

Political Science 15
Introduction to Research in Political Science
Lecture 3a: Introduction to Probability

Alice Lépiessier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Goals for this week

Learning outcomes

① Probability

- Distributions and thinking probabilistically
- Simulating the distribution of the mean
- Random variables, expectation, and variance
- Introducing 2 awesome theorems:

Law of Large Numbers


Central Limit Theorem

② Introducing bivariate regression (OLS)

Your to-do list for Week 2

- Read *Real Stats* Appendix A-G (for the Monday-Tuesday lectures)
- Read *Real Stats* Chapter 3 (for the Wednesday-Thursday lectures)
- Read a cautionary tale in *Real Stats* Chapter 2 (i.e. why we train you to have good data science practices!)
- Hand in Problem Set 2 (due Friday before midnight)

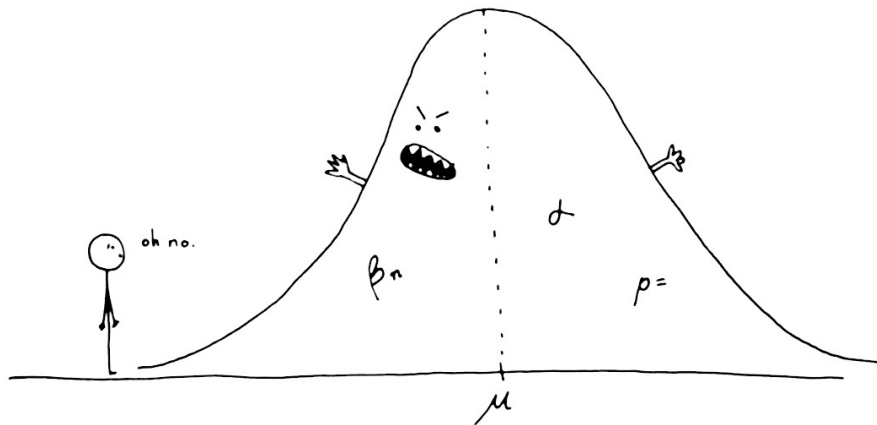
How to approach this week

 There is math!

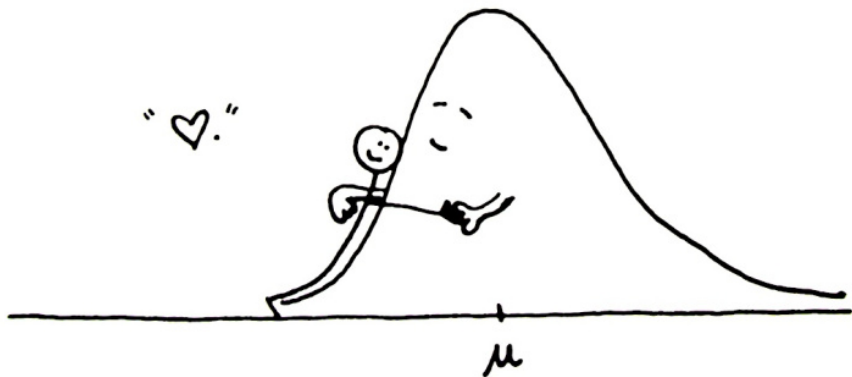
The math is here to help. It allows us to write down our ideas about probability in a simpler and more precise way.

Do not be put off by the symbols! The **trick is to read slow**, and come back to the material often. Do not skip the math - work through it.

Maybe you feel like this now



But hopefully you will feel like this soon



Thinking probabilistically

Today we'll start building a more theoretical foundation that we need in order to go farther.

One of the hardest things about statistics is understanding **distributions** of outcomes: the idea that **you could observe different outcomes with different probabilities, even when you only observe one outcome.**

Statistics has to do with understanding properties of that distribution of potential outcomes, not just the data you have.

Today we'll work on some **probability theory** and repeatedly think about distributions of potential outcomes.

Key idea

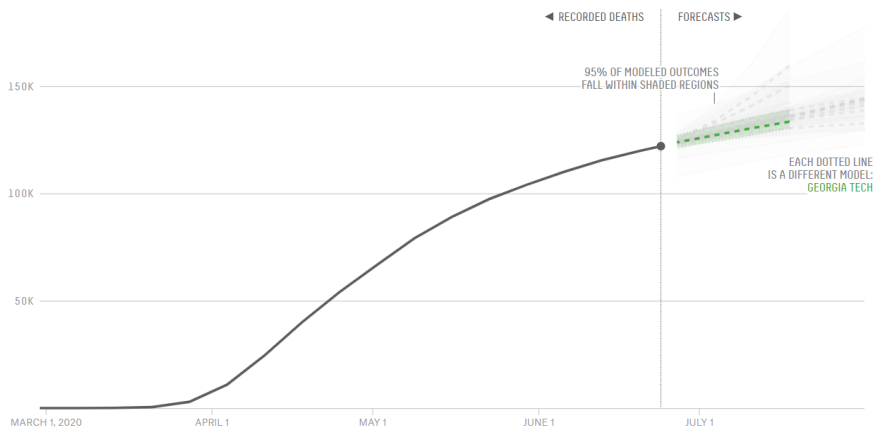
Thinking probabilistically means thinking not just about the 1 outcome that you observed, but about what different outcomes you *might* have observed and how likely those would have been.

Thinking probabilistically

One way to think about it is to think of the **multiverse**.

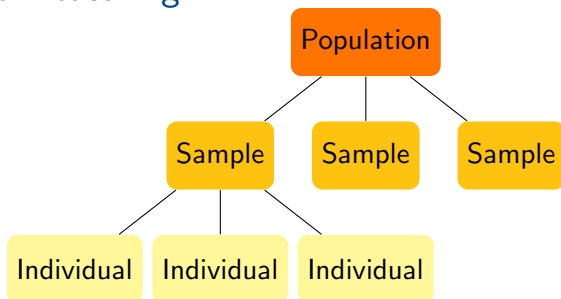
Say you observe one outcome (e.g. a candidate won with 60% of the vote). This is the *realized outcome*. But other outcomes could have been possible in a parallel universe. Probability theory allows us to think *how likely* those outcomes would have been to occur instead.

Modelling COVID-19 deaths with uncertainty



Source: <https://projects.fivethirtyeight.com/covid-forecasts/>

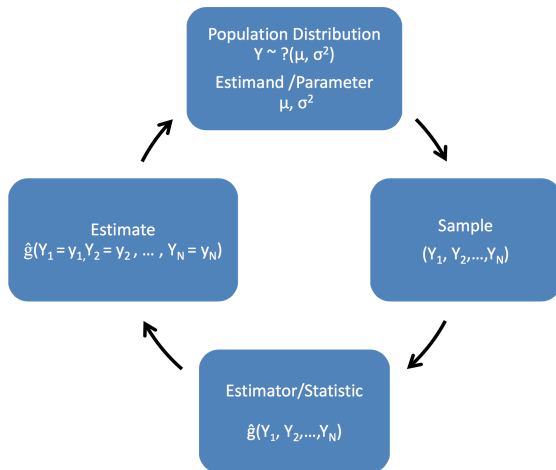
Who are we measuring?



- **Population:** the collection of individuals, beyond just the sample, for which we would like to understand patterns/trends.
- **Sample:** the collection of individuals on which statistical analyses are performed in order to **infer general trends for the population**.
- **Individual:** also called “observation” or “unit”, a single data point contributing to the sample (i.e. the subscripted i in the regression equation).

The Statistical Inference Process

This is the key workflow when doing statistical analysis.
(Should make sense at the end of the lecture).



Building Blocks: Random Variables

Random variables are a fundamental building block of probability theory.

Informal definition

A random variable is a variable whose values depend on outcomes of a random process.

Examples: flip of a coin, neck length of a random giraffe in Santa Barbara zoo, height of random UCSB student, etc.

Two types of random variables:

- 1 **Discrete:** can take on distinct or separate values, e.g. {heads, tail}, {1,2,3,4,5,6}
- 2 **Continuous:** can take on any value in an interval

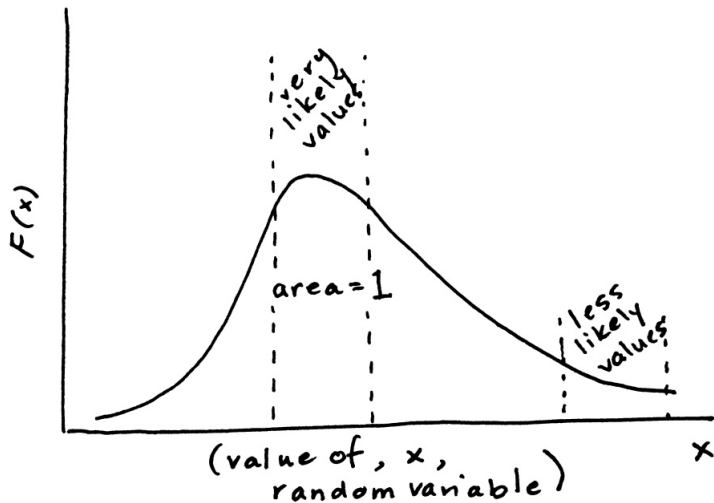
Building Blocks: Random Variables

For any random experiment, the thing we are measuring is the **random variable**.

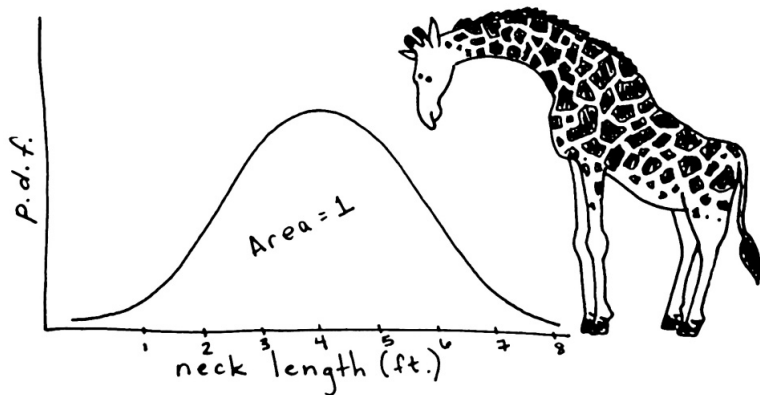
Properties of random variables

- They can take on **different possible values** as the result of a random experiment, e.g. discrete RV denoting whether race was won: $\{0, 1\}$; continuous RV denoting giraffe neck lengths: $[1, 8]$.
- Those outcomes have different probabilities of occurring. These are characterized by a **distribution**.
- You won't know which of those values the random variable will take until you do the random experiment (also called random draw).

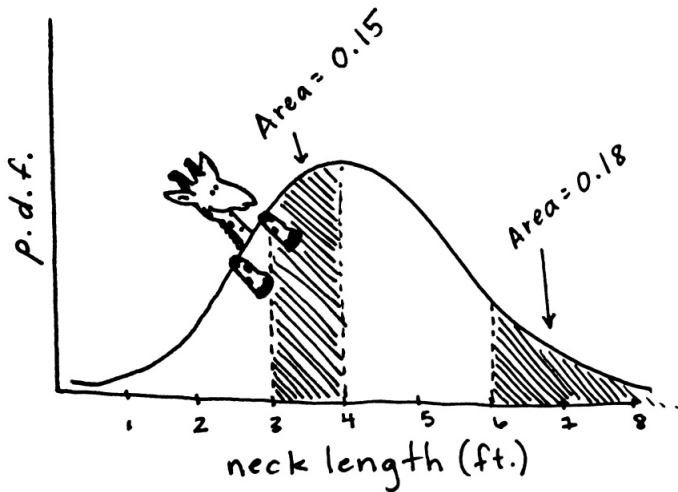
Reading a Probability Density Function



Reading a Probability Density Function

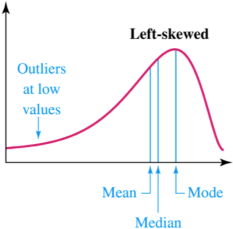


Reading a Probability Density Function

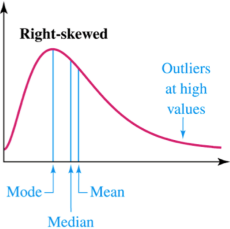


Measures of centrality

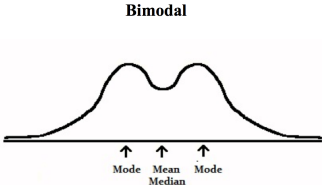
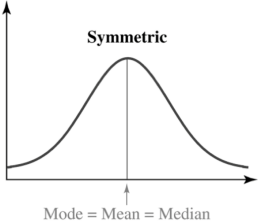
Measures of **central tendency** such as the mean, median, and mode tell us what the “typical” outcome is.



(a)

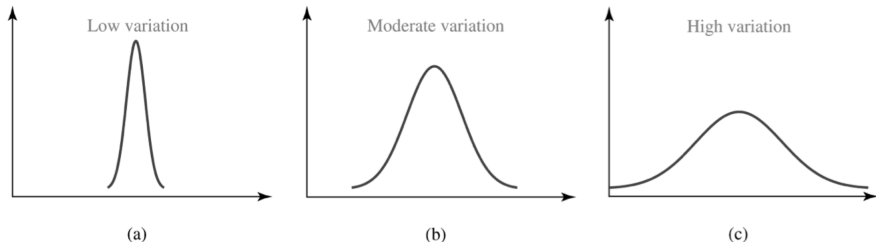


(b)



Measures of dispersion

Measures of **dispersion** such as the variance and the standard deviation tell us how “scattered” our data is, and how far away from the “typical” outcome some outcomes will lie.



REMEMBER THIS

1. A useful first step toward understanding data is to review sample size, mean, standard deviation, and minimum and maximum for each variable.
2. Plotting data is useful for identifying patterns and anomalies in data.

Political Science 15
Introduction to Research in Political Science
Lecture 3b: Distribution of the Mean

Alice Lépiessier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Goal: Understanding the ‘distribution of the mean’

Two related goals:

- **Goal 1:** Understand what is meant by the “distribution of the mean”.
- **Goal 2:** Understand how to figure out the distribution of a mean in theory, given you only have a *sample* of *individuals* and can’t actually see the *distribution of the mean*.

The first part we can see conceptually from simulations and graphs.

The second part will require some statistical theory.

Goal 1: Distribution of the mean, using simulation

Here we explore the concept of a distribution of the mean using **simulation**.

Suppose you can draw (sample) some quantity as many times as you want:

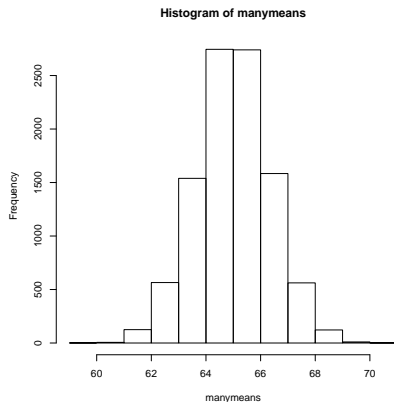
- Example: You gather data on the height of people in this class, pretending you never run out of people. All the people are from the general UCSB population.
- Take the mean (average) of, say, 20 peoples' height in inches.
- Do that over and over again.

Sample code:

```
getmean = function( N, mean, sd ){  
  height = rnorm( n = N, mean = mean, sd = sd )  
  meanheight = mean( height )  
  return( meanheight )  
}
```

```
manymeans = replicate( n = 10^4, getmean(N = 20, mean = 65, sd = 6) )  
hist(manymeans)
```

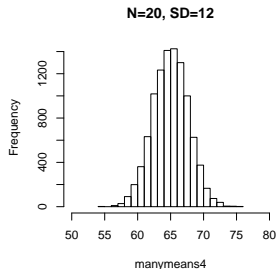
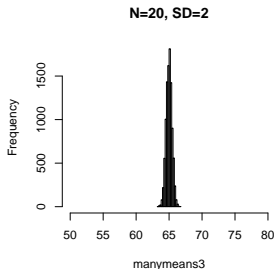
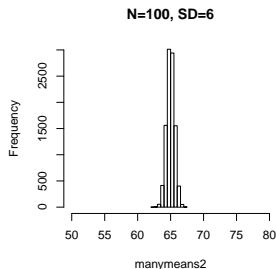
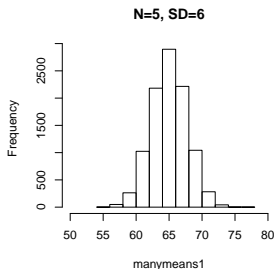
Goal 1: Distribution of the mean, using simulation



How do you think this picture changes if we change the number of people we are averaging together ($N = 5, 20$ or 100)?

How do you think this changes if the natural variation of height (sd) in the population was much larger or much smaller ($sd = 2, 6,$ or 12)?

Goal 1: Distribution of the mean, using simulation



What to note about distribution of the mean

Remember, the mean you get from a given sample is one number, but we're talking about the distribution governing what it *could have* looked like.

When working with *actual data*, you only see the mean once, and you can't do this simulation exercise directly. (You can't collect the data over and over again – you only have one sample in the data you have).

But, it turns out that the *only* thing the distribution of the mean depends on is the **sample size** N , the **mean** (\bar{X}), and the **variance** (SD^2) of the individuals.

- This will allow us to construct **null distributions** involving the mean very easily without simulation (more on that later in the course).
- It will also give us **confidence intervals** telling us how different the mean could have been if we had a slightly different sample (because of noise).

Come back to this slide at the end and see if it makes sense to you!

Goal 2: Distribution of the mean, using theory

We just simulated empirically what the distribution of the mean would look like. It looks pretty normal. Now, can we prove this will always be the case using **statistical theory**?

Recall the concept of **random variable**.

For any random experiment/process, the random variable will take on the values depending on the outcome of the random process. It quantifies the outcome for a random process.

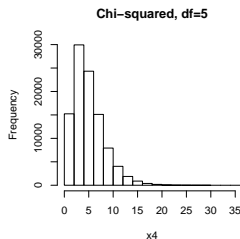
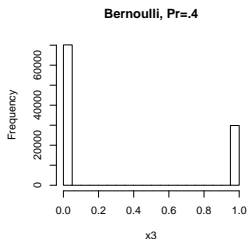
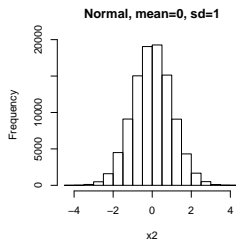
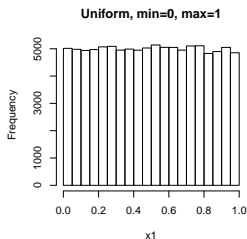
Random variables are usually denoted by capital letters. Examples:

- X = height of a random UCSB student
- Y = outcome of a coin flip
- Z = sum of the roll of 15 dices

Random variables have probability distributions which give the probabilities for the different possible outcomes of the RV.

Examples of Probability Distributions

Distributions are just ways of describing how often you get each possible value. Some examples of how variables look when drawn from 4 different distributions:



Expectation and Variance

Remember that random variables are measurements drawn from a distribution, whether we know that distribution or not.

Two important features of any distributions are **expectation** and **variance**.

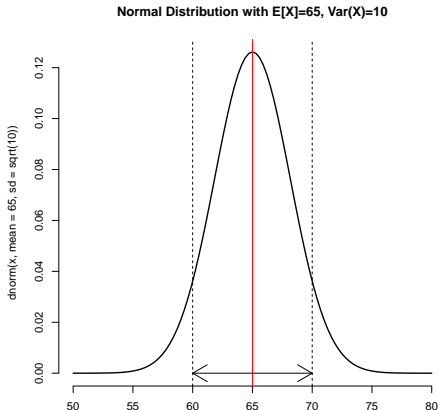
- **Expectation**: the best guess about what number will be drawn from the distribution.
- **Variance**: how far the numbers you draw tend to be from that best guess.

Turns out, you can **characterize a normal distribution** entirely from its expectation and its variance. This means that if you know the expectation and the variance, then you can draw that normal distribution!

A Normally Distributed Variable

Say we are interested in the height of the next person to watch this video module after you.

We will assume this random variable is distributed normally with expectation 65 inches and variance of 10 inches:



A bit more on Expectation and Variance

The **expectation**, or $\mathbb{E}[X]$ for a random variable, X :

- is a different concept from the average or mean, because it is something that exists for any random variable, not something you get from the data
- but, you can think of it loosely as the “average” of what you could get from the distribution
- if you could draw repeatedly from a distribution and take an average, it would get closer and closer to the expectation

A bit more on Expectation and Variance

Variance is a measure of spread, or how far you expect a random draw to be from the expectation:

- Technically: $\mathbb{E}[(X - \mathbb{E}[X])^2]$
- Informally: if you drew a bunch of numbers from same distribution, squared the difference from each to the mean, and averaged those.

We will talk about how to *estimate* $\mathbb{E}[X]$ and $\text{Var}(X)$ using a *sample* of values drawn repeatedly from the same distribution, but we delay this here to emphasize that the $\mathbb{E}[X]$ and $\text{Var}(X)$ are **properties of the distribution of X** , not of your data.

Final words on normal distributions

We'll soon see why the normal distribution appears so often.

Remember that a normal distribution is completely described by *only* its expectation and variance.

For some random variable X from a normal distribution, we write:

$$X \sim N(\mathbb{E}[X], \text{Var}(X))$$

Actually, we usually write

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

where μ is the expectation and σ^2 is the variance. Get used to this!

Political Science 15

Introduction to Research in Political Science

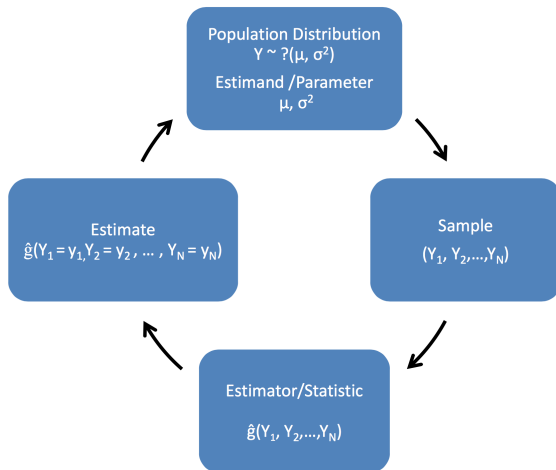
Lecture 3c: Sample Mean and Law of Large Numbers

Alice Lépiessier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

The Statistical Inference Process

Recall what we are trying to accomplish here.
(Should make sense at the end of the lecture).



Samples as Collections of Random Variables

Typically we have a *sample*: multiple RVs all with the same distribution.
E.g.

- person 1 has $height_1$, a RV drawn from $N(\mu, \sigma^2)$
- person 2 has $height_2$, a separate RV drawn from $N(\mu, \sigma^2)$
- ...
- person N has $height_N$, yet another RV drawn from $N(\mu, \sigma^2)$

Even though each observation is a different RV, since they all come from the same distribution, we leverage the sample to learn something about that distribution.

The Sample Mean

The sample mean of X_1, X_2, \dots, X_N is:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i = \frac{1}{N} [X_1 + X_2 + \dots + X_N]$$

Important notes:

- This is an example of an **estimator**: a function you can compute with data. You can use these estimators in R, for example `mean(data$variable)`.
- **The mean is itself a random variable!** It takes on some particular value in your study, but it could have taken different values.
- We are trying to understand what the *distribution of the mean* must look like, despite seeing the mean only once.

Goal 2: Distribution of the mean, using theory

Recall that the mean has a distribution

- you saw this notionally using simulations...
- but in the real world, you can't draw a bunch more data to get new means as we did in the simulation
- turns out that mathematically we can say a lot about the distribution of the mean despite only seeing it once!

The Law of Large Numbers

How can we use a sample of random variables drawn from the same distribution to learn about that distribution?

We leverage this theorem.

The Law of Large Numbers (LLN)

For RVs X_1, X_2, \dots, X_N , the mean \bar{X} gets closer and closer to $\mathbb{E}[X]$ as N grows.

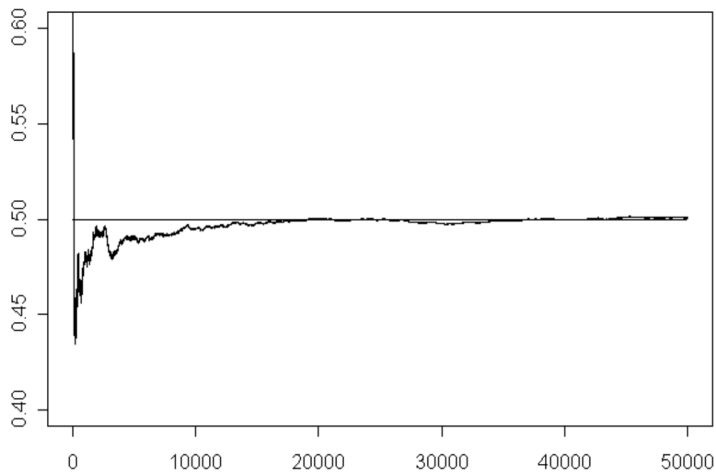
That is,

- As N becomes larger and larger (formally $N \rightarrow \infty$),
- then $\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \rightarrow \mathbb{E}[X]$
- works even if distribution of X is not normal!

Bottom line: while we make a distinction between $\mathbb{E}[X]$ and the mean from some sample, because of the law of large numbers, the latter is a good estimate of the former, and gets better as N grows. This is why we call the mean an **estimator**.

The Law of Large Numbers - Graphically

Law of Large Numbers for Tosses of a Fair Coin



Properties of the Distribution of the Mean

Property 1: The sample mean's distribution is centered around $\mathbb{E}[X]$.

- You only get a mean once, but you know it takes a value from a distribution centered on $\mathbb{E}[X]$.
- We don't know $\mathbb{E}[X]$ (though we know \bar{X} gets closer and closer as N grows).
- A taste of the math (don't worry about this)

$$\begin{aligned}\mathbb{E}[\bar{X}] &= \mathbb{E}\left[\frac{1}{N}(X_1 + X_2 + \dots + X_N)\right] \\ &= \frac{1}{N}\left[\mathbb{E}[X_1] + \mathbb{E}[X_2] + \dots + \mathbb{E}[X_N]\right] \\ &= \frac{1}{N}N * \mathbb{E}[X] \\ &= \mathbb{E}[X]\end{aligned}$$

Properties of the Distribution of the Mean

Property 2: The variance of the mean is $Var(\bar{X}) = \frac{Var(X)}{N}$.

Moreover, we estimate $Var(X)$ from the sample as follows:

$$\widehat{Var}(X) = \frac{1}{N-1} \sum_{i=1}^N (X_i - \bar{X})^2$$

So then we can estimate $Var(\bar{X}) = \frac{\widehat{Var}(X)}{N}$

Square root of this is **standard deviation of the mean**, often called the “standard error”.

Properties of the Distribution of the Mean (math aside)

A taste of the math (again, don't worry)

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{N}(X_1 + X_2 + \dots + X_N)\right) \\ &= \frac{1}{N^2} \left[\text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_N) \right] \\ &= \frac{1}{N^2} N * \text{Var}(X) \\ &= \frac{\text{Var}(X)}{N} \end{aligned}$$

Review so far...

The sample mean is something you get from a sample...

But it is a random variable, drawn from some distribution...

And we can say something about the (unseen) *distribution* from which our particular mean was drawn. So far, we've said:

- 1 Our mean, \bar{X} is a random variable with expectation $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$
- 2 The variance of \bar{X} is the variance of X divided by N .
 - We estimate the variance of X by

$$\widehat{\text{Var}}(X) = \frac{1}{N-1} \sum_{i=1}^N (X - \bar{X})^2$$

- Thus, we estimate the variance of \bar{X} by:

$$\widehat{\text{Var}}(\bar{X}) = \frac{1}{(N-1)N} \sum_{i=1}^N (X - \bar{X})^2$$

- We call $\sqrt{\widehat{\text{Var}}}$ the **standard deviation**, and for the mean, we often call the standard deviation the “standard error of the mean”.

Political Science 15
Introduction to Research in Political Science
Lecture 3d: Central Limit Theorem

Alice Lépiessier
University of California Santa Barbara

Special thanks to Chad Hazlett and Allison Horst for select slides and images, used with permission.

Recap: Properties of the Distribution of the Mean

So far, we have shown that:

- 1 The distribution of the sample mean \bar{X} is centered around $\mathbb{E}[X]$.
- 2 The variance of the sample mean is $Var(\bar{X}) = \frac{Var(X)}{N}$.

How is the Mean Distributed?

There is one more piece to understanding the distribution of the mean: we know its expectation and variance, but what about the shape?

The shape must depend on the distribution of the underlying X , you would think.

Fortunately, it does not...

Central Limit Theorem (CLT)

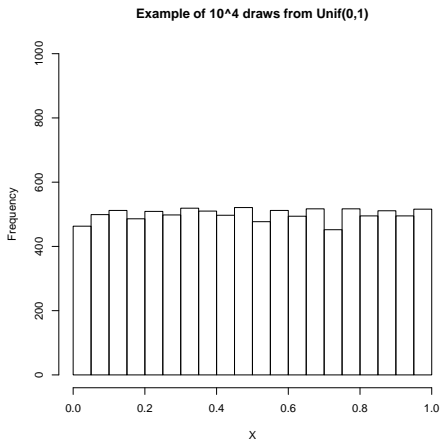
The distribution of the mean tends toward a normal distribution.

- This is magical: regardless of how the original X is distributed, when you take the mean of multiple RVs drawn from the same distribution, it starts to look normal.
- You do need N to be big enough for this to work, but that's often not a problem.
- We will discuss some rules for deciding if N is big enough, and adjustments to use when it is not.

Example of the CLT in action

Suppose X is distributed uniformly between 0 and 1.

Let's see what happens to the mean of samples drawn from this distribution, as N increases. This is the distribution of the uniform random variable (so X).

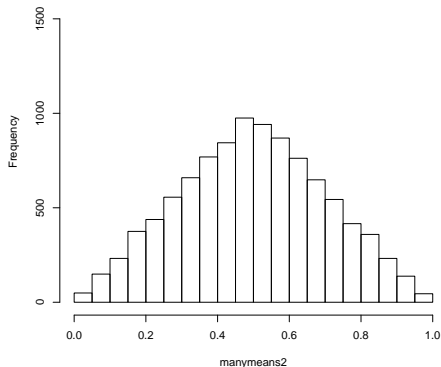


Example of the CLT in action

Suppose X is distributed uniformly between 0 and 1.

This is the distribution of the mean of samples of size $N = 2$ (so of \bar{X}).

Simulated distribution of means with N=2

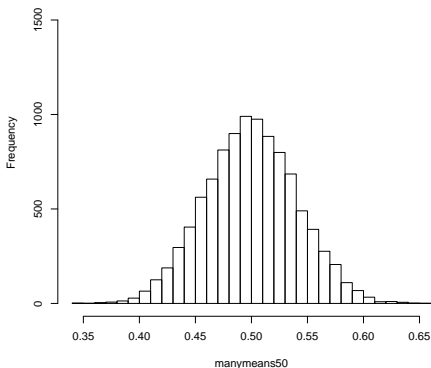


Example of the CLT in action

Suppose X is distributed uniformly between 0 and 1.

This is the distribution of the mean of samples of size $N = 50$. The distribution is starting to look more normal!

Simulated distribution of means with $N=50$



Example of the CLT in action

So that's a simulated example, but let's see how well it matches our theoretical understanding of the distribution:

- Take the $N = 50$ case
- We can estimate the center using \bar{X} , which is 0.5 (because we are drawing uniform RVs between 0 and 1)
- We can estimate the variance of \bar{X} using $\frac{1}{N} \widehat{\text{Var}}(X)$

```
>varmean=var(X)/N  
>SE=sqrt(varmean)
```

- We **know** the shape of the distribution of the mean is normal (from the CLT).

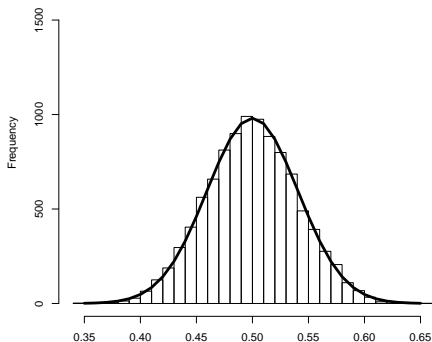
Example of the CLT in action

Using all this, we estimate that the mean should be distributed

$$\bar{X} \sim N(\mu = 0.5, \sigma^2 = \text{Var}(X)/N)$$

Let's superimpose our theoretical estimate of the distribution of the mean over the simulated one:

Simulated distribution of means with N=50



Why have we done all this?!

Remember, the big idea was that even when we only observe a mean once we can say a lot about how the mean would be *distributed* (as if we could observe it over and over again).

In particular, our key-aways are:

Properties of the distribution of the mean

The distribution of the sample mean is:

- 1 normal
- 2 centered around $\mathbb{E}[X]$
- 3 with variance $\text{Var}(X)/N$

Why have we done all this?!

A huge part of the statistical machinery depends on this finding. Next time we'll see how this allows us to:

- Construct **confidence intervals** around your estimated mean to characterize our uncertainty.
- Do **hypothesis tests**, get p-values, etc. for one-sample and two-samples tests involving means.
- Other hypothesis tests for categorical outcomes, binary outcomes, counts, etc. will follow similar patterns.

On a final note, this is tough stuff but paves the way for much of what we will do, so go over these slides.