identically distributed (iid) random variables of a Bernoulli distribution
then $X = \Sigma_{i=1}^n X_i$ is a binomial random variable
The PMF:

$P(X = x) = \binom{n}{x}p^x(1 - p)^{n-x}$

Note on notation:
$\binom{n}{x}$
(read as "n choose x") counts the number of ways of selecting x items out of n without replacement disregarding the order of the items
$\binom{n}{x} = \frac{n!}{x!(n-x)!}$
$\binom{n}{0} = \frac{n!}{0!(n-0)!} = \frac{n!}{1*n!} = 1$
likewise:
$\binom{n}{n} = \frac{n!}{n!(n-n)!} = \frac{n!}{n!*1} = 1$
since $0! = 1$

Example

Suppose a friend has 8 children, 7 of which are girls. If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?
$\binom{8}{7}0.5^7(1 - 0.5)^1 + \binom{8}{8}0.5^8(1 - 0.5)^0$

$$\frac{8!}{7!(8-7)!} * 0.5^7 * 0.5 + 1 * 0.5^8 * 1$$

$$\frac{8!}{7!*1} * 0.5^8 + 0.5^8 = 8 * 0.5^8 + 0.5^8$$

$$(8+1) * 0.5^8 = \frac{9}{2^8} \approx 0.035$$

In R:

```
#Explicitly doing the formula:
choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
```

[1] 0.03515625

```
#Using binom function:
pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
```

[1] 0.03515625

Normal Distribution

- Also called a **Gaussian distribution** with a mean μ and variance σ^2
    - Sometimes expressed as: $X \sim N(\mu, \sigma^2)$
- It's density is the bell curve, represented by the formula:
  $(2\pi\sigma^2)^{-1/2}e^{-(x-\mu)^2/2\sigma^2}$

Standard Normal Distribution

Normal Distribution when $\mu = 0$ and $\sigma = 1$
Standard normal random variables are often labeled z, which will come up when looking at *z-scores*
* There are a number of properties that are found using the standard normal distributions, since the values 1, 2, and 3 respectively represent 1σ, 2σ, and 3σ.
+ Any Normal distribution can have these properties too just being centered at the mean (i.e. $\mu + 1\sigma$, $\mu + 2\sigma$ …)

Emperical Rule

- The area within z * σ from the mean represents _% of the mean:
    - 1 : 68%

    - 2 : 95% (Really 1.96, but often rounded up)

    - 3 : 99.7%

Z scores

- A z-score represents how many σ a given point is from the mean

- Converting

21

- If $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma}$ $N(0,1)$

- If Z is standard normal then $X = \mu + Z\sigma$ $N(\mu, \sigma^2)$
  Notable percentiles:

- **X** is the XXth percentile
  - Negatives
    -2.33 : 1st
    -1.96 : 2.5^{th}
    -1.645 : 5th
    -1.28 : 10th

  - Positives
    1.28 : 90th
    1.645 : 95th
    1.96 : 97.5^{th} **2.33** : 99th

- Remembering one set will essentially allow you to reference the other set since SND is symetric
  - Some off-hand mnumonics I came up with
    * 1.28 relates to 10% since 2+8 = 10

    * 1.64**5** relates to **5**%, aside from that I have the golden ratio memorized and 5% is like the statistical "golden standard",so this starts simular, **1.6**18, then half of *8* is **4** giving the **1.645** (Not the best but meh)

    * 97.5% is like a step up from 95%, so it's number is 1.**96**

    * 99th percentile has to be less than *3* standard deviations, so it must start with **2**, then $99/3 = $ **33** giving the **2.33**

Using R

What is the 95th percentile of a $N(\mu, \sigma^2)$ distribution?

```
#v- Rough estimate of IQ data
mu <- 100
std <- 18
#Using known info from above, the 95th percentile should be
#1.645sd from the mean
mu+1.645*std
```

```
## [1] 129.61
```

```
qnorm(0.95, mean = mu, sd = std)
```

```
## [1] 129.6074
```

```
#Probability that an IQ is larger than 140?
pnorm(140, mu, std, lower.tail = FALSE)
```

```
## [1] 0.01313415
```

```
#or
1 - pnorm(140, mu, std)
```

```
## [1] 0.01313415
```

So it can be seen the 1.645 estimate is pretty close to the truth

Poisson

- Used to model counts over a time duration
  PMF:
  $P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$

The mean **and** variance of this distribution is λ
Poisson random variables are used to model rates
* $X \sim Poisson(\lambda t)$ where
+ $\lambda = E[X/t]$ is the expected count per unit of time
+ t is the total monitoring time

When we use Poisson

- Modeling count data (especially if there is no upper bound)

- Modeling event-time or survival data
  - Example: Modeling reoccurance of a system during the duration of a study

- Modeling contingency tables
  - Ex: Hair color by race

- Approximating binomials when n is large and p is small
  - Common in epidemiology

Example

The number of people that show up at a bus stop is Poisson witha mean of 2.5 per hour.
If watching the bus stop for 4 hours, what is the probability that 3 or fewer people show up for the whole time?

```
#Using the formula
l <- 2.5*4
((l^3)*exp(1)^(-l))/factorial(3) +
  ((l^2)*exp(1)^(-l))/factorial(2) +
  ((l^1)*exp(1)^(-l))/factorial(1) +
  ((l^0)*exp(1)^(-l))/factorial(0)
```

```
## [1] 0.01033605
```

```
#Using the function
ppois(3, lambda = 2.5*4)
```

[1] 0.01033605

Poisson Approximation to the Binomial

- When n is large and p is small the Poisson distribution is an accurate approximation to the binomial distribution

- Notation
    - $X \sim Binomial(n, p)$

    - $\lambda = np$

    - n gets large

    - p gets small

Example:
We flip a coin with success probability 0.01 five hundred times.
What's the probability of 2 or fewer successes?

```
pbinom(2, size = 500, prob = 0.01)
```

[1] 0.1233858

```
ppois(2, lambda = 500*0.01)
```

[1] 0.124652

```
err <- abs(pbinom(2, size = 500, prob = 0.01) -
           ppois(2, lambda = 500*0.01))
paste0("This has an absolute error of approximately ", round(err, 6))
```

[1] "This has an absolute error of approximately 0.001266"

Lesson with `swirl()`: CommonDistros

(No new content)

Asymptotics and the Law of Large Numbers (LLN)

- Asymptotics refers to the behavior of estimators as the sample size goes to infinity.
    - In statistics we'll be concerned with as the sample size tends to infinity

- Asymptotics are useful for simple statistical inference and approximations

- Asymptotics form the baiss for frequency interpretation of probabilities
  - If you flip a coin, the proportion of heads should approach .5 as your sample size approaches infinity
    * **Count Buffon** tossed a coin 4040 times. Result: 2048 heads, a proportion of 0.5069 for heads

    * **Karl Pearson** *heroically* tossed a coin 24000 times. Result: 12012 heads, a proportion of 0.5005

    * **John Kerrich**, while imprisoned by Germans durring WW2 tossed a coin 10000 times. Result: 5067 heads, a proportion of 0.5067

Law of Large Numbers

- The average limits to what its estiamting, the population

- Example $\bar{X}_n$ could be the average of the result of n coin flips

- As we flip a fair coin over and over, it eventually converges to the true probability of a head

Examples

```
#SND
set.seed(161803)
n <- 1000
means <- cumsum(rnorm(n))/(1:n)
plot <- ggplot(data.frame(x = 1 : n, y = means), aes(x,y)) +
  geom_line(size = 2, col = "#FF0000") + geom_hline(yintercept = 0)
plot + labs(x = "Number of obs.", y = "Cumulative mean")
```
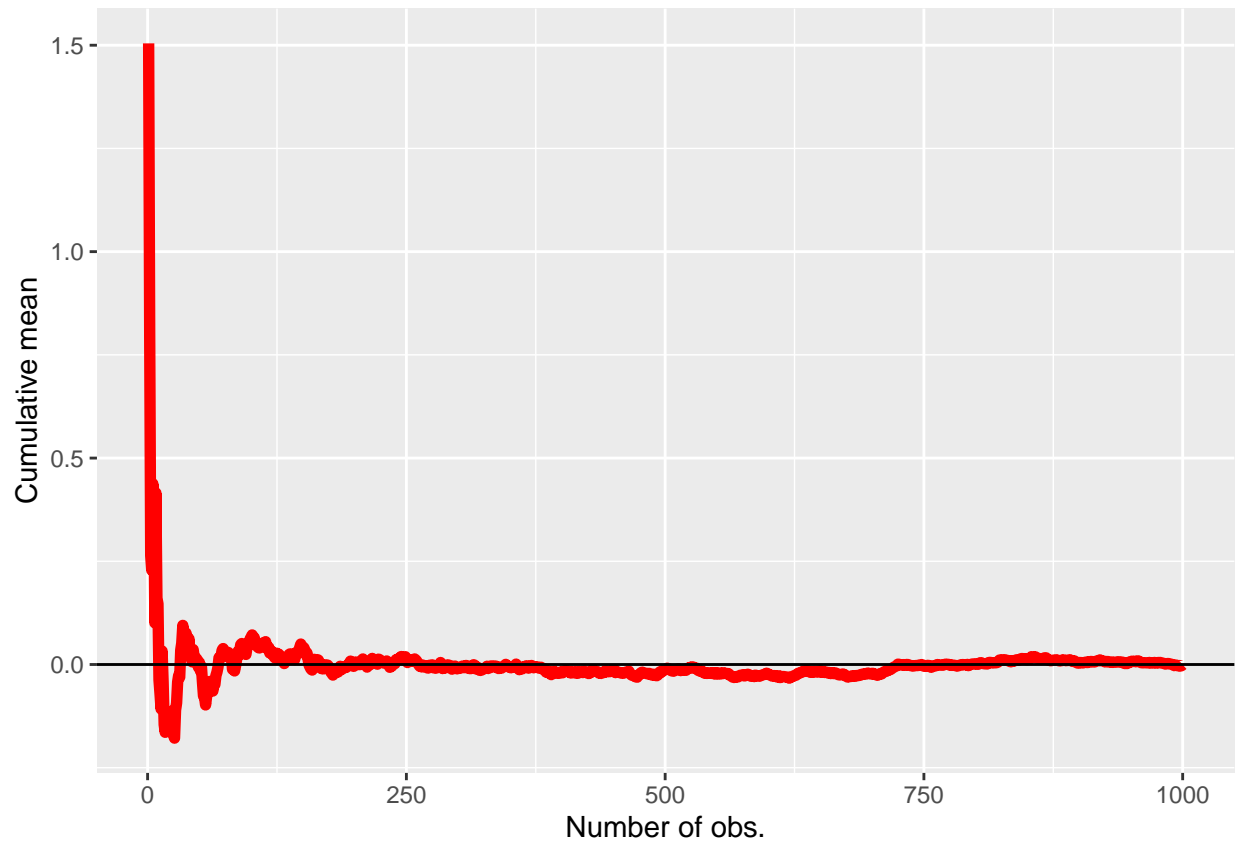
- It can be seen here that with smaller observations the cum. mean varries greatly, but as the observations increase the mean starts to trend near 0, the expected value of a SND.

```r
#Flipping a coin
set.seed(16)
n <- 1000
means <- cumsum(sample(0:1, n, replace = TRUE))/(1:n)
plot <- ggplot(data.frame(x = 1 : n, y = means), aes(x,y)) +
  geom_line(size = 2, col = "#FF0000") + geom_hline(yintercept = 0.5)
plot + labs(x = "Number of obs.", y = "Cumulative mean")
```
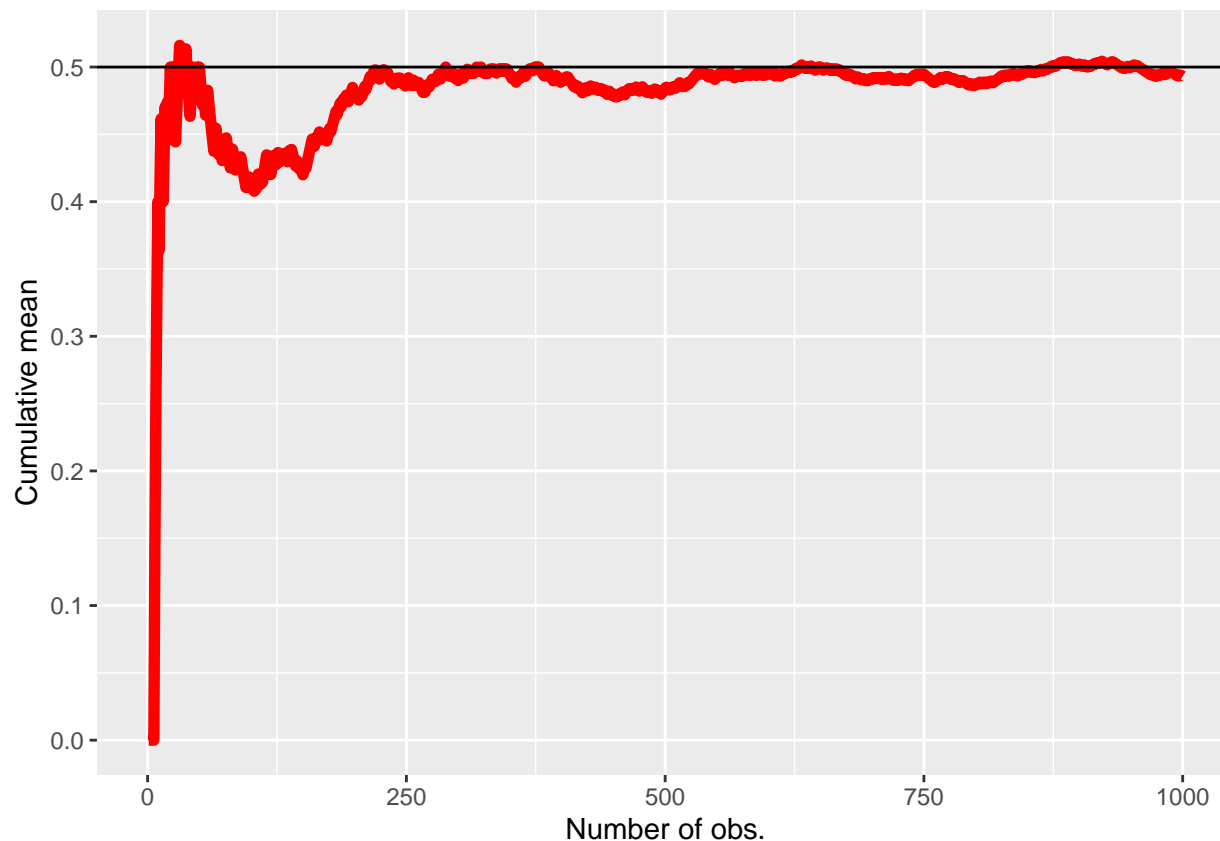
- Again here, we can see large variablity in small observations then it approaches the assymptote of 0.5 as the observations increase

Summary

- An estimator is **consistent** if it converges to what you want to estimate

- The LLN says that the sample mean of iid samples is consistent for the population mean

- Typically, good estimators are consistent; if we collect an infinite amount of data we would get the right answer

Asymptotics and the Central Limit Theorem (CLT)

- CLT states that the distribution of averages is often normal, even if the distribution that the data is being sampled from is very non-normal
    - This helps us create stategies for creating statistical inferences when we're not willing to assume much about the generating mechanism of our data.

CLT as a formula:
$\frac{\bar{X}_n - \mu}{\sigma / \sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}$

$$\frac{Estimate - Mean\ of\ estimate}{Std.\ Err.\ of\ estimate}$$

... will have a distribution like that of a standard normal for large n.

* A way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$

Example

```r
##Flipping an unfair coin
set.seed(1618)
nosim <- 1000
pHeads <- 0.9
cfunc <- function(x, n) {sqrt(n)*(mean(x)-pHeads)}/ sqrt(pHeads*(1-pHeads))
data <- data.frame(
  x = c(apply(matrix(sample(0:1, prob = c(1 - pHeads, pHeads), nosim * 10,
                            replace = TRUE), nosim), 1, cfunc, 10),
        apply(matrix(sample(0:1, prob = c(1 - pHeads, pHeads), nosim * 20,
                            replace = TRUE), nosim), 1, cfunc, 20),
        apply(matrix(sample(0:1, prob = c(1 - pHeads, pHeads), nosim * 30,
                            replace = TRUE), nosim), 1, cfunc, 30)
        ),
  size = factor(rep(c(10, 20, 30), rep(nosim, 3))))

plot <- ggplot(data, aes(x, fill = size)) +
  geom_histogram(binwidth = 0.3, colour = "#000000", aes(y = ..density..)) +
  facet_grid(. ~ size)
plot + stat_function(fun = dnorm, size = 2)
```
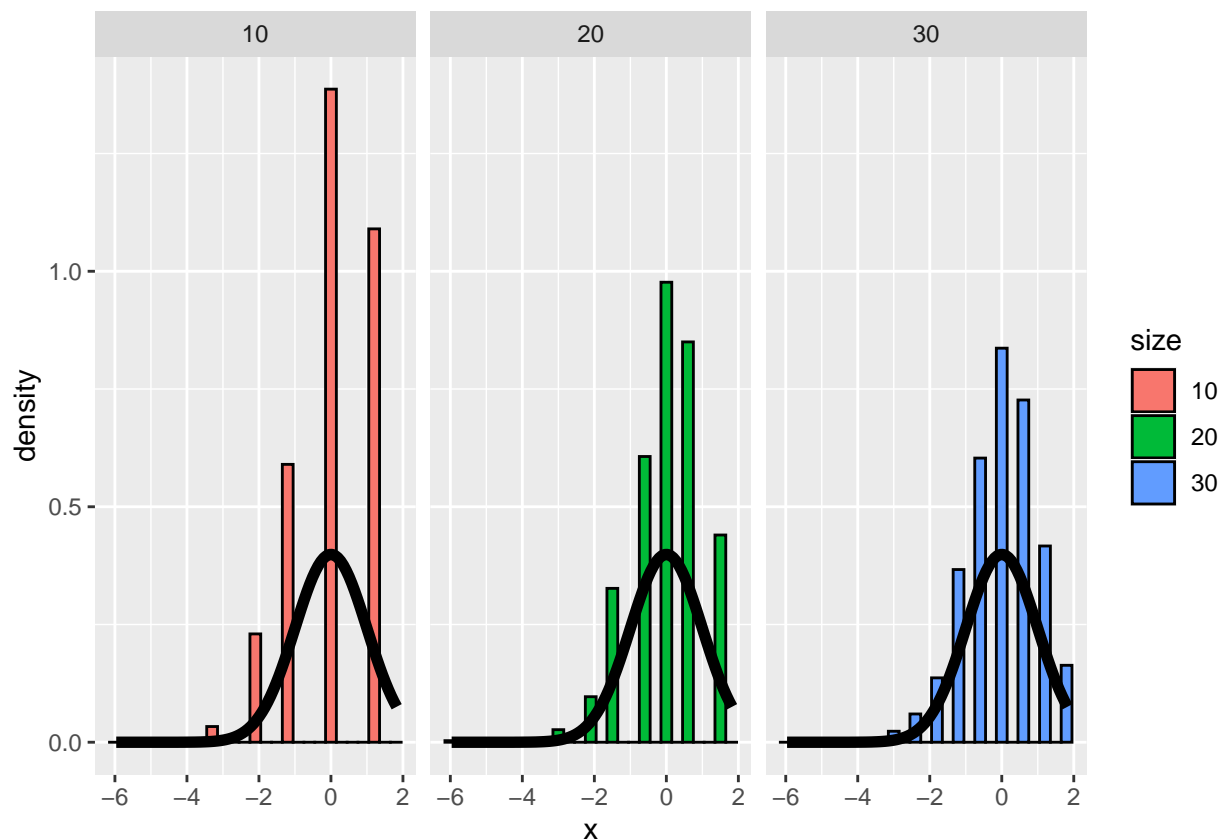
* In the smaller groups some of the discrete-ness of data can be seen, nevertheless as the samples grow in size it can be seen that the distributions more closely match the underlying normal distribution

- The **Galton Board** simulates the central limit theorem with a binomial distribution.

Asymptotics and Confidence Intervals

- The probability $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or maller than $\mu - 2\sigma/\sqrt{n}$ is 5%; in other words, the probability $\bar{X}$ is within that range is 95%

- As such, a 95% interval for μ is $\bar{X} \pm 2\sigma/\sqrt{n}$
  - If we were to repeatitly collect a sample in about 95% of the cases this interval would contain μ

Example

```
library(UsingR)
data(father.son)

x <- father.son$sheight
#Mean \pm the 97.5% quantile*sd/\sqrt{n}
```

```r
interval <- (mean(x) + c(-1,1)*qnorm(0.975)*sd(x)/sqrt(length(x)))
#Convert result to feet
feetint <- round(interval*1/12, 3)
feetint
```

```
## [1] 5.710 5.738
```

- we could say that we are 95% confident that the population mean height of sons from the population the data is drawn from is between 5.71 and 5.738 feet.

In a random sample of 100 likely voters, 56 intent to vote for your canidate
+ Can you relax? Do you have this race "in the bag"?

```r
#Executing formula
0.56 + c(-1,1) * qnorm(0.975) * sqrt(0.56 * 0.44/100)
```

```
## [1] 0.4627099 0.6572901
```

```r
#Using R function
binom.test(56, 100)$conf.int
```

```
## [1] 0.4571875 0.6591640
## attr(,"conf.level")
## [1] 0.95
```

- Since the interval goes below 0.50 you can't relax and ought to do more campaigning

- `binom.test` is exact, as it does some additional computations that aren't normally done by hand. Also tends to have wider intervals
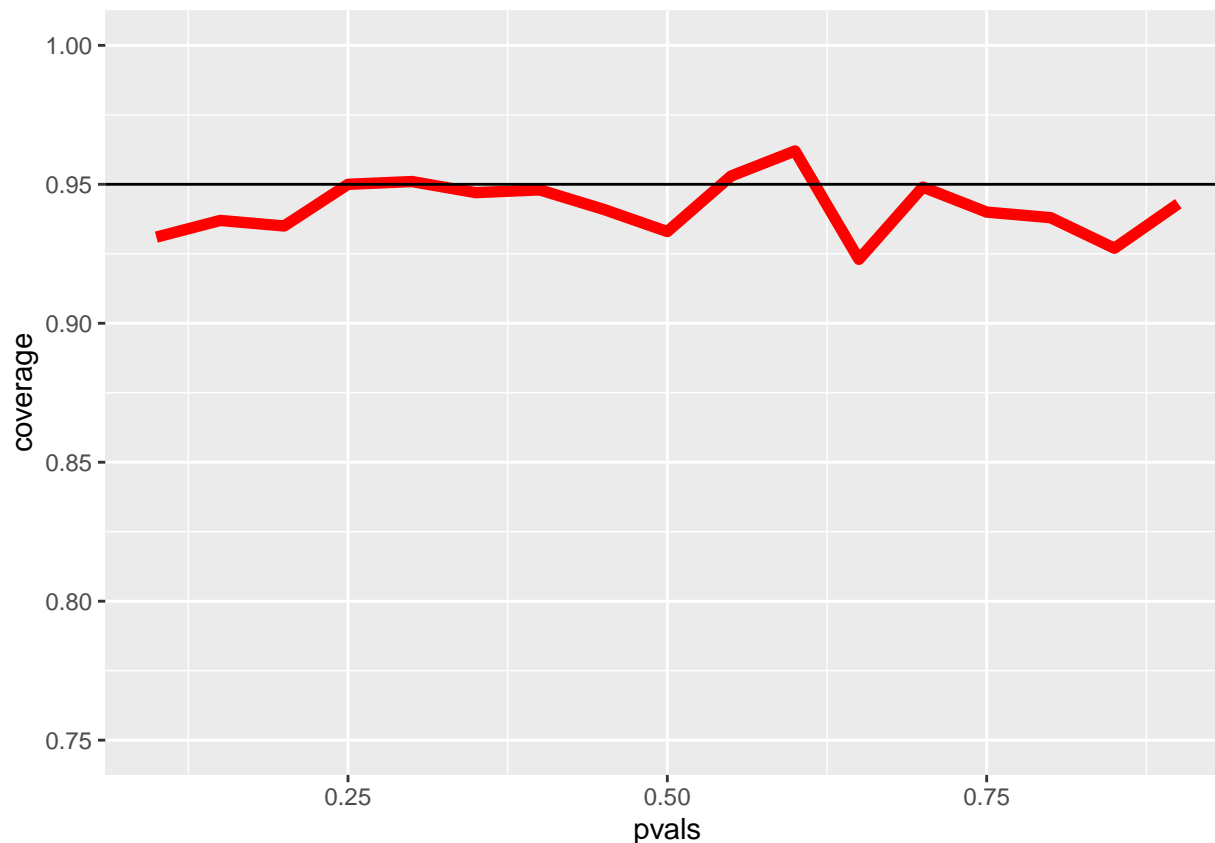
Simulation to check confidence intervals

```r
set.seed(1618)
n <- 20 #Flips per simulation
pvals <- seq(0.1, 0.9, by = 0.05) #'p's we'll check with
nosim <- 1000
coverage <- sapply(pvals, function(p){
  phats <- rbinom(nosim, prob = p, size = n)/n #point estimates
  ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  mean(ll < p & ul > p) #Check proportion of times the true p is in the range
})
ggplot(data.frame(pvals, coverage), aes(pvals, coverage)) +
  geom_line(size = 2, color = "#FF0000") +
  geom_hline(yintercept = 0.95) +
  ylim(0.75, 1)
```
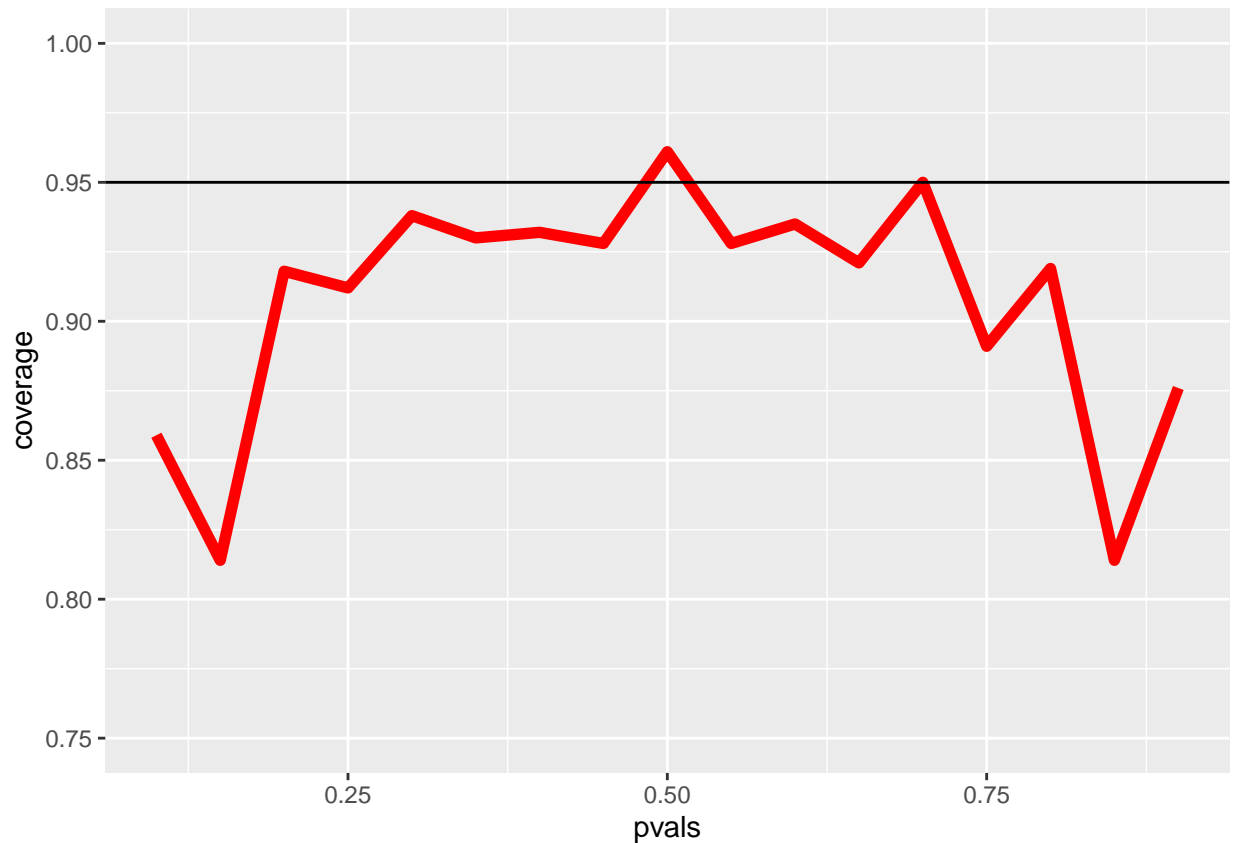
* Since **n** is small the coverage is often giving lower than 95% confidence

```r
#Larger n value
set.seed(1618033)
n <- 100 #Flips per simulation
pvals <- seq(0.1, 0.9, by = 0.05) #'p's we'll check with
nosim <- 1000
coverage <- sapply(pvals, function(p){
  phats <- rbinom(nosim, prob = p, size = n)/n #point estimates
  ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  mean(ll < p & ul > p) #Check proportion of times the true p is in the range
})
ggplot(data.frame(pvals, coverage), aes(pvals, coverage)) +
  geom_line(size = 2, color = "#FF0000") +
  geom_hline(yintercept = 0.95) +
  ylim(0.75, 1)
```

Agresti/Coull interval

- A quick fix for when n isn't large enough for the CLT to be applicable for many of the values of p

- Add two successes and failures $\frac{X+2}{n+4}$

- An advantage of this procedure is that it does not strongly depend upon the value of n and/or p, and was recommended by Agresti and Coull for virtually all combinaitons of n and p

- **Further Reading**

Looking at the $n = 20$ example again

```
set.seed(1618)
n <- 20 #Flips per simulation
pvals <- seq(0.1, 0.9, by = 0.05) #'p's we'll check with
nosim <- 1000
coverage <- sapply(pvals, function(p){
  phats <- (rbinom(nosim, prob = p, size = n) + 2)/(n + 4) #point estimates
  ll <- phats - qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  ul <- phats + qnorm(0.975) * sqrt(phats * (1 - phats)/n)
  mean(ll < p & ul > p) #Check proportion of times the true p is in the range
})
```

```r
ggplot(data.frame(pvals, coverage), aes(pvals, coverage)) +
  geom_line(size = 2, color = "#FF0000") +
  geom_hline(yintercept = 0.95) +
  ylim(0.75, 1)
```
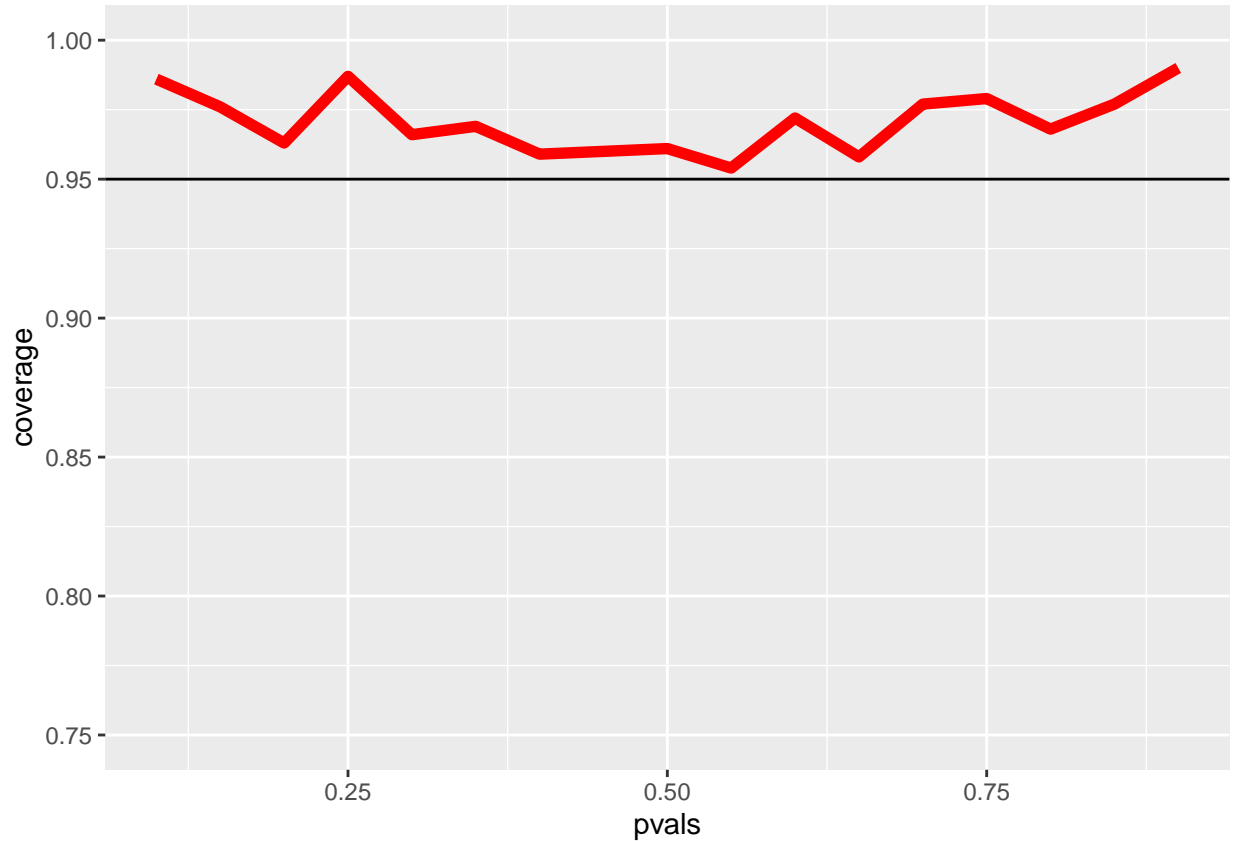

* In this simulation, after applying the Agresti/Coull interval, there are no pvals that result in the CI dipping below 95%

Poisson interval

A nuclear pump failed 5 times out of 94.32 days, give a 95% confidence interval for the failure rate per day
* $X \sim Poisson(\lambda t)$
* Estimate $\hat{\lambda} = X/t$
* $Var(\hat{\lambda}) = \lambda/t$
* $\hat{\lambda}/t$ is our variance estimate

```r
x <- 5
t <- 94.32
lambda <- x/t

#Using the formula
round(lambda + c(-1, 1) * qnorm(0.975) * sqrt(lambda/t), 5)
```

```
## [1] 0.00655 0.09948
```

```r
#Using R function
round(poisson.test(x, T = 94.32)$conf, 5)
```

```
## [1] 0.01721 0.12371
## attr(,"conf.level")
## [1] 0.95
```

- The function gives a more conservative estimate as it's more exact from doing some additional computations to guarentee the coverage

Simulating the Poisson Coverage Rate

```r
set.seed(1618)
lambdavals <- seq(0.005, 0.1, by = 0.01)
nosim <- 1000
t <- 100
coverage <- sapply(lambdavals, function(lambda) {
  lhats <- rpois(nosim, lambda = lambda*t)/t
  ll <- lhats - qnorm(0.975) * sqrt(lhats/t)
  ul <- lhats + qnorm(0.975) * sqrt(lhats/t)
  mean(ll < lambda & ul > lambda)
})
ggplot(data.frame(lambdavals, coverage), aes(lambdavals, coverage)) +
  geom_line(size = 2, colour = "#FF0000") +
  geom_hline(yintercept = 0.95) +
  ylim(0, 1)
```
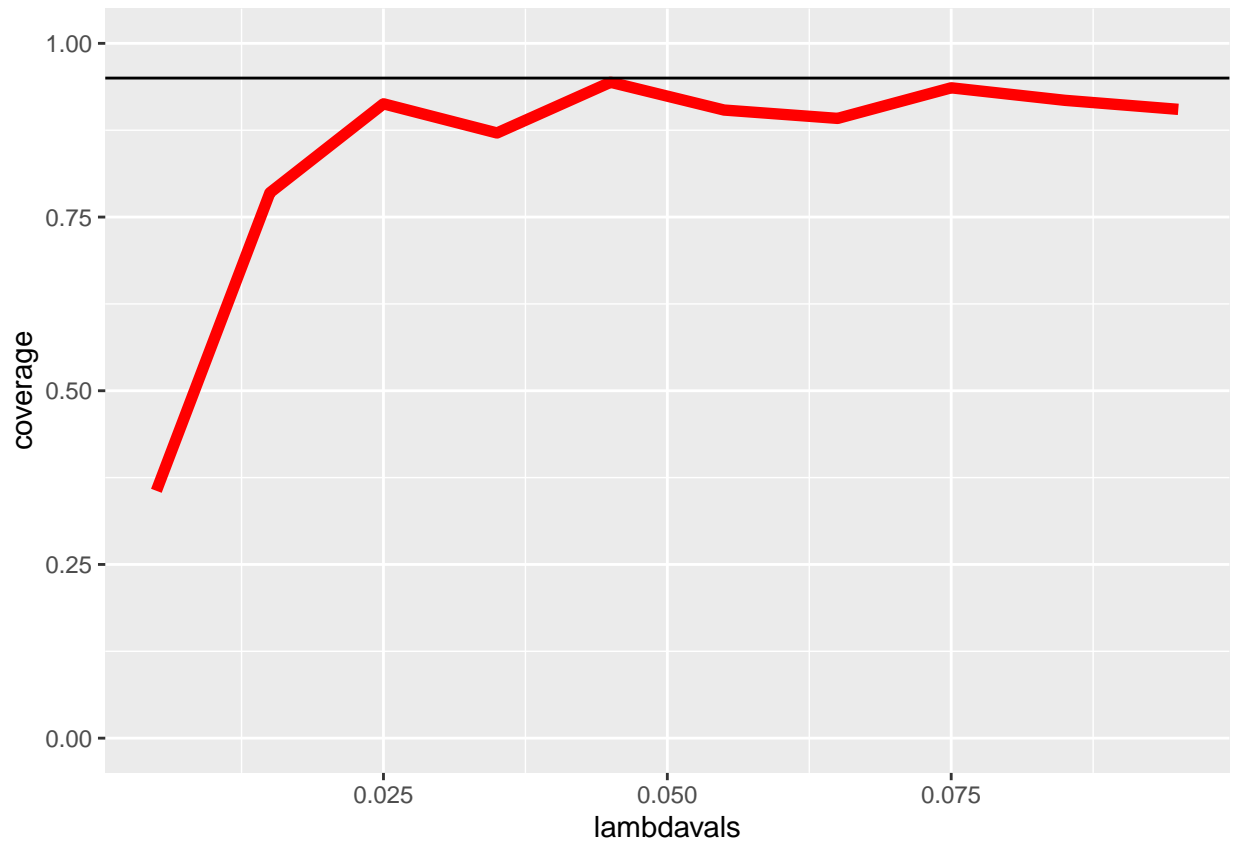
* It can be seen above that the coverage is really bad for smaller values of lambda

When t is larger the coverage is better, but still really bad for smalelr values of lambda

```r
set.seed(1618)
lambdavals <- seq(0.005, 0.1, by = 0.01)
nosim <- 1000
t <- 1000
coverage <- sapply(lambdavals, function(lambda) {
  lhats <- rpois(nosim, lambda = lambda*t)/t
  ll <- lhats - qnorm(0.975) * sqrt(lhats/t)
  ul <- lhats + qnorm(0.975) * sqrt(lhats/t)
  mean(ll < lambda & ul > lambda)
})
ggplot(data.frame(lambdavals, coverage), aes(lambdavals, coverage)) +
  geom_line(size = 2, colour = "#FF0000") +
  geom_hline(yintercept = 0.95) +
  ylim(0, 1)
```
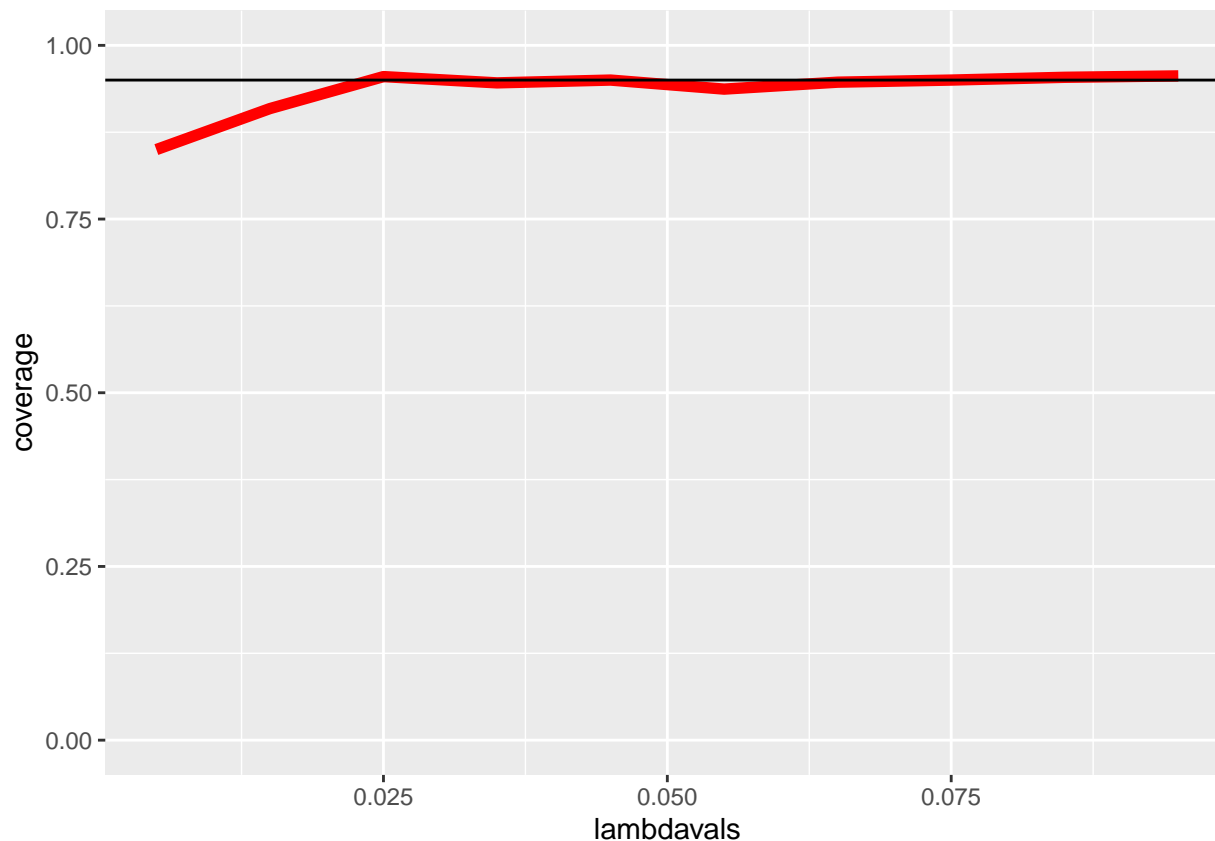
Summary

- The LLN states that averages of iid samples converge tot eh popualiton means that they are estimating

- The CLT states that averages are approximately normal, with
  - distributions centered at the population mean

  - standard deviation equal to the standard error of the mean

- CLT gives no guarantee that n is large enough

- Taking the mean and $\pm$ the relevant normal quantile times the standard error yields a confidence interval for the mean
  - $\pm$ 2 standard errors works for 95% intervals

- Confidence intervals get wider as the coverage increases
  - Being more confident that the mean is included means more values are included in the interval

- The Poisson and binomial cases have exact intervals that don't require the CLT
  - But a quick fix for small sample size binomial calculations si to add 2 successes and

failures

Lesson with `swirl()`: Asymptotics

(No new content)

Quiz 2

2. Suppose that diastolic blood pressures (DBPs) for men aged 35-44 are normally distributed with a mean of 80 (mm Hg) and a standard deviation of 10. About what is the probability that a random 35-44 year old has a DBP less than 70?

```
pnorm(70, 80, 10)
```

```
## [1] 0.1586553
```

3. Brain volume for adult women is normally distributed with a mean of about 1,100 cc for women with a standard deviation of 75 cc. What brain volume represents the 95th percentile?

```
qnorm(.95, 1100, 75)
```

```
## [1] 1223.364
```

4. Refer to the previous question. Brain volume for adult women is about 1,100 cc for women with a standard deviation of 75 cc. Consider the sample mean of 100 random adult women from this population. What is the 95th percentile of the distribution of that sample mean?

```
qnorm(.95, 1100, 75/sqrt(100))
```

```
## [1] 1112.336
```

5. You flip a fair coin 5 times, about what's the probability of getting 4 or 5 heads?

```
pbinom(3, 5, 0.5, lower.tail = FALSE)
```

```
## [1] 0.1875
```

6. The respiratory disturbance index (RDI), a measure of sleep disturbance, for a specific population has a mean of 15 (sleep events per hour) and a standard deviation of 10. They are not normally distributed. Give your best estimate of the probability that a sample mean RDI of 100 people is between 14 and 16 events per hour?

```
pnorm(16, 15, 10/sqrt(100)) - pnorm(14, 15, 10/sqrt(100))
```

```
## [1] 0.6826895
```

7. Consider a standard uniform density. The mean for this density is .5 and the variance is 1 / 12. You sample 1,000 observations from this distribution and take the sample mean, what value would you expect it to be near?

```
qnorm(.5, .5, (1/12)/sqrt(1000))
```

```
## [1] 0.5
```

8. The number of people showing up at a bus stop is assumed to be Poisson with a mean of 5 5 5 people per hour. You watch the bus stop for 3 hours. About what's the probability of viewing 10 or fewer people?

```
ppois(10, 5*3)
```

[1] 0.1184644

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