Gov 50: 16. Sampling Distributions

Matthew Blackwell

Harvard University

Roadmap

1. Poll example
2. Random variables and probability distributions
3. Sampling distribution
4. Normal variables and the Central Limit Theorem

1/ Poll example

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
  - Sept 1st-16th

How popular is Joe Biden?

- What proportion of the public approves of Biden’s job as president?
- Latest Gallup poll:
    - Sept 1st-16th
    - 812 adult Americans

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
    - Sept 1st-16th
    - 812 adult Americans
    - Telephone interviews

How popular is Joe Biden?

- What proportion of the public approves of Biden's job as president?
- Latest Gallup poll:
  - Sept 1st-16th
  - 812 adult Americans
  - Telephone interviews
  - Approve (42%), Disapprove (56%)

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.
- **Population parameter**: population proportion of all US adults that approve of Biden.

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.
- **Population parameter**: population proportion of all US adults that approve of Biden.
    - Census: not possible.

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.
- **Population parameter**: population proportion of all US adults that approve of Biden.
    - Census: not possible.
- **Sample:** random digit dialing phone numbers (cell and landline).

Poll in our framework

- **Population**: adults 18+ living in 50 US states and DC.
- **Population parameter**: population proportion of all US adults that approve of Biden.
  - Census: not possible.
- **Sample:** random digit dialing phone numbers (cell and landline).
- **Point estimate**: sample proportion that approve of Biden

2/ Random variables and probability distributions

Random variables

Random variables are numerical summaries of chance processes:

$$X_i = \begin{cases} 1 & \text{if respondent } i \text{ supports Biden,} \\ 0 & \text{otherwise} \end{cases}$$

Random variables

Random variables are numerical summaries of chance processes:

$$X_i = \begin{cases} 1 & \text{if respondent } i \text{ supports Biden,} \\ 0 & \text{otherwise} \end{cases}$$

With a simple random sample, chance of $X_i = 1$ is equal to the population proportion of people that support Biden.

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
    - Number of heads in 5 coin flips

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
    - Number of heads in 5 coin flips
    - Sampled senator is a woman ($X = 1$) or not ($X = 0$)

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
    - Number of heads in 5 coin flips
    - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
    - Number of battle deaths in a civil war

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
    - Number of heads in 5 coin flips
    - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
    - Number of battle deaths in a civil war
- **Continuous**: X can take any real value (usually within an interval).

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
  - Number of heads in 5 coin flips
  - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
  - Number of battle deaths in a civil war
- **Continuous**: X can take any real value (usually within an interval).
  - GDP per capita (average income) in a country.

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
  - Number of heads in 5 coin flips
  - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
  - Number of battle deaths in a civil war
- **Continuous**: X can take any real value (usually within an interval).
  - GDP per capita (average income) in a country.
  - Share of population that approves of Biden.

Types of random variables

- **Discrete**: X can take a finite (or countably infinite) number of values.
    - Number of heads in 5 coin flips
    - Sampled senator is a woman ($X = 1$) or not ($X = 0$)
    - Number of battle deaths in a civil war
- **Continuous**: X can take any real value (usually within an interval).
    - GDP per capita (average income) in a country.
    - Share of population that approves of Biden.
    - Amount of time spent on a website.

Probability distributions

Probability distributions tell us the chances of different values of a r.v. occurring

Probability distributions

Probability distributions tell us the chances of different values of a r.v. occurring

Discrete variables: like a frequency barplot for the population distribution.

Probability distributions

Probability distributions tell us the chances of different values of a r.v. occurring

Discrete variables: like a frequency barplot for the population distribution.

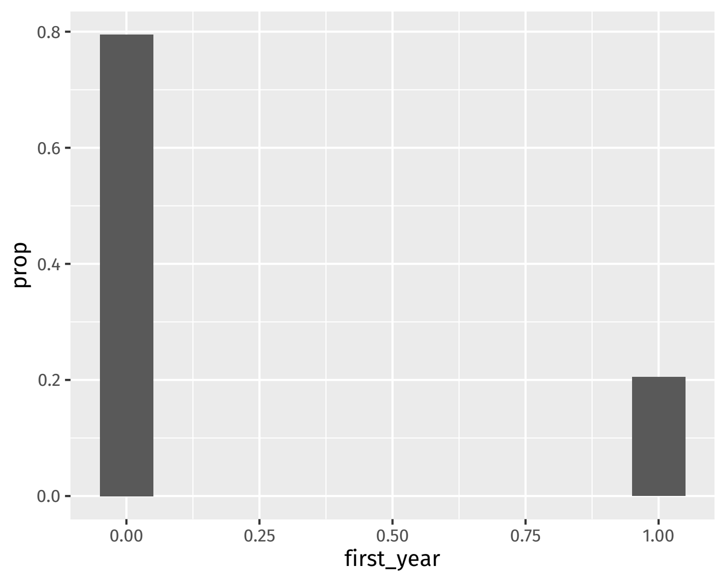
Continuous variables: like a continuous version of population histogram.

Discrete probability distribution

We can use the `y = ..prop..` aesthetic to get a barplot with proportions instead of count to show us the chance/probability of selecting a first-year student:

```
library(gov50data)
class_years |>
  mutate(first_year = as.numeric(year == "First-Year")) |>
  ggplot(aes(x = first_year)) +
  geom_bar(mapping = aes(y = ..prop..), width = 0.1)
```

Discrete probability distribution



Midwest data

```
library(ggplot2)
midwest
```

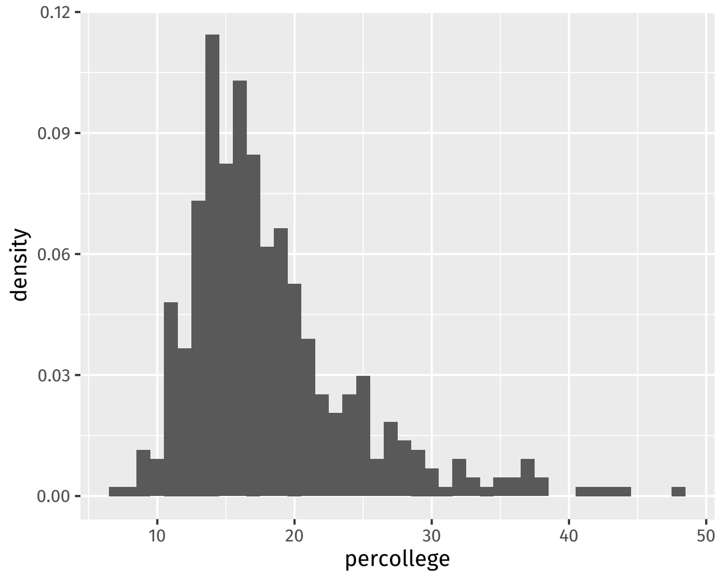
```
## # A tibble: 437 x 28
##      PID county  state  area popto~1 popde~2 popwh~3 popbl~4
##    <int> <chr>   <chr> <dbl>   <int>   <dbl>   <int>   <int>
##  1   561 ADAMS   IL    0.052   66090   1271.   63917    1702
##  2   562 ALEXAN~ IL    0.014   10626    759     7054    3496
##  3   563 BOND    IL    0.022   14991    681.   14477     429
##  4   564 BOONE   IL    0.017   30806   1812.   29344     127
##  5   565 BROWN   IL    0.018    5836    324.    5264     547
##  6   566 BUREAU  IL    0.05    35688    714.   35157      50
##  7   567 CALHOUN IL    0.017    5322    313.    5298       1
##  8   568 CARROLL IL    0.027   16805    622.   16519     111
##  9   569 CASS    IL    0.024   13437    560.   13384      16
## 10   570 CHAMPA~ IL    0.058  173025   2983.  146506   16559
## # ... with 427 more rows, 20 more variables:
## #   popamerindian <int>, popasian <int>, popother <int>,
## #   percwhite <dbl>, percblack <dbl>, percamerindan <dbl>,
## #   percasian <dbl>, percother <dbl>, popadults <int>,
## #   perchsd <dbl>, percollege <dbl>, percprof <dbl>,
## #   poppovertyknown <int>, percpovertyknown <dbl>,
```

Continuous probability distribution

We can use the `y = ..density..` to create a **density histogram** instead of a count histogram so that the area of the histogram boxes are equal to the chance of randomly selecting a unit in that bin:

```
midwest |>
  ggplot(aes(x = percollege)) +
  geom_histogram(aes(y = ..density..), binwidth = 1)
```

Continuous probability distribution



Why density?

Histograms with **density** on the y-axis are drawn so that the area of each box is equal to the proportion of units in the sample in that horizontal bin.

Why density?

Histograms with **density** on the y-axis are drawn so that the area of each box is equal to the proportion of units in the sample in that horizontal bin.

Easier to compare distributions across sample sizes.

Why density?

Histograms with **density** on the y-axis are drawn so that the area of each box is equal to the proportion of units in the sample in that horizontal bin.

Easier to compare distributions across sample sizes.

Sum up all the area = 1 (but heights can go above 1)

3/ Sampling distribution

Key properties of sums and means

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from a population distribution with mean μ ("mu") and variance σ^2 ("sigma squared")

Key properties of sums and means

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from a population distribution with mean μ ("mu") and variance σ^2 ("sigma squared")

Sample mean: $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\overline{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

. . .

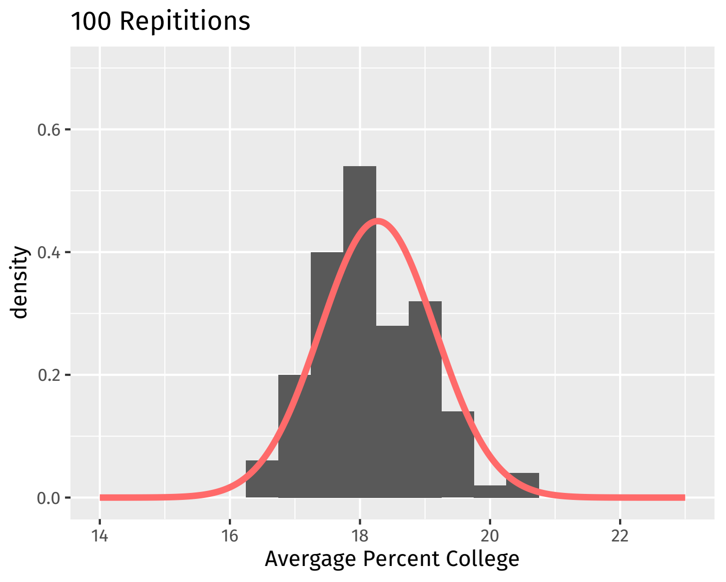
$\overline{X}_n$ is a random variable with a distribution!!

Sample means/proportions distribution

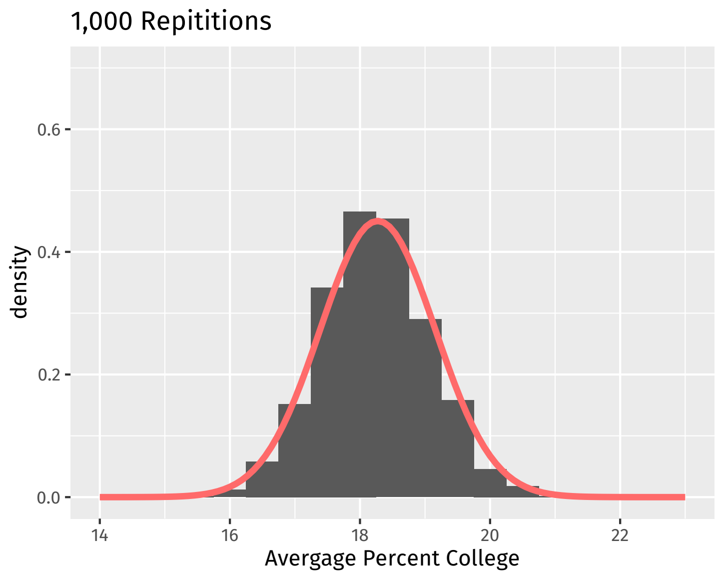
Sampling distributions are the probability distributions of an estimator like $\overline{X}_n$

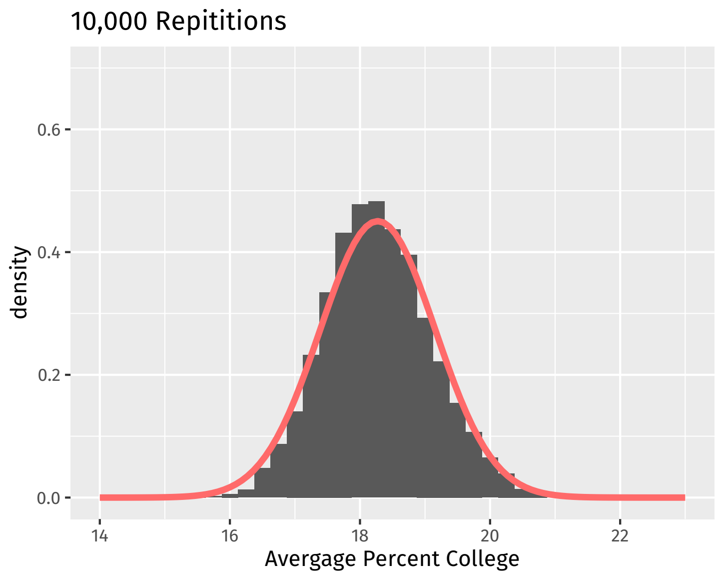
When we have access to the full population, we can approximate the sampling distribution with repeated sampling.

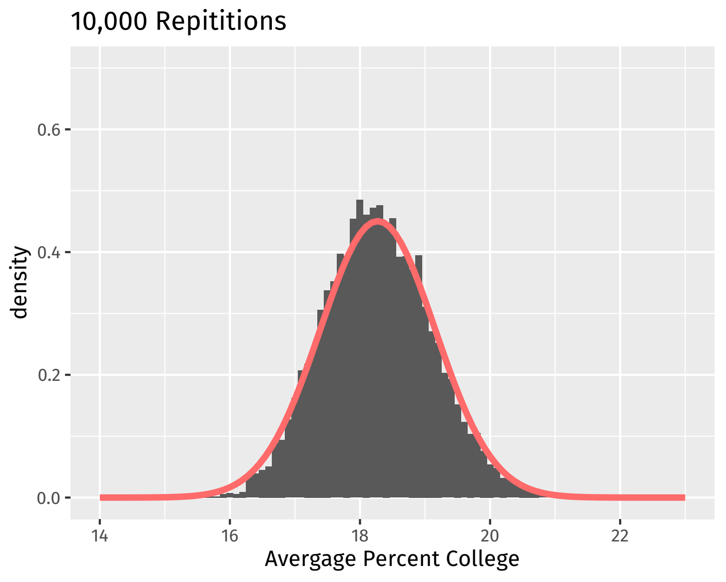
```
library(infer)
midwest |>
  rep_slice_sample(n = 50, reps = 100) |>
  group_by(replicate) |>
  summarize(`Avergage Percent College` = mean(percollege)) |>
  ggplot(aes(x = `Avergage Percent College`)) +
  geom_histogram(mapping = aes(y = ..density..), binwidth = 0.5) +
  coord_cartesian(xlim = c(14, 23), ylim = c(0, 0.7)) +
  labs(title = "100 Repititions") +
  stat_function(fun = dnorm, args = c(mean(midwest$percollege), sd(midwest$
                color = "indianred1", size = 1.5, xlim = c(14, 23))
```



100 Repititions
density
0.0
0.2
0.4
0.6
14
16
18
20
22
Avergage Percent College


1,000 Repititions
density
0.0
0.2
0.4
0.6
14
16
18
20
22
Avergage Percent College


10,000 Repititions
density
0.0
0.2
0.4
0.6
14
16
18
20
22
Avergage Percent College


10,000 Repititions
0.6
0.4
0.2
0.0
density
14
16
18
20
22
Avergage Percent College

Sampling distribution of the sample mean

Suppose $X_1, X_2, \dots, X_n$ is a simple random sample from a population distribution with mean μ and variance σ^2.

Sampling distribution of the sample mean

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from a population distribution with mean μ and variance σ^2.

Expected value of the distribution of $\overline{X}_n$ is the population mean, μ.

Sampling distribution of the sample mean

Suppose $X_1, X_2, \ldots, X_n$ is a simple random sample from a population distribution with mean μ and variance σ^2.

Expected value of the distribution of $\overline{X}_n$ is the population mean, μ.

Standard error of the distribution of $\overline{X}_n$ is approximately $\sigma/\sqrt{n}$:

$$SE \approx \frac{\text{population standard deviation}}{\sqrt{\text{sample size}}}$$

Unbiasedness

An estimator is **unbiased** when its expected value across repeated samples equals the population parameter of interest.

Unbiasedness

An estimator is **unbiased** when its expected value across repeated samples equals the population parameter of interest.

Sample mean of a simple random sample is **unbiased** for the population mean, $\mathbb{E}[\overline{X}_n] = \mu$

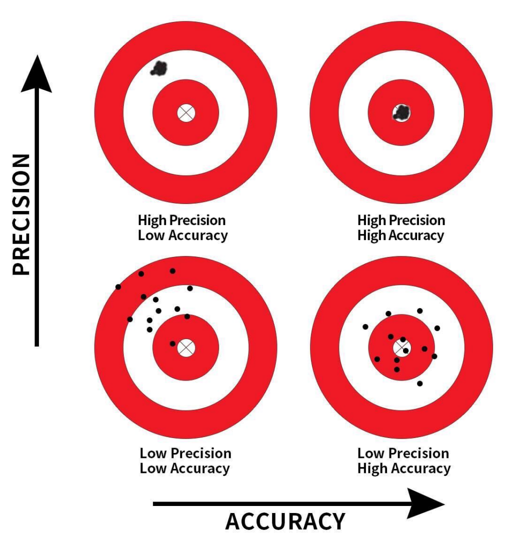
Unbiasedness

An estimator is **unbiased** when its expected value across repeated samples equals the population parameter of interest.

Sample mean of a simple random sample is **unbiased** for the population mean, $\mathbb{E}[\overline{X}_n] = \mu$

An estimator that isn't unbiased is called **biased**.

Precision vs accuracy



Law of large numbers

Law of large numbers

Let $X_1, \dots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

Law of large numbers

Law of large numbers

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

- Probability of $\overline{X}_n$ being "far away" from μ goes to 0 as n gets big.

Law of large numbers

Law of large numbers

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

- Probability of $\overline{X}_n$ being “far away” from μ goes to 0 as n gets big.
- The distribution of sample mean “collapses” to population mean.

Law of large numbers

Law of large numbers

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

- Probability of $\overline{X}_n$ being "far away" from μ goes to 0 as n gets big.
- The distribution of sample mean "collapses" to population mean.
- Can see this from the SE of $\overline{X}_n$: $SE = \sigma/\sqrt{n}$.

Law of large numbers

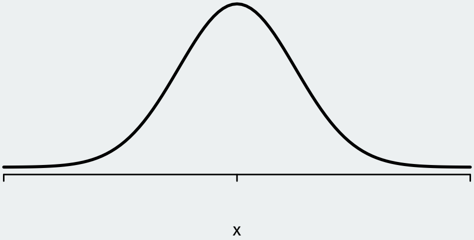
Law of large numbers

Let $X_1, \dots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ converges to μ as n gets large.

- Probability of $\overline{X}_n$ being "far away" from μ goes to 0 as n gets big.
- The distribution of sample mean "collapses" to population mean.
- Can see this from the SE of $\overline{X}_n$: $SE = \sigma/\sqrt{n}$.
- Not necessarily true with a biased sample!

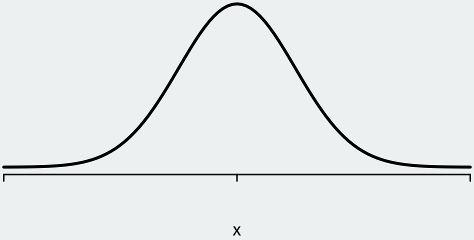
4/ Normal variables and the Central Limit Theorem

Normal random variable



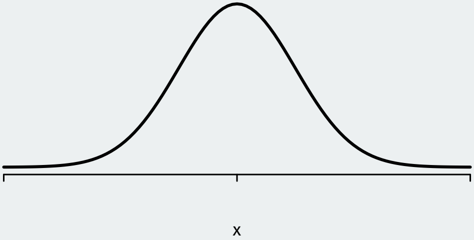
- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.

Normal random variable



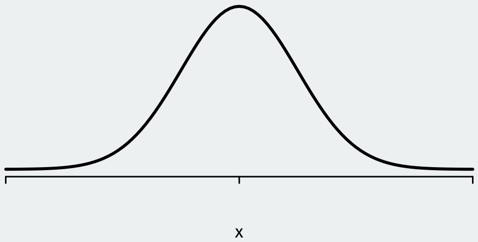
- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.
    - Extremely ubiquitous in statistics.

Normal random variable



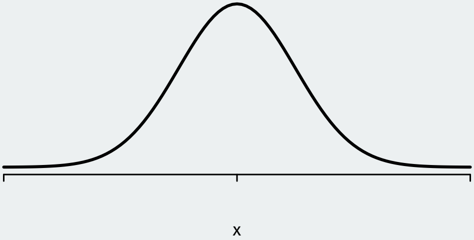
- A **normal distribution** has a PDF that is the classic "bell-shaped" curve.
    - Extremely ubiquitous in statistics.
    - An r.v. is more likely to be in the center, rather than the tails.

Normal random variable



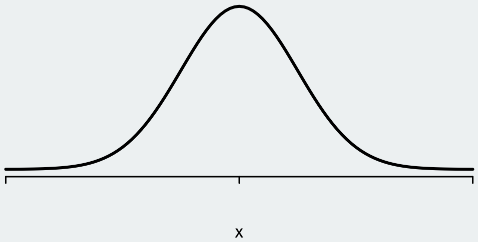
- A **normal distribution** has a PDF that is the classic "bell-shaped" curve.
    - Extremely ubiquitous in statistics.
    - An r.v. is more likely to be in the center, rather than the tails.
- Three key properties of this PDF:

Normal random variable



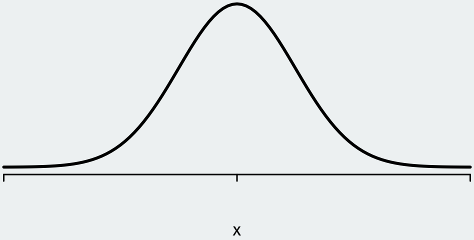
- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.
  - Extremely ubiquitous in statistics.
  - An r.v. is more likely to be in the center, rather than the tails.
- Three key properties of this PDF:
  - **Unimodal**: one peak at the mean.

Normal random variable



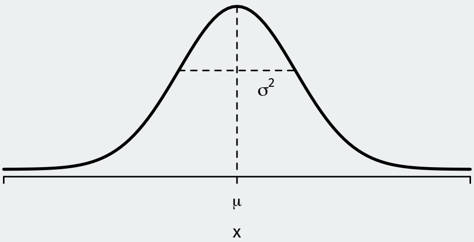
- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.
    - Extremely ubiquitous in statistics.
    - An r.v. is more likely to be in the center, rather than the tails.
- Three key properties of this PDF:
    - **Unimodal**: one peak at the mean.
    - **Symmetric** around the mean.

Normal random variable



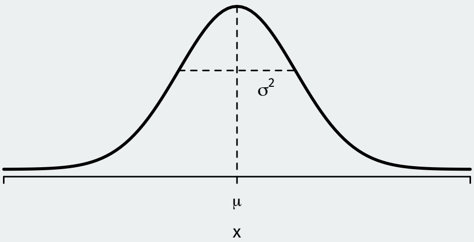
- A **normal distribution** has a PDF that is the classic “bell-shaped” curve.
    - Extremely ubiquitous in statistics.
    - An r.v. is more likely to be in the center, rather than the tails.
- Three key properties of this PDF:
    - **Unimodal**: one peak at the mean.
    - **Symmetric** around the mean.
    - **Everywhere positive**: any real value can possibly occur.

Normal distribution



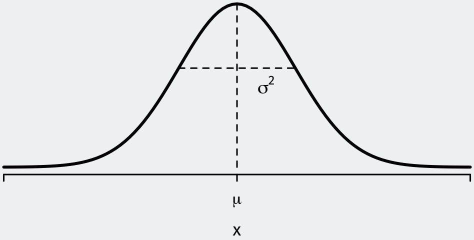
- A normal distribution can be affect by two values:

Normal distribution



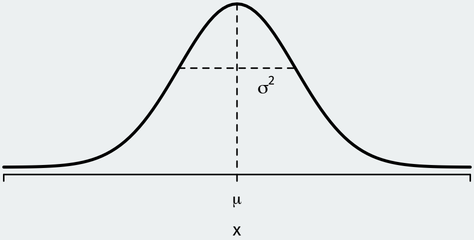
- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ

Normal distribution



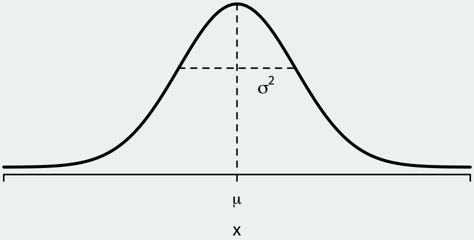
- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ
    - **variance** written as σ^2 (standard deviation is σ)

Normal distribution



- A normal distribution can be affect by two values:
    - **mean/expected value** usually written as μ
    - **variance** written as σ^2 (standard deviation is σ)
    - Written $X \sim N(\mu, \sigma^2)$.

Normal distribution



- A normal distribution can be affect by two values:
  - **mean/expected value** usually written as μ
  - **variance** written as σ^2 (standard deviation is σ)
  - Written $X \sim N(\mu, \sigma^2)$.
- **Standard normal distribution**: mean 0 and standard deviation 1.

Central limit theorem

Central limit theorem

Let $X_1, \dots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

Central limit theorem

Central limit theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

- "Sample means tend to be normally distributed as samples get large."

Central limit theorem

Central limit theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

- "Sample means tend to be normally distributed as samples get large."
- $\rightsquigarrow$ we know (an approx. of) the entire probability distribution of $\overline{X}_n$

Central limit theorem

Central limit theorem

Let $X_1, \dots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

- "Sample means tend to be normally distributed as samples get large."
- $\rightsquigarrow$ we know (an approx. of) the entire probability distribution of $\overline{X}_n$
  - Approximation is better as n goes up.

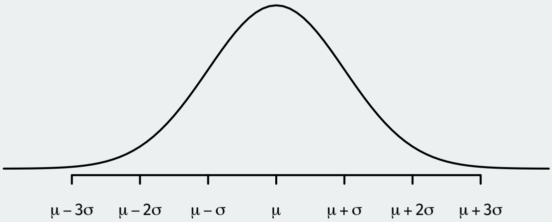
Central limit theorem

Central limit theorem

Let $X_1, \ldots, X_n$ be a simple random sample from a population with mean μ and finite variance σ^2. Then, $\overline{X}_n$ will be approximately distributed $N(\mu, \sigma^2/n)$ in large samples.

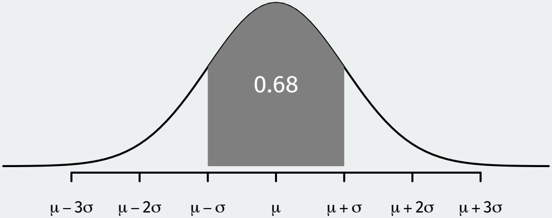
- "Sample means tend to be normally distributed as samples get large."
- $\rightsquigarrow$ we know (an approx. of) the entire probability distribution of $\overline{X}_n$
    - Approximation is better as n goes up.
    - Does not depend on the distribution of X_i!

Empirical Rule for the Normal Distribution



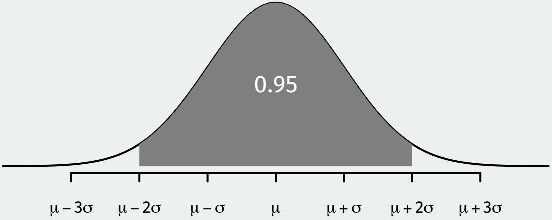
- If $X \sim N(\mu, \sigma^2)$, then:

Empirical Rule for the Normal Distribution



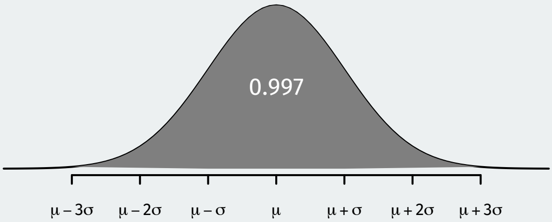
- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.

Empirical Rule for the Normal Distribution



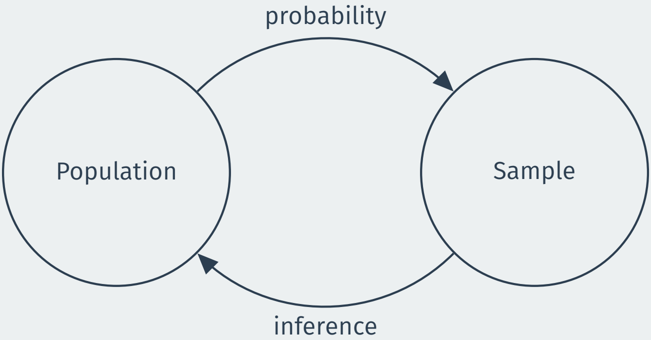
- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.
  - $\approx$ 95% of the distribution of X is within 2 SDs of the mean.

Empirical Rule for the Normal Distribution



- If $X \sim N(\mu, \sigma^2)$, then:
  - $\approx$ 68% of the distribution of X is within 1 SD of the mean.
  - $\approx$ 95% of the distribution of X is within 2 SDs of the mean.
  - $\approx$ 99.7% of the distribution of X is within 3 SDs of the mean.
- CLT + empirical rule: we'll know the rough distribution of estimation errors we should expect.

Where are we going?



We only get 1 sample. Can we learn about the population from that sample?