About this book

This book is written as a companion book to the Statistical Inference Coursera class as part of the Data Science Specialization. However, if you do not take the class, the book 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 have 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.

Figure 1: Creative Commons

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 Chines pinball setup.

Introduction

Before beginning

This book is designed as a companion to the Statistical Inference Coursera class as part of the Data Science Specialization, 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, 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.

The entirety of the book is on github here. Please submit pull requests if you find errata!

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

Statistical inference defined

Statistical inference is the process of drawing formal conclusions from data.

In our class, we wil define formal statistical inference as settings where one wants to infer facts about a population using noisy statistical data where uncertainty must be accounted for.

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: is hormone replacement therapy effective?

A large clinical trial (the Women's Health Initiative) published results in 2002 that contradicted prior evidence on the efficacy of hormone replacement therapy for post menopausal women and suggested a negative impact of HRT for several key health outcomes. **Based on a statistically based protocol, the study was stopped early due an excess number of negative events.**

Here's there's two inferential problems.

1. Is HRT effective?
2. How long should we continue the trial in the presence of contrary evidence?

See WHI writing group paper JAMA 2002, Vol 288:321 - 333. for the paper and Steinkellner et al. Menopause 2012, Vol 19:616 621 for a discussion of the long term impacts

Motivating example

Brain activation

Summary

- 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 uknown 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.

• Example goals of inference

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 polution levels, will asthma rates decline?")
5. Talk about the probability that something occurs.

• Example tools of the trade

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 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 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.

- **In this class**

- In this class, we will primarily focus on basic sampling models, basic probability models and frequency style analyses to create standard inferences.

- 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.

Where to learn more on the topics not covered

1. Explicit use of random sampling in inferences: look in references on "finite population statistics". Used heavily in polling and sample surveys.
2. Explicit use of randomization in inferences: look in references on "causal inference" especially in clinical trials.
3. Bayesian probability and Bayesian statistics: look for basic itroductory books (there are many).
4. Missing data: well covered in biostatistics and econometric references; look for references to "multiple imputation", a popular tool for addressing missing data.
5. Study design: consider looking in the subject matter area that you are interested in; some examples with rich histories in design:
6. The epidemiological literature is very focused on using study design to investigate public health.
7. The classical development of study design in agriculture broadly covers design and design principles.
8. The industrial quality control literature covers design thoroughly.

Probability

Watch this video before beginning.

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 philosphical 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 occuring without apparent deterministic paterns. 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. Probabilitity for us will be the long 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. In addition, there's the actual Coursera course that I run periodically (this is the first Coursera class that I ever taught). In addition there are a set of (notes on github)[http://github.com/bcaffo/Caffo-Coursera]. Finally, there's a followup class, uninspiringly named Mathematical Biostatistics Boot Camp 2, that is more devoted to biostatistical topics that has an associated YouTube playlist, Coursera Class and GitHub notes.

Kolmogorov's Three Rules

Watch this lecture before beginning

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, 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 Kolmorogov already did it for us). Instead here's a list of some of the consequences of Kolmogorov's rules that are often useful.

- The probability that nothing occurs is 0 The probability that something

- occurs is 1 The probability of something is 1 minus the probability that the
- opposite occurs The probability of at least one of two (or more) things
- that can not simultaneously occur (mutually exclusive) is the sum of
- their respective probabilities 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) 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

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 in the same sense 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 variable 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 line 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'll use these examples a lot to help us build intuition. However, they aren't interesting in the sense of lacking any context. 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

$$\theta$$

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

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 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.

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

{title='Code for plotting the density', line-numbers=off,lang=r} ~ x <- c(-0.5, 0, 1, 1, 1.5) y <- c(0, 0, 2, 0, 0) plot(x, y, lwd = 3,frame = FALSE, type = "l") ~

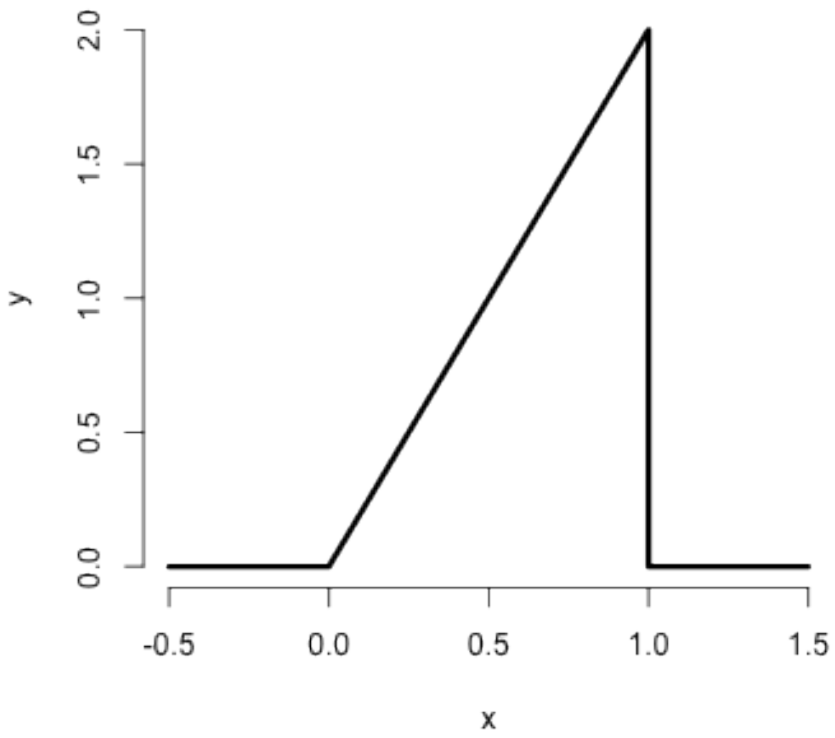


Figure 2: Help call density

The result of the code is given below.

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.

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.

{line-numbers=off,lang=r} ~~ > 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.

{line-numbers=off,lang=r} ~~

```
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

$$x$$

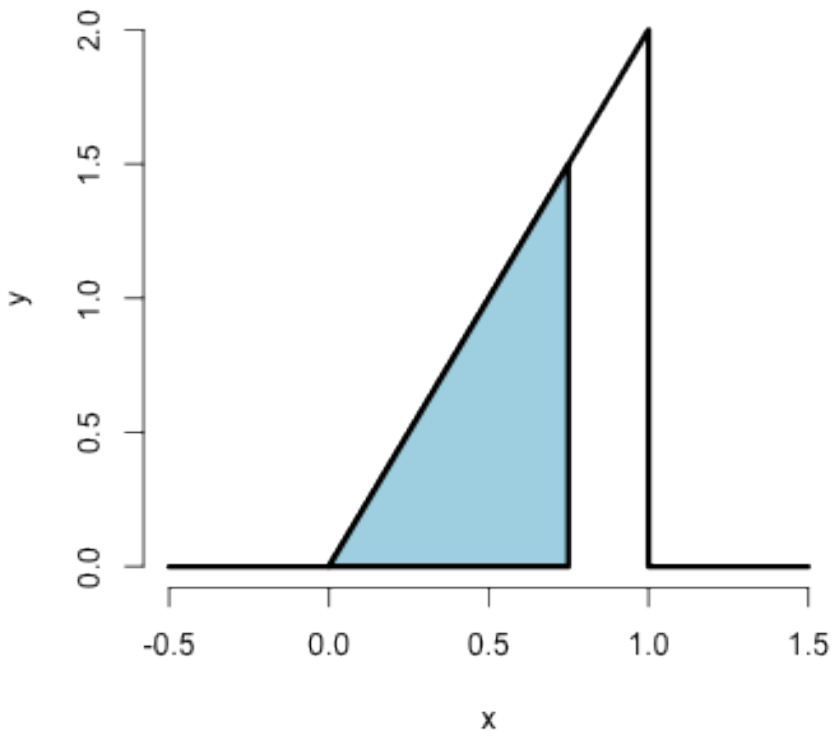


Figure 3: Help call density

. 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

$$F(.6)$$

{line-numbers=off,lang=r} ~ > 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

$$\alpha^{th}$$

quantile of a distribution with distribution function

$$F$$

is the point

$$x_\alpha$$

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

$$\alpha$$

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

{line-numbers=off,lang=r} ~ > 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

{line-numbers=off,lang=r} ~ > qbeta(0.5, 2, 1)

[1] 0.7071 ~

Exercises

Conditional probability

Conditional probability, motivation

Watch this video before beginning

Conditioning 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*, (disccussed 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.

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

Baye's rule is a famous result in statistics and probability. It forms the foundation for large branches of statistical thinking. Baye's 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?".

Baye's rule allows us to switch the conditioning event, provided a little bit of extra information. Formally Baye's 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 Baye's rule in practice, let's go over them in greater detail. (In addition, understanding Baye's 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,

$$P(+ \mid D)$$

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 pre test 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

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:

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

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 are 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

Expected values

Watch this video before beginning

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. 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 of a distribution. Specifically,

$$E[X]$$

represents the center of mass of a collection of locations and weights,

$$\{x, p(x)\}$$

.

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

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

.

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 the data

{title="Loading in and displaying the Galton data", line-numbers=off,lang=r} ~ 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.., fill = variable), binwidth=1, colour = "black") + geom_density(size = 2) g <- g + facet_grid(. ~ variable) g ~

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.

{title="Using manipulate to explore the mean", line-numbers=off,lang=r} ~ library(manipulate) myHist <- function(mu){ g <- ggplot(galton, aes(x = child)) g <- g + geom_histogram(fill = "salmon", binwidth=1, aes(y = ..density..), colour = "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)) ~

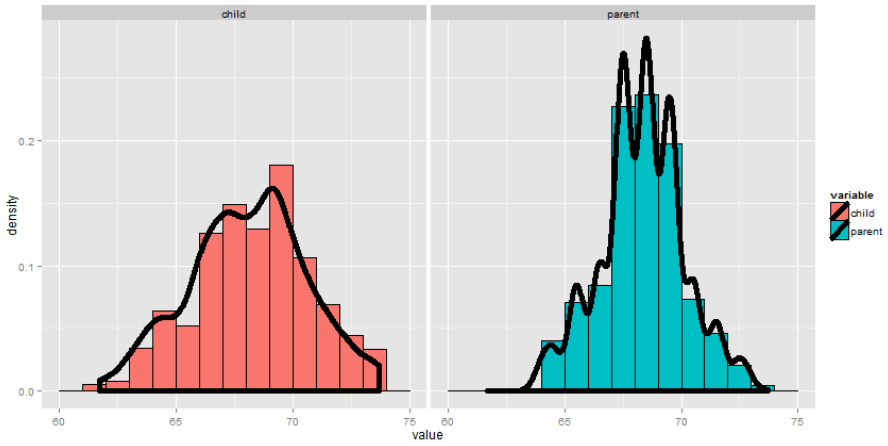


Figure 4: Galton's Data

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

Example of a population mean, a fair coin

Watch the video before beginning here.

Suppose a coin is flipped and

$$X$$

is declared 0 or 1 corresponding to a head or a tail, respectively. 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.

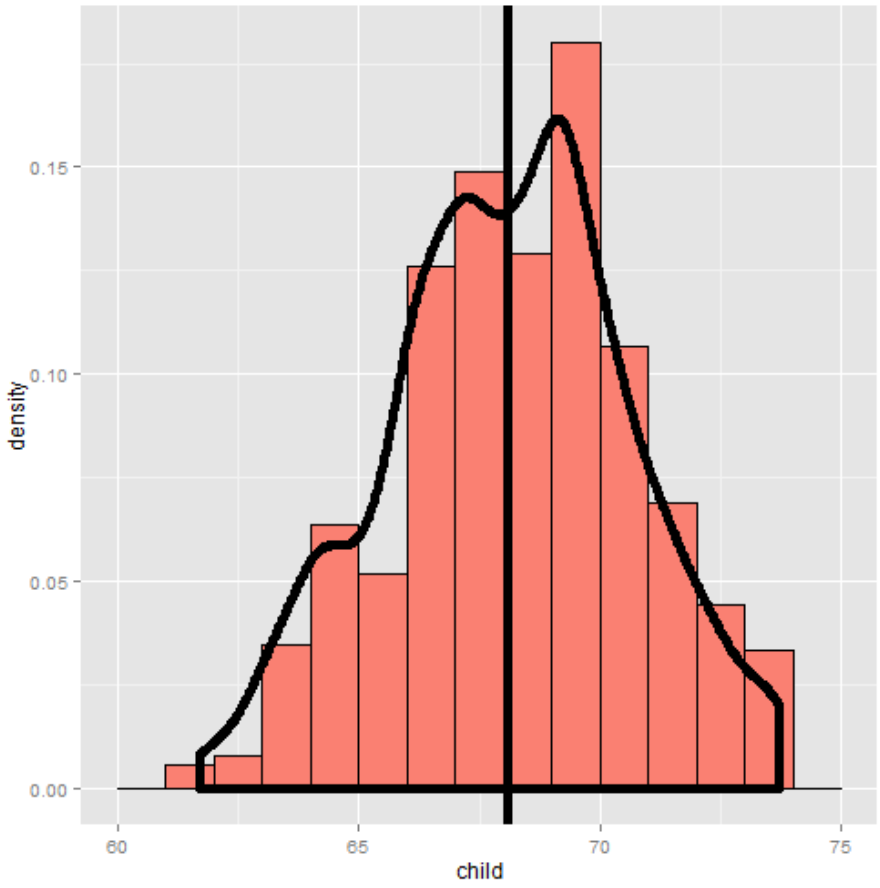


Figure 5: Histogram illustration

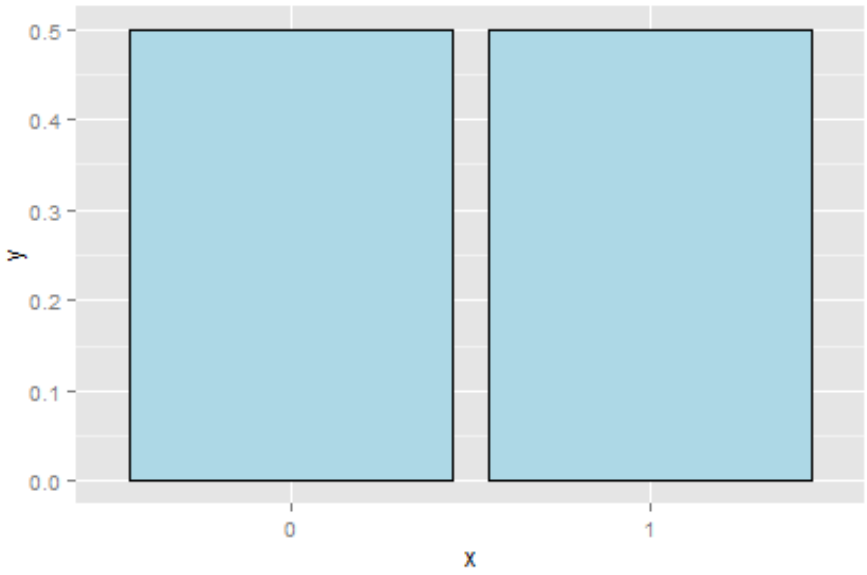


Figure 6: Fair coin mass function

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.

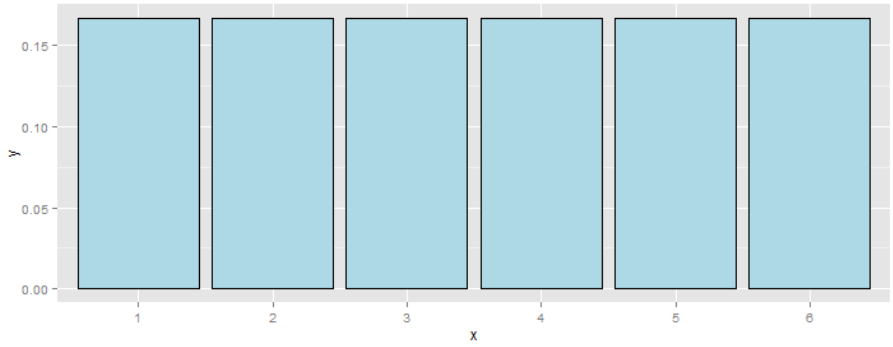


Figure 7: Bar graph of die probabilities

Continuous random variables

Watch this video before beginning

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?

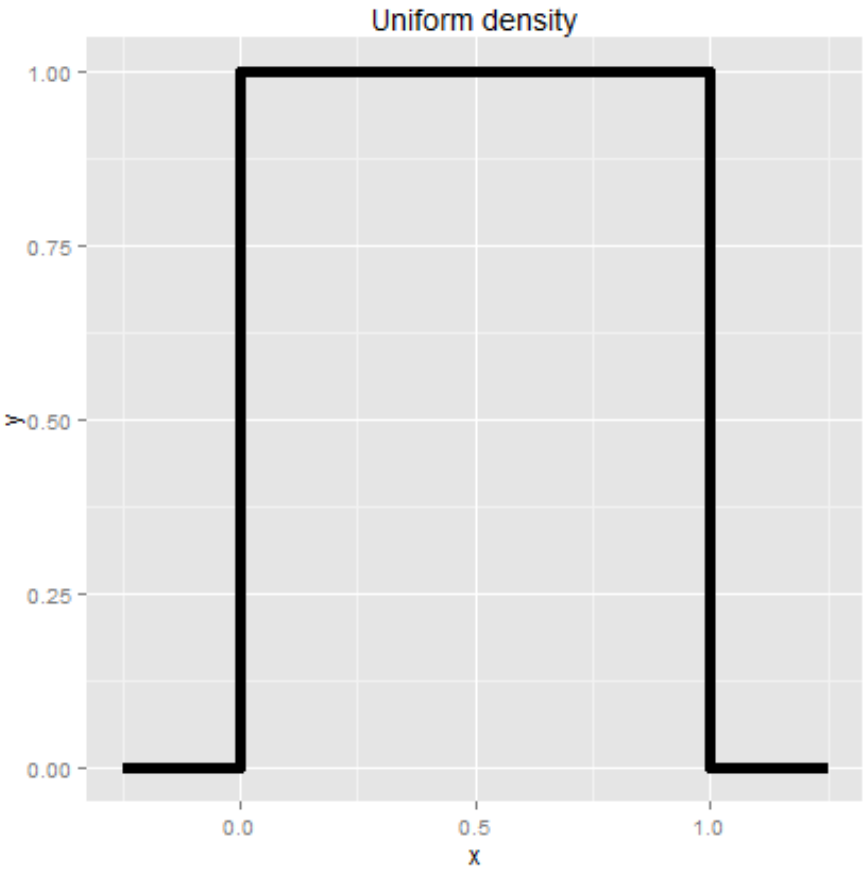


Figure 8: Uniform Density

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.

Interestingly, the average of ten randomly sampled people's height is itself of random variable, in the same way that the average of ten die roll 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.

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).

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 concetrated around 3.5.

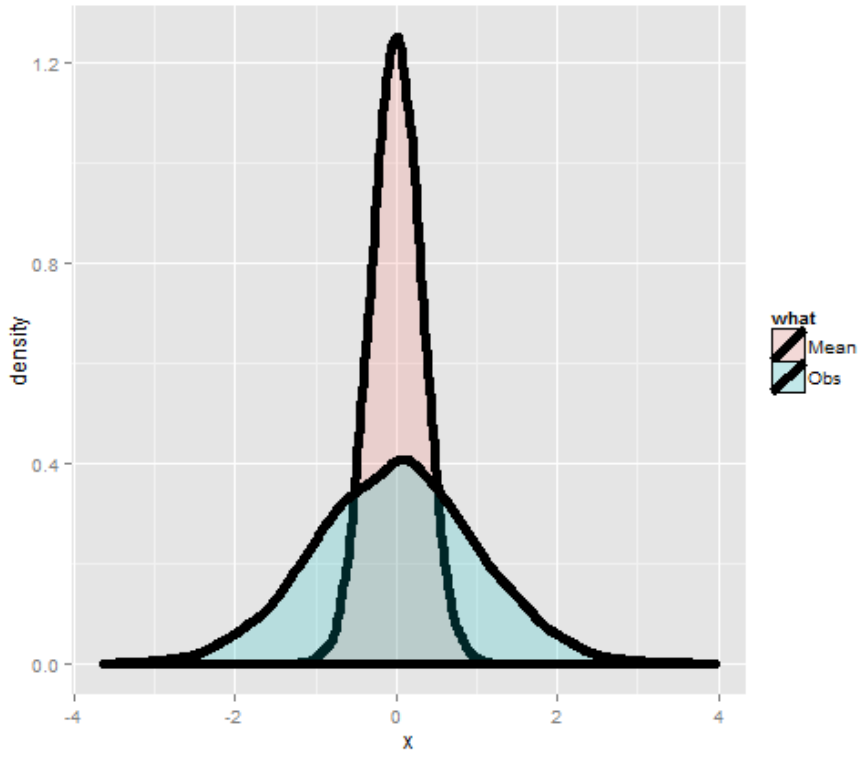


Figure 9: Simulation of normals

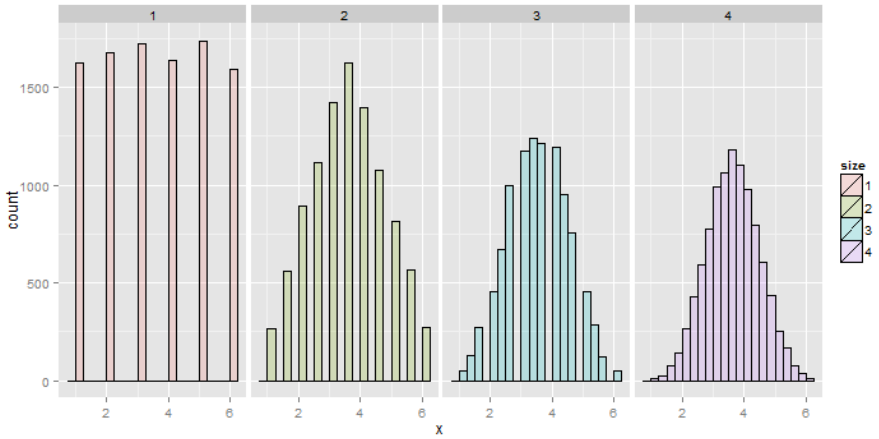


Figure 10: Simulation of die rolls

Averages of x coin flips

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

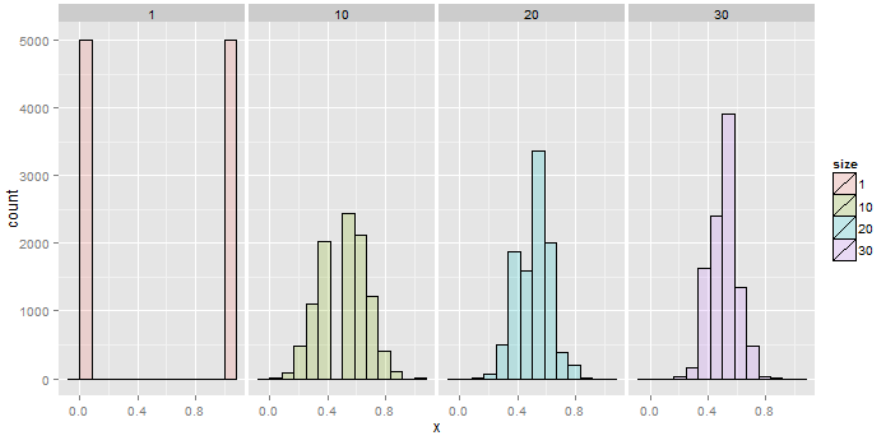


Figure 11: Simulation of coin flips

Summary

- 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

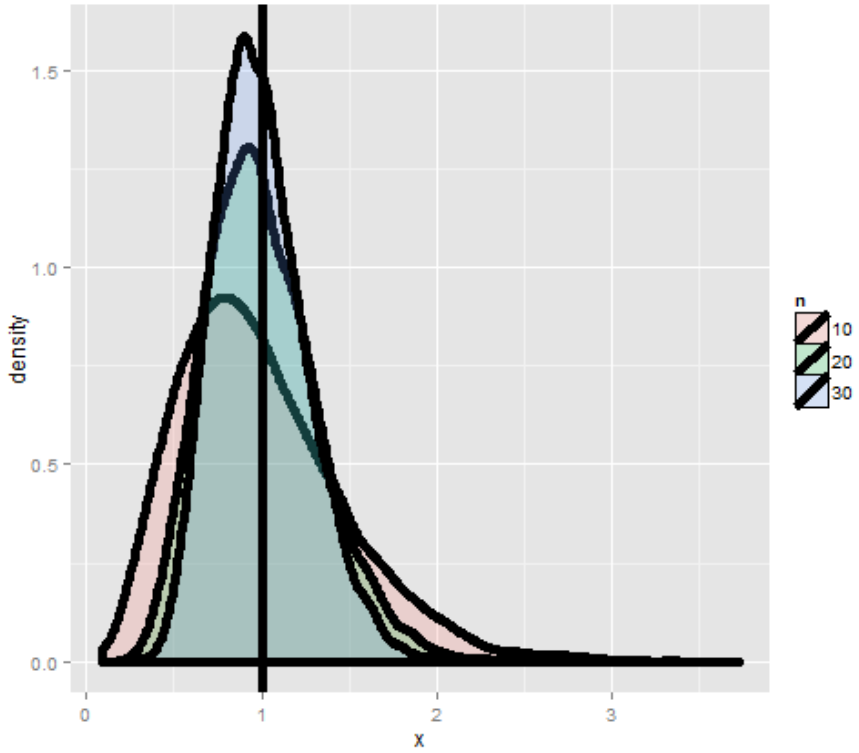
Exercises

Variation

Watch this video before beginning.

The variance

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.



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 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

$$\mu$$

, 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 idea. 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

$$p = 0.5$$

.

{title="Plotting the binomial variance", line-numbers=off,lang=r} ~ p = seq(0 , 1, length = 1000) y = p * (1 - p) plot(p, y, type = "l", lwd = 3, frame = FALSE) ~

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.

The sample variance is almost, but not quite, the average squared deviation from the sample mean since we divide by

$$n-1$$

instead of

$$n$$

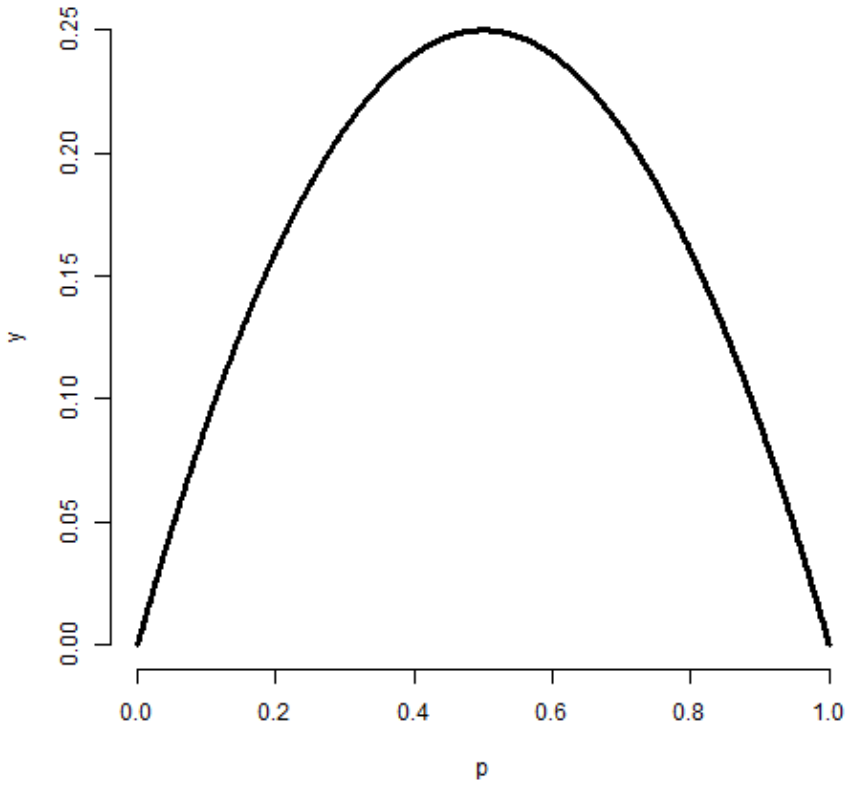


Figure 12: Plot of the binomial variance

. 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

$$n$$

.

Moreover, as we collect more data, the distribution of the sample variance gets more concentrated around the population variance that it's estimating.

Simulation experiments

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.

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.

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, the 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 in a sample, when we only get to observe one realization. Recall that the average of random sample from a population is itself a random

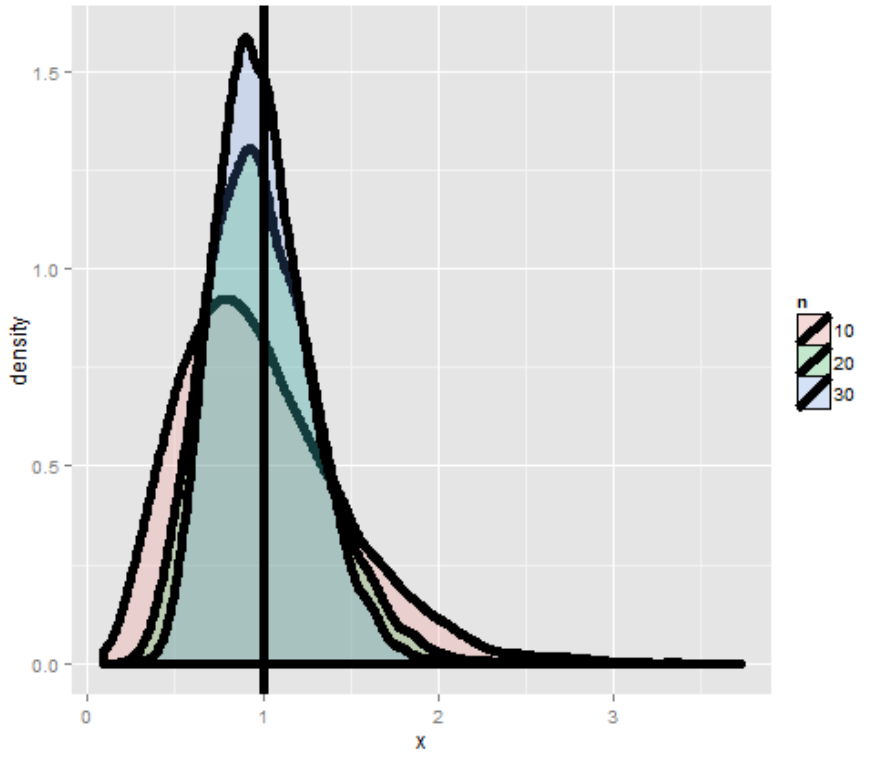


Figure 13: Simulation of variances of samples of standard normals

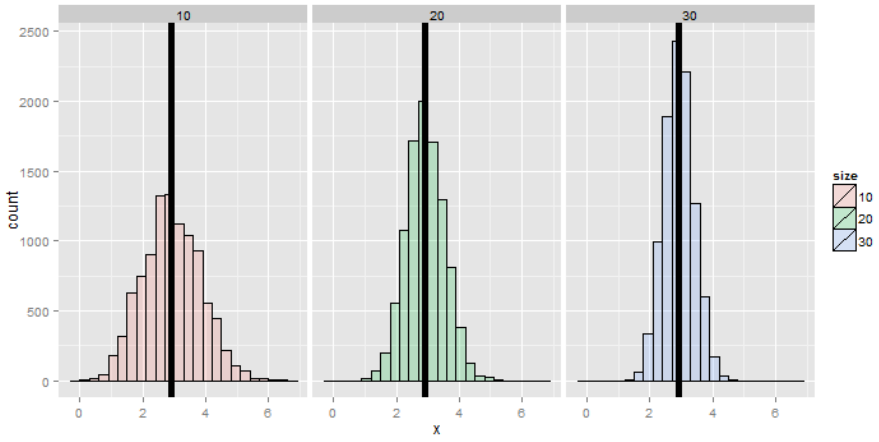


Figure 14: Simulated distributions of variances of dies

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

$$\sigma^2$$

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

$$\sigma^2$$

via the sample variance. So, we can get a good estimate of the variability of the mean, even though we only get to observer 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.

Summing up

- The sample variance,

  $$S^2$$

  , estimates the population variance,

  $$\sigma^2$$

- The distribution of the sample variance is centered around

  $$\sigma^2$$

- The variance of the sample mean is

  $$\sigma^2/n$$

- 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

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

{title="Simulating means of random normals", line-numbers=off,lang=r} ~ > 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 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.

{title="Simulating means of uniforms", line-numbers=off,lang=r} ~ > 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) 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.

{title="Simulating means of Poisson variates", line-numbers=off,lang=r} ~ > 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 variance 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

$$1/(2\sqrt{n})$$

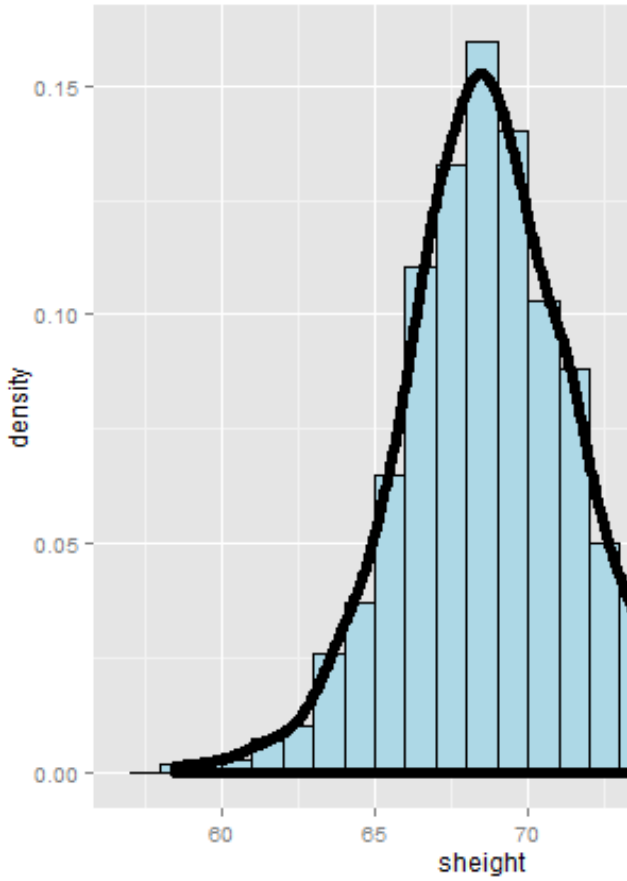
. Let's try it.

{title="Simulating means of coin flips", line-numbers=off,lang=r} ~ > 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

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.

{title="Loading the data", line-numbers=off,lang=r} ~ library(UsingR); data(father.son); x <- father.son$sheight n<-length(x) ~



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

{title="Loading the data", line-numbers=off,lang=r} ~ >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 variances of the sons' heights. Therefore, 7.92 tells us exactly how variable sons' heights were in the data and (under the assumption that these sons are a random sample from the population) 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.

Summarizing what we know about variances

- 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