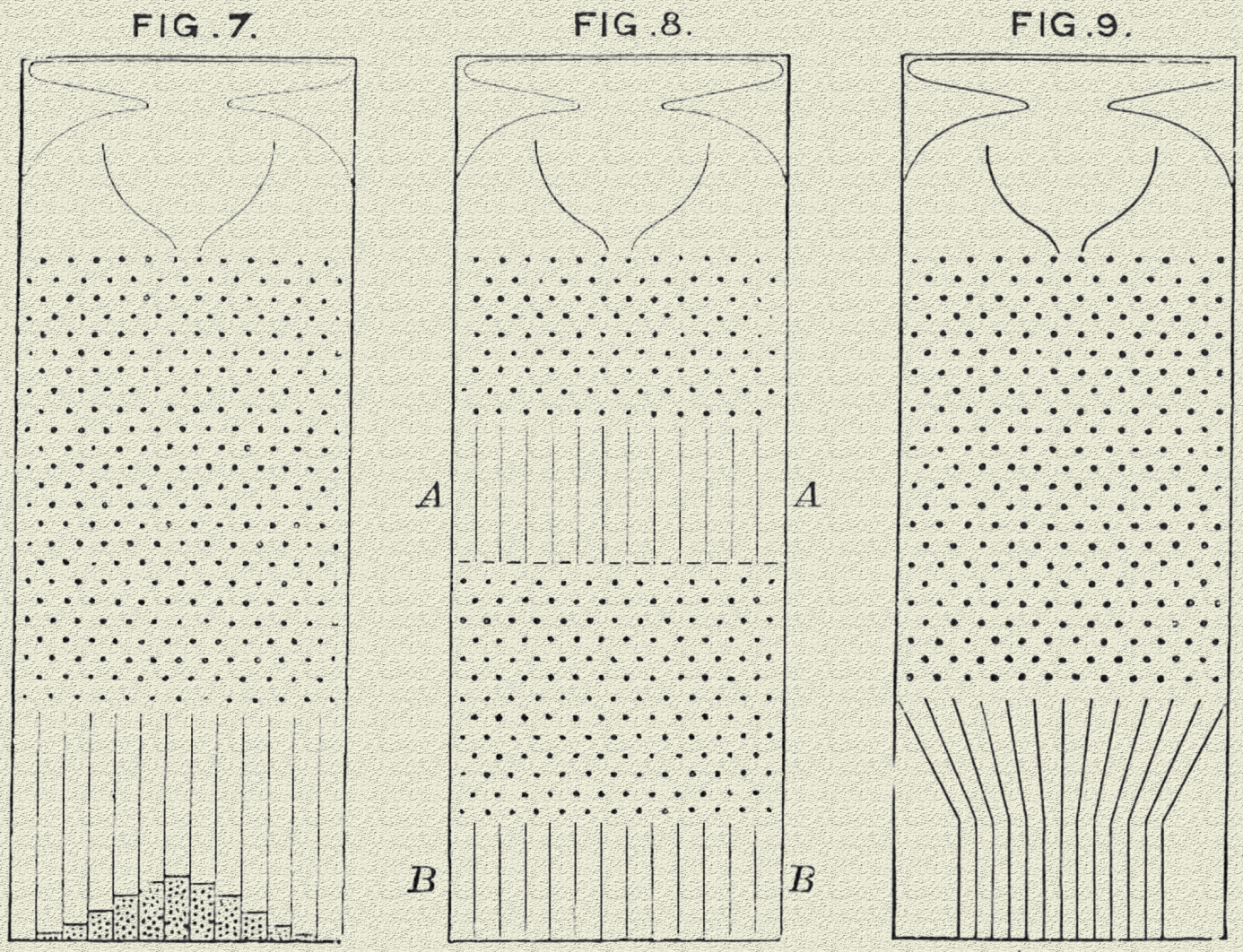
Statistical inference for data science



Brian Caffo

Statistical inference for data science

A companion to the Coursera Statistical Inference Course

Brian Caffo

This book is for sale at http://leanpub.com/LittleInferenceBook

This version was published on 2016-05-24

Also By Brian Caffo

Regression Models for Data Science in R

Executive Data Science

Advanced Linear Models for Data Science

Developing Data Products in R

To Kerri, Penelope and Scarlett

Contents

About this book

This book is written as a companion book to the Statistical Inference[1] Coursera class as part of the Data Science Specialization[2]. However, if you do not take the class, the book mostly stands on its own. A useful component of the book is a series of YouTube videos that comprise the Coursera class.

The book is intended to be a low cost introduction to the important field of statistical inference. The intended audience are students who are numerically and computationally literate, who would like to put those skills to use in Data Science or Statistics. The book is offered for free as a series of markdown documents on github and in more convenient forms (epub, mobi) on LeanPub and retail outlets.

About the picture on the cover

The picture on the cover is a public domain image taken from Wikipedia's article on Francis Galton's quincunx. Francis Galton was an 19th century polymath who invented many of key concepts of statistics. The quincunx was an ingenious invention for illustrating the central limit theorem using a pinball setup.

[1] https://www.coursera.org/course/statinference

[2] https://www.coursera.org/specialization/jhudatascience/1?utm_medium=courseDescripTop

[3] http://creativecommons.org/licenses/by-nc-sa/4.0/

1. Introduction

Before beginning

This book is designed as a companion to the Statistical Inference[1] Coursera class as part of the Data Science Specialization[2], a ten course program offered by three faculty, Jeff Leek, Roger Peng and Brian Caffo, at the Johns Hopkins University Department of Biostatistics.

The videos associated with this book can be watched in full here[3], though the relevant links to specific videos are placed at the appropriate locations throughout.

Before beginning, we assume that you have a working knowledge of the R programming language. If not, there is a wonderful Coursera class by Roger Peng, that can be found here[4].

The entirety of the book is on GitHub here[5]. Please submit pull requests if you find errata! In addition the course notes can be found also on GitHub here[6]. While most code is in the book, *all* of the code for every figure and analysis in the book is in the R markdown files files (.Rmd) for the respective lectures.

Finally, we should mention `swirl` (statistics with interactive R programming). `swirl` is an intelligent tutoring system developed by Nick Carchedi, with contributions by Sean Kross and Bill and Gina Croft. It offers a way to learn R in R. Download `swirl` here[7]. There's a swirl module for this course![8]. Try it out, it's probably the most effective way to learn.

Statistical inference defined

Watch this video before beginning.[9]

We'll define statistical inference as the process of generating conclusions about a population from a noisy sample. Without statistical inference we're simply living within our data. With statistical inference, we're trying to generate new knowledge.

Knowledge and parsimony, (using simplest reasonable models to explain complex phenomena), go hand in hand. Probability models will serve as our parsimonious description of the world. The use

[1] https://www.coursera.org/course/statinference
[2] https://www.coursera.org/specialization/jhudatascience/1?utm_medium=courseDescripTop
[3] https://www.youtube.com/watch?v=WkOinijQmPU&list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ
[4] https://www.coursera.org/course/rprog
[5] https://github.com/bcaffo/LittleInferenceBook
[6] https://github.com/bcaffo/courses/tree/master/06_StatisticalInference
[7] http://swirlstats.com
[8] https://github.com/swirldev/swirl_courses#swirl-courses
[9] http://youtu.be/WkOinijQmPU?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

of probability models as the connection between our data and a populations represents the most effective way to obtain inference.

Motivating example: who's going to win the election?

In every major election, pollsters would like to know, ahead of the actual election, who's going to win. Here, the target of estimation (the estimand) is clear, the percentage of people in a particular group (city, state, county, country or other electoral grouping) who will vote for each candidate.

We can not poll everyone. Even if we could, some polled may change their vote by the time the election occurs. How do we collect a reasonable subset of data and quantify the uncertainty in the process to produce a good guess at who will win?

Motivating example, predicting the weather

When a weatherman tells you the probability that it will rain tomorrow is 70%, they're trying to use historical data to predict tomorrow's weather - and to actually attach a probability to it. That probability refers to population.

Motivating example, brain activation

An example that's very close to the research I do is trying to predict what areas of the brain activate when a person is put in the fMRI scanner. In that case, people are doing a task while in the scanner. For example, they might be tapping their finger. We'd like to compare when they are tapping their finger to when they are not tapping their finger and try to figure out what areas of the brain are associated with the finger tapping.

Summary notes

These examples illustrate many of the difficulties of trying to use data to create general conclusions about a population.

Paramount among our concerns are:

- Is the sample representative of the population that we'd like to draw inferences about?
- Are there known and observed, known and unobserved or unknown and unobserved variables that contaminate our conclusions?
- Is there systematic bias created by missing data or the design or conduct of the study?
- What randomness exists in the data and how do we use or adjust for it? Here randomness can either be explicit via randomization or random sampling, or implicit as the aggregation of many complex unknown processes.
- Are we trying to estimate an underlying mechanistic model of phenomena under study?

Statistical inference requires navigating the set of assumptions and tools and subsequently thinking about how to draw conclusions from data.

The goals of inference

You should recognize the goals of inference. Here we list five examples of inferential goals.

1. Estimate and quantify the uncertainty of an estimate of a population quantity (the proportion of people who will vote for a candidate).
2. Determine whether a population quantity is a benchmark value ("is the treatment effective?").
3. Infer a mechanistic relationship when quantities are measured with noise ("What is the slope for Hooke's law?")
4. Determine the impact of a policy? ("If we reduce pollution levels, will asthma rates decline?")
5. Talk about the probability that something occurs.

The tools of the trade

Several tools are key to the use of statistical inference. We'll only be able to cover a few in this class, but you should recognize them anyway.

1. *Randomization*: concerned with balancing unobserved variables that may confound inferences of interest.
2. *Random sampling*: concerned with obtaining data that is representative of the population of interest.
3. *Sampling models*: concerned with creating a model for the sampling process, the most common is so called "iid".
4. *Hypothesis testing*: concerned with decision making in the presence of uncertainty.
5. *Confidence intervals*: concerned with quantifying uncertainty in estimation.
6. *Probability models*: a formal connection between the data and a population of interest. Often probability models are assumed or are approximated.
7. *Study design*: the process of designing an experiment to minimize biases and variability.
8. *Nonparametric* bootstrapping: the process of using the data to, with minimal probability model assumptions, create inferences.
9. *Permutation*, randomization and exchangeability testing: the process of using data permutations to perform inferences.

Different thinking about probability leads to different styles of inference

We won't spend too much time talking about this, but there are several different styles of inference. Two broad categories that get discussed a lot are:

1. *Frequency probability*: is the long run proportion of times an event occurs in independent, identically distributed repetitions.
2. *Frequency style inference*: uses frequency interpretations of probabilities to control error rates. Answers questions like "What should I decide given my data controlling the long run proportion of mistakes I make at a tolerable level."
3. *Bayesian probability*: is the probability calculus of beliefs, given that beliefs follow certain rules.
4. *Bayesian style inference*: the use of Bayesian probability representation of beliefs to perform inference. Answers questions like "Given my subjective beliefs and the objective information from the data, what should I believe now?"

Data scientists tend to fall within shades of gray of these and various other schools of inference. Furthermore, there are so many shades of gray between the styles of inferences that it is hard to pin down most modern statisticians as either Bayesian or frequentist. In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences. This is the most popular style of inference by far.

Being data scientists, we will also consider some inferential strategies that
rely heavily on the observed data, such as permutation testing and bootstrapping. As probability modeling will be our starting point, we first build up basic probability as our first task.

Exercises

1. The goal of statistical inference is to?
    - Infer facts about a population from a sample.
    - Infer facts about the sample from a population.
    - Calculate sample quantities to understand your data.
    - To torture Data Science students.
2. The goal of randomization of a treatment in a randomized trial is to?
    - It doesn't really do anything.
    - To obtain a representative sample of subjects from the population of interest.
    - Balance unobserved covariates that may contaminate the comparison between the treated and control groups.
    - To add variation to our conclusions.
3. Probability is a?
    - Population quantity that we can potentially estimate from data.
    - A data quantity that does not require the idea of a population.

2. Probability

Watch this video before beginning.[1]

Probability forms the foundation for almost all treatments of statistical inference. In our treatment, probability is a law that assigns numbers to the long run occurrence of random phenomena after repeated unrelated realizations.

Before we begin discussing probability, let's dispense with some deep philosophical questions, such as "What is randomness?" and "What is the fundamental interpretation of probability?". One could spend a lifetime studying these questions (and some have). For our purposes, randomness is any process occurring without apparent deterministic patterns. Thus we will treat many things as if they were random when, in fact they are completely deterministic. In my field, biostatistics, we often model disease outcomes as if they were random when they are the result of many mechanistic components whose aggregate behavior appears random. Probability for us will be the long run proportion of times some occurs in repeated unrelated realizations. So, think of the proportion of times that you get a head when flipping a coin.

For the interested student, I would recommend the books and work by Ian Hacking to learn more about these deep philosophical issues. For us data scientists, the above definitions will work fine.

Where to get a more thorough treatment of probability

In this lecture, we will cover the fundamentals of probability at low enough of a level to have a basic understanding for the rest of the series. For a more complete treatment see the class Mathematical Biostatistics Boot Camp 1, which can be viewed on YouTube here[2]. In addition, there's the actual Coursera course[3] that I run periodically (this is the first Coursera class that I ever taught). Also there are a set of notes on GitHub[4]. Finally, there's a follow up class, uninspiringly named Mathematical Biostatistics Boot Camp 2, that is more devoted to biostatistical topics that has an associated YouTube playlist[5], Coursera Class[6] and GitHub notes[7].

[1] http://youtu.be/oTERv_vrmJM?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

[2] Youtube:www.youtube.com/playlist?list=PLpl-gQkQivXhk6qSyiNj51qamjAtZISJ-

[3] Coursera:www.coursera.org/course/biostats

[4] http://github.com/bcaffo/Caffo-Coursera

[5] http://www.youtube.com/playlist?list=PLpl-gQkQivXhwOsKPQ4fbCBYOWjvdzrSM

[6] https://www.coursera.org/course/biostats2

[7] https://github.com/bcaffo/MathematicsBiostatisticsBootCamp2

Kolmogorov's Three Rules

Watch this lecture before beginning.[8]

Given a random experiment (say rolling a die) a probability measure is a population quantity that summarizes the randomness. The brilliant discovery of the father of probability, the Russian mathematician Kolmogorov[9], was that to satisfy our intuition about how probability should behave, only three rules were needed.

Consider an experiment with a random outcome. Probability takes a possible outcome from an experiment and:

1. assigns it a number between 0 and 1
2. requires that the probability that something occurs is 1
3. required that the probability of the union of any two sets of outcomes that have nothing in common (mutually exclusive) is the sum of their respective probabilities.

From these simple rules all of the familiar rules of probability can be developed. This all might seem a little odd at first and so we'll build up our intuition with some simple examples based on coin flipping and die rolling.

I would like to reiterate the important definition that we wrote out: *mutually exclusive*. Two events are mutually exclusive if they cannot both simultaneously occur. For example, we cannot simultaneously get a 1 and a 2 on a die. Rule 3 says that since the event of getting a 1 and 2 on a die are mutually exclusive, the probability of getting at least one (the union) is the sum of their probabilities. So if we know that the probability of getting a 1 is 1/6 and the probability of getting a 2 is 1/6, then the probability of getting a 1 or a 2 is 2/6, the sum of the two probabilities since they are mutually exclusive.

Consequences of The Three Rules

Let's cover some consequences of our three simple rules. Take, for example, the probability that something occurs is 1 minus the probability of the opposite occurring. Let A be the event that we get a 1 or a 2 on a rolled die. Then A^c is the opposite, getting a 3, 4, 5 or 6. Since A and A^c cannot both simultaneously occur, they are mutually exclusive. So the probability that either A or A^c is $P(A) + P(A^c)$. Notice, that the probability that either occurs is the probability of getting a 1, 2, 3, 4, 5 or 6, or in other words, the probability that something occurs, which is 1 by rule number 2. So we have that $1 = P(A) + P(A^c)$ or that $P(A) = 1 - P(A^c)$.

We won't go through this tedious exercise (since Kolmogorov already did it for us). Instead here's a list of some of the consequences of Kolmogorov's rules that are often useful.

[8] http://youtu.be/Shzt9uZ8BII?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

[9] http://en.wikipedia.org/wiki/Andrey_Kolmogorov

1. The probability that nothing occurs is 0
2. The probability that something occurs is 1
3. The probability of something is 1 minus the probability that the opposite occurs
4. The probability of at least one of two (or more) things that can not simultaneously occur (mutually exclusive) is the sum of their respective probabilities
5. For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection.

This last rules states that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ shows what is the issue with adding probabilities that are not mutually exclusive. If we do this, we've added the probability that both occur in twice! (Watch the video where I draw a Venn diagram to illustrate this).

Example of Implementing Probability Calculus

The National Sleep Foundation (www.sleepfoundation.org[10]) reports that around 3% of the American population has sleep apnea. They also report that around 10% of the North American and European population has restless leg syndrome. Does this imply that 13% of people will have at least one sleep problems of these sorts? In other words, can we simply add these two probabilities?

Answer: No, the events can simultaneously occur and so are not mutually exclusive. To elaborate let:

$$\begin{aligned} A_1 &= \{\text{Person has sleep apnea}\} \\ A_2 &= \{\text{Person has RLS}\} \end{aligned}$$

Then

$$\begin{aligned} P(A_1 \cup A_2) &= P(A_1) + P(A_2) - P(A_1 \cap A_2) \\ &= 0.13 - \text{Probability of having both} \end{aligned}$$

Given the scenario, it's likely that some fraction of the population has both. This example serves as a reminder *don't add probabilities unless the events are mutually exclusive*. We'll have a similar rule for multiplying probabilities and independence.

Random variables

Watch this video before reading this section[11]

[10] http://www.sleepfoundation.org/

[11] http://youtu.be/Shzt9uZ8BII?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Probability calculus is useful for understanding the rules that probabilities must follow. However, we need ways to model and think about probabilities for numeric outcomes of experiments (broadly defined). Densities and mass functions for random variables are the best starting point for this. You've already heard of a density since you've heard of the famous "bell curve", or Gaussian density. In this section you'll learn exactly what the bell curve is and how to work with it.

Remember, everything we're talking about up to at this point is a population quantity, not a statement about what occurs in our data. Think about the fact that 50% probability for head is a statement about the coin and how we're flipping it, not a statement about the percentage of heads we obtained in a particular set of flips. This is an important distinction that we will emphasize over and over in this course. Statistical inference is about describing populations using data. Probability density functions are a way to mathematically characterize the population. In this course, we'll assume that our sample is a random draw from the population.

So our definition is that a **random variable** is a numerical outcome of an experiment. The random variables that we study will come in two varieties, **discrete** or **continuous**. Discrete random variables are random variables that take on only a countable number of possibilities. Mass functions will assign probabilities that they take specific values. Continuous random variable can conceptually take any value on the real line or some subset of the real line and we talk about the probability that they lie within some range. Densities will characterize these probabilities.

Let's consider some examples of measurements that could be considered random variables. First, familiar gambling experiments like the tossing of a coin and the rolling of a die produce random variables. For the coin, we typically code a tail as a 0 and a head as a 1. (For the die, the number facing up would be the random variable.) We will use these examples a lot to help us build intuition. However, they aren't interesting in the sense of seeming very contrived. Nonetheless, the coin example is particularly useful since many of the experiments we consider will be modeled as if tossing a biased coin. Modeling any binary characteristic from a random sample of a population can be thought of as a coin toss, with the random sampling performing the roll of the toss and the population percentage of individuals with the characteristic is the probability of a head. Consider, for example, logging whether or not subjects were hypertensive in a random sample. Each subject's outcome can be modeled as a coin toss. In a similar sense the die roll serves as our model for phenomena with more than one level, such as hair color or rating scales.

Consider also the random variable of the number of web hits for a site each day. This variable is a count, but is largely unbounded (or at least we couldn't put a specific reasonable upper limit). Random variables like this are often modeled with the so called Poisson distribution.

Finally, consider some continuous random variables. Think of things like lengths or weights. It is mathematically convenient to model these as if they were continuous (even if measurements were truncated liberally). In fact, even discrete random variables with lots of levels are often treated as continuous for convenience.

For all of these kinds of random variables, we need convenient mathematical functions to model the probabilities of collections of realizations. These functions, called mass functions and densities, take possible values of the random variables, and assign the associated probabilities. These entities

describe the population of interest. So, consider the most famous density, the normal distribution. Saying that body mass indices follow a normal distribution is a statement about the population of interest. The goal is to use our data to figure out things about that normal distribution, where it's centered, how spread out it is and even whether our assumption of normality is warranted!

Probability mass functions

A probability mass function evaluated at a value corresponds to 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

Let X be the result of a coin flip where $X = 0$ represents tails and $X = 1$ represents heads. $p(x) = (1/2)^x(1/2)^{1-x}$ for $x = 0, 1$. Suppose that we do not know whether or not the coin is fair; Let θ be the probability of a head expressed as a proportion (between 0 and 1). $p(x) = \theta^x(1-\theta)^{1-x}$ for $x = 0, 1$

Probability density functions

Watch this video before beginning.[12]

A probability density function (pdf), is a function associated with a continuous random variable. Because of the peculiarities of treating measurements as having been recorded to infinite decimal expansions, we need a different set of rules. This leads us to the central dogma of probability density functions:

Areas under PDFs correspond to probabilities for that random variable

Therefore, when one says that intelligence quotients (IQ) in population follows a bell curve, they are saying that the probability of a randomly selected person from this population having an IQ between two values is given by the area under the bell curve.

Not every function can be a valid probability density function. For example, if the function dips below zero, then we could have negative probabilities. If the function contains too much area underneath it, we could have probabilities larger than one. The following two rules tell us when a function is a valid probability density function.

Specifically, to be a valid pdf, a function must satisfy

1. It must be larger than or equal to zero everywhere.
2. The total area under it must be one.

[12] http://youtu.be/mPe0Us4VYDM?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

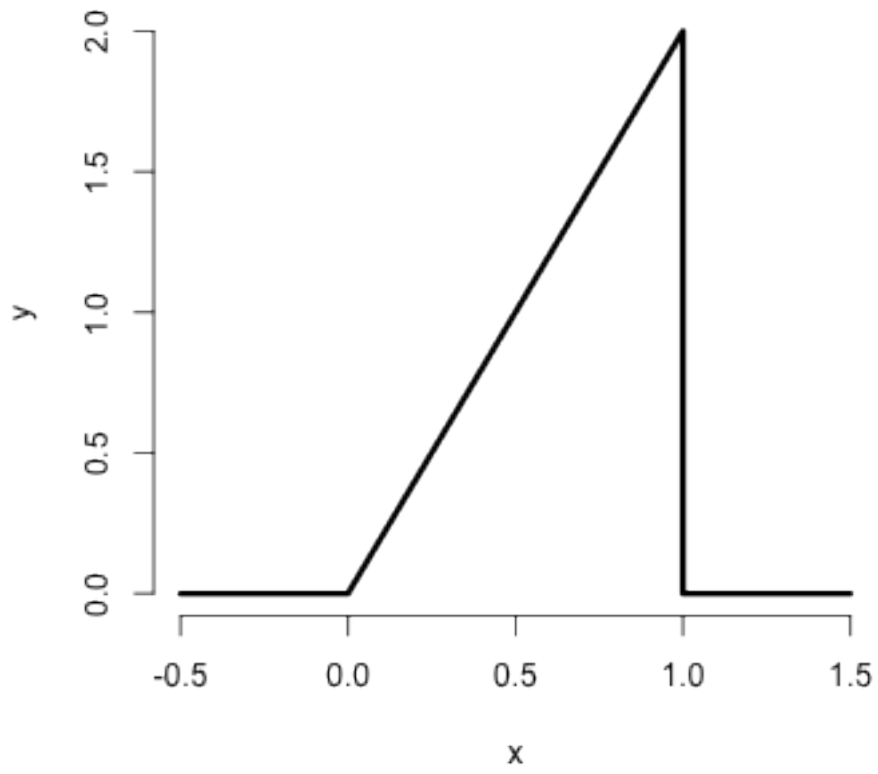
Example

Suppose that the proportion of help calls that get addressed in a random day by a help line is given by $f(x) = 2x$ for $0 < x < 1$. The R code for plotting this density is

Code for plotting the density

```
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")
```

The result of the code is given below.

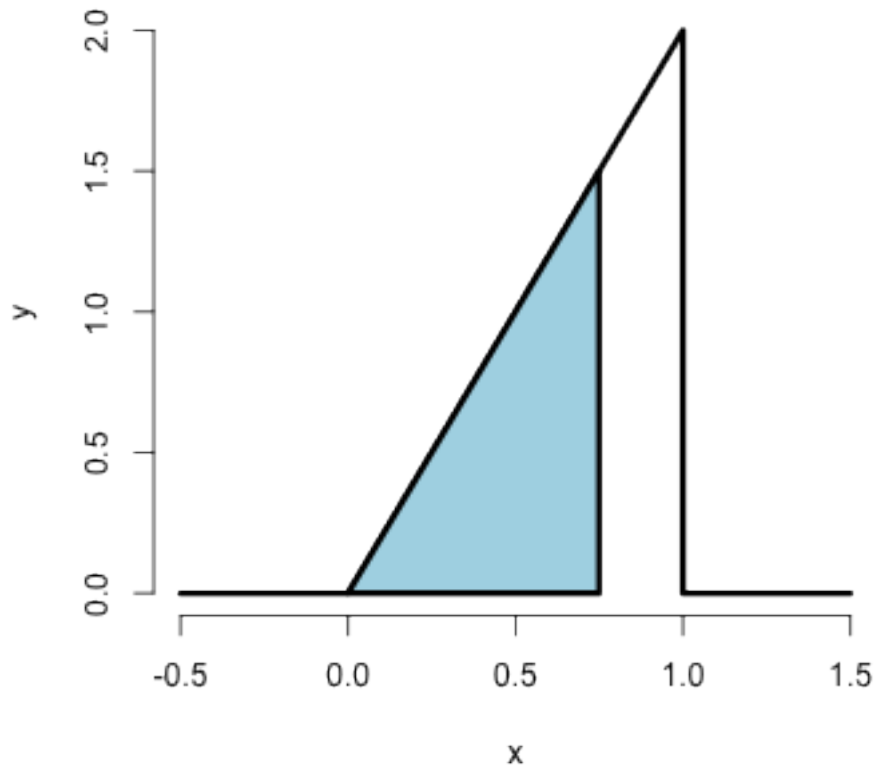


Help call density

Is this a mathematically valid density? To answer this we need to make sure it satisfies our two conditions. First it's clearly nonnegative (it's at or above the horizontal axis everywhere). The area

is similarly easy. Being a right triangle in the only section of the density that is above zero, we can calculate it as 1/2 the area of the base times the height. This is $\frac{1}{2} \times 1 \times 2 = 1$

Now consider answering the following question. What is the probability that 75% or fewer of calls get addressed? Remember, for continuous random variables, probabilities are represented by areas underneath the density function. So, we want the area from 0.75 and below, as illustrated by the figure below.



Help call density

This again is a right triangle, with length of the base as 0.75 and height 1.5. The R code below shows the calculation.

```
> 1.5 * 0.75/2
```

```
[1] 0.5625
```

Thus, the probability of 75% or fewer calls getting addressed in a random day for this help line is 56%. We'll do this a lot throughout this class and work with more useful densities. It should be noted

that this specific density is a special case of the so called *beta* density. Below I show how to use R's built in evaluation function for the beta density to get the probability.

```
> pbeta(0.75, 2, 1)
```

```
[1] 0.5625
```

Notice the syntax `pbeta`. In R, a prefix of `p` returns probabilities, `d` returns the density, `q` returns the quantile and `r` returns generated random variables. (You'll learn what each of these does in subsequent sections.)

CDF and survival function

Certain areas of PDFs and PMFs are so useful, we give them names. The **cumulative distribution function** (CDF) of a random variable, X, returns the probability that the random variable is less than or equal to the value x. Notice the (slightly annoying) convention that we use an upper case X to denote a random, unrealized, version of the random variable and a lowercase x to denote a specific number that we plug into. (This notation, as odd as it may seem, dates back to Fisher and isn't going anywhere, so you might as well get used to it. Uppercase for unrealized random variables and lowercase as placeholders for numbers to plug into.) So we could write the following to describe the distribution function F:

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

This definition applies regardless of
whether the random variable is discrete or continuous. 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 $S(x) = 1 - F(x)$, since the survival function evaluated at a particular value of x is calculating the probability of the opposite event (greater than as opposed to less than or equal to). The survival function is often preferred in biostatistical applications while the distribution function is more generally used (though both convey the same information.)

Example

What are the survival function and CDF from the density considered before?

$$F(x) = P(X \leq x) = \frac{1}{2} Base \times Height = \frac{1}{2}(x) \times (2x) = x^2,$$

for $1 \geq x \geq 0$. Notice that calculating the survival function is now trivial given that we've already calculated the distribution function.

$$S(x) = 1 = F(x) = 1 - x^2$$

Again, R has a function that calculates the distribution function for us in this case, `pbeta`. Let's try calculating $F(.4)$, $F(.5)$ and $F(.6)$

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

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

Notice, of course, these are simply the numbers squared. By default the prefix `p` in front of a density in R gives the distribution function (`pbeta`, `pnorm`, `pgamma`). If you want the survival function values, you could always subtract by one, or give the argument `lower.tail = FALSE` as an argument to the function, which asks R to calculate the upper area instead of the lower.

Quantiles

You've heard of sample quantiles. If you were the 95th percentile on an exam, you know that 95% of people scored worse than you and 5% scored better. These are sample quantities. But you might have wondered, what are my sample quantiles estimating? In fact, they are estimating the population quantiles. Here we define these population analogs.

The α^{th} **quantile** of a distribution with distribution function F is the point x_α so that

$$F(x_\alpha) = \alpha$$

So the 0.95 quantile of a distribution is the point so that 95% of the mass of the density lies below it. Or, in other words, the point so that the probability of getting a randomly sampled point below it is 0.95. This is analogous to the sample quantiles where the 0.95 sample quantile is the value so that 95% of the data lies below it.

A **percentile** is simply a quantile with α expressed as a percent rather than a proportion. The (population) **median** is the 50^{th} percentile. Remember that percentiles are not probabilities! Remember that quantiles have units. So the population median height is the height (in inches say) so that the probability that a randomly selected person from the population is shorter is 50%. The sample, or empirical, median would be the height so in a sample so that 50% of the people in the sample were shorter.

Example

What is the median of the distribution that we were working with before? We want to solve $0.5 = F(x) = x^2$, resulting in the solution

```
> sqrt(0.5)
```

```
[1] 0.7071
```

Therefore, 0.7071 of calls being answered on a random day is the median. Or, the probability that 70% or fewer calls get answered is 50%.

R can approximate quantiles for you for common distributions with the prefix `q` in front of the distribution name

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

```
[1] 0.7071
```

Exercises

1. Can you add the probabilities of any two events to get the probability of at least one occurring?
2. I define a PMF, p so that for $x = 0$ and $x = 1$ we have $p(0) = -0.1$ and $p(1) = 1.1$. Is this a valid PMF?
3. What is the probability that 75% or fewer calls get answered in a randomly sampled day from the population distribution from this chapter?
4. The 97.5th percentile of a distribution is?
5. Consider influenza epidemics for two parent heterosexual families. Suppose that the probability is 15% that at least one of the parents has contracted the disease. The probability that the father has contracted influenza is 10% while that the mother contracted the disease is 9%. What is the probability that both contracted influenza expressed as a whole number percentage? Watch a video solution to this problem.[13] and see a written out solution.[14]
6. A random variable, X, is uniform, a box from 0 to 1 of height 1. (So that it's density is $f(x) = 1$ for $0 \leq x \leq 1$.) What is it's median expressed to two decimal places? Watch a video solution to this problem here[15] and see written solutions here[16].
7. If a continuous density that never touches the horizontal axis is symmetric about zero, can we say that its associated median is zero? Watch a worked out solution to this problem here[17] and see the question and a typed up answer here[18]

[13] http://youtu.be/CvnmoCuIN08?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[14] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#3
[15] http://youtu.be/UXcarD-1xAM?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[16] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#4
[17] http://youtu.be/sn48CGH_TXI?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[18] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#9

3. Conditional probability

Conditional probability, motivation

Watch this video before beginning.[1]

Conditioning is a central subject in statistics. If we are given information about a random variable, it changes the probabilities associated with it. For example, the probability of getting a one when rolling a (standard) die is usually assumed to be one sixth. If you were given the extra information that the die roll was an odd number (hence 1, 3 or 5) then *conditional on this new information*, the probability of a one is now one third.

This is the idea of conditioning, taking away the randomness that we know to have occurred. Consider another example, such as the result of a diagnostic imaging test for lung cancer. What's the probability that a person has cancer given a positive test? How does that probability change under the knowledge that a patient has been a lifetime heavy smoker and both of their parents had lung cancer? *Conditional* on this new information, the probability has increased dramatically.

Conditional probability, definition

We can formalize the definition of conditional probability so that the mathematics matches our intuition.

Let B be an event so that $P(B) > 0$. Then the conditional probability of an event A given that B has occurred is:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

If A and B are unrelated in any way, or in other words *independent*, (discussed more later in the lecture), then

$$P(A \mid B) = \frac{P(A)P(B)}{P(B)} = P(A)$$

That is, if the occurrence of B offers no information about the occurrence of A - the probability conditional on the information is the same as the probability without the information, we say that the two events are independent.

[1] http://youtu.be/u6AH6qsSVA4?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Example

Consider our die roll example again. Here we have that $B = \{1, 3, 5\}$ and $A = \{1\}$

$$P(\text{one given that roll is odd}) = P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$

Which exactly mirrors our intuition.

Bayes' rule

Watch this video before beginning[2]

Bayes' rule is a famous result in statistics and probability. It forms the foundation for large branches of statistical thinking. Bayes' rule allows us to reverse the conditioning set provided that we know some marginal probabilities.

Why is this useful? Consider our lung cancer example again. It would be relatively easy for physicians to calculate the probability that the diagnostic method is positive for people with lung cancer and negative for people without. They could take several people who are already known to have the disease and apply the test and conversely take people known not to have the disease. However, for the collection of people with a positive test result, the reverse probability is more of interest, "given a positive test what is the probability of having the disease?", and "given a given a negative test what is the probability of not having the disease?".

Bayes' rule allows us to switch the conditioning event, provided a little bit of extra information. Formally Bayes' rule is:

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)}.$$

Diagnostic tests

Since diagnostic tests are a really good example of Bayes' rule in practice, let's go over them in greater detail. (In addition, understanding Bayes' rule will be helpful for your own ability to understand medical tests that you see in your daily life). We require a few definitions first.

Let $+$ and $-$ be the events that the result of a diagnostic test is positive or negative respectively Let D and D^c be the event that the subject of the test has or does not have the disease respectively

The **sensitivity** is the probability that the test is positive given that the subject actually has the disease, $P(+ \mid D)$

[2] http://youtu.be/TfeaZ_26iQk?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

The **specificity** is the probability that the test is negative given that the subject does not have the disease, $P(- \mid D^c)$

So, conceptually at least, the sensitivity and specificity are straightforward to estimate. Take people known to have and not have the disease and apply the diagnostic test to them. However, the reality of estimating these quantities is quite challenging. For example, are the people known to have the disease in its later stages, while the diagnostic will be used on people in the early stages where it's harder to detect? Let's put these subtleties to the side and assume that they are known well.

The quantities that we'd like to know are the predictive values.

The **positive predictive value** is the probability that the subject has the disease given that the test is positive, $P(D \mid +)$

The **negative predictive value** is the probability that the subject does not have the disease given that the test is negative, $P(D^c \mid -)$

Finally, we need one last thing, the **prevalence of the disease** - which is the marginal probability of disease, $P(D)$. Let's now try to figure out a PPV in a specific setting.

Example

A study comparing the efficacy of HIV tests, reports on an experiment which concluded that HIV antibody tests have a sensitivity of 99.7% and a specificity of 98.5% Suppose that a subject, from a population with a .1% prevalence of HIV, receives a positive test result. What is the positive predictive value?

Mathematically, we want $P(D \mid +)$ given the sensitivity, $P(+ \mid D) = .997$, the specificity, $P(- \mid D^c) = .985$ and the prevalence $P(D) = .001$.

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)} \\ &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + \{1 - P(- \mid D^c)\}\{1 - P(D)\}} \\ &= \frac{.997 \times .001}{.997 \times .001 + .015 \times .999} \\ &= .062 \end{aligned}$$

In this population a positive test result only suggests a 6% probability that the subject has the disease, (the positive predictive value is 6% for this test). If you were wondering how it could be so low for this test, the low positive predictive value is due to low prevalence of disease and the somewhat modest specificity

Suppose it was known that the subject was an intravenous drug user and routinely had intercourse with an HIV infected partner? Our prevalence would change dramatically, thus increasing the PPV. You might wonder if there's a way to summarize the evidence without appealing to an often unknowable prevalence? Diagnostic likelihood ratios provide this for us.

Diagnostic Likelihood Ratios

The diagnostic likelihood ratios summarize the evidence of disease given a positive or negative test. They are defined as:

The **diagnostic likelihood ratio of a positive test**, labeled DLR_+, is $P(+ \mid D)/P(+ \mid D^c)$, which is the $sensitivity/(1 - specificity)$.

The **diagnostic likelihood ratio of a negative test**, labeled DLR_-, is $P(- \mid D)/P(- \mid D^c)$, which is the $(1 - sensitivity)/specificity$.

How do we interpret the DLRs? This is easiest when looking at so called **odds ratios**. Remember that if p is a probability, then $p/(1 - p)$ is the odds. Consider now the odds in our setting:

Using Bayes rule, we have

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}$$

and

$$P(D^c \mid +) = \frac{P(+ \mid D^c)P(D^c)}{P(+ \mid D)P(D) + P(+ \mid D^c)P(D^c)}.$$

Therefore, dividing these two equations we have:

$$\frac{P(D \mid +)}{P(D^c \mid +)} = \frac{P(+ \mid D)}{P(+ \mid D^c)} \times \frac{P(D)}{P(D^c)}$$

In other words, the post test odds of disease is the pretest odds of disease times the DLR_+. Similarly, DLR_- relates the decrease in the odds of the disease after a negative test result to the odds of disease prior to the test.

So, the DLRs are the factors by which you multiply your pretest odds to get your post test odds. Thus, if a test has a DLR_+ of 6, regardless of the prevalence of disease, the post test odds is six times that of the pretest odds.

HIV example revisited

Let's reconsider our HIV antibody test again.
Suppose a subject has a positive HIV test

$$DLR_+ = .997/(1 - .985) = 66$$

The result of the positive test is that the odds of disease is now 66 times the pretest odds. Or, equivalently, the hypothesis of disease is 66 times more supported by the data than the hypothesis of no disease

Suppose instead that a subject has a negative test result

$$DLR_{-} = (1 - .997)/.985 = .003$$

Therefore, the post-test odds of disease is now 0.3% of the pretest odds given the negative test. Or, the hypothesis of disease is supported .003 times that of the hypothesis of absence of disease given the negative test result

Independence

Watch this video before beginning.[3]

Statistical independence of events is the idea that the events are unrelated. Consider successive coin flips. Knowledge of the result of the first coin flip tells us nothing about the second. We can formalize this into a definition.

Two events A and B are **independent** if

$$P(A \cap B) = P(A)P(B)$$

Equivalently if $P(A \mid B) = P(A)$. Note that since A is independent of B we know that A^c is independent of B A is independent of B^c A^c is independent of B^c.

While this definition works for sets, remember that random variables are really the things that we are interested in. Two random variables, X and Y are independent if for any two sets A and B $P([X \in A] \cap [Y \in B]) = P(X \in A)P(Y \in B)$

We will almost never work with these definitions. Instead, the important principle is that probabilities of independent things multiply! This has numerous consequences, including the idea that we shouldn't multiply non-independent probabilities.

Example

Let's cover a very simple example: "What is the probability of getting two consecutive heads?". Then we have that A is the event of getting a head on flip 1 $P(A) = 0.5$ B is the event of getting a head on flip 2 $P(B) = 0.5$ $A \cap B$ is the event of getting heads on flips 1 and 2. Then independence would tell us that:

[3] http://youtu.be/MY1EfrR1ZUs?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

$$P(A \cap B) = P(A)P(B) = 0.5 \times 0.5 = 0.25$$

This is exactly what we would have intuited of course. But, it's nice that the mathematics mirrors our intuition. In more complex settings, it's easy to get tripped up. Consider the following famous (among statisticians at least) case study.

Case Study

Volume 309 of Science reports on a physician who was on trial for expert testimony in a criminal trial. Based on an estimated prevalence of sudden infant death syndrome (SIDS) of 1 out of 8,543, a physician testified that that the probability of a mother having two children with SIDS was $(1/8,543)^2$. The mother on trial was convicted of murder.

Relevant to this discussion, the principal mistake was to *assume* that the events of having SIDs within a family are independent. That is, $P(A_1 \cap A_2)$ is not necessarily equal to $P(A_1)P(A_2)$. This is because biological processes that have a believed genetic or familiar environmental component, of course, tend to be dependent within families. Thus, we can't just multiply the probabilities to obtain the result.

There are many other interesting aspects to the case. For example, the idea of a low probability of an event representing evidence against a plaintiff. (Could we convict all lottery winners of fixing the lotter since the chance that they would win is so small.)

IID random variables

Now that we've introduced random variables and independence, we can introduce a central modeling assumption made in statistics. Specifically the idea of a random sample. Random variables are said to be independent and identically distributed (*iid*) if they are independent and all are drawn from the same population. The reason iid samples are so important is that they are a model for random samples. This is a default starting point for most statistical inferences.

The idea of having a random sample is powerful for a variety of reasons. Consider that in some study designs, such as in election polling, great pains are made to make sure that the sample is randomly drawn from a population of interest. The idea is to expend a lot of effort on design to get robust inferences. In these settings assuming that the data is iid is both natural and warranted.

In other settings, the study design is far more opaque, and statistical inferences are conducted under the assumption that the data arose from a random sample, since it serves as a useful benchmark. Most studies in the fields of epidemiology and economics fall under this category. Take, for example, studying how policies impact countries gross domestic product by looking at countries before and after enacting the policies. The countries are not a random sample from the set of countries. Instead, conclusions must be made under the assumption that the countries are a random sample and the interpretation of the strength of the inferences adapted in kind.

Exercises

1. I pull a card from a deck and do not show you the result. I say that the resulting card is a heart. What is the probability that it is the queen of hearts?
2. The odds associated with a probability, p, are defined as?
3. The probability of getting two sixes when rolling a pair of dice is?
4. The probability that a manuscript gets accepted to a journal is 12% (say). However, given that a revision is asked for, the probability that it gets accepted is 90%. Is it possible that the probability that a manuscript has a revision asked for is 20%? Watch a video of this problem getting solved[4] and see the worked out solutions here[5].
5. Suppose 5% of housing projects have issues with asbestos. The sensitivity of a test for asbestos is 93% and the specificity is 88%. What is the probability that a housing project has no asbestos given a negative test expressed as a percentage to the nearest percentage point? Watch a video solution here[6] and see the worked out problem here[7].

[4] http://youtu.be/E4kE4M1J15s?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[5] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#3
[6] https://www.youtube.com/watch?v=rbI97tSvGvQ&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=11
[7] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#5

4. Expected values

Watch this video before beginning.[1]

Expected values characterize a distribution. The most useful expected value, the mean, characterizes the center of a density or mass function. Another expected value summary, the variance, characterizes how spread out a density is. Yet another expected value calculation is the skewness, which considers how much a density is pulled toward high or low values.

Remember, in this lecture we are discussing population quantities. It is convenient (and of course by design) that the names for all of the sample analogs estimate the associated population quantity. So, for example, the sample or empirical mean estimates the population mean; the sample variance estimates the population variance and the sample skewness estimates the population skewness.

The population mean for discrete random variables

The **expected value** or (population) **mean** of a random variable is the center of its distribution. For discrete random variable X with PMF $p(x)$, it is defined as follows:

$$E[X] = \sum_x xp(x).$$

where the sum is taken over the possible values of x. Where did they get this idea from? It's taken from the physical idea of the center of mass. Specifically, $E[X]$ represents the center of mass of a collection of locations and weights, $\{x, p(x)\}$. We can exploit this fact to quickly calculate population means for distributions where the center of mass is obvious.

The sample mean

It is important to contrast the population mean (the estimand) with the sample mean (the estimator). The sample mean estimates the population mean. Not coincidentally, since the population mean is the center of mass of the population distribution, the sample mean is the center of mass of the data. In fact, it's exactly the same equation:

$$\bar{X} = \sum_{i=1}^{n} x_i p(x_i),$$

where $p(x_i) = 1/n$.

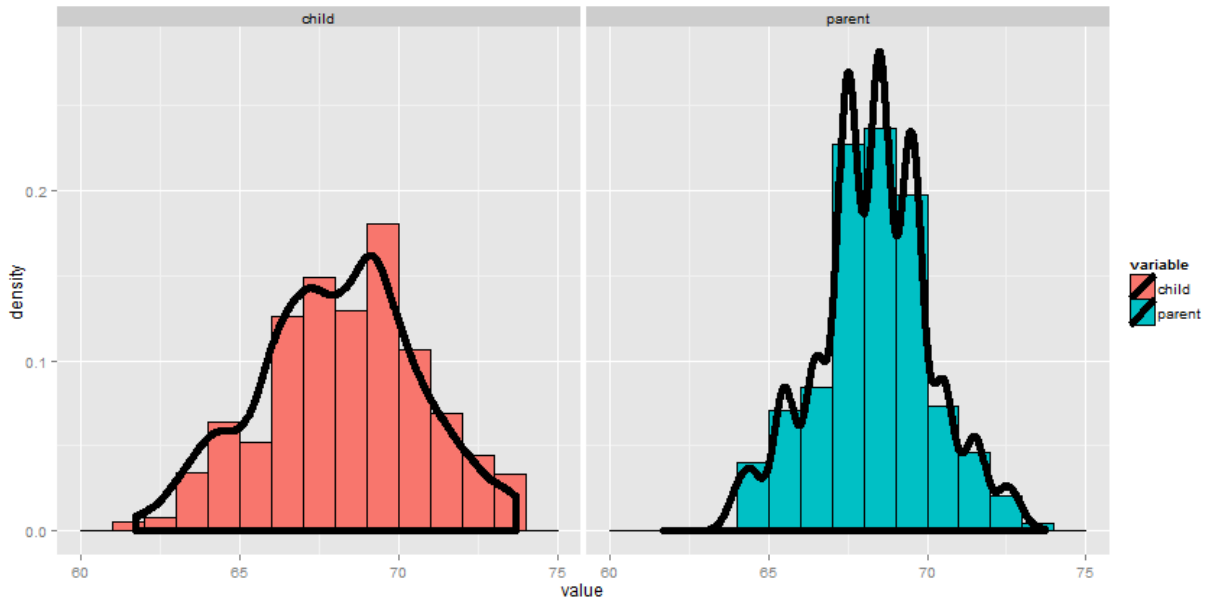
[1] http://youtu.be/zljxRbu6jyc?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Example Find the center of mass of the bars

Let's go through an example of illustrating how the sample mean is the center of mass of observed data. Below we plot Galton's fathers and sons data:

Loading in and displaying the Galton data

```
library(UsingR); data(galton); library(ggplot2); library(reshape2)
longGalton <- melt(galton, measure.vars = c("child", "parent"))
g <- ggplot(longGalton, aes(x = value)) + geom_histogram(aes(y = ..density..,  f\
ill = variable), binwidth=1, color = "black") + geom_density(size = 2)
g <- g + facet_grid(. ~ variable)
g
```



Galton's Data

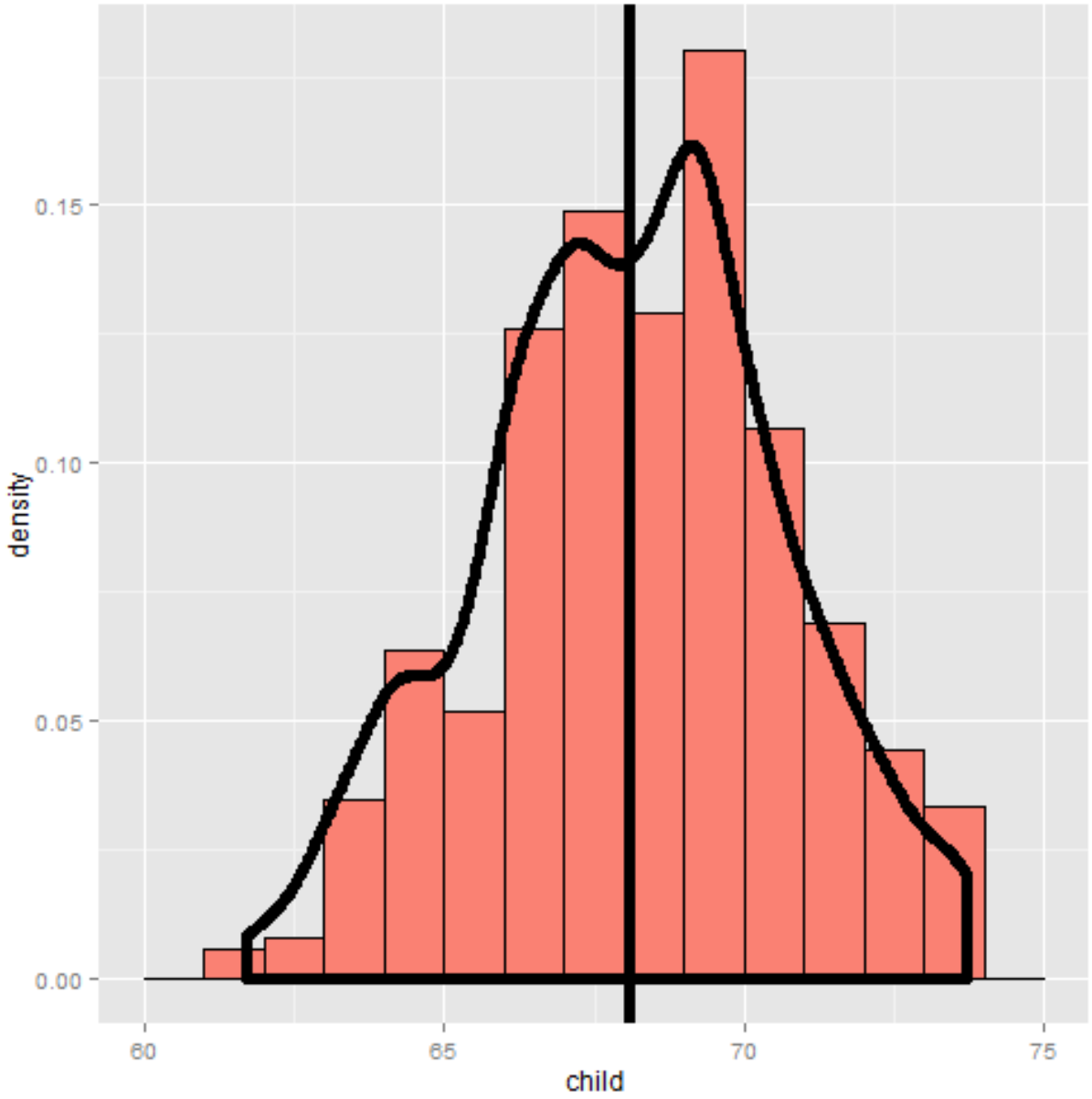
Using rStudio's `manipulate` package, you can try moving the histogram around and see what value balances it out. Be sure to watch the video to see this in action.

Using manipulate to explore the mean

```
library(manipulate)
myHist <- function(mu){
    g <- ggplot(galton, aes(x = child))
    g <- g + geom_histogram(fill = "salmon",
      binwidth=1, aes(y = ..density..), color = "black")
    g <- g + geom_density(size = 2)
    g <- g + geom_vline(xintercept = mu, size = 2)
    mse <- round(mean((galton$child - mu)^2), 3)
    g <- g + labs(title = paste('mu = ', mu, ' MSE = ', mse))
    g
}
manipulate(myHist(mu), mu = slider(62, 74, step = 0.5))
```

Going through this exercise, you find that the point that balances out the histogram is the empirical mean. (Note there's a small distinction here that comes about from rounding with the histogram bar widths, but ignore that for the time being.) If the bars of the histogram are from the observed data, the point that balances it out is the empirical mean; if the bars are the true population probabilities (which we don't know of course) then the point is the population mean. Let's now go through some examples of mathematically calculating the population mean.

The center of mass is the empirical mean



Histogram illustration

Example of a population mean, a fair coin

Watch the video before beginning here.[2]

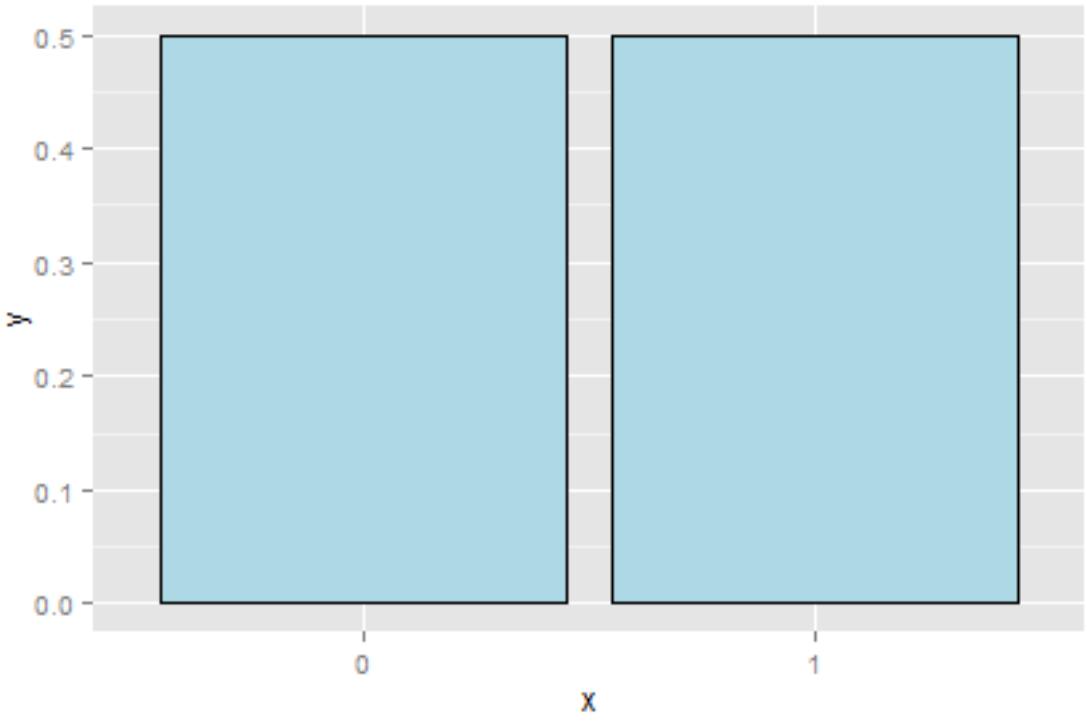
Suppose a coin is flipped and X is declared 0 or 1 corresponding to a head or a tail, respectively.

[2] http://youtu.be/F4XMuD_axN8?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

What is the expected value of X?

$$E[X] = .5 \times 0 + .5 \times 1 = .5$$

Note, if thought about geometrically, this answer is obvious; if two equal weights are spaced at 0 and 1, the center of mass will be 0.5.



Fair coin mass function

What about a biased coin?

Suppose that a random variable, X , is so that $P(X = 1) = p$ and $P(X = 0) = (1 - p)$ (This is a biased coin when $p \neq 0.5$.) What is its expected value?

$$E[X] = 0 * (1 - p) + 1 * p = p$$

Notice that the expected value isn't a value that the coin can take in the same way that the sample proportion of heads will also likely be neither 0 nor 1.

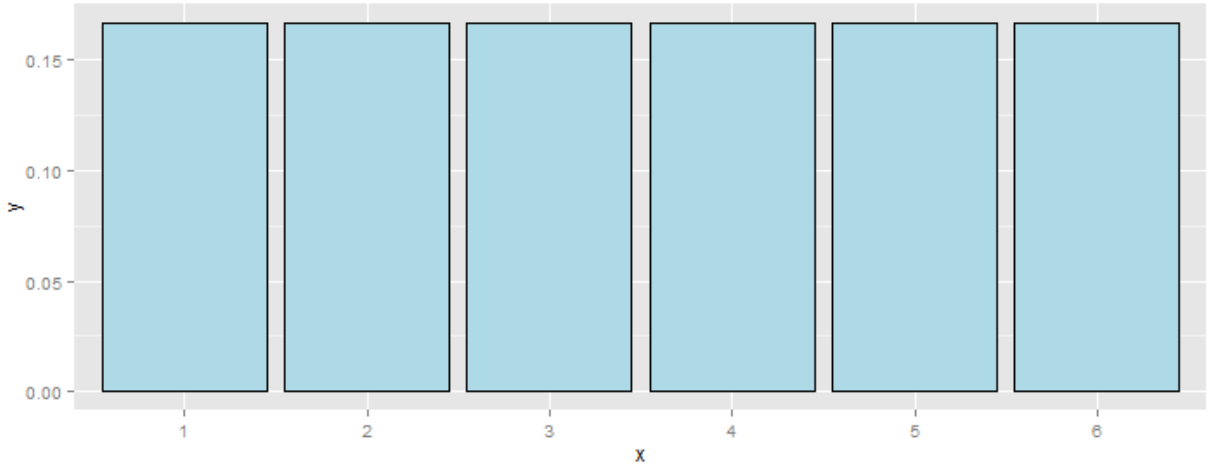
This coin example is not exactly trivial as it serves as the basis for a random sample of any population for a binary trait. So, we might model the answer from an election polling question as if it were a coin flip.

Example Die Roll

Suppose that a die is rolled and X is the number face up. What is the expected value of X?

$$E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$$

Again, the geometric argument makes this answer obvious without calculation.



Bar graph of die probabilities

Continuous random variables

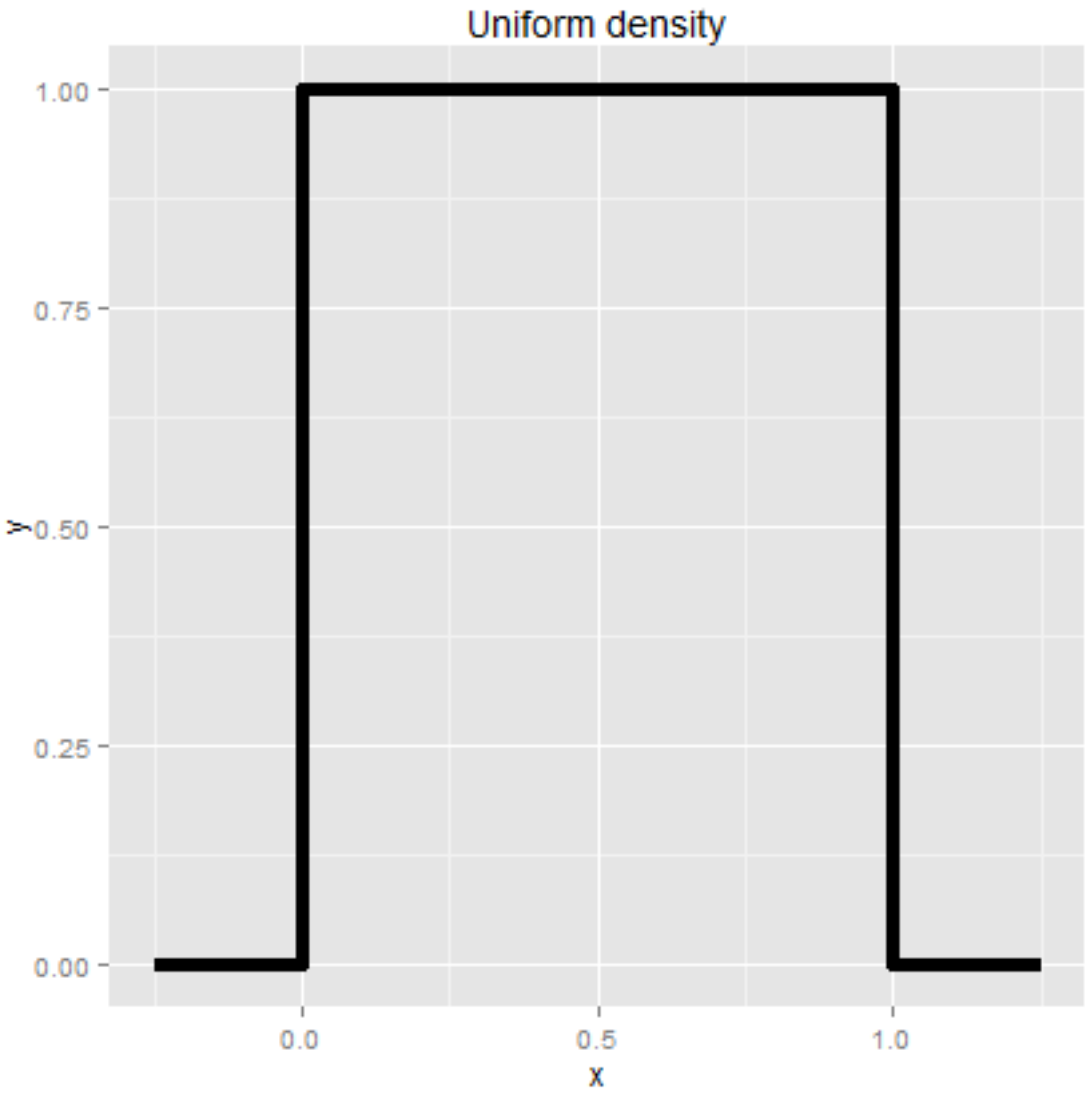
Watch this video before beginning.[3]

For a continuous random variable, X, with density, f, the expected value is again exactly the center of mass of the density. Think of it like cutting the continuous density out of a thick piece of wood and trying to find the point where it balances out.

Example

Consider a density where $f(x) = 1$ for x between zero and one. Suppose that X follows this density; what is its expected value?

[3] http://youtu.be/YS5EIKsamXI?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ



Uniform Density

The answer is clear since the density looks like a box, it would balance out exactly in the middle, 0.5.

Facts about expected values

Recall that expected values are properties of population distributions. The expected value, or mean, height is the center of the population density of heights.

Of course, the average of ten randomly sampled people's height is itself a random variable, in the same way that the average of ten die rolls is itself a random number. Thus, the distribution of heights

gives rise to the distribution of averages of ten heights in the same way that distribution associated with a die roll gives rise to the distribution of the average of ten dice.

An important question to ask is: "What does the distribution of averages look like?". This question is important, since it tells us things about averages, the best way to estimate the population mean, when we only get to observe one average.

Consider the die rolls again. If wanted to know the distribution of averages of 100 die rolls, you could (at least in principle) roll 100 dice, take the average and repeat that process. Imagine, if you could only roll the 100 dice once. Then we would have direct information about the distribution of die rolls (since we have 100 of them), but we wouldn't have any direct information about the distribution of the average of 100 die rolls, since we only observed one average.

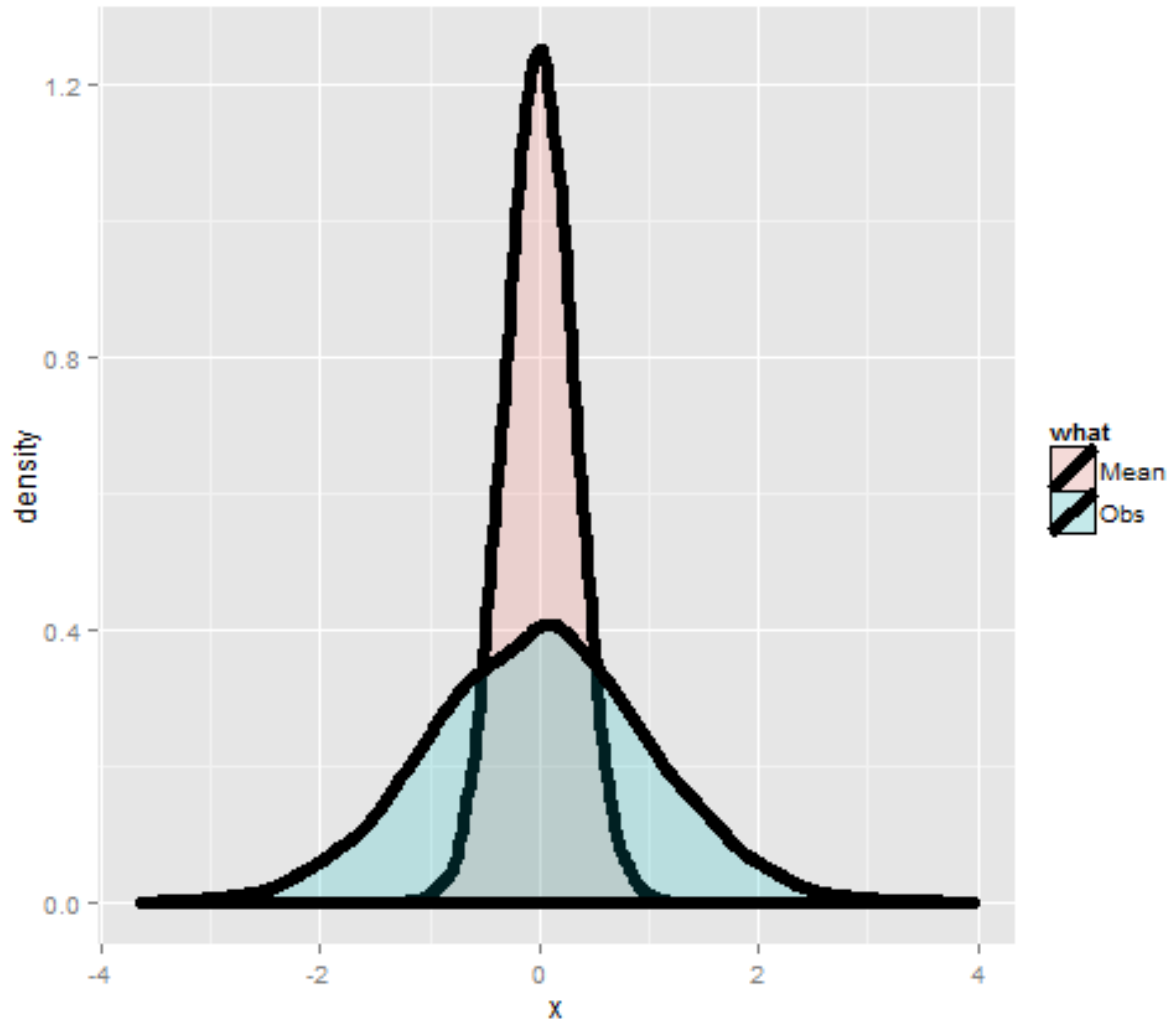
Fortunately, the mathematics tells us about that distribution. Notably, it's centered at the same spot as the original distribution! Thus, the distribution of the estimator (the sample mean) is centered at the distribution of what it's estimating (the population mean). When the expected value of an estimator is what its trying to estimate, we say that the estimator is **unbiased**.

Let's go through several simulation experiments to see this more fully.

Simulation experiments

Standard normals

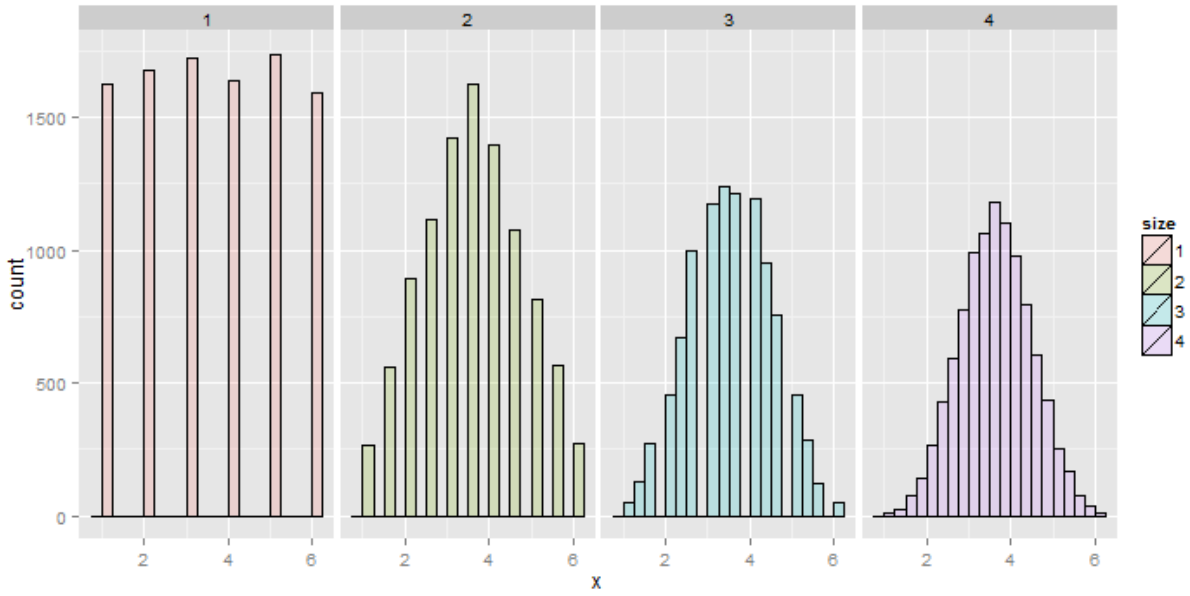
Consider simulating a lot of standard normals and plotting a histogram (the blue density). Now consider simulating lots of averages of 10 standard normals and plotting their histogram (the salmon colored density). Notice that they're centered in the same spot! It's also more concentrated around that point. (We'll discuss that more in the next lectures).



Simulation of normals

Averages of x die rolls

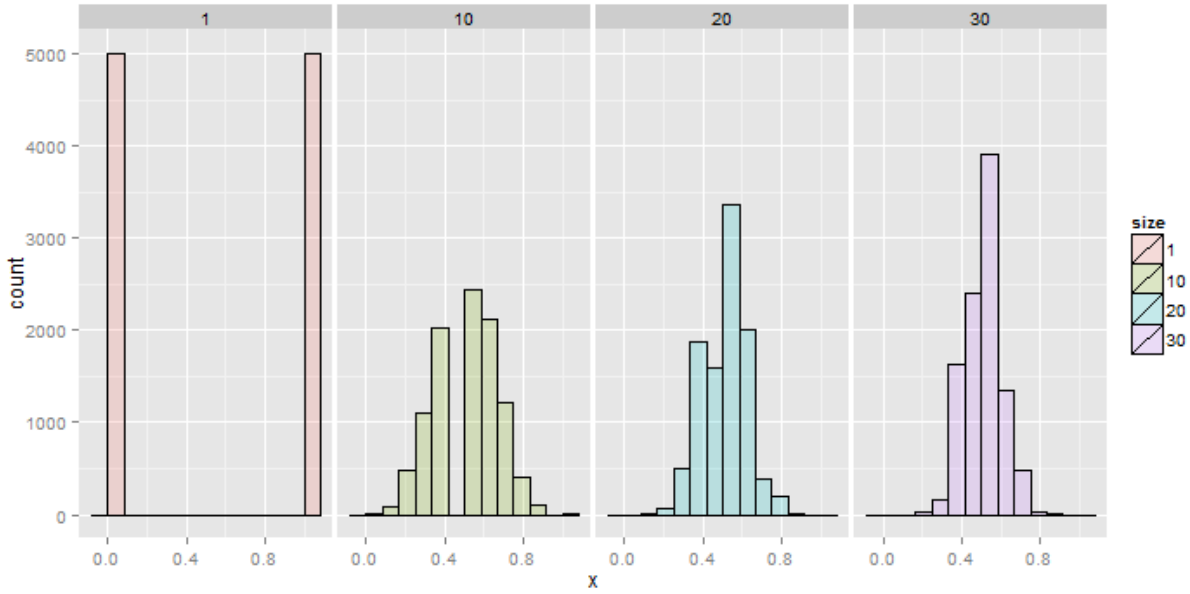
Consider rolling a die a lot of times and taking a histogram of the results, that's the left most plot. The bars are equally distributed at the six possible outcomes and thus the histogram is centered around 3.5. Now consider simulating lots of averages of 2 dice. Its histogram is also centered at 3.5. So is it for 3 and 4. Notice also the distribution gets increasing Gaussian looking (like a bell curve) and increasingly concentrated around 3.5.



Simulation of die rolls

Averages of x coin flips

For the coin flip simulation exactly the same occurs. All of the distributions are centered around 0.5.



Simulation of coin flips

Summary notes

- Expected values are properties of distributions.
- 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.

Exercises

1. A standard die takes the values 1, 2, 3, 4, 5, 6 with equal probability. What is the expected value?
2. Consider a density that is uniform from -1 to 1. (I.e. has height equal to 1/2 and looks like a box starting at -1 and ending at 1). What is the mean of this distribution?
3. If a population has mean μ, what is the mean of the distribution of averages of 20 observations from this distribution?
4. 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 odds that the coin is heads is d. What is your expected earnings? Watch a video of the solution to this problem[4] and look at the problem and the solution here.[5].
5. If you roll ten standard dice, take their average, then repeat this process over and over and construct a histogram what would it be centered at? Watch a video solution here[6] and see the original problem here[7].

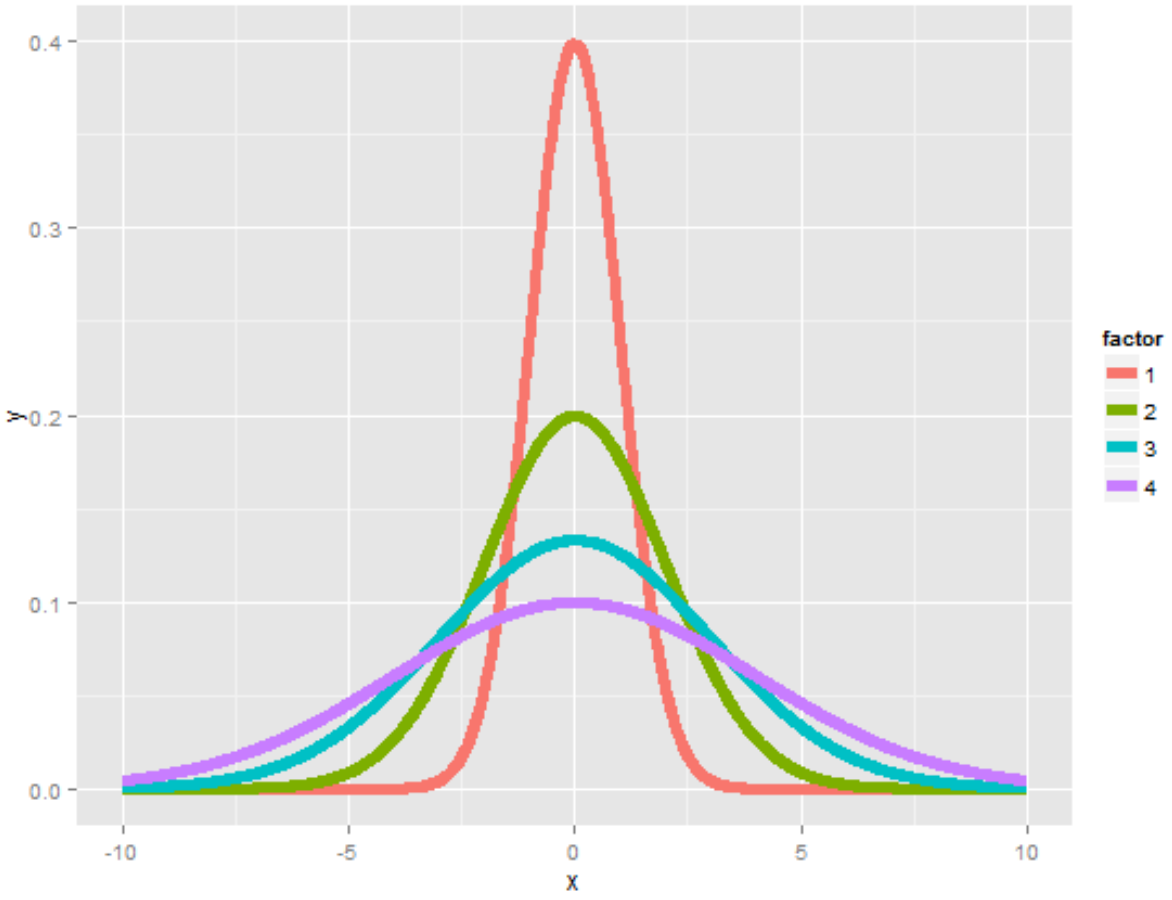
[4] http://youtu.be/5J88Zq0q81o?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[5] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#5
[6] https://www.youtube.com/watch?v=ia3n2URiJaw&index=16&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[7] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#11

5. Variation

The variance

Watch this video before beginning.[1]

Recall that the mean of distribution was a measure of its center. The variance, on the other hand, is a measure of *spread*. To get a sense, the plot below shows a series of increasing variances.



Distributions with increasing variance

We saw another example of how variances changed in the last chapter when we looked at the distribution of averages; they were always centered at the same spot as the original distribution, but

[1] http://youtu.be/oLQVU-VRiHo?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

are less spread out. Thus, it is less likely for sample means to be far away from the population mean than it is for individual observations. (This is why the sample mean is a better estimate than the population mean.)

If X is a random variable with mean μ, the variance of X is defined as

$$Var(X) = E[(X - \mu)^2] = E[X^2] - E[X]^2.$$

The rightmost equation is the shortcut formula that is almost always used for calculating variances in practice.
Thus the variance is the expected (squared) distance from the mean. Densities with a higher variance are more spread out than densities with a lower variance. The square root of the variance is called the **standard deviation**. The main benefit of working with standard deviations is that they have the same units as the data, whereas the variance has the units squared.

In this class, we'll only cover a few basic examples for calculating a variance. Otherwise, we're going to use the ideas without the formalism. Also remember, what we're talking about is the population variance. It measures how spread out the population of interest is, unlike the sample variance which measures how spread out the observed data are. Just like the sample mean estimates the population mean, the sample variance will estimate the population variance.

Example

What's the variance from the result of a toss of a die? First recall that $E[X] = 3.5$, as we discussed in the previous lecture. Then let's calculate the other bit of information that we need, $E[X^2]$.

$$E[X^2] = 1^2 \times \frac{1}{6} + 2^2 \times \frac{1}{6} + 3^2 \times \frac{1}{6} + 4^2 \times \frac{1}{6} + 5^2 \times \frac{1}{6} + 6^2 \times \frac{1}{6} = 15.17$$

Thus now we can calculate the variance as:

$$Var(X) = E[X^2] - E[X]^2 \approx 2.92.$$

Example

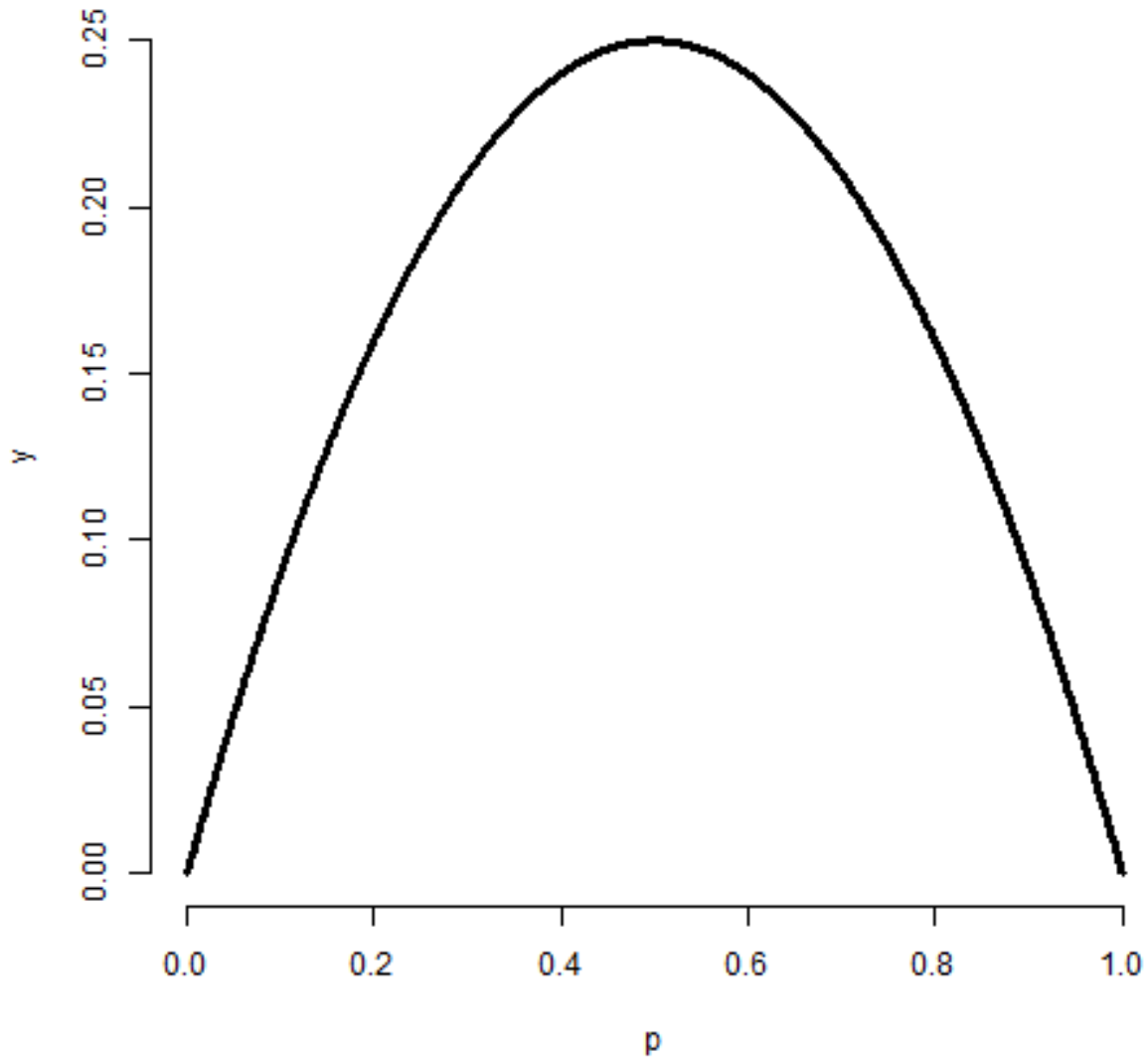
What's the variance from the result of the toss of a (potentially biased) coin with probability of heads (1) of p? First recall that $E[X] = 0 \times (1 - p) + 1 \times p = p$. Secondly, recall that since X is either 0 or 1, $X^2 = X$. So we know that:

$$E[X^2] = E[X] = p.$$

Thus we can now calculate the variance of a coin flip as $Var(X) = E[X^2] - E[X]^2 = p - p^2 = p(1 - p)$. This is a well known formula, so it's worth committing to memory. It's interesting to note that this function is maximized at $p = 0.5$. The plot below shows this by plotting $p(1 - p)$ by p.

Plotting the binomial variance

```
p = seq(0 , 1, length = 1000)
y = p * (1 - p)
plot(p, y, type = "l", lwd = 3, frame = FALSE)
```



Plot of the binomial variance

The sample variance

The sample variance is the estimator of the population variance. Recall that the population variance is the expected squared deviation around the population mean. The sample variance is (almost) the average squared deviation of observations around the sample mean. It is given by

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

The sample standard deviation is the square root of the sample variance. Note again that the sample variance is almost, but not quite, the average squared deviation from the sample mean since we divide by $n-1$ instead of n. Why do we do this you might ask? To answer that question we have to think in the terms of simulations. Remember that the sample variance is a random variable, thus it has a distribution and that distribution has an associated population mean. That mean is the population variance that we're trying to estimate if we divide by $(n-1)$ rather than n.

It is also nice that as we collect more data the distribution of the sample variance gets more concentrated around the population variance that it's estimating.

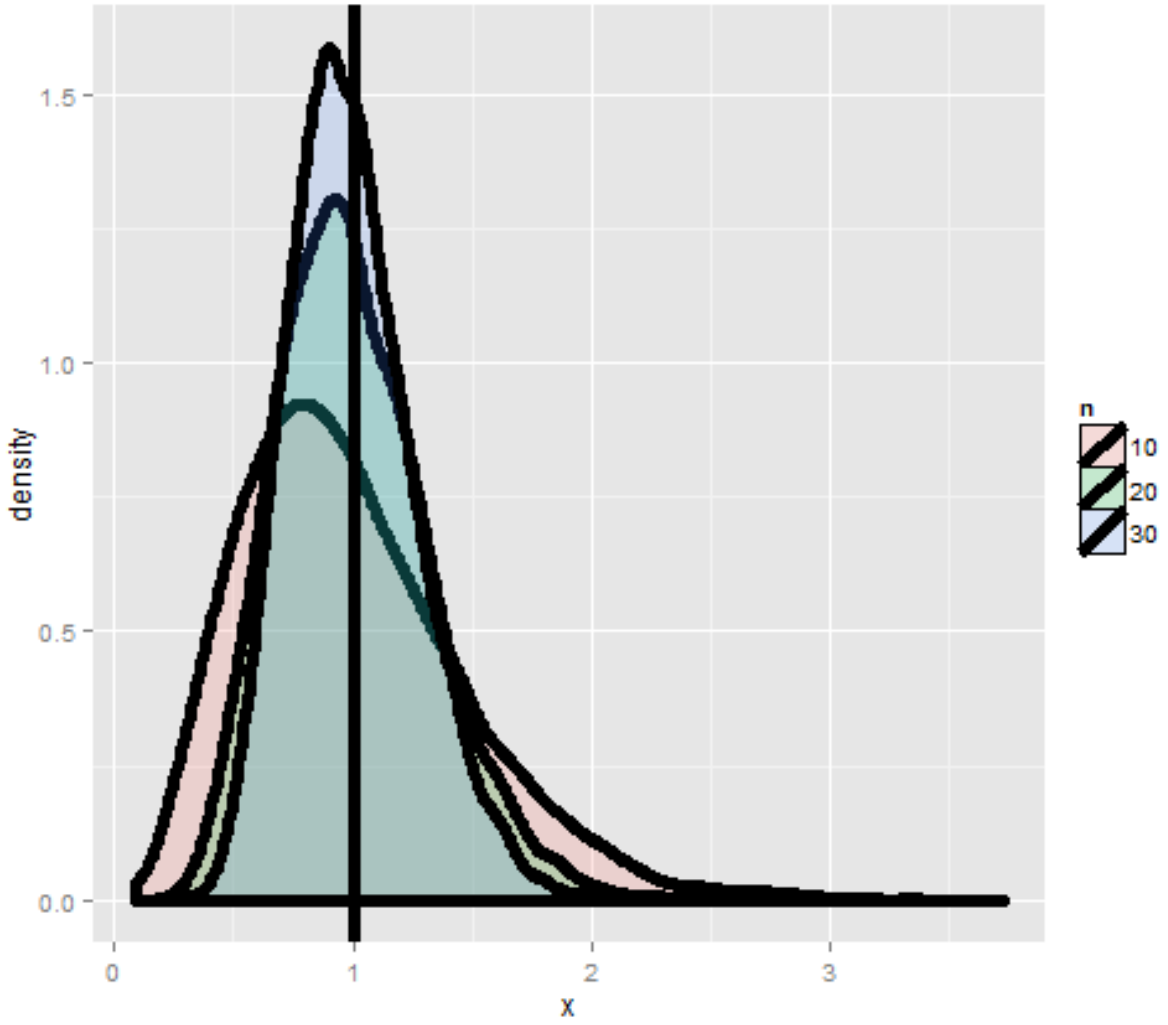
Simulation experiments

Watch this video before beginning.[2]

Simulating from a population with variance 1

Let's try simulating collections of standard normals and taking the variance. If we repeat this over and over, we get a sense of the distribution of sample variances variances.

[2] http://youtu.be/uPjHB9JjGKI?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

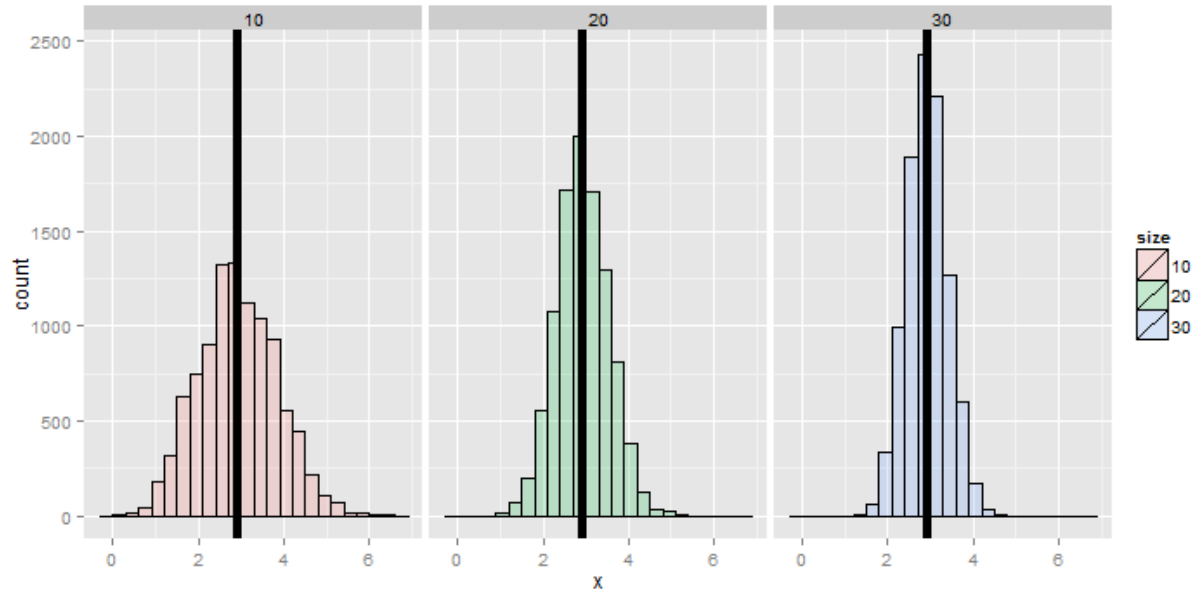


Simulation of variances of samples of standard normals

Notice that these histograms are always centered in the same spot, 1. In other words, the sample variance is an unbiased estimate of the population variances. Notice also that they get more concentrated around the 1 as more data goes into them. Thus, sample variances comprised of more observations are less variable than sample variances comprised of fewer.

Variances of x die rolls

Let's try the same thing, now only with die rolls instead of simulating standard normals. In this experiment, we simulated samples of die rolls, took the variance and then repeated that process over and over. What is plotted are histograms of the collections of sample variances.



Simulated distributions of variances of dies

Recall that we calculated the variance of a die roll as 2.92 earlier on in this chapter. Notice each of the histograms are centered there. In addition, they get more concentrated around 2.92 as more the variances are comprised of more dice.

The standard error of the mean

At last, we finally get to a perhaps very surprising (and useful) fact: how to estimate the variability of the mean of a sample, when we only get to observe one realization. Recall that the average of random sample from a population is itself a random variable having a distribution, which in simulation settings we can explore by repeated sampling averages. We know that this distribution is centered around the population mean, $E[\bar{X}] = \mu$. We also know the variance of the distribution of means of random samples.

The variance of the sample mean is: $Var(\bar{X}) = \sigma^2/n$ where σ^2 is the variance of the population being sampled from.

This is very useful, since we don't have repeat sample means to get its variance directly using the data. We already know a good estimate of σ^2 via the sample variance. So, we can get a good estimate of the variability of the mean, even though we only get to observe 1 mean.

Notice also this explains why in all of our simulation experiments the variance of the sample mean kept getting smaller as the sample size increased. This is because of the square root of the sample size in the denominator.

Often we take the square root of the variance of the mean to get the standard deviation of the mean. We call the standard deviation of a statistic its standard error.

Summary notes

- The sample variance, S^2, estimates the population variance, σ^2.
- The distribution of the sample variance is centered around σ^2.
- The variance of the sample mean is σ^2/n.
  - Its logical estimate is s^2/n.
  - The logical estimate of the standard error is $S/\sqrt{n}$.
- S, the standard deviation, talks about how variable the population is.
- $S/\sqrt{n}$, the standard error, talks about how variable averages of random samples of size n from the population are.

Simulation example 1: standard normals

Watch this video before beginning.[3]

Standard normals have variance 1. Let's try sampling means of n standard normals. If our theory is correct, they should have standard deviation $1/\sqrt{n}$

Simulating means of random normals

```
> nosim <- 1000
> n <- 10
## simulate nosim averages of 10 standard normals
> sd(apply(matrix(rnorm(nosim * n), nosim), 1, mean))
[1] 0.3156
## Let's check to make sure that this is sigma / sqrt(n)
> 1 / sqrt(n)
[1] 0.3162
```

So, in this simulation, we simulated 1000 means of 10 standard normals. Our theory says the standard deviation of averages of 10 standard normals must be $1/\sqrt{n}$. Taking the standard deviation of the 10000 means yields nearly exactly that. (Note that it's only close, 0.3156 versus 0.31632. To get it to be exact, we'd have to simulate infinitely many means.)

Simulation example 2: uniform density

Standard uniforms have variance $1/12$. Our theory mandates that means of random samples of n uniforms have sd $1/\sqrt{12 \times n}$. Let's try it with a simulation.

[3] http://youtu.be/uPjHB9JjGKI?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Simulating means of uniforms

```
> nosim <- 1000
> n <- 10
> sd(apply(matrix(runif(nosim * n), nosim), 1, mean))
[1] 0.09017
> 1 / sqrt(12 * n)
[1] 0.09129
```

Simulation example 3: Poisson

Poisson(4) random variables have variance 4. Thus means of random samples of n Poisson(4) should have standard deviation $2/\sqrt{n}$. Again let's try it out.

Simulating means of Poisson variates

```
> nosim <- 1000
> n <- 10
> sd(apply(matrix(rpois(nosim * n, 4), nosim), 1, mean))
[1] 0.6219
> 2 / sqrt(n)
[1] 0.6325
```

Simulation example 4: coin flips

Our last example is an important one. Recall that the variance of a coin flip is $p(1-p)$. Therefore the standard deviation of the average of n coin flips should be $\sqrt{\frac{p(1-p)}{n}}$.

Let's just do the simulation with a fair coin. Such coin flips have variance 0.25. Thus means of random samples of n coin flips have sd $1/(2\sqrt{n})$. Let's try it.

Simulating means of coin flips

```
> nosim <- 1000
> n <- 10
> sd(apply(matrix(sample(0 : 1, nosim * n, replace = TRUE),
                  nosim), 1, mean))
[1] 0.1587
> 1 / (2 * sqrt(n))
[1] 0.1581
```

Data example

Watch this before beginning.[4]

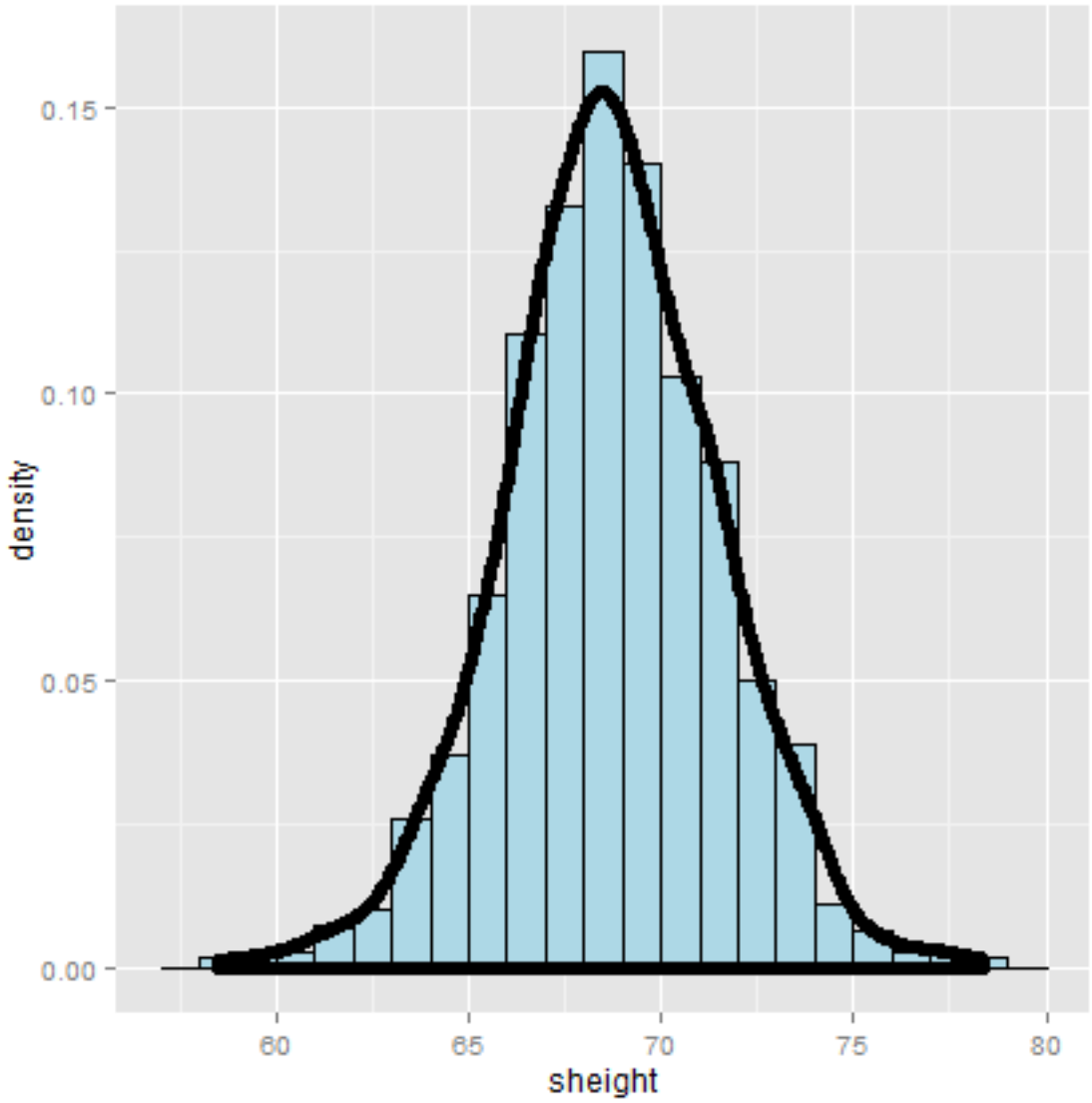
Now let's work through a data example to show how the standard error of the mean is used in practice. We'll use the `father.son` height data from Francis Galton.

Loading the data

```
library(UsingR); data(father.son);
x <- father.son$sheight
n<-length(x)
```

Here's a histogram of the sons' heights from the dataset. Let' calculate different variances and interpret them in this context.

[4] http://youtu.be/Lm2DMVyZVxk?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ



Histogram of the sons' heights

Loading the data

```
>round(c(var(x), var(x) / n, sd(x), sd(x) / sqrt(n)),2)
[1] 7.92 0.01 2.81 0.09
```

The first number, 7.92, and its square root, 2.81, are the estimated variance and standard deviation of the sons' heights. Therefore, 7.92 tells us exactly how variable sons' heights were in the data and estimates how variable sons' heights are in the population. In contrast 0.01, and the square root 0.09, estimate how variable averages of n sons' heights are.

Therefore, the smaller numbers discuss the precision of our estimate of the mean of sons' heights. The larger numbers discuss how variable sons' heights are in general.

Summary notes

- The sample variance estimates the population variance.
- The distribution of 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.
  - The square root is the standard error.
- It turns out that we can say a lot about the distribution of averages from random samples, even though we only get one to look at in a given data set.

Exercises

1. If I have a random sample from a population, the sample variance is an estimate of?
   - The population standard deviation.
   - The population variance.
   - The sample variance.
   - The sample standard deviation.
2. The distribution of the sample variance of a random sample from a population is centered at what?
   - The population variance.
   - The population mean.
3. I keep drawing samples of size n from a population with variance σ^2 and taking their average. I do this thousands of times. If I were to take the variance of the collection of averages, about what would it be?
4. You get a random sample of n observations from a population and take their average. You would like to estimate the variability of averages of $$n$$ observations from this population to better understand how precise of an estimate it is. Do you need to repeated collect averages to do this?
   - No, we can multiply our estimate of the population variance by $1/n$ to get a good estimate of the variability of the average.
   - Yes, you have to get repeat averages.
5. A random variable takes the value -4 with probability .2 and 1 with probability .8. What is the variance of this random variable? Watch a video solution to this problem.[5] and look at a version with a worked out solution.[6]
6. If $\bar{X}$ and $\bar{Y}$ are comprised of n iid random variables arising from distributions having means μ_x and μ_y, respectively and common variance σ^2 what is the variance $\bar{X} - \bar{Y}$? Watch a video solution to this problem here[7] and see a typed up solution here[8]

[5] http://youtu.be/Em-xJeQO1rc?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[6] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#6
[7] http://youtu.be/7zJhPzX6jns?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[8] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#7

7. Let X be a random variable having standard deviation σ. What can be said about the variance of X/σ? Watch a video solution to this problem here[9] and typed up solutions here[10].
8. Consider the following pmf given in R by the code `p <- c(.1, .2, .3, .4)` and ‘`x <- 2:5`‘. What is the variance? Watch a video solution to this problem here[11] and here is the problem worked out[12].
9. If you roll ten standard dice, take their average, then repeat this process over and over and construct a histogram, what would be its variance expressed to 3 decimal places? Watch a video solution here[13] and see the text here[14].

[9] http://youtu.be/0WUj18_BUPA?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[10] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#8
[11] http://youtu.be/HSn8n4DsGSg?list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[12] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw1.html#10
[13] https://www.youtube.com/watch?v=MLfo9zz1zX4&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=17
[14] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#12

6. Some common distributions

The Bernoulli distribution

The **Bernoulli distribution** arises as the result of a binary outcome, such as a coin flip. Thus, Bernoulli random variables take (only) the values 1 and 0 with probabilities of (say) p and $1-p$, respectively. Recall that the PMF for a Bernoulli random variable X is $P(X=x) = p^x(1-p)^{1-x}$.

The mean of a Bernoulli random variable is p and the variance is $p(1-p)$. If we let X be a Bernoulli random variable, it is typical to call $X=1$ as a "success" and $X=0$ as a "failure".

If a random variable follows a Bernoulli distribution with success probability p we write that $X \sim$ Bernoulli(p).

Bernoulli random variables are commonly used for modeling any binary trait for a random sample. So, for example, in a random sample whether or not a participant has high blood pressure would be reasonably modeled as Bernoulli.

Binomial trials

The **binomial random variables** are obtained as the sum of iid Bernoulli trials. So if a Bernoulli trial is the result of a coin flip, a binomial random variable is the total number of heads.

To write it out as mathematics, let $X_1, \ldots, X_n$ be iid Bernoulli(p), then $X = \sum_{i=1}^n X_i$ is a binomial random variable. We write out that $X \sim$ Binomial(n,p). The binomial mass function is

$$P(X=x) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x (1-p)^{n-x}$$

where $x = 0, \ldots, n$. Recall that the notation

$$\left(\begin{array}{c} n \\ x \end{array} \right) = \frac{n!}{x!(n-x)!}$$

(read "n choose x") counts the number of ways of selecting x items out of n without replacement disregarding the order of the items. It turns out that n choose 0 is 1 and n choose 1 and n choose $n-1$ are both n.

Example

Suppose a friend has 8 children, 7 of which are girls and none are twins. 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?

$$\left(\begin{array}{c} 8 \\ 7 \end{array} \right) .5^7(1-.5)^1 + \left(\begin{array}{c} 8 \\ 8 \end{array} \right) .5^8(1-.5)^0 \approx 0.04.$$

Simulating means of coin flips

```
> choose(8, 7) * 0.5^8 + choose(8, 8) * 0.5^8
[1] 0.03516
> pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
[1] 0.03516
```

The normal distribution

Watch this video before beginning[1]

The normal distribution is easily the handiest distribution in all of statistics. It can be used in an endless variety of settings. Moreover, as we'll see later on in the course, sample means follow normal distributions for large sample sizes.

Remember the goal of probability modeling. We are assuming a probability distribution for our population as a way of parsimoniously characterizing it. In fact, the normal distribution only requires two numbers to characterize it. Specifically, a random variable is said to follow a **normal** or **Gaussian** distribution with mean μ and variance σ^2 if the associated density is:

$$(2\pi\sigma^2)^{-1/2} e^{-(x-\mu)^2/2\sigma^2}.$$

If X is a RV with this density then $E[X] = \mu$ and $Var(X) = \sigma^2$. That is, the normal distribution is characterized by the mean and variance. We write $X \sim N(\mu, \sigma^2)$ to denote a normal random variable. When $\mu = 0$ and $\sigma = 1$ the resulting distribution is called **the standard normal distribution**. Standard normal RVs are often labeled Z

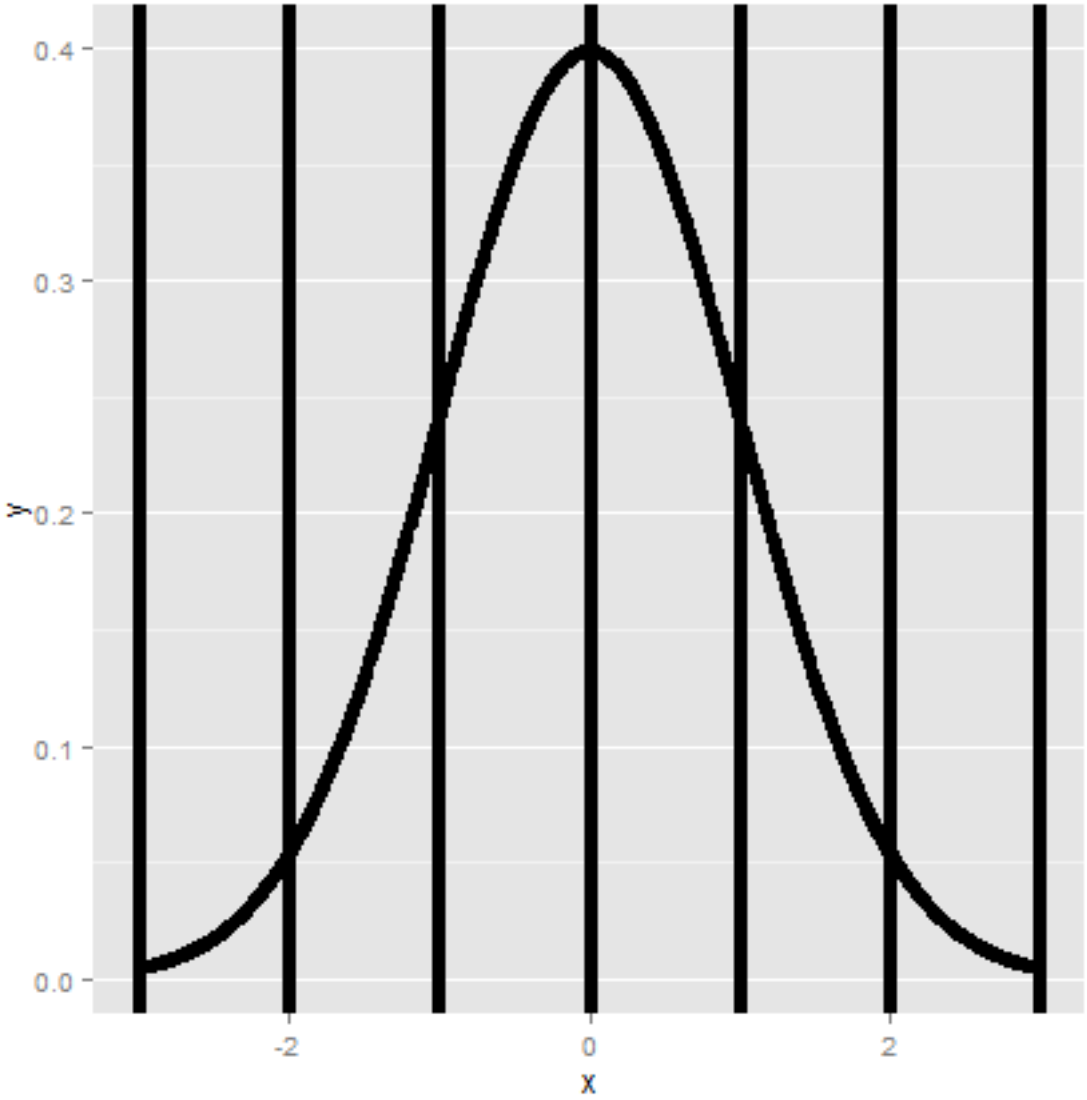
Consider an example, if we say that intelligence quotients are normally distributed with a mean of 100 and a standard deviation of 15. Then, we are saying that if we randomly sample a person from this population, the probability that they have an IQ of say 120 or larger, is governed by a normal distribution with a mean of 100 and a variance of 15^2.

Taken another way, if we know that the population is normally distributed then to estimate everything about the population, we need only estimate the population mean and variance. (Estimated by the sample mean and the sample variance.)

[1] http://youtu.be/dUTWvKa0Leo?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Reference quantiles for the standard normal

The normal distribution is so important that it is useful to memorize reference probabilities and quantiles. The image below shows reference lines at 0, 1, 2 and 3 standard deviations above and below 0. This is for the standard normal; however, all of the rules apply to non standard normals as 0, 1, 2 and 3 standard deviations above and below μ, the population mean.



Standard normal reference lines.

The most relevant probabilities are.

1. Approximately 68\%, 95\% and 99\% of the normal density lies within 1, 2 and 3 standard deviations from the mean, respectively.

2. -1.28, -1.645, -1.96 and -2.33 are the 10^{th}, 5^{th}, 2.5^{th} and 1^{st} percentiles of the standard normal distribution, respectively.
3. By symmetry, 1.28, 1.645, 1.96 and 2.33 are the 90^{th}, 95^{th}, 97.5^{th} and 99^{th} percentiles of the standard normal distribution, respectively.

Shifting and scaling normals

Since the normal distribution is characterized by only the mean and variance, which are a shift and a scale, we can transform normal random variables to be standard normals and vice versa. For example If $X \sim N(\mu, \sigma^2)$ then:

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

If Z is standard normal

$$X = \mu + \sigma Z \sim N(\mu, \sigma^2)$$

then X is $X \sim N(\mu, \sigma^2)$. We can use these facts to answer questions about non-standard normals by relating them back to the standard normal.

Example

What is the 95^{th} percentile of a $N(\mu, \sigma^2)$ distribution? Quick answer in R `qnorm(.95, mean = mu, sd = sigma)`. Alternatively, because we have the standard normal quantiles memorized, and we know that 1.645 is its 95th percentile, the answer has to be $\mu + \sigma 1.645$.

In general, $\mu + \sigma z_0$ where z_0 is the appropriate standard normal quantile.

To put some context on our previous setting, population mean BMI for men is reported as[2] 29 kg/mg^2 with a standard deviation of 4.73. Assuming normality of BMI, what is the population 95^{th} percentile? The answer is then:

$$29 + 4.73 \times 1.645 = 36.78.$$

Or alternatively, we could simply type `r qnorm(.95, 29, 4.73)` in R.

Now let's reverse the process. Imaging asking what's the probability that a randomly drawn subject from this population has a BMI less than 24.27? Notice that

$$\frac{24.27 - 29}{4.73} = -1.$$

[2] http://www.ncbi.nlm.nih.gov/pubmed/23675464

Therefore, 24.27 is 1 standard deviation below the mean. We know that 16% lies below or above 1 standard deviation from the mean. Thus 16% lies below. Alternatively, `pnorm(24.27, 29, 4.73)` yields the result.

Example

Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. What's the probability of getting more than 1,160 clicks in a day? Notice that:

$$\frac{1160 - 1020}{50} = 2.8$$

Therefore, 1,160 is 2.8 standard deviations above the mean. We know from our standard normal quantiles that the probability of being larger than 2 standard deviation is 2.5% and 3 standard deviations is far in the tail. Therefore, we know that the probability has to be smaller than 2.5% and should be very small. We can obtain it exactly as `r pnorm(1160, 1020, 50, lower.tail = FALSE)` which is 0.3%. Note that we can also obtain the probability as `r pnorm(2.8, lower.tail = FALSE)`.

Example

Consider the previous example again. What number of daily ad clicks would represent the one where 75% of days have fewer clicks (assuming days are independent and identically distributed)? We can obtain this as:

Finding a normal quantile

```
> qnorm(0.75, mean = 1020, sd = 50)
[1] 1054
```

The Poisson distribution

Watch this video before beginning.[3]

The Poisson distribution is used to model counts. It is perhaps only second to the normal distribution usefulness. In fact, the Bernoulli, binomial and multinomial distributions can all be modeled by clever uses of the Poisson.

The Poisson distribution is especially useful for modeling unbounded counts or counts per unit of time (rates). Like the number of clicks on advertisements, or the number of people who show up at a

[3] http://youtu.be/ZPLZg7qz4xE?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

bus stop. (While these are in principle bounded, it would be hard to actually put an upper limit on it.) There is also a deep connection between the Poisson distribution and popular models for so-called event-time data. In addition, the Poisson distribution is the default model for so-called contingency table data, which is simply tabulations of discrete characteristics. Finally, when n is large and p is small, the Poisson is an accurate approximation to the binomial distribution.

The Poisson mass function is:

$$P(X = x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}$$

for $x = 0, 1, \ldots$. The mean of this distribution is λ. The variance of this distribution is also λ. Notice that x ranges from 0 to ∞. Therefore, the Poisson distribution is especially useful for modeling unbounded counts.

Rates and Poisson random variables

The Poisson distribution is useful for rates, counts that occur over units of time. Specifically, if $X \sim Poisson(\lambda t)$ where $\lambda = E[X/t]$ is the expected count per unit of time and t is the total monitoring time.

Example

The number of people that show up at a bus stop is Poisson with a 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?

Finding a normal quantile

```
> ppois(3, lambda = 2.5 * 4)
[1] 0.01034
```

Therefore, there is about a 1% chance that 3 or fewer people show up. Notice the multiplication by four in the function argument. Since lambda is specified as events per *hour* we have to multiply by four to consider the number of events that occur in 4 hours.

Poisson approximation to the binomial

When n is large and p is small the Poisson distribution is an accurate approximation to the binomial distribution. Formally, if $X \sim$ Binomial(n, p) then X is approximately Poisson where $\lambda = np$ provided that n is large p is small.

Example, Poisson approximation to the binomial

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

Finding a normal quantile

```
> pbinom(2, size = 500, prob = 0.01)
[1] 0.1234
> ppois(2, lambda = 500 * 0.01)
[1] 0.1247
```

So we can see that the probabilities agree quite well. This approximation is often done as the Poisson model is a more convenient model in many respects.

Exercises

1. Your friend claims that changing the font to comic sans will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic sans. If it was really a coin flip for these 10 sites, what's the probability of getting 9 or 10 out of 10 with more revenue for the new font?
2. A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, whats the probability of fewer than 5 documentation errors?
3. The number of search entries entered at a web site is Poisson at a rate of 9 searches per minute. The site is monitored for 5 minutes. What is the probability of 40 or fewer searches in that time frame?
4. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What's the probability that a given day has fewer than 93 hits per day expressed as a percentage to the nearest percentage point? Watch a video solution[4] and see the problem[5].
5. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. What number of web hits per day represents the number so that only 5% of days have more hits? Watch a video solution[6] and see the problem and solution[7].
6. Suppose that the number of web hits to a particular site are approximately normally distributed with a mean of 100 hits per day and a standard deviation of 10 hits per day. Imagine taking a random sample of 50 days. What number of web hits would be the point so that only 5% of averages of 50 days of web traffic have more hits? Watch a video solution[8] and see the problem and solution[9].

[4] https://www.youtube.com/watch?v=E-ancc7iTho&index=10&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[5] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#4
[6] https://www.youtube.com/watch?v=rv48_5C8gx4&index=12&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[7] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#6
[8] https://www.youtube.com/watch?v=c_B2AuOhdzg&index=13&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[9] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#7

7. You don't believe that your friend can discern good wine from cheap. Assuming that you're right, in a blind test where you randomize 6 paired varieties (Merlot, Chianti, ...) of cheap and expensive wines. What is the change that she gets 5 or 6 right? Watch a video solution[10] and see the original problem[11].
8. The number of web hits to a site is Poisson with mean 16.5 per day. What is the probability of getting 20 or fewer in 2 days? Watch a video solution[12] and see a written solution[13].

[10] https://www.youtube.com/watch?v=ILm2OUl6p_w&index=14&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[11] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#8
[12] https://www.youtube.com/watch?v=PMPFbwtpp1k&index=18&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[13] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#12

7. Asymptopia

Asymptotics

Watch this video before beginning.[1]

Asymptotics is the term for the behavior of statistics as the sample size limits to infinity. Asymptotics are incredibly useful for simple statistical inference and approximations. Asymptotics often make hard problems easy and difficult calculations simple. We will not cover the philosophical considerations in this book, but is true nonetheless, that asymptotics often lead to nice understanding of procedures. In fact, the ideas of asymptotics are so important form the basis for frequency interpretation of probabilities by considering the long run proportion of times an event occurs.

Some things to bear in mind about the seemingly magical nature of asymptotics. There's no free lunch and unfortunately, asymptotics generally give no assurances about finite sample performance.

Limits of random variables

We'll only talk about the limiting behavior of one statistic, the sample mean. Fortunately, for the sample mean there's a set of powerful results. These results allow us to talk about the large sample distribution of sample means of a collection of iid observations.

The first of these results we intuitively already know. It says that the average limits to what its estimating, the population mean. This result is called the Law of Large Numbers. It simply says that if you go to the trouble of collecting an infinite amount of data, you estimate the population mean perfectly. Note there's sampling assumptions that have to hold for this result to be true. The data have to be iid.

A great example of this comes from coin flipping. Imagine if $\bar{X}_n$ is the average of the result of n coin flips (i.e. the sample proportion of heads). The Law of Large Numbers states that as we flip a coin over and over, it eventually converges to the true probability of a head.

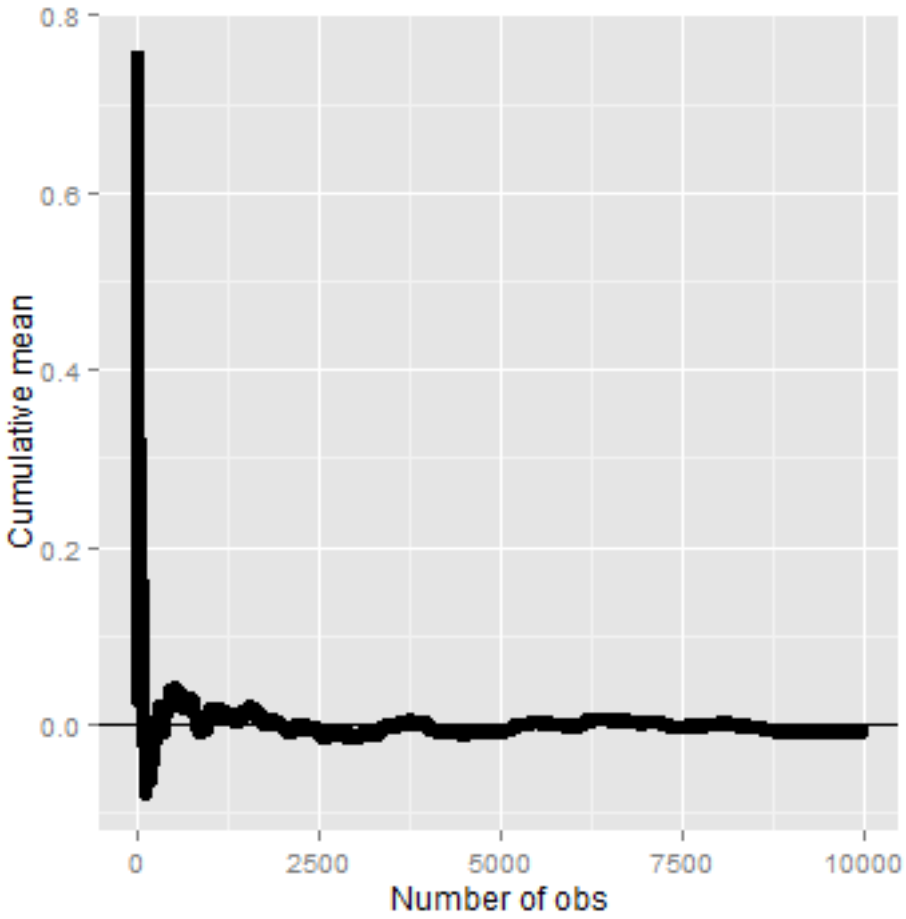
Law of large numbers in action

Let's try using simulation to investigate the law of large numbers in action. Let's simulate a lot of standard normals and plot the cumulative means. If the LLN is correct, the line should converge to 0, the mean of the standard normal distribution.

[1] http://youtu.be/WRUgUEBIYZY?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Finding a normal quantile

```
n <- 10000
means <- cumsum(rnorm(n))/(1:n)
library(ggplot2)
g <- ggplot(data.frame(x = 1:n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0) + geom_line(size = 2)
g <- g + labs(x = "Number of obs", y = "Cumulative mean")
g
```



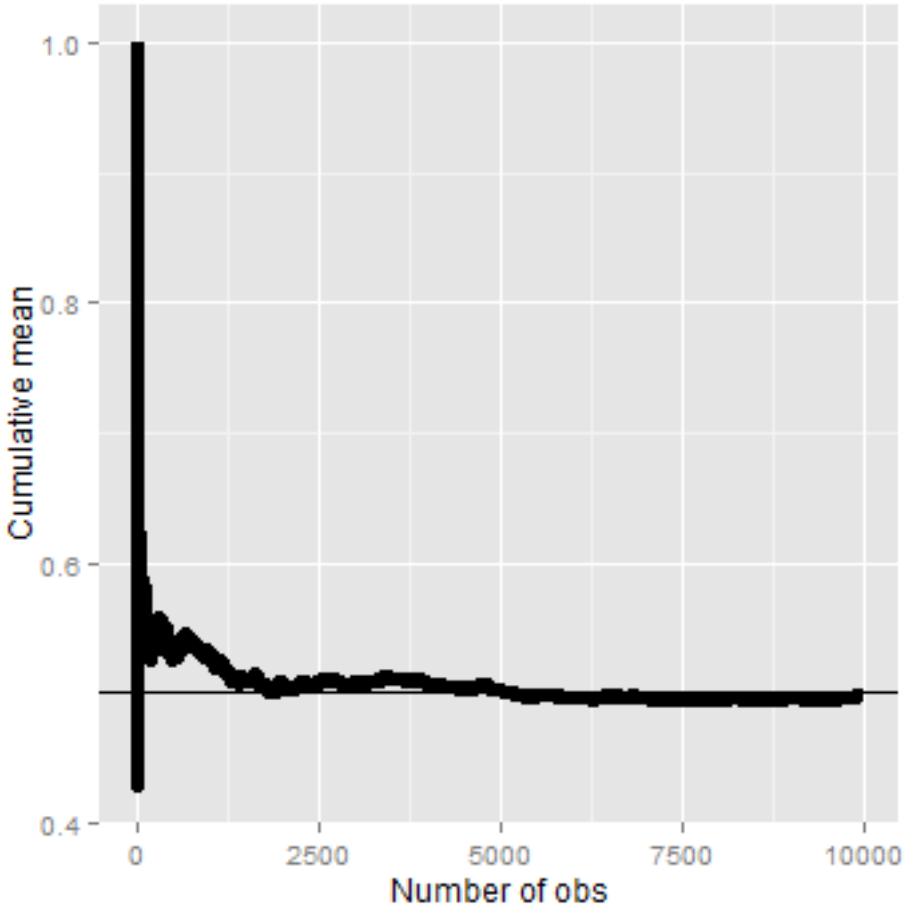
Cumulative average from standard normal simulations.

Law of large numbers in action, coin flip

Let's try the same thing, but for a fair coin flip. We'll simulate a lot of coin flips and plot the cumulative proportion of heads.

Finding a normal quantile

```
means <- cumsum(sample(0:1, n, replace = TRUE))/(1:n)
g <- ggplot(data.frame(x = 1:n, y = means), aes(x = x, y = y))
g <- g + geom_hline(yintercept = 0.5) + geom_line(size = 2)
g <- g + labs(x = "Number of obs", y = "Cumulative mean")
g
```



Cumulative proportion of heads from a sequence of coin flips.

Discussion

An estimator is called **consistent** if it converges to what you want to estimate. Thus, the LLN says that the sample mean of iid sample is consistent for the population mean. Typically, good estimators are consistent; it's not too much to ask that if we go to the trouble of collecting an infinite amount of data that we get the right answer. The sample variance and the sample standard deviation of iid random variables are consistent as well.

The Central Limit Theorem

Watch this video before beginning.[2]

The **Central Limit Theorem** (CLT) is one of the most important theorems in statistics. For our purposes, the CLT states that the distribution of averages of iid variables becomes that of a standard normal as the sample size increases. Consider this fact for a second. We already know the mean and standard deviation of the distribution of averages from iid samples. The CLT gives us an approximation to the full distribution! Thus, for iid samples, we have a good sense of distribution of the average event though: (1) we only observed one average and (2) we don't know what the population distribution is. Because of this, the CLT applies in an endless variety of settings and is one of the most important theorems ever discovered.

The formal result is that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

has a distribution like that of a standard normal for large n. Replacing the standard error by its estimated value doesn't change the CLT.

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

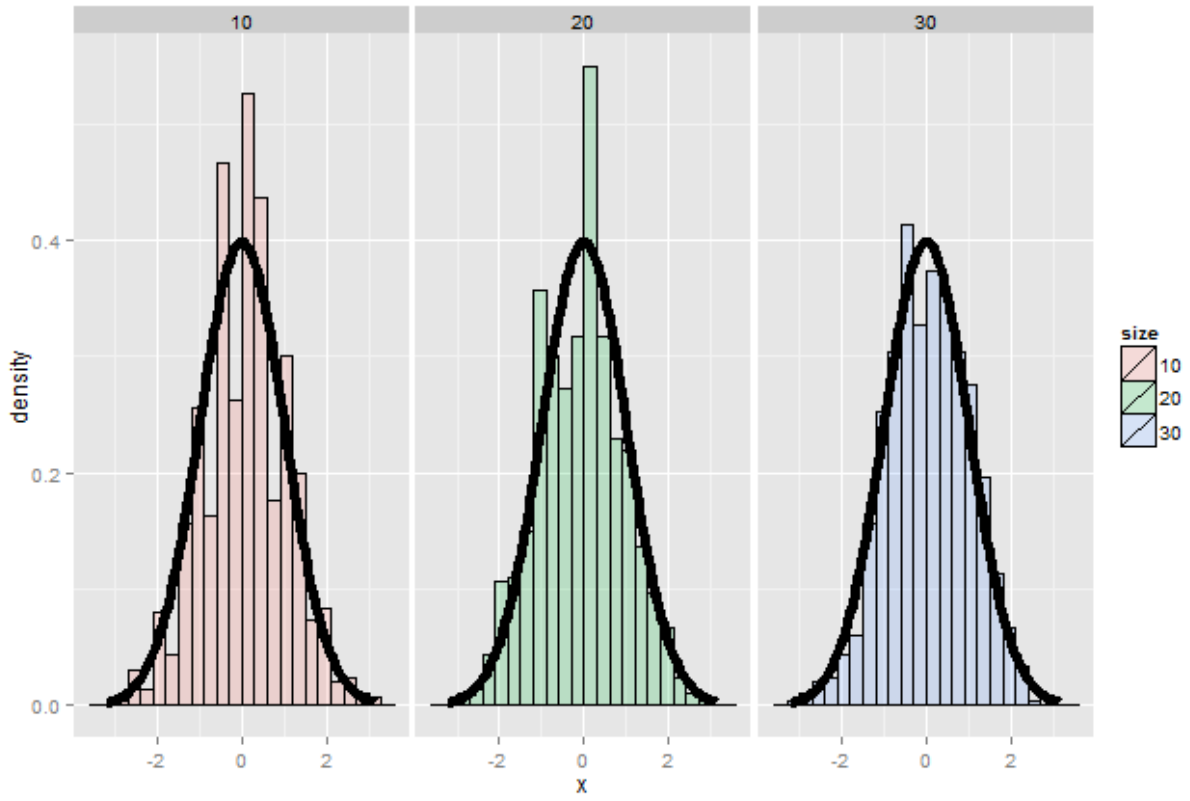
CLT simulation experiments

Let's try simulating lots of averages from various distributions and showing that the resulting distribution looks like a bell curve.

Die rolling

- Simulate a standard normal random variable by rolling n (six sided) dice.
- Let X_i be the outcome for die i.
- Then note that $\mu = E[X_i] = 3.5$.
- Recall also that $Var(X_i) = 2.92$.
- SE $\sqrt{2.92/n} = 1.71/\sqrt{n}$.
- Lets roll n dice, take their mean, subtract off 3.5, and divide by $1.71/\sqrt{n}$ and repeat this over and over.

[2] http://youtu.be/FAIyVHmniK0?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ



Result of coin CLT simulation.

It's pretty remarkable that the approximation works so well with so few rolls of the die. So, if you're stranded on an island, and need to simulate a standard normal without a computer, but you do have a die, you can get a pretty good approximation with 10 rolls even.

Coin CLT

In fact the oldest application of the CLT is to the idea of flipping coins (by de Moivre)[3]. Let X_i be the 0 or 1 result of the i^{th} flip of a possibly unfair coin. The sample proportion, say $\hat{p}$, is the average of the coin flips. We know that:

- $E[X_i] = p$,
- $Var(X_i) = p(1-p)$,
- $\sqrt{Var(\hat{p})} = \sqrt{p(1-p)/n}$.

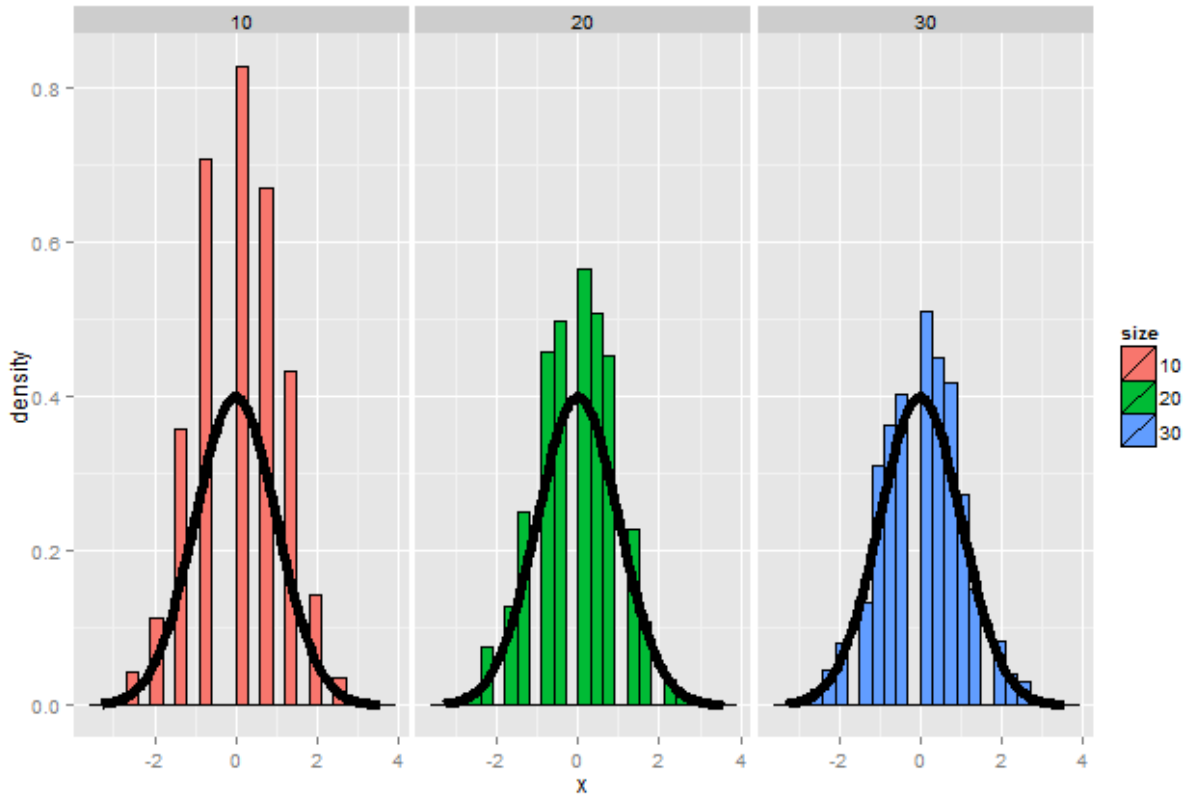
Furthermore, because of the CLT, we also know that:

[3] http://en.wikipedia.org/wiki/De_Moivre%E2%80%93Laplace_theorem

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

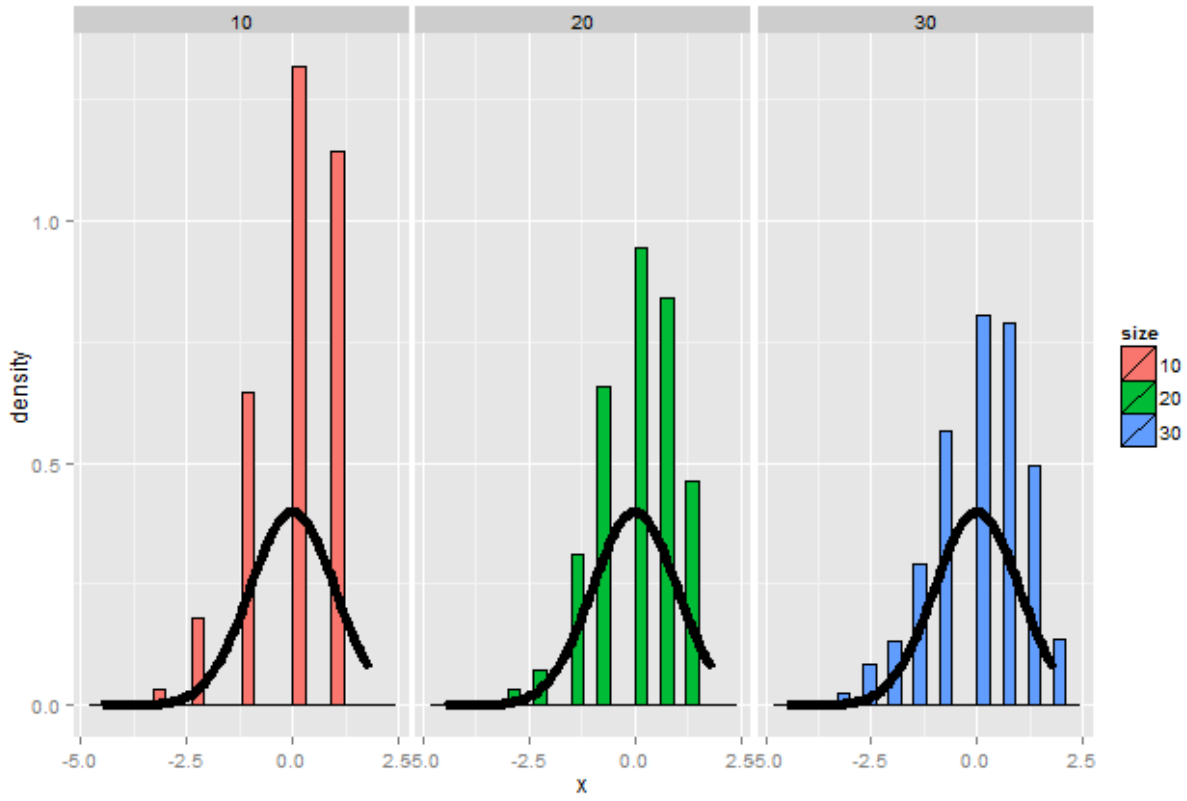
will be approximately normally distributed.

Let's test this by flipping a coin n times, taking the sample proportion of heads, subtract off 0.5 and multiply the result by $2\sqrt{n}$ (divide by $1/(2\sqrt{n})$).



Results of the coin CLT simulation.

This convergence doesn't look quite as good as the die, since the coin has fewer possible outcomes. In fact, among coins of various degrees of bias, the convergence to normality is governed by how far from 0.5 p is. Let's redo the simulation, now using $p = 0.9$ instead of $p = 0.5$ like we did before.



Results of the simulation when p=0.9

Notice that the convergence to normality is quite poor. Thus, be careful when using CLT approximations for sample proportions when your proportion is very close to 0 or 1.

Confidence intervals

Watch this video before beginning.[4]

Confidence intervals are methods for quantifying uncertainty in our estimates. The fact that the interval has width characterizes that there is randomness that prevents us from getting a perfect estimate. Let's go through how a confidence interval using the CLT is constructed.

According to the CLT, the sample mean, $\bar{X}$, is approximately normal with mean μ and standard deviation $\sigma/\sqrt{n}$. Furthermore,

$$\mu + 2\sigma/\sqrt{n}$$

is pretty far out in the tail (only 2.5% of a normal being larger than 2 sds in the tail). Similarly,

[4] http://youtu.be/u85aQ0mtiZ8?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

$$\mu - 2\sigma/\sqrt{n}$$

is pretty far in the left tail (only 2.5% chance of a normal being smaller than 2 standard deviations in the tail). So the probability $\bar{X}$ is bigger than $\mu + 2\sigma/\sqrt{n}$ or smaller than $\mu - 2\sigma/\sqrt{n}$ is 5%. Or equivalently, the probability that these limits contain μ is 95%. The quantity:

$$\bar{X} \pm 2\sigma/\sqrt{n}$$

is called a 95% interval for μ. The 95% refers to the fact that if one were to repeatedly get samples of size n, about 95% of the intervals obtained would contain μ. The 97.5th quantile is 1.96 (so I rounded to 2 above). If instead of a 95% interval, you wanted a 90% interval, then you want (100 - 90) / 2 = 5% in each tail. Thus your replace the 2 with the 95th percentile, which is 1.645.

Example CI

Give a confidence interval for the average height of sons in Galton's data.

Finding a confidence interval.

```
> library(UsingR)
> data(father.son)
> x <- father.son$sheight
> (mean(x) + c(-1, 1) * qnorm(0.975) * sd(x)/sqrt(length(x)))/12
[1] 5.710 5.738
```

Here we divided by 12 to get our interval in feet instead of inches. So we estimate the average height of the sons as 5.71 to 5.74 with 95% confidence.

Example using sample proportions

In the event that each X_i is 0 or 1 with common success probability p then $\sigma^2 = p(1-p)$. The interval takes the form:

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}}.$$

Replacing p by $\hat{p}$ in the standard error results in what is called a Wald confidence interval for p. Remember also that $p(1-p)$ is maximized at 1/4. Plugging this in and setting our Z quantile as 2 (which is about a 95% interval) we find that a quick and dirty confidence interval is:

$$\hat{p} \pm \frac{1}{\sqrt{n}}.$$

This is useful for doing quick confidence intervals for binomial proportions in your head.

Example

Your campaign advisor told you that in a random sample of 100 likely voters, 56 intent to vote for you. Can you relax? Do you have this race in the bag? Without access to a computer or calculator, how precise is this estimate?

```
> 1/sqrt(100)
[1] 0.1
```

so a back of the envelope calculation gives an approximate 95% interval of `(0.46, 0.66)`.

Thus, since the interval contains 0.5 and numbers below it, there's not enough votes for you to relax; better go do more campaigning!

The basic rule of thumb is then, $1/\sqrt{n}$ gives you a good estimate for the margin of error of a proportion. Thus, $n = 100$
for about 1 decimal place, 10,000 for 2, 1,000,000 for 3.

```
> round(1/sqrt(10^(1:6)), 3)
[1] 0.316 0.100 0.032 0.010 0.003 0.001
```

We could very easily do the full Wald interval, which is less conservative (may provide a narrower interval). Remember the Wald interval for a binomial proportion is:

$$\hat{p} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Here's the R code for our election setting, both coding it directly and using `binom.test`.

```
> 0.56 + c(-1, 1) * qnorm(0.975) * sqrt(0.56 * 0.44/100)
[1] 0.4627 0.6573
> binom.test(56, 100)$conf.int
[1] 0.4572 0.6592
```

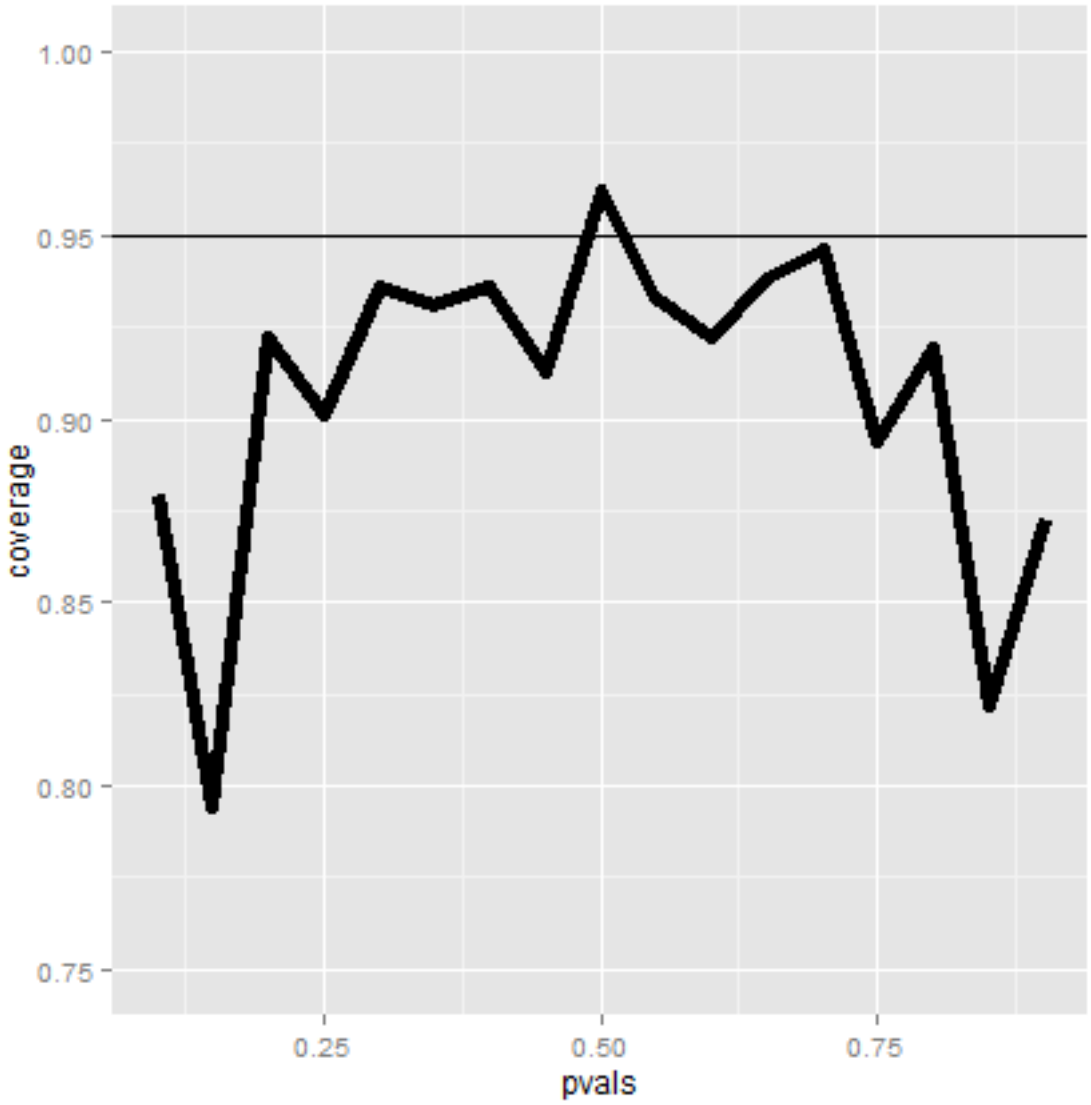
Simulation of confidence intervals

It is interesting to note that the coverage of confidence intervals describes an aggregate behavior. In other words the confidence interval describes the percentage of intervals that would cover the parameter being estimated if we were to repeat the experiment over and over. So, one can not technically say that the interval contains the parameter with probability 95%, say. So called Bayesian credible intervals address this issue at the expense (or benefit depending on who you ask) of adopting a Bayesian framework.

For our purposes, we're using confidence intervals and so will investigate their frequency performance over repeated realizations of the experiment. We can do this via simulation. Let's consider different values of p and look at the Wald interval's coverage when we repeatedly create confidence intervals.

Code for investigating Wald interval coverage

```
n <- 20
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage <- sapply(pvals, function(p) {
    phats <- rbinom(nosim, prob = p, size = n)/n
    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)
})
```



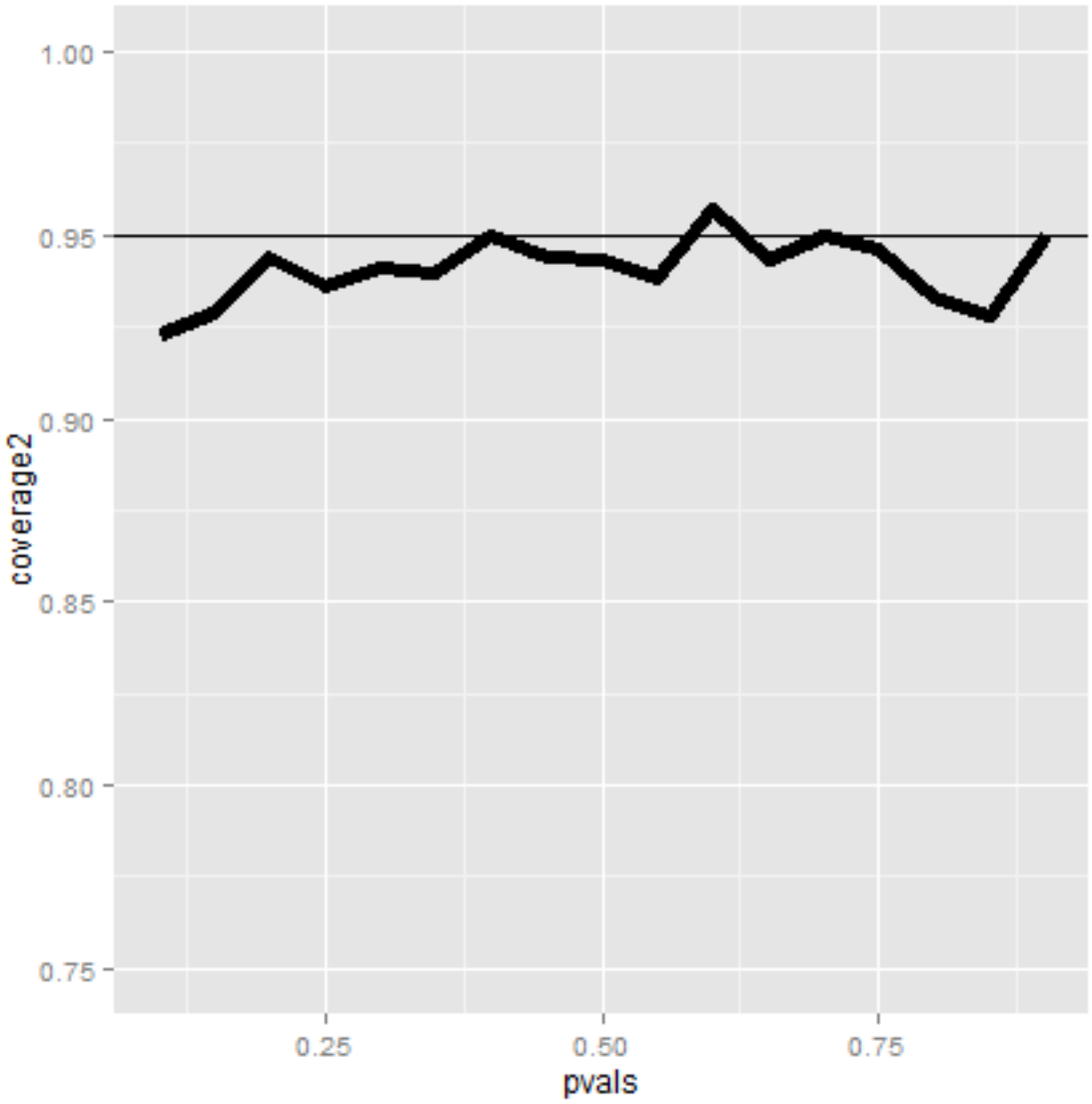
Plot of Wald interval coverage.

The figure shows that if we were to repeatedly try experiments for any fixed value of p, it's rarely the case that our intervals will cover the value that they're trying to estimate in 95% of them. This is bad, since covering the parameter that its estimating 95% of the time is the confidence interval's only job!

So what's happening? Recall that the CLT is an approximation. In this case n isn't large enough for the CLT to be applicable for many of the values of p. Let's see if the coverage improves for larger n.

Code for investigating Wald interval coverage

```
n <- 100
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage2 <- sapply(pvals, function(p) {
    phats <- rbinom(nosim, prob = p, size = n)/n
    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)
})
```



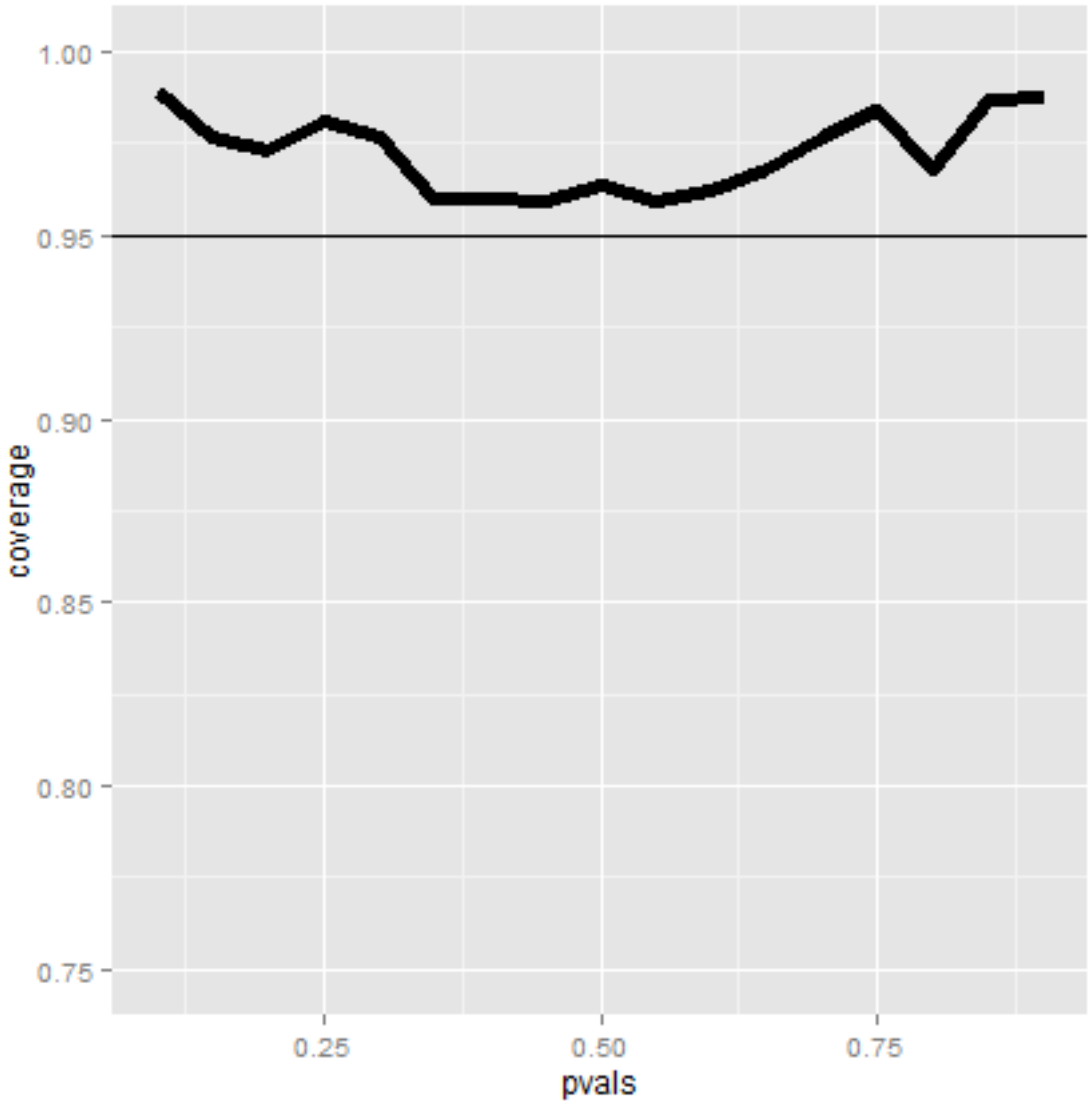
Output of simulation with $n = 100$.

Now it looks much better. Of course, increasing our sample size is rarely an option. There's exact fixes to make this interval work better for small sample sizes.

However, for a quick fix is to take your data and add two successes and two failures. So, for example, in our election example, we would form our interval with 58 votes out of 104 sampled (disregarding that the actual numbers were 56 and 100). This interval is called the Agresti/Coull interval. This interval has much better coverage. Let's show it via a simulation.

Code for investigating Agresti/Coull interval coverage when n=20.

```
n <- 20
pvals <- seq(0.1, 0.9, by = 0.05)
nosim <- 1000
coverage <- sapply(pvals, function(p) {
    phats <- (rbinom(nosim, prob = p, size = n) + 2)/(n + 4)
    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)
})
```



Coverage of the Agresti/Coull interval with $n = 20$

The coverage is better, if maybe a little conservative in the sense of being over the 95% line most of the time. If the interval is too conservative, it's likely a little too wide. To see this clearly, imagine if we made our interval $-\infty$ to ∞. Then we would always have 100% coverage in any setting, but the interval wouldn't be useful. Nonetheless, the Agrestic/Coull interval gives a much better trade off between coverage and width than the Wald interval.

In general, one should use the add two successes and failures method for binomial confidence intervals with smaller n. For very small n consider using an exact interval (not covered in this class).

Poisson interval

Since the Poisson distribution is so central for data science, let's do a Poisson confidence interval. Remember that if $X \sim \text{Poisson}(\lambda t)$ then our estimate of λ is $\hat{\lambda} = X/t$. Furthermore, we know that $Var(\hat{\lambda}) = \lambda/t$ and so the natural estimate is $\hat{\lambda}/t$. While it's not immediate how the CLT applies in this case, the interval is of the familiar form

$$\text{Estimate} \pm Z_{1-\alpha/2}\text{SE}.$$

So our Poisson interval is:

$$\hat{\lambda} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$$

Example

A nuclear pump failed 5 times out of 94.32 days. Give a 95% confidence interval for the failure rate per day.

Code for asymptotic Poisson confidence interval

```
> x <- 5
> t <- 94.32
> lambda <- x/t
> round(lambda + c(-1, 1) * qnorm(0.975) * sqrt(lambda/t), 3)
[1] 0.007 0.099
```

A non-asymptotic test, one that guarantees coverage, is also available. But, it has to be evaluated numerically.

Code for exact Poisson confidence interval

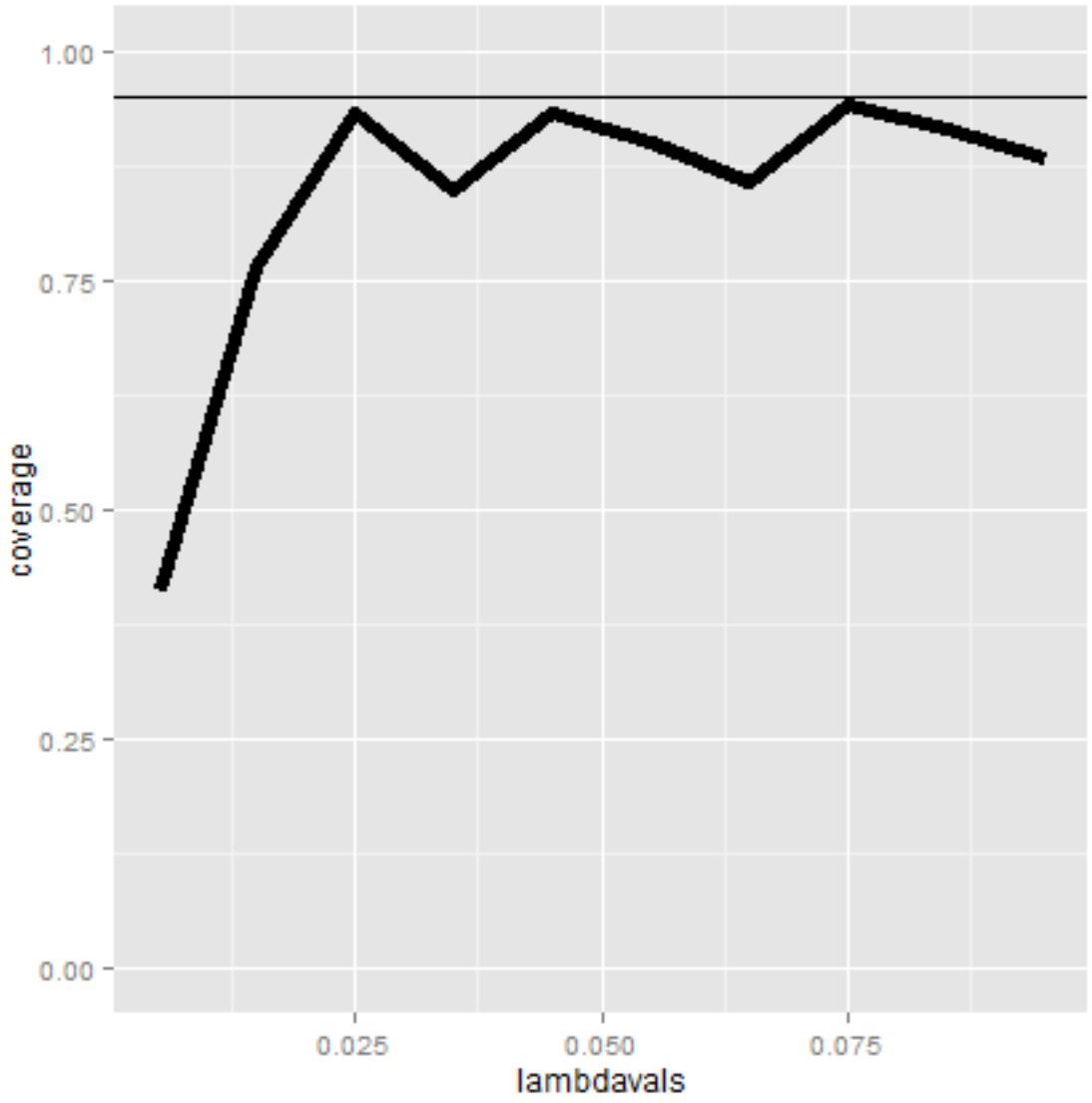
```
> poisson.test(x, T = 94.32)$conf
[1] 0.01721 0.12371
```

Simulating the Poisson coverage rate

Let's see how the asymptotic interval performs for lambda values near what we're estimating.

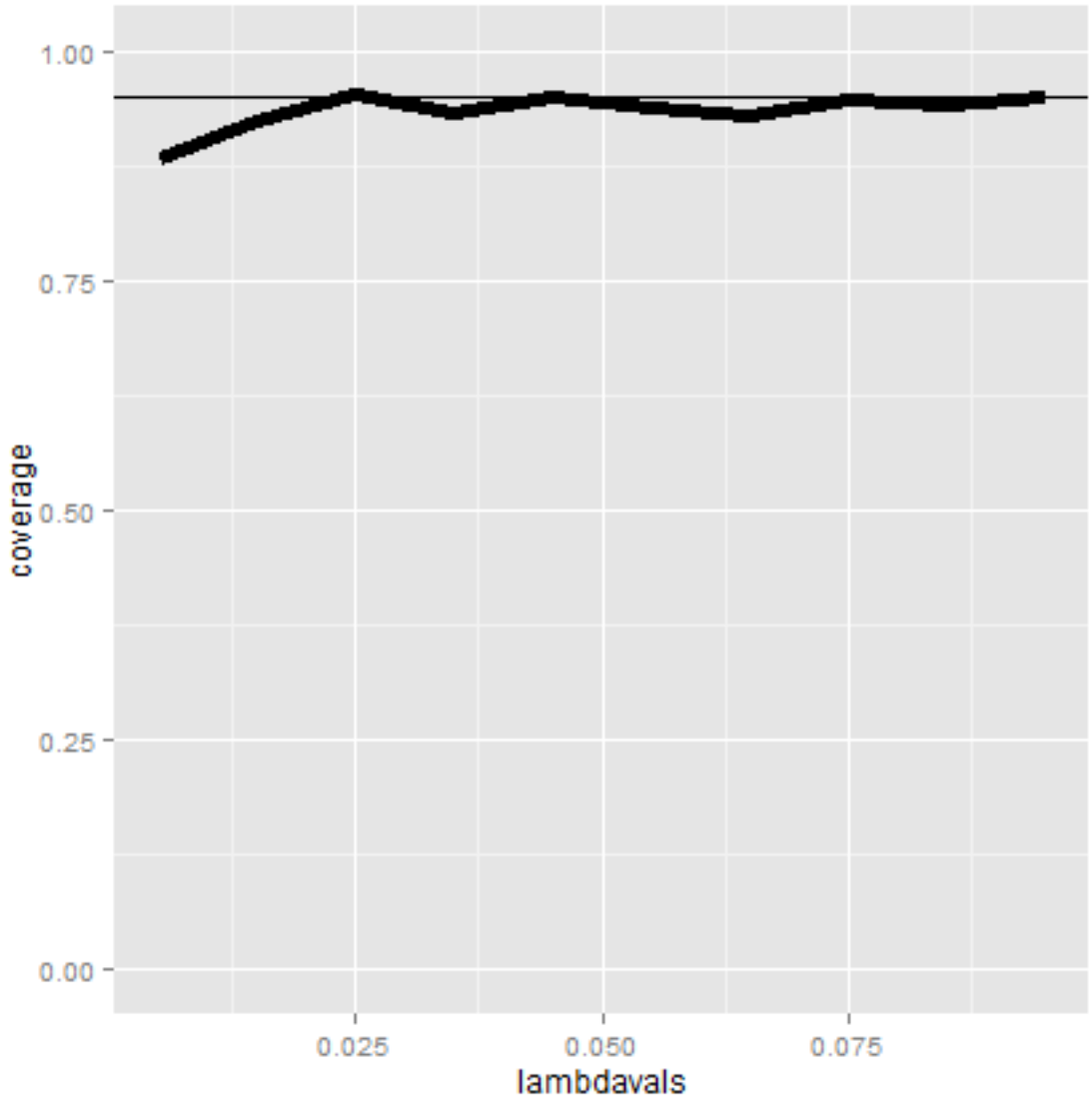
Code for evaluating the coverage of the asymptotic Poisson confidence interval

```
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)
})
```



Coverage of Poisson intervals for various values of lambda

The coverage can be low for low values of lambda. In this case the asymptotics works as we increase the monitoring time, t. Here's the coverage if we increase t to 1,000.



Coverage of Poisson intervals for various values of lambda and t=1000

Summary notes

- The LLN states that averages of iid samples. converge to the population means that they are estimating.
- The CLT states that averages are approximately normal, with distributions.
  - centered at the population mean.
  - with standard deviation equal to the standard error of the mean.
  - CLT gives no guarantee that n is large enough.
- Taking the mean and adding and subtracting the relevant. normal quantile times the SE yields a confidence interval for the mean.

  - Adding and subtracting 2 SEs works for 95% intervals.
- Confidence intervals get wider as the coverage increases.
- Confidence intervals get narrower with less variability or larger sample sizes.
- The Poisson and binomial case have exact intervals that don't require the CLT.
  - But a quick fix for small sample size binomial calculations is to add 2 successes and failures.

Exercises

1. I simulate 1,000,000 standard normals. The LLN says that their sample average must be close to?
2. About what is the probability of getting 45 or fewer heads out 100 flips of a fair coin? (Use the CLT, not the exact binomial calculation).
3. Consider the father.son data. Using the CLT and assuming that the fathers are a random sample from a population of interest, what is a 95% confidence mean height in inches?
4. The goal of a a confidence interval having coverage 95% is to imply that:
   - If one were to repeated collect samples and reconstruct the intervals, around 95% percent of them would contain the true mean being estimated.
   - The probability that the sample mean is in the interval is 95%.
5. The rate of search entries into a web site was 10 per minute when monitoring for an hour. Use R to calculate the exact Poisson interval for the rate of events per minute?
6. Consider a uniform distribution. If we were to sample 100 draws from a a uniform distribution (which has mean 0.5, and variance 1/12) and take their mean, $\bar{X}$. What is the approximate probability of getting as large as 0.51 or larger? Watch this video solution[5] and see the problem and solution here.[6].

[5] https://www.youtube.com/watch?v=JsiLK0g3IZ4&index=15&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L

[6] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw2.html#9

8. *t* Confidence intervals

Small sample confidence intervals

Watch this video before beginning.[1]

In the previous lecture, we discussed creating a confidence interval using the CLT. Our intervals took the form:

$$Est \pm Z \times SE_{Est}.$$

In this lecture, we discuss some methods for small samples, notably Gosset's *t* distribution and *t* confidence intervals.

These intervals are of the form:

$$Est \pm t \times SE_{Est}.$$

So the only change is that we've replaced the Z quantile now with a *t* quantile. These are some of the handiest of intervals in all of statistics. If you want a rule between whether to use a *t* interval or normal interval, just always use the *t* interval.

Gosset's *t* distribution

The *t* distribution was invented by William Gosset (under the pseudonym "Student") in 1908. Fisher provided further mathematical details about the distribution later. This distribution has thicker tails than the normal. It's indexed by a degrees of freedom and it gets more like a standard normal as the degrees of freedom get larger. It assumes that the underlying data are iid Gaussian with the result that

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows Gosset's *t* distribution with $n - 1$ degrees of freedom. (If we replaced s by σ the statistic would be exactly standard normal.) The interval is

$$\bar{X} \pm t_{n-1} S/\sqrt{n},$$

where t_{n-1} is the relevant quantile from the *t* distribution.

[1] http://youtu.be/pHXrDMjzyYg?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Code for manipulate

You can use rStudio's manipulate function to to compare the *t* and Z distributions.

Code for investigating *t* and Z densities.

```
k <- 1000
xvals <- seq(-5, 5, length = k)
myplot <- function(df){
  d <- data.frame(y = c(dnorm(xvals), dt(xvals, df)),
                  x = xvals,
                  dist = factor(rep(c("Normal", "T"), c(k,k))))
  g <- ggplot(d, aes(x = x, y = y))
  g <- g + geom_line(size = 2, aes(color = dist))
  g
}
manipulate(myplot(mu), mu = slider(1, 20, step = 1))
```

The difference is perhaps easier to see in the tails. Therefore, the following code plots the upper quantiles of the Z distribution by those of the *t* distribution.

Code for investigating the upper quantiles of the *t* and Z densities.

```
pvals <- seq(.5, .99, by = .01)
myplot2 <- function(df){
  d <- data.frame(n= qnorm(pvals),t=qt(pvals, df),
                  p = pvals)
  g <- ggplot(d, aes(x= n, y = t))
  g <- g + geom_abline(size = 2, col = "lightblue")
  g <- g + geom_line(size = 2, col = "black")
  g <- g + geom_vline(xintercept = qnorm(0.975))
  g <- g + geom_hline(yintercept = qt(0.975, df))
  g
}
manipulate(myplot2(df), df = slider(1, 20, step = 1))
```

Summary notes

In this section, we give an overview of important facts about the *t* distribution.

- The *t* interval technically assumes that the data are iid normal, though it is robust to this assumption.

- It works well whenever the distribution of the data is roughly symmetric and mound shaped.
- Paired observations are often analyzed using the *t* interval by taking differences.
- For large degrees of freedom, *t* quantiles become the same as standard normal quantiles; therefore this interval converges to the same interval as the CLT yielded.
- For skewed distributions, the spirit of the *t* interval assumptions are violated.
  - Also, for skewed distributions, it doesn't make a lot of sense to center the interval at the mean.
  - In this case, consider taking logs or using a different summary like the median.
- For highly discrete data, like binary, other intervals are available.

Example of the *t* interval, Gosset's sleep data

Watch this video before beginning.[2]

In R typing `r data(sleep)` brings up the sleep data originally analyzed in Gosset's Biometrika paper, which shows the increase in hours for 10 patients on two soporific drugs. R treats the data as two groups rather than paired.

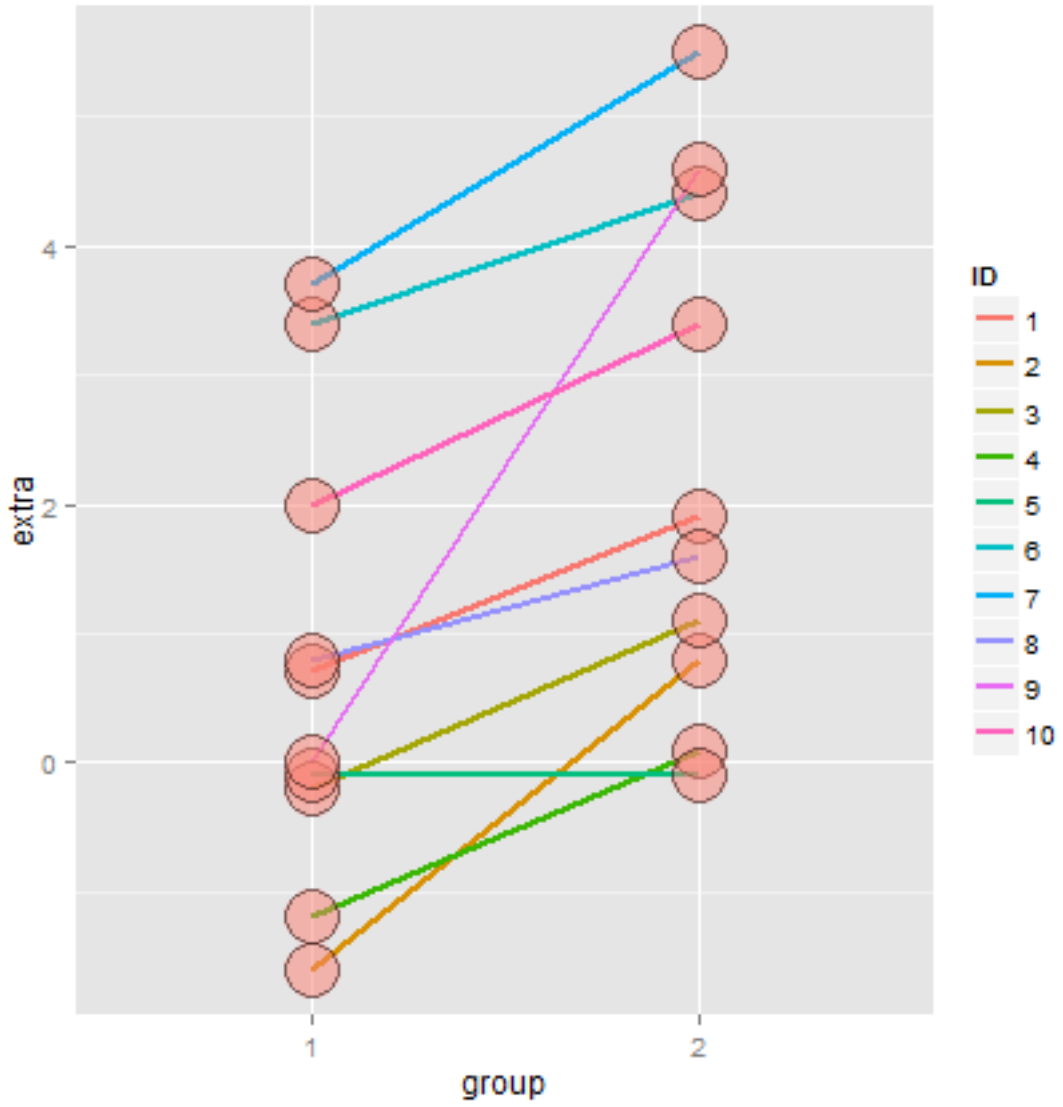
The data

Loading Galton's data.

```
> data(sleep)
> head(sleep)
  extra group ID
1   0.7     1  1
2  -1.6     1  2
3  -0.2     1  3
4  -1.2     1  4
5  -0.1     1  5
6   3.4     1  6
```

Here's a plot of the data. In this plot paired observations are connected with a line.

[2] http://youtu.be/2L41xqPvPso?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ



A plot of the pairs of observations from Galton's sleep data.

Now let's calculate the *t* interval for the differences from baseline to follow up. Below we give four different ways for calculating the interval.

Loading Galton's data.

```
g1 <- sleep$extra[1 : 10]; g2 <- sleep$extra[11 : 20]
difference <- g2 - g1
mn <- mean(difference); s <- sd(difference); n <- 10
## Calculating directly
mn + c(-1, 1) * qt(.975, n-1) * s / sqrt(n)
## using R's built in function
t.test(difference)
## using R's built in function, another format
t.test(g2, g1, paired = TRUE)
## using R's built in function, another format
t.test(extra ~ I(relevel(group, 2)), paired = TRUE, data = sleep)
## Below are the results (after a little formatting)
       [,1] [,2]
 [1,] 0.7001 2.46
 [2,] 0.7001 2.46
 [3,] 0.7001 2.46
 [4,] 0.7001 2.46
```

Therefore, since our interval doesn't include 0, our 95% confidence interval estimate for the mean change (follow up - baseline) is 0.70 to 2.45.

Independent group *t* confidence intervals

Watch this video before beginning.[3]

Suppose that we want to compare the mean blood pressure between two groups in a randomized trial; those who received the treatment to those who received a placebo. The randomization is useful for attempting to balance unobserved covariates that might contaminate our results. Because of the randomization, it would be reasonable to compare the two groups without considering further variables.

We cannot use the paired *t* interval that we just used for Galton's data, because the groups are independent. Person 1 from the treated group has no relationship with person 1 from the control group. Moreover, the groups may have different sample sizes, so taking paired differences may not even be possible even if it isn't advisable in this setting.

We now present methods for creating confidence intervals for comparing independent groups.

[3] http://youtu.be/J1XqN0yumEQ?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Confidence interval

A $(1-\alpha)\times 100\%$ confidence interval for the mean difference between the groups, $\mu_y - \mu_x$ is:

$$\bar{Y} - \bar{X} \pm t_{n_x+n_y-2,1-\alpha/2} S_p \left(\frac{1}{n_x} + \frac{1}{n_y}\right)^{1/2}.$$

The notation $t_{n_x+n_y-2,1-\alpha/2}$ means a t quantile with $n_x + n_y - 2$ degrees of freedom. The pooled variance estimator is:

$$S_p^2 = \{(n_x - 1)S_x^2 + (n_y - 1)S_y^2\}/(n_x + n_y - 2).$$

This variance estimate is used if one is willing to assume a constant variance across the groups. It is a weighted average of the group-specific variances, with greater weight given to whichever group has the larger sample size.

If there is some doubt about the constant variance assumption, assume a different variance per group, which we will discuss later.

Mistakenly treating the sleep data as grouped

Let's first go through an example where we treat paired data as if it were independent. Consider Galton's sleep data from before. In the code below, we do the R code for grouped data directly, and using the `r t.test` function.

Galton's data treated as grouped and independent.

```
n1 <- length(g1); n2 <- length(g2)
sp <- sqrt( ((n1 - 1) * sd(x1)^2 + (n2-1) * sd(x2)^2) / (n1 + n2-2))
md <- mean(g2) - mean(g1)
semd <- sp * sqrt(1 / n1 + 1/n2)
rbind(
  md + c(-1, 1) * qt(.975, n1 + n2 - 2) * semd,
  t.test(g2, g1, paired = FALSE, var.equal = TRUE)$conf,
  t.test(g2, g1, paired = TRUE)$conf
)
```

The results are:

```
       [,1]  [,2]
[1,] -0.2039 3.364
[2,] -0.2039 3.364
[3,]  0.7001 2.460
```

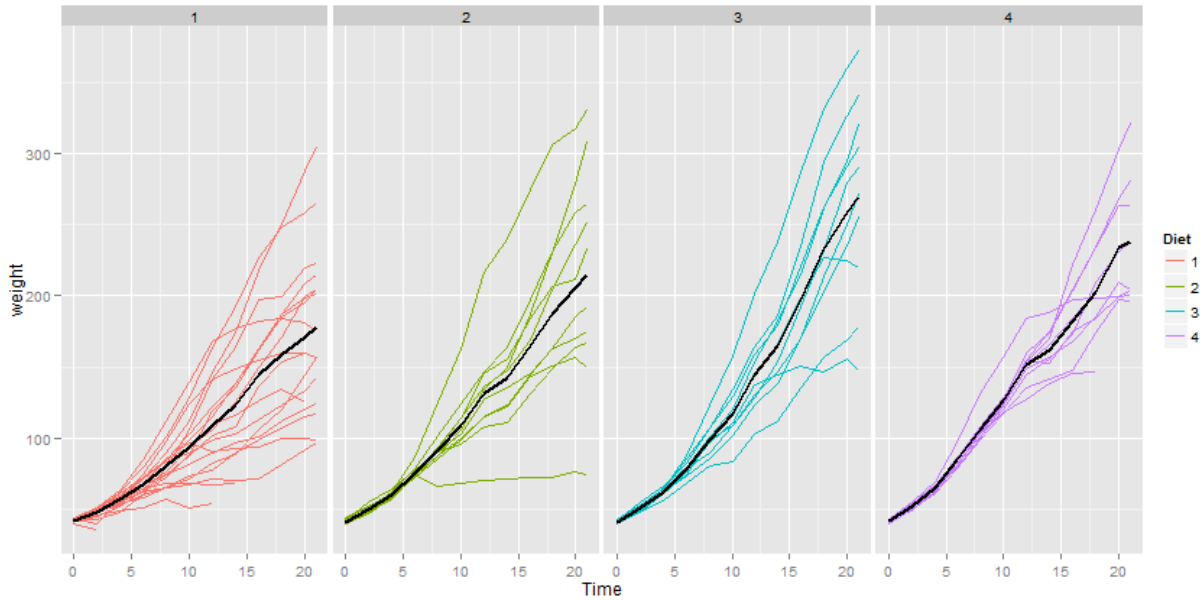
Notice that the paired interval (the last row) is entirely above zero. The grouped interval (first two rows) contains zero. Thus, acknowledging the pairing explains variation that would otherwise be absorbed into the variation for the group means. As a result, treating the groups as independent results in wider intervals. Even if it didn't result in a shorter interval, the paired interval would be correct as the groups are not statistically independent!

`ChickWeight` data in R

Now let's try an example with actual independent groups. Load in the `ChickWeight` data in R. We are also going to manipulate the dataset to have a total weight gain variable using dplyr.

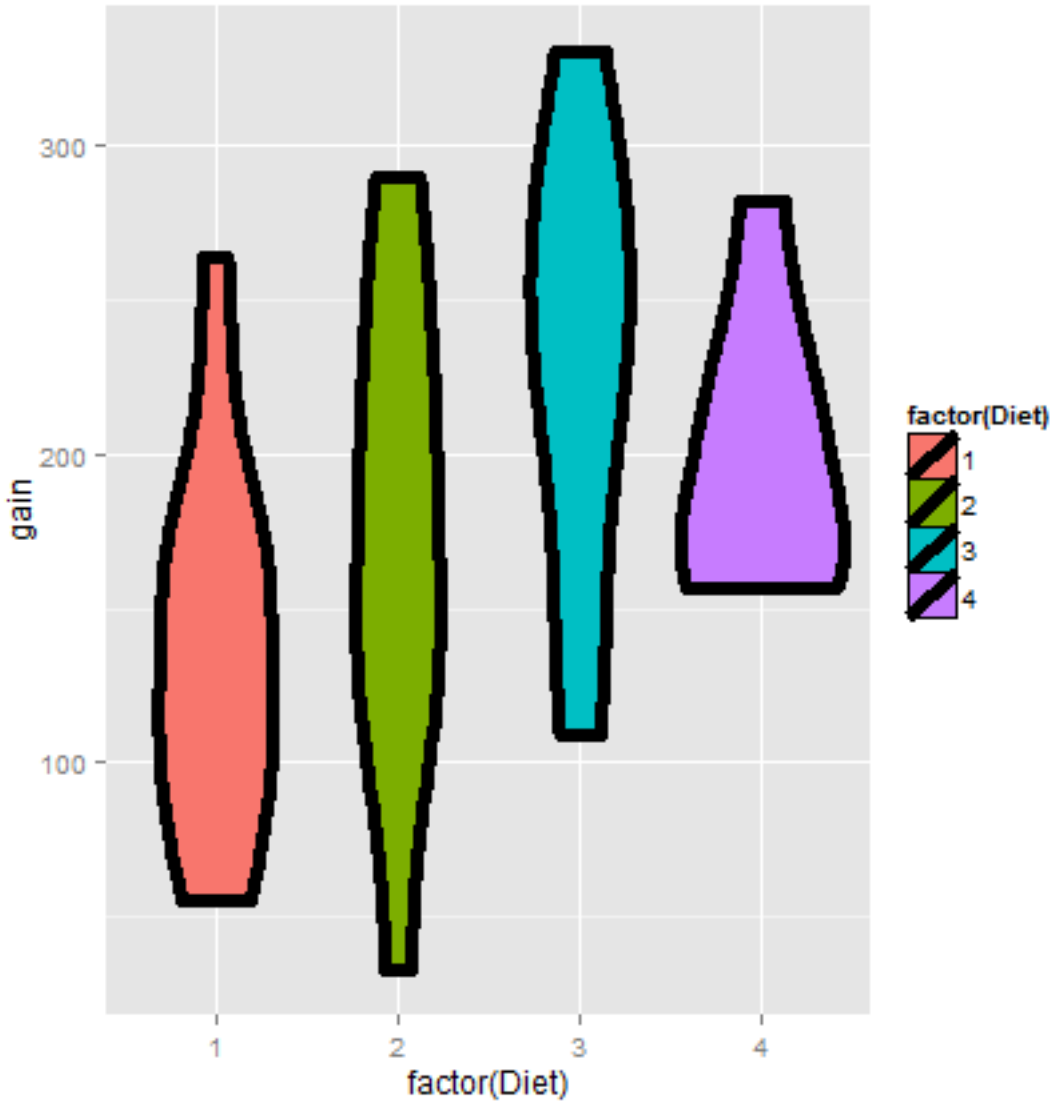
```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Here's a plot of the data.



Chickweight data over time.

Here's a plot only of the weight gain for the diets.



Violin plots of chickweight data by weight gain (final minus baseline) by diet.

Now let's do a *t* interval comparing groups 1 and 4. We'll show the two intervals, one assuming that the variances are equal and one assuming otherwise.

Code for t interval of the chickWeight data

```
wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
rbind(
  t.test(gain ~ Diet, paired = FALSE, var.equal = TRUE, data = wideCW14)$conf,
  t.test(gain ~ Diet, paired = FALSE, var.equal = FALSE, data = wideCW14)$conf
)
```

```
     [,1]   [,2]
[1,] -108.1 -14.81
[2,] -104.7 -18.30
```

For the time being, let's interpret the equal variance interval. Since the interval is entirely below zero it suggest that group 1 had less weight gain than group 4 (at 95% confidence).

Unequal variances

Watch this video before beginning.[4]

Under unequal variances our t interval becomes:

$$\bar{Y} - \bar{X} \pm t_{df} \times \left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^{1/2}$$

where t_{df} is the t quantile calculated with degrees of freedom:

$$df = \frac{\left(S_x^2/n_x + S_y^2/n_y\right)^2}{\left(\frac{S_x^2}{n_x}\right)^2/(n_x - 1) + \left(\frac{S_y^2}{n_y}\right)^2/(n_y - 1)}$$

which will be approximately a 95% interval. This works really well. So when in doubt, just assume unequal variances. Also, we present the formula for completeness. In practice, it's easy to mess up, so make sure to do `t.test`.

Referring back to the previous `ChickWeight` example, the violin plots suggest that considering unequal variances would be wise. Recall the code is

[4] https://www.youtube.com/watch?v=CVDdbR4VuOE&list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ&index=24

```
> t.test(gain ~ Diet, paired = FALSE, var.equal = FALSE, data = wideCW14)$conf
[2,] -104.7 -18.30
```

This interval is remains entirely below zero. However, it is wider than the equal variance interval.

Summary notes

- The *t* distribution is useful for small sample size comparisons.
- It technically assumes normality, but is robust to this assumption within limits.
- The *t* distribution gives rise to *t* confidence intervals (and tests, which we will see later)

For other kinds of data, there are preferable small and large sample intervals and tests.

- For binomial data, there's lots of ways to compare two groups.
  - Relative risk, risk difference, odds ratio.
  - Chi-squared tests, normal approximations, exact tests.
- For count data, there's also Chi-squared tests and exact tests.
- We'll leave the discussions for comparing groups of data for binary and count data until covering glms in the regression class.
- In addition, Mathematical Biostatistics Boot Camp 2 covers many special cases relevant to biostatistics.

Exercises

1. For iid Gaussian data, the statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ must follow a:
   - Z distribution
   - *t* distribution
2. Paired differences T confidence intervals are useful when:
   - Pairs of observations are linked, such as when there is subject level matching or in a study with baseline and follow up measurements on all participants.
   - When there was randomization of a treatment between two independent groups.
3. The assumption that the variances are equal for the independent group T interval means that:
   - The sample variances have to be nearly identical.
   - The population variances are identical, but the sample variances may be different.
4. Load the data set `mtcars` in the `datasets` R package. Calculate a 95% confidence interval to the nearest MPG for the variable `mpg`. Watch a video solution[5] and see written solutions[6].

[5] https://www.youtube.com/watch?v=5BPY6JqRLbE&index=19&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L

[6] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#3

5. Suppose that standard deviation of 9 paired differences is 1. What value would the average difference have to be so that the lower endpoint of a 95% students t confidence interval touches zero? Watch a video solution here[7] and see the text here[8].
6. An independent group Student's T interval is used instead of a paired T interval when:
   - The observations are paired between the groups.
   - The observations between the groups are naturally assumed to be statistically independent.
   - As long as you do it correctly, either is fine.
   - More details are needed to answer this question. watch a discussion of this problem[9] and see the text.[10]
7. Consider the `mtcars` dataset. Construct a 95% T interval for MPG comparing 4 to 6 cylinder cars (subtracting in the order of 4 - 6) assume a constant variance. Watch a video solution[11] and see the text[12]. 10.
8. If someone put a gun to your head and said "Your confidence interval must contain what it's estimating or I'll pull the trigger", what would be the smart thing to do?
   - Make your interval as wide as possible.
   - Make your interval as small as possible.
   - Call the authorities. Watch the video solution[13] and see the text.[14]
9. Refer back to comparing MPG for 4 versus 6 cylinders (question 7). What do you conclude?
   - The interval is above zero, suggesting 6 is better than 4 in the terms of MPG.
   - The interval is above zero, suggesting 4 is better than 6 in the terms of MPG.
   - The interval does not tell you anything about the hypothesis test; you have to do the test.
   - The interval contains 0 suggesting no difference. Watch a video solution[15] and see the text.[16]
10. Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects' body mass indices (BMIs) were measured at a baseline and again after having received the treatment or placebo for four weeks. The average difference from follow-up to the baseline (followup - baseline) was 3 kg/m2 for the treated group and 1 kg/m2 for the placebo group. The corresponding standard deviations of the differences was 1.5 kg/m2 for the treatment group and 1.8 kg/m2 for the placebo group. The study aims to answer whether the change in BMI over the four week period appear to differ between the treated and placebo

[7] https://www.youtube.com/watch?v=ioDwUPCy508&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=20
[8] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#4
[9] https://www.youtube.com/watch?v=zJWJljxJ7Zk&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=21
[10] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#5
[11] https://www.youtube.com/watch?v=QfuMgsUlu_w&index=23&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[12] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#6
[13] https://www.youtube.com/watch?v=8zM1RV4Rb7A&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=24
[14] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#7
[15] https://www.youtube.com/watch?v=zUVoueHLPdo&index=25&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[16] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#8

groups. What is the pooled variance estimate? Watch a video solution here[17] and see the text here.[18]

[17] https://www.youtube.com/watch?v=kzRzrrDWTRQ&index=26&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[18] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw3.html#9

9. Hypothesis testing

Hypothesis testing is concerned with making decisions using data.

Hypothesis testing

Watch this video before beginning.[1]

To make decisions using data, we need to characterize the kinds of conclusions we can make. Classical hypothesis testing is concerned with deciding between two decisions (things get much harder if there's more than two). The first, a null hypothesis is specified that represents the status quo. This hypothesis is usually labeled, H_0. This is what we assume by default. The alternative or research hypothesis is what we require evidence to conclude. This hypothesis is usually labeled, H_a or sometimes H_1 (or some other number other than 0).

So to reiterate, the null hypothesis is assumed true and statistical evidence is required to reject it in favor of a research or alternative hypothesis

Example

A respiratory disturbance index (RDI) of more than 30 events / hour, say, is considered evidence of severe sleep disordered breathing (SDB). Suppose that in a sample of 100 overweight subjects with other risk factors for sleep disordered breathing at a sleep clinic, the mean RDI was 32 events / hour with a standard deviation of 10 events / hour.

We might want to test the hypothesis that

$$H_0 : \mu = 30$$

versus the hypothesis

$$H_a : \mu > 30$$

where μ is the population mean RDI. Clearly, somehow we must figure out a way to decide between these hypotheses using the observed data, particularly the sample mean.

Before we go through the specifics, let's set up the central ideas.

[1] http://youtu.be/Wqvx6_12ZMs?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Types of errors in hypothesis testing

The alternative hypotheses are typically of the form of the true mean being $<$, $>$ or $\neq$ to the hypothesized mean, such as $H_a : \mu > 30$ from our example. The null typically sharply specifies the mean, such as $H_0 : \mu = 30$ in our example. More complex null hypotheses are possible, but are typically covered in later courses.

Note that there are four possible outcomes of our statistical decision process:

Truth	Decide	Result
H_0	H_0	Correctly accept null
H_0	H_a	Type I error
H_a	H_a	Correctly reject null
H_a	H_0	Type II error

We will perform hypothesis testing by forcing the probability of a Type I error to be small. This approach consequences, which we can discuss with an analogy to court cases.

Discussion relative to court cases

Consider a court of law and a criminal case. The null hypothesis is that the defendant is innocent. The rules requires a standard on the available evidence to reject the null hypothesis (and the jury to convict). The standard is specified loosely in this case, such as convict if the defendant appears guilty "Beyond reasonable doubt". In statistics, we can be mathematically specific about our standard of evidence.

Note the consequences of setting a standard. If we set a low standard, say convicting only if there circumstantial or greater evidence, then we would increase the percentage of innocent people convicted (type I errors). However, we would also increase the percentage of guilty people convicted (correctly rejecting the null).

If we set a high standard, say the standard of convicting if the jury has "No doubts whatsoever", then we increase the the percentage of innocent people let free (correctly accepting the null) while we would also increase the percentage of guilty people let free (type II errors).

Building up a standard of evidence

Watch this video before beginning.[2]

Consider our sleep example again. A reasonable strategy would reject the null hypothesis if the sample mean, $\bar{X}$, was larger than some constant, say C. Typically, C is chosen so that the probability

[2] http://youtu.be/obNxIau2zrs?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

of a Type I error, labeled α, is 0.05 (or some other relevant constant) To reiterate, $\alpha =$ Type I error rate = Probability of rejecting the null hypothesis when, in fact, the null hypothesis is correct

Let's see if we can figure out what C has to be. The standard error of the mean is $10/\sqrt{100} = 1$. Furthermore, under H_0 we know that $\bar{X} \sim N(30, 1)$ (at least approximately) via the CLT. We want to chose C so that:

$$P(\bar{X} > C; H_0) = 0.05.$$

The 95th percentile of a normal distribution is 1.645 standard deviations from the mean. So, if C is set 1.645 standard deviations from the mean, we should be set since the probability of getting a sample mean that large is only 5%. The 95th percentile from a $N(30, 1)$ is:

$$C = 30 + 1 \times 1.645 = 31.645.$$

So the rule "Reject H_0 when $\bar{X} \geq 31.645$" has the property that the probability of rejection is 5% when H_0 is true.

In general, however, we don't convert C back to the original scale. Instead, we calculate how many standard errors the observed mean is from the hypothesized mean

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}.$$

This is called a Z-score. We can compare this statistic to standard normal quantiles.

To reiterate, the Z-score is how many standard errors the sample mean is above the hypothesized mean. In our example:

$$\frac{32 - 30}{10/\sqrt{100}} = 2$$

Since 2 is greater than 1.645 we would reject. Setting the rule "We reject if our Z-score is larger than 1.645" controls the Type I error rate at 5%. We could write out a general rule for this alternative hypothesis as reject whenever $\sqrt{n}(\bar{X} - \mu_0)/s > Z_{1-\alpha}$ where α is the desired Type I error rate.

Because the Type I error rate was controlled to be small, if we reject we know that one of the following occurred:

1. the null hypothesis is false,
2. we observed an unlikely event in support of the alternative even though the null is true,
3. our modeling assumptions are wrong.

The third option can be difficult to check and at some level all bets are off depending on how wrong we are about our basic assumptions. So for this course, we speak of our conclusions under the assumption that our modeling choices (such as the iid sampling model) are correct, but do so wide eyed acknowledging the limitations of our approach.

General rules

We developed our test for one possible alternatives. Here's some general rules for the three most important alternatives.

Consider the Z test for $H_0 : \mu = \mu_0$ versus: $H_1 : \mu < \mu_0$, $H_2 : \mu \neq \mu_0$, $H_3 : \mu > \mu_0$. Our test statistic

$$TS = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}.$$

We reject the null hypothesis when:

$H_1 : \quad TS \leq Z_\alpha = -Z_{1-\alpha}$,

$$H_2 : \quad |TS| \geq Z_{1-\alpha/2}$$

$H_3 : \quad TS \geq Z_{1-\alpha}$,

respectively.

Summary notes

- We have fixed α to be low, so if we reject H_0 (either our model is wrong) or there is a low probability that we have made an error.
- We have not fixed the probability of a type II error, β; therefore we tend to say "Fail to reject H_0" rather than accepting H_0.
- Statistical significance is no the same as scientific significance.
- The region of TS values for which you reject H_0 is called the rejection region.
- The Z test requires the assumptions of the CLT and for n to be large enough for it to apply.
- If n is small, then a Gosset's t test is performed exactly in the same way, with the normal quantiles replaced by the appropriate Student's t quantiles and $n - 1$ df.
- The probability of rejecting the null hypothesis when it is false is called **power**
- Power is a used a lot to calculate sample sizes for experiments.

Example reconsidered

Watch this video before beginning.[3]

Consider our example again. Suppose that $n = 16$ (rather than 100). The statistic

$$\frac{\bar{X} - 30}{s/\sqrt{16}},$$

follows a t distribution with 15 df under H_0.

Under H_0, the probability that it is larger that the 95th percentile of the t distribution is 5%. The 95th percentile of the T distribution with 15 df is 1.7531 (obtained via `r qt(.95, 15)`).

Assuming that everything but the sample size is the same, our test statistic is now $\sqrt{16}(32 - 30)/10 = 0.8$. Since 0.8 is not larger than 1.75, we now fail to reject.

Two sided tests

In many settings, we would like to reject if the true mean is *different* than the hypothesized, not just larger or smaller. I other words, we would reject the null hypothesis if in fact the sample mean was much larger or smaller than the hypothesized mean. In our example, we want to test the alternative $H_a : \mu \neq 30$.

We will reject if the test statistic, 0.8, is either too large or too small. Then we want the probability of rejecting under the null to be 5%, split equally as 2.5% in the upper tail and 2.5% in the lower tail.

Thus we reject if our test statistic is larger than `qt(.975, 15)` or smaller than `qt(.025, 15)`. This is the same as saying: reject if the absolute value of our statistic is larger than `qt(0.975, 15)` = 2.1314.

In this case, since our test statistic is 0.8, which is smaller than 2.1314, we fail to reject the two sided test (as well as the one sided test).

If you fail to reject the one sided test, then you would fail to reject the two sided test. Because of its larger rejection region, two sided tests are the norm (even in settings where a one sided test makes more sense).

T test in R

Let's try the t test on the pairs of fathers and sons in Galton's data.

[3] http://youtu.be/5iMMBTlOFTI?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Example of using the *t* test in R.

```
> library(UsingR); data(father.son)
> t.test(father.son$sheight - father.son$fheight)

        One Sample t-test

data:  father.son$sheight - father.son$fheight
t = 11.79, df = 1077, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.831 1.163
sample estimates:
mean of x
    0.997
```

Connections with confidence intervals

Consider testing $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$. Take the set of all possible values for which you fail to reject H_0, this set is a $(1 - \alpha)100\%$ confidence interval for μ.

The same works in reverse; if a $(1 - \alpha)100\%$ interval contains μ_0, then we *fail to* reject H_0.

In other words, two sided tests and confidence intervals agree.

Two group intervals

Doing group tests is now straightforward given that we've already covered independent group T intervals. Our rejection rules are the same, the only change is how the statistic is calculated. However, the form is familiar:

$$\frac{\text{Estimate} - \text{Hypothesized Value}}{\text{Standard Error}}.$$

Consider now testing $H_0 : \mu_1 = \mu_2$. Our statistic is

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_0)}{S_p\sqrt{\frac{1}{n1} + \frac{1}{n_2}}}.$$

For the equal variance case and and

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_0)}{\sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n_2}}}.$$

Let's just go through an example.

Example `chickWeight` data

Recall that we reformatted this data as follows

Reformatting the data.

```
library(datasets); data(ChickWeight); library(reshape2)
##define weight gain or loss
wideCW <- dcast(ChickWeight, Diet + Chick ~ Time, value.var = "weight")
names(wideCW)[-(1 : 2)] <- paste("time", names(wideCW)[-(1 : 2)], sep = "")
library(dplyr)
wideCW <- mutate(wideCW,
  gain = time21 - time0
)
```

Unequal variance T test comparing diets 1 and 4.

```
> wideCW14 <- subset(wideCW, Diet %in% c(1, 4))
> t.test(gain ~ Diet, paired = FALSE,
+        var.equal = TRUE, data = wideCW14)

          Two Sample t-test

  data:  gain by Diet
  t = -2.725, df = 23, p-value = 0.01207
  alternative hypothesis: true difference in means is not equal to 0
  95 percent confidence interval:
   -108.15  -14.81
  sample estimates:
  mean in group 1 mean in group 4
            136.2           197.7
```

Exact binomial test

Recall this problem. Suppose a friend has 8 children, 7 of which are girls and none are twins.

Perform the relevant hypothesis test. $H_0 : p = 0.5$ versus $H_a : p > 0.5$.

What is the relevant rejection region so that the probability of rejecting is (less than) 5%?

Rejection region	Type I error rate
$[0:8]$	1
$[1:8]$	0.9961
$[2:8]$	0.9648
$[3:8]$	0.8555
$[4:8]$	0.6367
$[5:8]$	0.3633
$[6:8]$	0.1445
$[7:8]$	0.0352
$[8:8]$	0.0039

Thus if we reject under 7 or 8 girls, we will have a less than 5% chance of rejecting under the null hypothesis.

It's impossible to get an exact 5% level test for this case due to the discreteness of the binomial. The closest is the rejection region $[7:8]$. Further note that an alpha level lower than 0.0039 is not attainable. For larger sample sizes, we could do a normal approximation.

Extended this test to two sided test isn't obvious. Given a way to do two sided tests, we could take the set of values of p_0 for which we fail to reject to get an exact binomial confidence interval (called the Clopper/Pearson interval, by the way). We'll cover two sided versions of this test when we cover P-values.

Exercises

1. Which hypothesis is typically assumed to be true in hypothesis testing?
   - The null.
   - The alternative.
2. The type I error rate controls what?
3. Load the data set `mtcars` in the `datasets` R package. Assume that the data set mtcars is a random sample. Compute the mean MPG, $\bar{x}$, of this sample. You want to test whether the true MPG is μ_0 or smaller using a one sided 5% level test. ($H_0 : \mu = \mu_0$ versus $H_a : \mu < \mu_0$). Using that data set and a Z test: Based on the mean MPG of the sample $\bar{x}$, and by using a Z test: what is the smallest value of μ_0 that you would reject for (to two decimal places)? Watch a video solution here[4] and see the text here[5].
4. Consider again the `mtcars` dataset. Use a two group t-test to test the hypothesis that the 4 and 6 cyl cars have the same mpg. Use a two sided test with unequal variances. Do you reject? Watch the video here[6] and see the text here[7]

[4] https://www.youtube.com/watch?v=gReR0uxLnIA&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=27

[5] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#3

[6] https://www.youtube.com/watch?v=Zo5TirzS9rU&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=28

[7] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#4

5. A sample of 100 men yielded an average PSA level of 3.0 with a sd of 1.1. What are the complete set of values that a 5% two sided Z test of $H_0 : \mu = \mu_0$ would fail to reject the null hypothesis for? Watch the video solution[8] and see the text[9].
6. You believe the coin that you're flipping is biased towards heads. You get 55 heads out of 100 flips. Do you reject at the 5% level that the coin is fair? Watch a video solution[10] and see the text[11].
7. Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid, perform a Z test of equivalence. Do you reject at the 5% level? Watch a video solution[12] and see the text.[13]
8. A confidence interval for the mean contains:
   - All of the values of the hypothesized mean for which we would fail to reject with $\alpha = 1 - Conf.Level$.
   - All of the values of the hypothesized mean for which we would fail to reject with $2\alpha = 1 - Conf.Level$.
   - All of the values of the hypothesized mean for which we would reject with $\alpha = 1 - Conf.Level$.
   - All of the values of the hypothesized mean for which we would reject with $2\alpha = 1 - Conf.Level$. Watch a video solution[14] and see the text[15].
9. In a court of law, all things being equal, if via policy you require a lower standard of evidence to convict people then
   - Less guilty people will be convicted.
   - More innocent people will be convicted.
   - More Innocent people will be not convicted. Watch a video solution[16] and see the text[17].

[8] https://www.youtube.com/watch?v=TooyEaVgLZc&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=29
[9] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#5
[10] https://www.youtube.com/watch?v=0sqOErsfhqo&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=30
[11] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#6
[12] https://www.youtube.com/watch?v=Or4ly4rOiaA&index=32&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[13] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#8
[14] https://www.youtube.com/watch?v=UiNV1mXQGLs&index=33&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[15] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#9
[16] https://www.youtube.com/watch?v=GlKPG24bZMI&index=36&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[17] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#12

10. P-values

Introduction to P-values

Watch this video before beginning.[1]

P-values are the most common measure of statistical significance. Their ubiquity, along with concern over their interpretation and use makes them controversial among statisticians. The following manuscripts are interesting reads about P-values.

- http://warnercnr.colostate.edu/~anderson/thompson1.html[2]
- Also see *Statistical Evidence: A Likelihood Paradigm* by Richard Royall[3]
- *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* by Steve Goodman[4]
- The hilariously titled: *The Earth is Round (p < .05)* by Cohen.[5]
- Some positive comments
  - simply statistics[6]
  - normal deviate[7]
  - Error statistics[8]

What is a P-value?

The central idea of a P-value is to assume that the null hypothesis is true and calculate how unusual it would be to see data (in the form of a test statistic) as extreme as was seen in favor of the alternative hypothesis. The formal definition is:

A **P-value** is the probability of observing a test statistic as or more extreme in favor of the alternative than was actually obtained, where the probability is calculated assuming that the null hypothesis is true.

A P-value then requires a few steps. 1. Decide on a statistic that evaluates support of the null or alternative hypothesis. 2. Decide on a distribution of that statistic under the null hypothesis (null

[1] http://youtu.be/Ky68x_7iK6c?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

[2] http://warnercnr.colostate.edu/~anderson/thompson1.html

[3] http://www.crcpress.com/product/isbn/9780412044113

[4] https://scholar.google.com/scholar?q=towards+evidence+based+medical+statistics+the+p-value+fallacy&hl=en&as_sdt=0&as_vis=1&oi=scholart&sa=X&ei=uOTjVNHdG4anggSMlYOwBQ&ved=0CBsQgQMwAA

[5] http://www.scopus.com/record/display.url?eid=2-s2.0-0039802908&origin=inward&txGid=BBE363C58BE8785BFF9E71AB60004733.ZmAySxCHIBxxTXbnsoe5w%3a2

[6] http://simplystatistics.org/2012/01/06/p-values-and-hypothesis-testing-get-a-bad-rap-but-we/

[7] http://normaldeviate.wordpress.com/2013/03/14/double-misunderstandings-about-p-values/

[8] http://errorstatistics.com/2013/06/14/p-values-cant-be-trusted-except-when-used-to-argue-that-p-values-cant-be-trusted/

distribution). 3. Calculate the probability of obtaining a statistic as or more extreme as was observed using the distribution in 2.

The way to interpret P-values is as follows. If the P-value is small, then either H_0 is true and we have observed a rare event or H_0 is false (or possibly the null model is incorrect).

Let's do a quick example. Suppose that you get a t statistic of 2.5 for 15 degrees of freedom testing $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$. What's the probability of getting a t statistic as large as 2.5?

P-value calculation in R.

```
> pt(2.5, 15, lower.tail = FALSE)
[1] 0.01225
```

Therefore, the probability of seeing evidence as extreme or more extreme than that actually obtained under H_0 is 0.0123. So, (assuming our model is correct) either we observed data that was pretty unlikely under the null, or the null hypothesis if false.

The attained significance level

Recall in a previous chapter that our test statistic was 2 for $H_0 : \mu_0 = 30$ versus $H_a : \mu > 30$ using a normal test (n was 100). Notice that we rejected the one sided test when $\alpha = 0.05$, would we reject if $\alpha = 0.01$, how about 0.001?

The smallest value for alpha that you still reject the null hypothesis is called the *attained significance level*. This is mathematically equivalent, but philosophically a little different from, the *P-value*. Whereas the P-value is interpreted in the terms of how probabilistically extreme our test statistic is under the null, the attained significance level merely conveys what the smallest level of α that one could reject at.

This equivalence makes P-values very convenient to convey. The reader of the results can perform the test at whatever α he or she choses. This is especially useful in multiple testing circumstances.

Here's the two rules for performing hypothesis tests with P-values. * If the P-value for a test is less than α you reject the null hypothesis * For two sided hypothesis test, double the smaller of the two one sided hypothesis test Pvalues

Binomial P-value example

Suppose a friend has 8 children, 7 of which are girls and none are twins. 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?

This calculation is a P-value where the statistic is the number of girls and the null distribution is a fair coin flip for each gender. We want to test $H_0 : p = 0.5$ versus $H_a : p > 0.5$, where p is the probability of having a girl for each birth.

Recall here's the calculation:

Example of a Binomial P-value calculation in R.

```
> pbinom(6, size = 8, prob = 0.5, lower.tail = FALSE)
[1] 0.03516
```

Since our P-value is less than 0.05 we would reject at a 5% error rate. Note, however, if we were doing a two sided test, we would have to double the P-value and thus would then fail to reject.

Poisson example

Watch this video before beginning.[9]

Suppose that a hospital has an infection rate of 10 infections per 100 person/days at risk (rate of 0.1) during the last monitoring period. Assume that an infection rate of 0.05 is an important benchmark.

Given a Poisson model, could the observed rate being larger than 0.05 be attributed to chance? We want to test $H_0 : \lambda = 0.05$ where λ is the rate of infections per person day so that 5 would be the rate per 100 days. Thus we want to know if 9 events per 100 person/days is unusual with respect to a Poisson distribution with a rate of 5 events per 100. Consider $H_a : \lambda > 0.05$.

Poisson P-value calculation.

```
> ppois(9, 5, lower.tail = FALSE)
[1] 0.03183
```

Again, since this P-value is less than 0.05 we reject the null hypothesis. The P-value would be 0.06 for two sided hypothesis (double) and so we would fail to reject in that case.

Exercises

1. P-values are probabilities that are calculated assuming which hypothesis is true?
    - the alternative
    - the null
2. You get a P-value of 0.06. Would you reject for a type I error rate of 0.05?
    - Yes you would reject the null

[9] http://youtu.be/Tcw2OVyEX3s?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

- No you would not reject the null
- It depends on information not given

3. The proposed procedure for getting a two sided P-value for the exact binomial test considered here is what?
   - Multiplying the one sided P-value by one half
   - Doubling the larger of the two one sided P-values
   - Doubling the smaller of the two one sided P-values
   - No procedure exists
4. Consider again the `mtcars` dataset. Use a two group t-test to test the hypothesis that the 4 and 6 cyl cars have the same mpg. Use a two sided test with unequal variances. Give a P-value. Watch the video here[10] and see the text here[11]
5. You believe the coin that you're flipping is biased towards heads. You get 55 heads out of 100 flips. Give an exact P-value for the hypothesis that the coin is fair. Watch a video solution[12] and see the text[13].
6. A web site was monitored for a year and it received 520 hits per day. In the first 30 days in the next year, the site received 15,800 hits. Assuming that web hits are Poisson. Give an exact one sided P-value to the hypothesis that web hits are up this year over last. Do you reject? Watch the video solutions[14] and see the problem text[15].
7. Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. Assuming that the groups are independent and that they days are iid, perform a Z test of equivalence. Give a P-value for the test? Watch a video solution[16] and see the text.[17]
8. Consider the `mtcars` data set.
   - Give the p-value for a t-test comparing MPG for 6 and 8 cylinder cars assuming equal variance, as a proportion to 3 decimal places.
   - Give the associated P-value for a z test.
   - Give the common standard deviation estimate for MPG across cylinders to 3 decimal places.
   - Would the t test reject at the two sided 0.05 level (0 for no 1 for yes)? Watch a video solution[18] and see the text[19].

[10] https://www.youtube.com/watch?v=Zo5TirzS9rU&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=28
[11] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#4
[12] https://www.youtube.com/watch?v=0sqOErsfhqo&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=30
[13] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#6
[14] https://www.youtube.com/watch?v=cE_88-Q7TX0&index=31&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[15] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#7
[16] https://www.youtube.com/watch?v=Or4ly4rOiaA&index=32&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[17] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#8
[18] https://www.youtube.com/watch?v=m0B5p0w2wJI&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L&index=37
[19] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#13

11. Power

Power

Watch this video before beginning.[1] and then watch this video as well.[2]

Power is the probability of rejecting the null hypothesis when it is false. Ergo, power (as its name would suggest) is a good thing; you want more power. A type II error (a bad thing, as its name would suggest) is failing to reject the null hypothesis when it's false; the probability of a type II error is usually called β. Note Power $= 1 - \beta$.

Let's go through an example of calculating power. Consider our previous example involving RDI. $H_0 : \mu = 30$ versus $H_a : \mu > 30$. Then power is:

$$P\left(\frac{\bar{X} - 30}{s/\sqrt{n}} > t_{1-\alpha,n-1} \; ; \; \mu = \mu_a\right).$$

Note that this is a function that depends on the specific value of μ_a! Further notice that as μ_a approaches 30 the power approaches α.

Pushing this example further, we reject if

$$Z = \frac{\bar{X} - 30}{\sigma/\sqrt{n}} > z_{1-\alpha}$$

Or, equivalently, if

$$\bar{X} > 30 + Z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$$

But, note that, under $H_0 : \bar{X} \sim N(\mu_0, \sigma^2/n)$. However, under $H_a : \bar{X} \sim N(\mu_a, \sigma^2/n)$.

So for this test we could calculate power with this R code:

[1] http://youtu.be/-TsBOLiW4rQ?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ
[2] http://youtu.be/GRS2b1aedmk?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

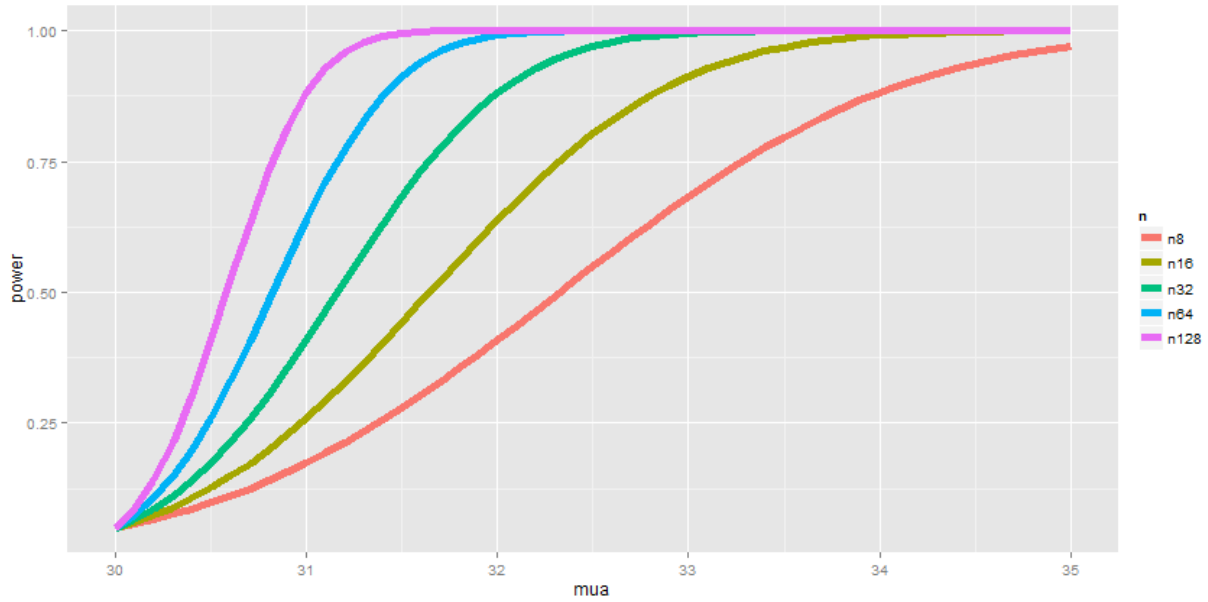
Power calculation for the sleep example in R

```
alpha = 0.05
z = qnorm(1 - alpha)
pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FALS\
E)
```

Let's plug in the specific numbers for our example where: $\mu_a = 32$, $\mu_0 = 30$, $n = 16$, $\sigma = 4$.

```
> mu0 = 30
> mua = 32
> sigma = 4
> n = 16
> z = qnorm(1 - alpha)
> pnorm(mu0 + z * sigma/sqrt(n), mean = mu0, sd = sigma/sqrt(n), lower.tail = FA\
LSE)
[1] 0.05
> pnorm(mu0 + z * sigma/sqrt(n), mean = mua, sd = sigma/sqrt(n), lower.tail = FA\
LSE)
[1] 0.6388
```

When we plug in μ_0, the value under the null hypothesis, we get that the probability of rejection is 5%, as the test was designed. However, when we plug in a value of 32, we get 64%. Therefore, the probability of rejection is 64% when the true value of μ is 32. We could create a curve of the power as a function of μ_a, as seen below. We also varied the sample size to see how the curve depends on that.



Plot of power as μ_a varies.

The code below shows how to use manipulate to investigate power as the various inputs change.

Code for investigating power.

```
library(manipulate)
mu0 = 30
myplot <- function(sigma, mua, n, alpha) {
    g = ggplot(data.frame(mu = c(27, 36)), aes(x = mu))
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mu0,
        sd = sigma/sqrt(n)), size = 2, col = "red")
    g = g + stat_function(fun = dnorm, geom = "line", args = list(mean = mua,
        sd = sigma/sqrt(n)), size = 2, col = "blue")
    xitc = mu0 + qnorm(1 - alpha) * sigma/sqrt(n)
    g = g + geom_vline(xintercept = xitc, size = 3)
    g
}
manipulate(myplot(sigma, mua, n, alpha), sigma = slider(1, 10, step = 1, initial\
 = 4),
    mua = slider(30, 35, step = 1, initial = 32), n = slider(1, 50, step = 1,
        initial = 16), alpha = slider(0.01, 0.1, step = 0.01, initial = 0.05))
```

Question

Watch this video before beginning.[3]

When testing $H_a : \mu > \mu_0$, notice if power is $1 - \beta$, then

$$1 - \beta = P\left(\bar{X} > \mu_0 + z_{1-\alpha}\frac{\sigma}{\sqrt{n}}; \mu = \mu_a\right)$$

where $\bar{X} \sim N(\mu_a, \sigma^2/n)$. The unknowns in the equation are: μ_a, σ, n, β and the knowns are: μ_0, α. Specify any 3 of the unknowns and you can solve for the remainder.

Notes

- The calculation for $H_a : \mu < \mu_0$ is similar
- For $H_a : \mu \neq \mu_0$ calculate the one sided power using $\alpha/2$ (this is only approximately right, it excludes the probability of getting a large TS in the opposite direction of the truth)
- Power goes up as α gets larger
- Power of a one sided test is greater than the power of the associated two sided test
- Power goes up as μ_1 gets further away from μ_0
- Power goes up as n goes up
- Power doesn't need μ_a, σ and n, instead only $\frac{\sqrt{n}(\mu_a - \mu_0)}{\sigma}$
  - The quantity $\frac{\mu_a - \mu_0}{\sigma}$ is called the *effect size*, the difference in the means in standard deviation units.
  - Being unit free, it has some hope of interpretability across settings.

T-test power

Watch this before beginning.[4]

Consider calculating power for a Gosset's *t* test for our example where we now assume that $n = 16$. The power is

$$P\left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} > t_{1-\alpha,n-1}\ ;\ \mu = \mu_a\right).$$

Calculating this requires the so-called non-central t distribution. However, fortunately for us, the R function `power.t.test` does this very well. Omit (exactly) any one of the arguments and it solves for it. Our t-test power again only relies on the effect size.

Let's do our example trying different options.

[3] http://youtu.be/3bWhP5MyuqI?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ
[4] http://youtu.be/1DiwutNpt5Y?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Example of using 'power.t.test' in R.

```
# omitting the power and getting a power estimate
> power.t.test(n = 16, delta = 2/4, sd = 1, type = "one.sample", alt = "one.side\
d")$power
[1] 0.604
# illustrating that it depends only on the effect size, delta/sd
> power.t.test(n = 16, delta = 2, sd = 4, type = "one.sample", alt = "one.sided"\
)$power
[1] 0.604
# same thing again
> power.t.test(n = 16, delta = 100, sd = 200, type = "one.sample", alt = "one.si\
ded")$power
[1] 0.604
# specifying the power and getting n
> power.t.test(power = 0.8, delta = 2/4, sd = 1, type = "one.sample", alt = "one\
.sided")$n
[1] 26.14
# again illustrating that the effect size is all that matters
power.t.test(power = 0.8, delta = 2, sd = 4, type = "one.sample", alt = "one.sid\
ed")$n
[1] 26.14
# again illustrating that the effect size is all that matters
> power.t.test(power = 0.8, delta = 100, sd = 200, type = "one.sample", alt = "o\
ne.sided")$n
[1] 26.14
```

Exercises

1. Power is a probability calculation assuming which is true:
    - The null hypothesis
    - The alternative hypothesis
    - Both the null and alternative
2. As your sample size gets bigger, all else equal, what do you think would happen to power?
    - It would get larger
    - It would get smaller
    - It would stay the same
    - It cannot be determined from the information given
3. What happens to power as μ_a gets further from μ_0?
    - Power decreases
    - Power increases

- Power stays the same
- Power oscillates

4. In the context of calculating power, the effect size is?
   - The null mean divided by the standard deviation
   - The alternative mean divided by the standard error
   - The difference between the null and alternative means divided by the standard deviation
   - The standard error divided by the null mean
5. Recall this problem "Suppose that in an AB test, one advertising scheme led to an average of 10 purchases per day for a sample of 100 days, while the other led to 11 purchases per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day." Assuming that 10 purchases per day is a benchmark null value, that days are iid and that the standard deviation is 4 purchases for day. Suppose that you plan on sampling 100 days. What would be the power for a one sided 5% Z mean test that purchases per day have increased under the alternative of $\mu = 11$ purchase per day? Watch a video solution[5] and see the text[6].
6. Researchers would like to conduct a study of healthy adults to detect a four year mean brain volume loss of .01 mm3. Assume that the standard deviation of four year volume loss in this population is .04 mm3. What is necessary sample size for the study for a 5% one sided test versus a null hypothesis of no volume loss to achieve 80% power? Watch the video solution[7] and see the text[8].

[5] https://www.youtube.com/watch?v=RiS6EFnPYY8&index=34&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[6] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#10
[7] https://www.youtube.com/watch?v=lrXyJrtatzk&index=35&list=PLpl-gQkQivXhHOcVeU3bSJg78zaDYbP9L
[8] http://bcaffo.github.io/courses/06_StatisticalInference/homework/hw4.html#11

12. The bootstrap and resampling

The bootstrap

Watch this video before beginning.[1]

The bootstrap is a tremendously useful tool for constructing confidence intervals and calculating standard errors for difficult statistics. For a classic example, how would one derive a confidence interval for the median? The bootstrap procedure follows from the so called bootstrap principle

To illustrate the bootstrap principle, imagine a die roll. The image below shows the mass function of a die roll on the left. On the right we show the empirical distribution obtained by repeatedly averaging 50 independent die rolls. By this simulation, without any mathematics, we have a good idea of what the distribution of averages of 50 die rolls looks like.

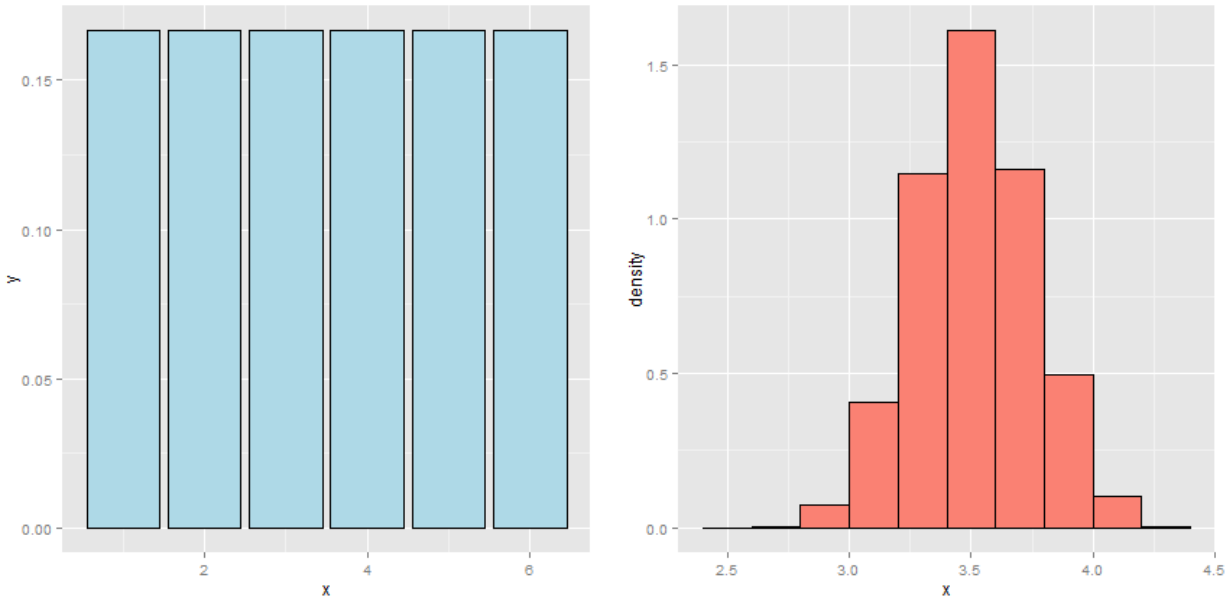


Image of true die roll distribution (left) and simulation of averages of 50 die rolls

Now imagine a case where we didn't know whether or not the die was fair. We have a sample of size 50 and we'd like to investigate the distribution of the average of 50 die rolls *where we're not allowed to roll the die anymore*. This is more like a real data analysis, we only get one sample from the population.

[1] http://youtu.be/0hNQx9nagq4?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

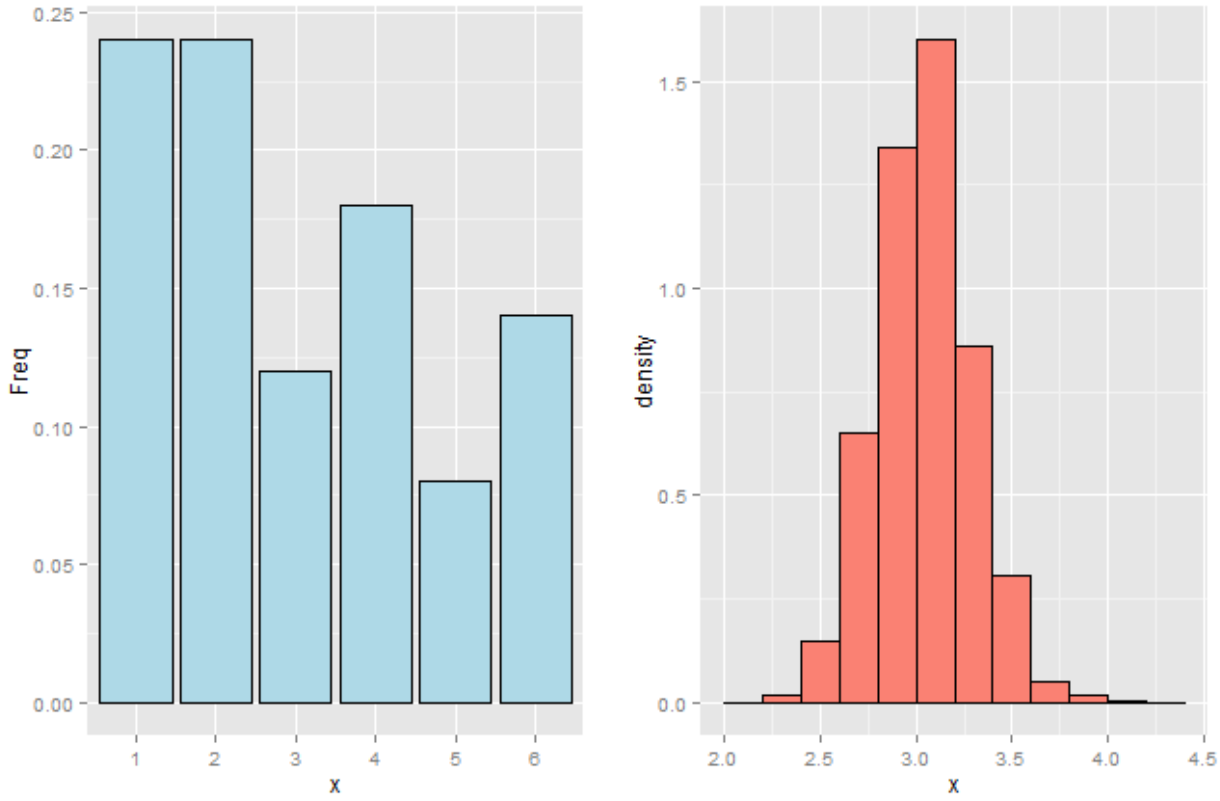


Image of empirical die roll distribution (left) and simulates of averages of 50 die rolls from this distribution

The bootstrap principle is to use the empirical mass function of the data to perform the simulation, rather than the true distribution. That is, we simulate averages of 50 samples from the histogram that we observe. With enough data, the empirical distribution should be a good estimate of the true distribution and this should result in a good approximation of the sampling distribution.

That's the bootstrap principle: investigate the sampling distribution of a statistic by simulating repeated realizations from the observed distribution.

If we could simulate from the true distribution, then we would know the exact sampling distribution of our statistic (if we ran our computer long enough.) However, since we only get to sample from that distribution once, we have to be content with using the empirical distribution. This is the clever idea of the bootstrap.

Example Galton's fathers and sons dataset

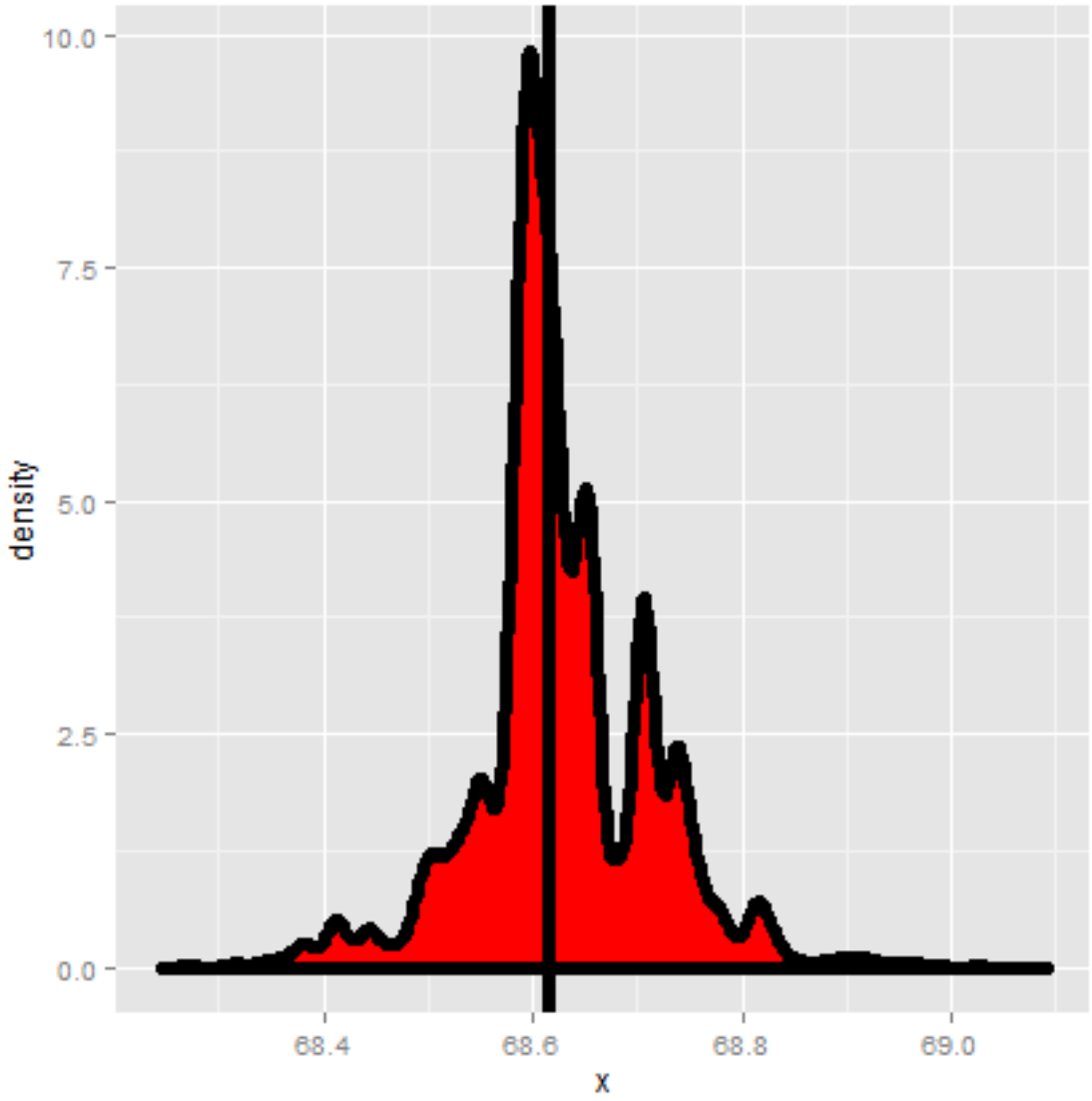
Watch this video before beginning.[2]

The code below creates resamples via draws of size n with replacement with the original data of the son's heights from Galton's data and plots a histogram of the median of each resampled dataset.

[2] http://youtu.be/yNTWcmbWvWg?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Bootstrapping example

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
B <- 10000
resamples <- matrix(sample(x,
                           n * B,
                           replace = TRUE),
                    B, n)
resampledMedians <- apply(resamples, 1, median)
```



Bootstrapping example for the median of sons' heights from Galton's

The bootstrap principle

Watch this video before beginning.[3]

Suppose that I have a statistic that estimates some population parameter, but I don't know its sampling distribution. The bootstrap principle suggests using the distribution defined by the data to approximate its sampling distribution

[3] http://youtu.be/BKrMjX7FBno?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

The bootstrap in practice

In practice, the bootstrap principle is always carried out using simulation. We will cover only a few aspects of bootstrap resampling. The general procedure follows by first simulating complete data sets from the observed data with replacement. This is approximately drawing from the sampling distribution of that statistic, at least as far as the data is able to approximate the true population distribution. Calculate the statistic for each simulated data set Use the simulated statistics to either define a confidence interval or take the standard deviation to calculate a standard error.

Nonparametric bootstrap algorithm example

Bootstrap procedure for calculating confidence interval for the median from a data set of n observations:

1. Sample n observations **with replacement** from the observed data resulting in one simulated complete data set.
2. Take the median of the simulated data set
3. Repeat these two steps B times, resulting in B simulated medians
4. These medians are approximately drawn from the sampling distribution of the median of n observations; therefore we can:
    - Draw a histogram of them
    - Calculate their standard deviation to estimate the standard error of the median
    - Take the 2.5^{th} and 97.5^{th} percentiles as a confidence interval for the median

For the general bootstrap, just replace the median with whatever statistic that you're investigating.

Example code

Consider our father/son data from before. Here is the relevant code for doing the resampling.

```
B <- 10000
resamples <- matrix(sample(x,
                           n * B,
                           replace = TRUE),
                    B, n)
medians <- apply(resamples, 1, median)
```

And here is some results.

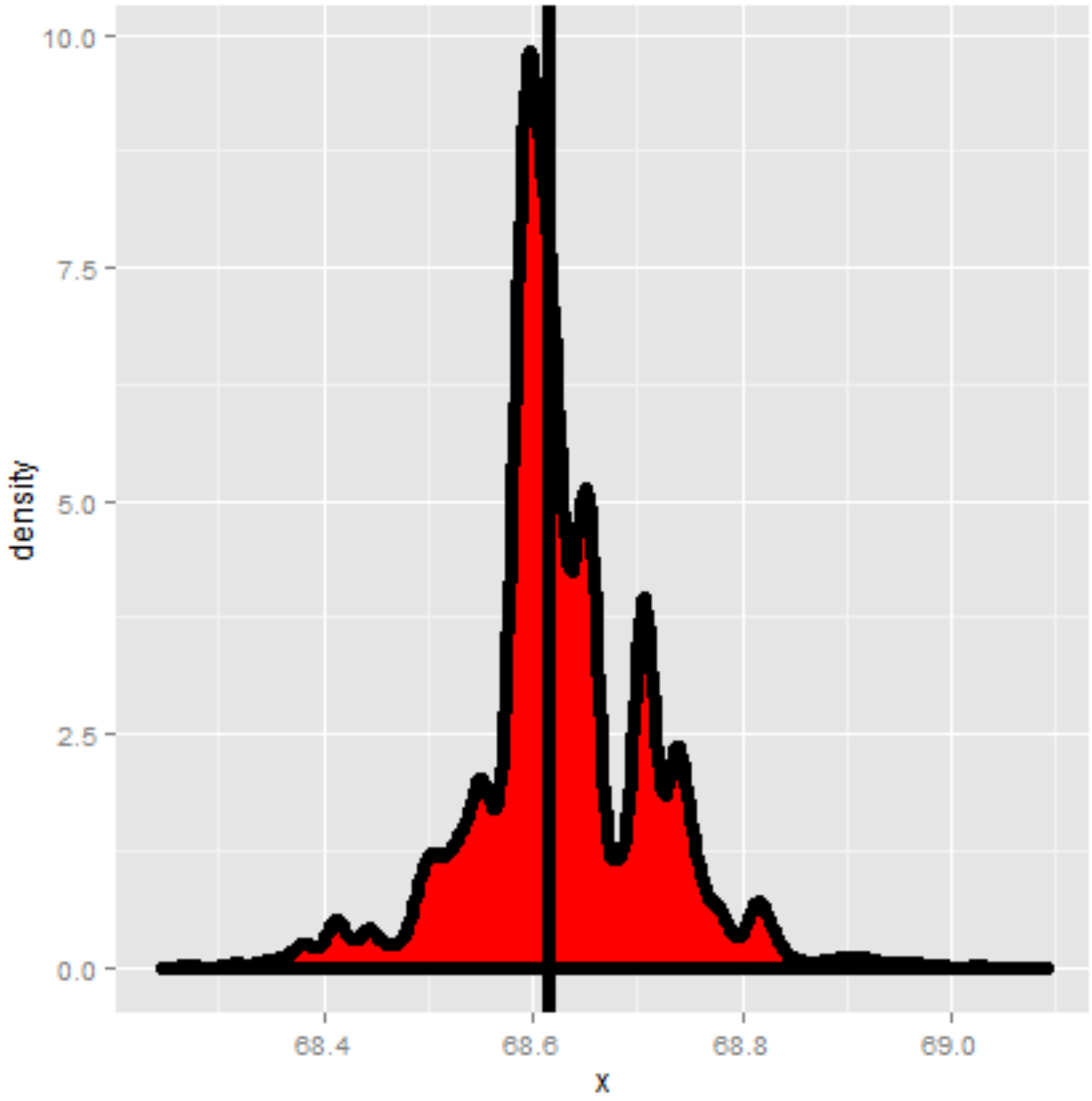
```
> sd(medians)
[1] 0.08424
```

Thus, 0.084 estimates the standard error of the median for this data set. It did this by repeatedly sampling medians from the observed distribution and taking the standard deviation of the resulting collection of medians. Taking the 2.5 and 97.5 percentiles gives us a bootstrap 95% confidence interval for the median.

```
> quantile(medians, c(.025, .975))
 2.5% 97.5%
68.43 68.81
```

We also always want to plot a histogram or density estimate of our simulated statistic.

```
g = ggplot(data.frame(medians = medians), aes(x = medians))
g = g + geom_histogram(color = "black", fill = "lightblue", binwidth = 0.05)
g
```



Bootstrapping example for the median of sons' heights from Galton's

Summary notes on the bootstrap

- The bootstrap is non-parametric.
- Better percentile bootstrap confidence intervals correct for bias.
- There are lots of variations on bootstrap procedures; the book An Introduction to the Bootstrap[4] by Efron and Tibshirani is a great place to start for both bootstrap and jackknife information.

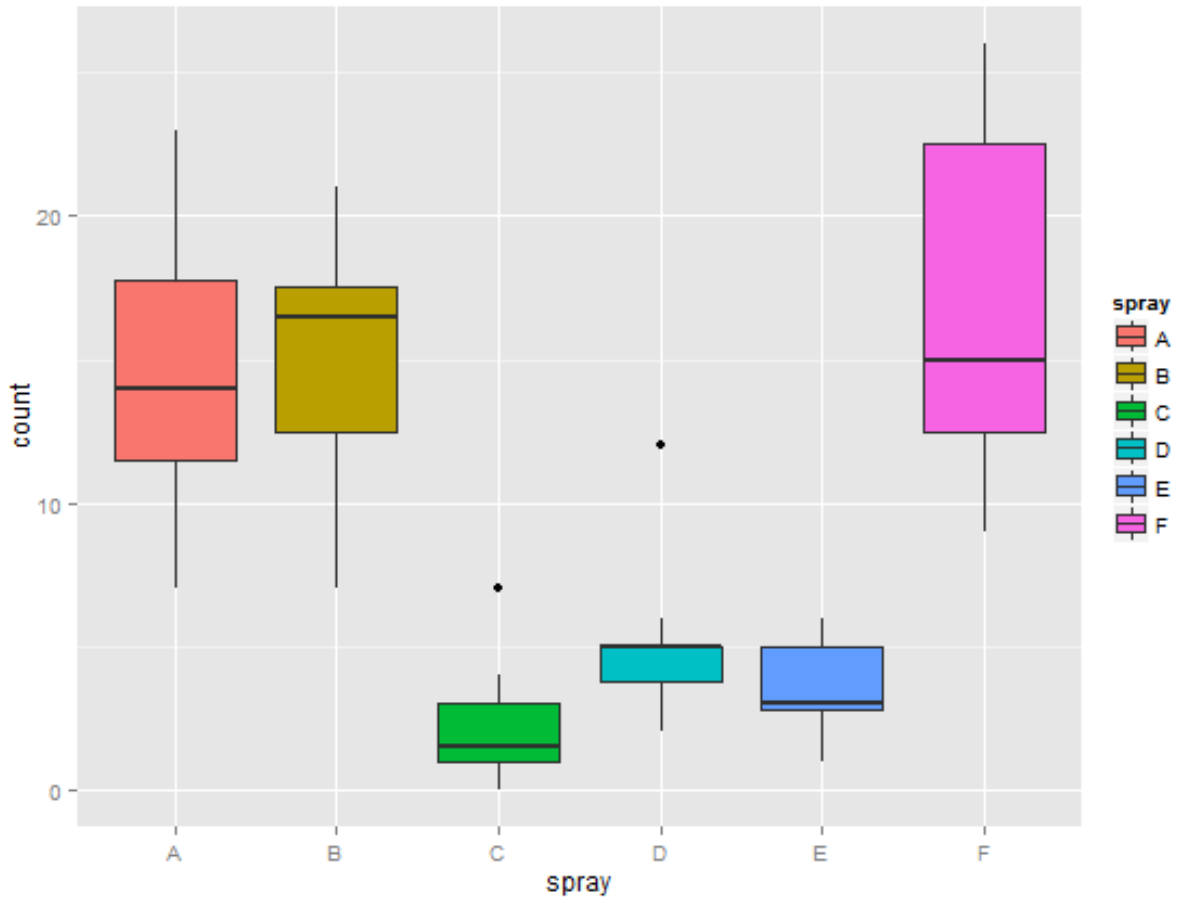
[4] http://www.crcpress.com/product/isbn/9780412042317

Group comparisons via permutation tests

Watch this video before beginning.[5]

Consider comparing two independent groups. Example, comparing sprays B and C.

```
data(InsectSprays)
g = ggplot(InsectSprays, aes(spray, count, fill = spray))
g = g + geom_boxplot()
g
```



Comparison of insect spray.

[5] http://youtu.be/nn1t9Kk7nn8?list=PLpl-gQkQivXiBmGyzLrUjzsblmQsLtkzJ

Permutation tests

Consider comparing means between the group. However, let's use the calculate the distribution of our statistic under a null hypothesis that the labels are irrelevant (exchangeable). This is a handy way to create a null distribution for our test statistic by simply permuting the labels over and over and seeing how extreme our data are with respect to this permuted distribution.

The procedure would be as follows:

1. consider a data from with count and spray,
2. permute the spray (group) labels,
3. recalculate the statistic (such as the difference in means),
4. calculate the percentage of simulations where the simulated statistic was more extreme (toward the alternative) than the observed.

Variations on permutation testing

This idea of exchangeability of the group labels is so powerful, that it's been reinvented several times in statistic. The table below gives three famous tests that are obtained by permuting group labels.

Data type	Statistic	Test name
Ranks	rank sum	rank sum test
Binary	hypergeometric prob	Fisher's exact test
Raw data		permutation test

Also, so-called *randomization tests* are exactly permutation tests, with a different motivation. In that case, think of the permutation test as replicating the random assignment over and over.

For matched or paired data, it wouldn't make sense to randomize the group labels, since that would break the association between the pairs. Instead, one can randomize the signs of the pairs. For data that has been replaced by ranks, you might of heard of this test before as the the signed rank test.

Again we won't cover more complex examples, but it should be said that permutation strategies work for regression as well by permuting a regressor of interest (though this needs to be done with care). These tests work very well in massively multivariate settings.

Permutation test B v C

Let's create some code for our example. Our statistic will be the difference in the means in each group.

Permutation distribution for the insect sprays dataset.

```
subdata <- InsectSprays[InsectSprays$spray %in% c("B", "C"),]
y <- subdata$count
group <- as.character(subdata$spray)
testStat <- function(w, g) mean(w[g == "B"]) - mean(w[g == "C"])
observedStat <- testStat(y, group)
permutations <- sapply(1 : 10000, function(i) testStat(y, sample(group)))
```

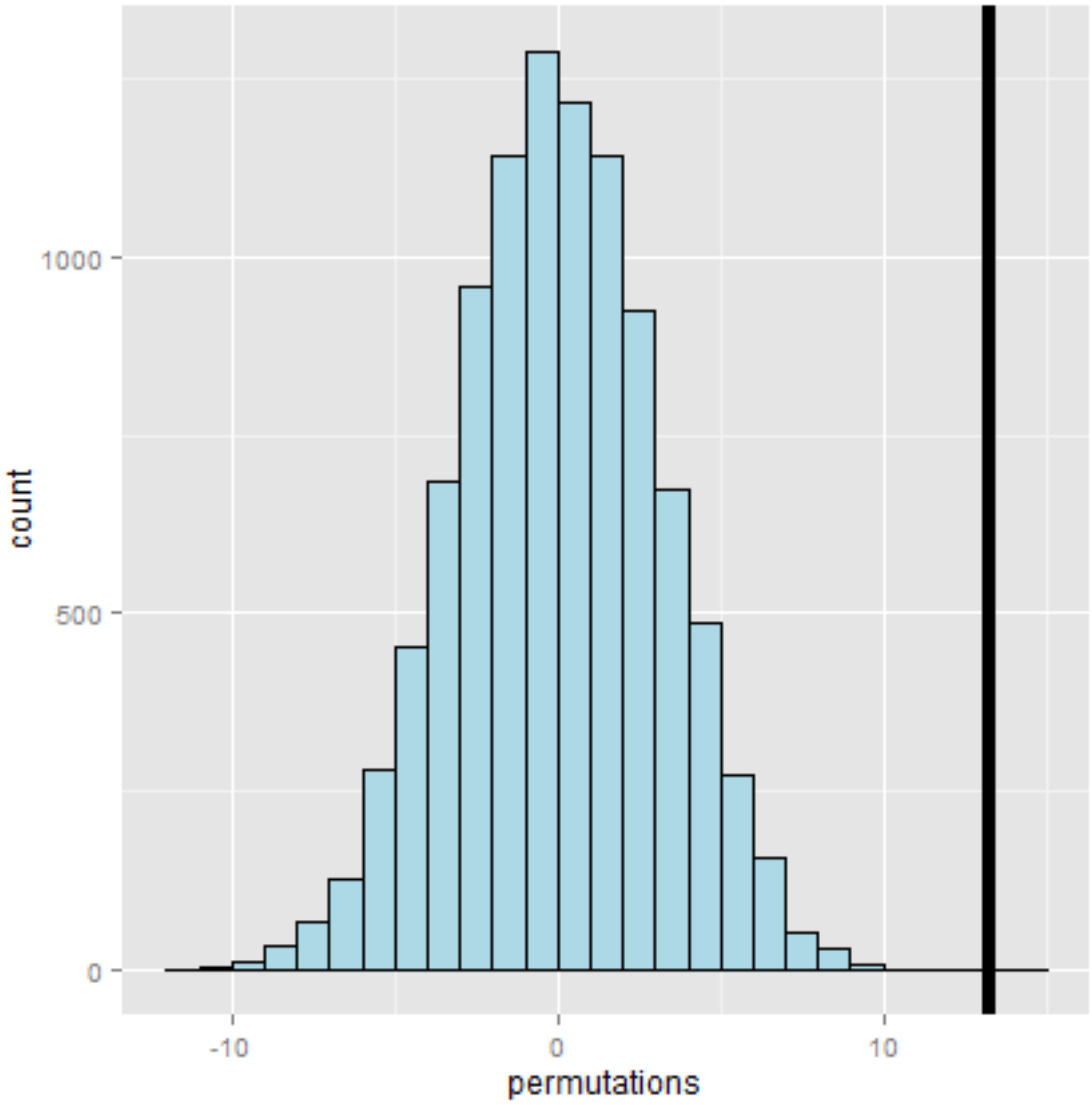
Let's look at some of the results. First let's look at the observed statistic.

```
> observedStat
[1] 13.25
```

Now let's see what proportion of times we got a simulated statistic larger than our observed statistic.

```
mean(permutations > observedStat)
[1] 0
```

Since this is 0, our estimate of the P-value is 0 (i.e. we strongly reject the NULL). It's useful to look at a histogram of permuted statistics with a vertical line drawn at the observed test statistic for reference.



Permutation distribution from the insectsprays dataset

Exercises

1. The bootstrap uses what to estimate the sampling distribution of a statistic?
    - The true population distribution
    - The empirical distribution that puts probability 1/n for each observed data point
2. When performing the bootstrap via Monte Carlo resampling for a data set of size n which is true? Assume that you're going to do 10,000 bootstrap resamples?
    - You sample n complete data sets of size 10,000 with replacement
    - You sample 10,000 complete data sets of size n without replacement

- You sample 10,000 complete data sets of size n with replacement
- You sample n complete data sets of size 10,000 without replacement

3. Permutation tests do what?
    - Creates a null distribution for a hypothesis test by permuting a predictor variable.
    - Creates a null distribution by resampling from the response with replacement.
    - Creates an alternative distribution by permuting group labels.
    - Creates confidence intervals by resampling with replacement.