STAT011 Statistical Methods I

Lecture 10 Central Limit Theorem

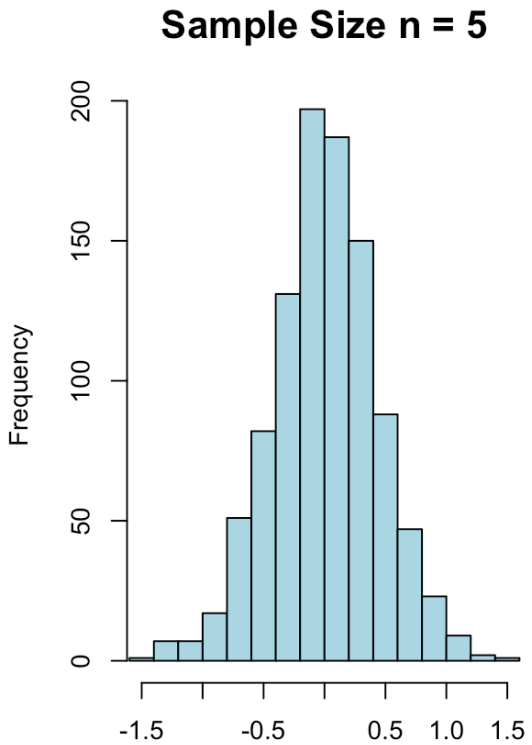
Lu Chen
Swarthmore College
2/21/2019

Review

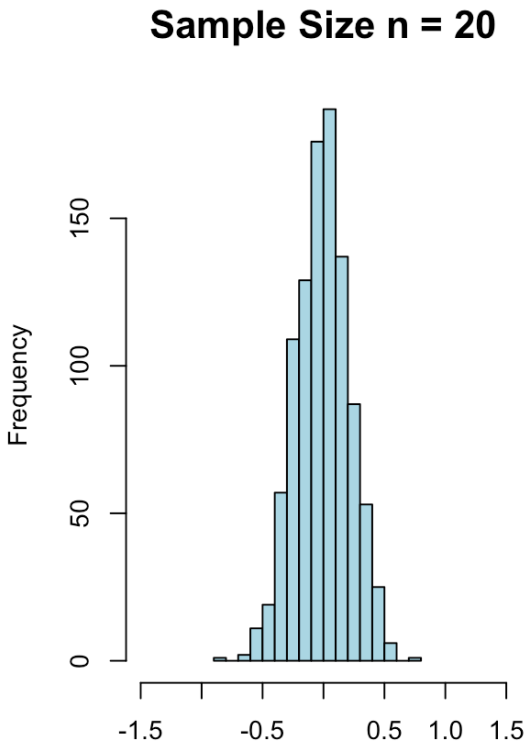
- Population, sample, parameter and statistic
  - We use sample statistics (change from sample to sample) to estimate population parameters (fixed and unknown)
- Statistical inference uses a fact about a sample to estimate the truth about the whole population.
- Sampling variability
  - The value of a statistic varies in repeated random sampling.
- Sampling distribution
- Bias and variability
  - Bias concerns the center; variability concerns the spread.
  - To reduce bias, *use random sampling*; to reduce variability, *increase sample size*.
- Sampling distribution of a sample mean
  - If population distribution is $X \sim N(\mu, \sigma)$, the sampling distribution of the sample mean is $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

Review - Sampling distribution by simulation

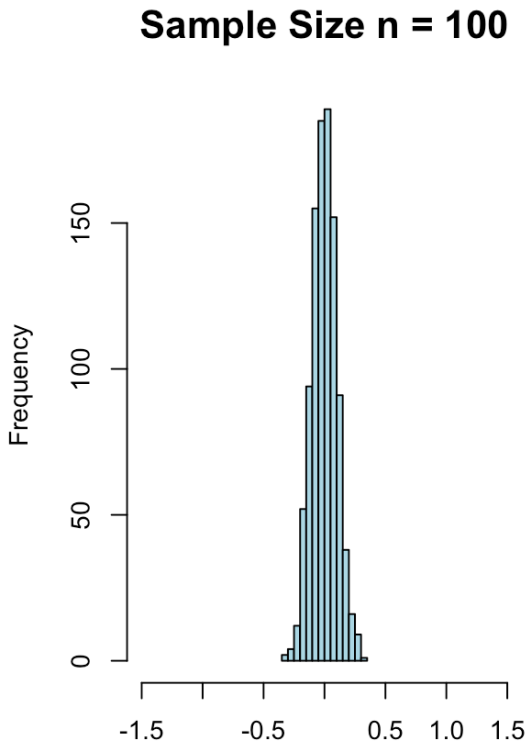
Simulation: 1000 samples



Mean: 0.010; SD: 0.436



Mean: -0.005; SD: 0.219



Mean: -0.001; SD: 0.101

Outline

- Sampling distribution of a sample mean
  - When population distribution is Normal
  - When population distribution is not Normal
- Central Limit Theorem (CLT)
  - Definition
  - Examples
    - Sampling distribution of a proportion
      - Normal approximation
      - Examples

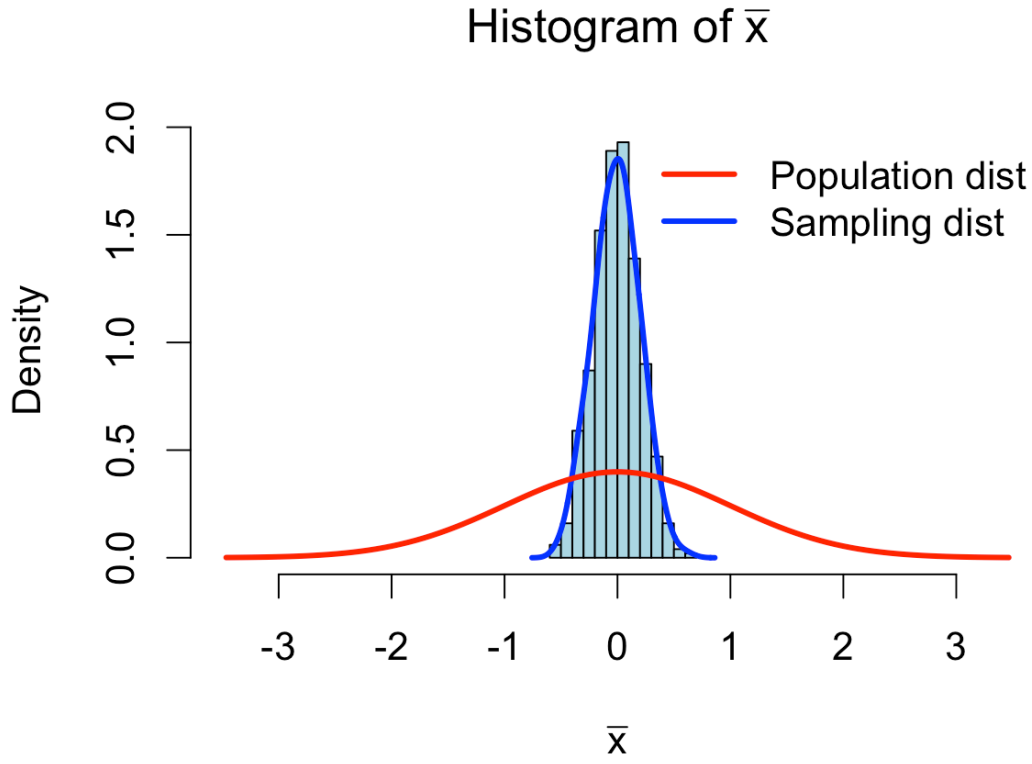
Sampling distribution of a sample mean

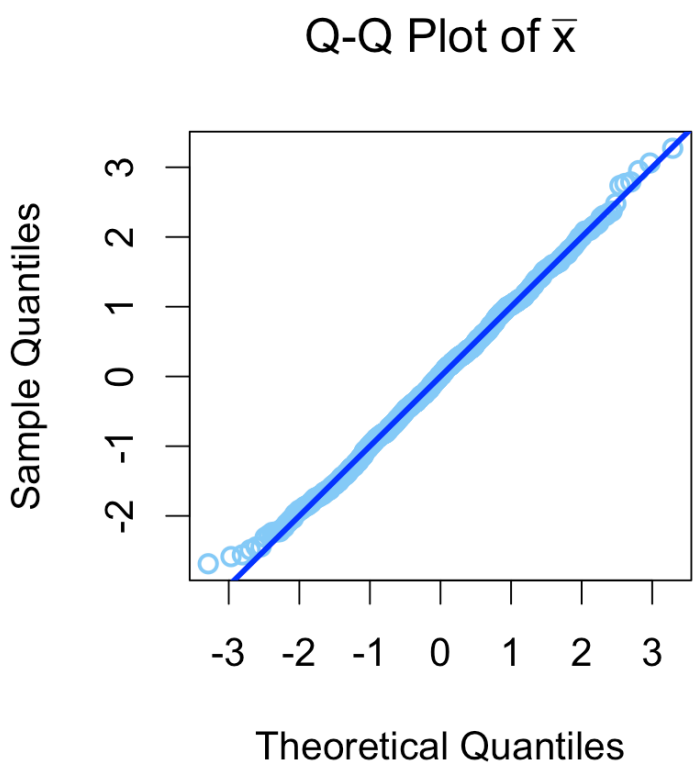
- Population distribution: $X \sim N(0, 1)$.
- 1000 samples with size 25 are generated from this population. Mean $\bar{x}$ is calculated for each sample.
- Sampling distribution of $\bar{x}$ is?

Sampling distribution simulation

Sampling distribution of a sample mean

$X \sim N(0, 1)$. $n = 25$. 1000 samples.





Sampling distribution of a sample mean

Let $\bar{x}$ be the mean of an SRS of size n from a population having Normal distribution with mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

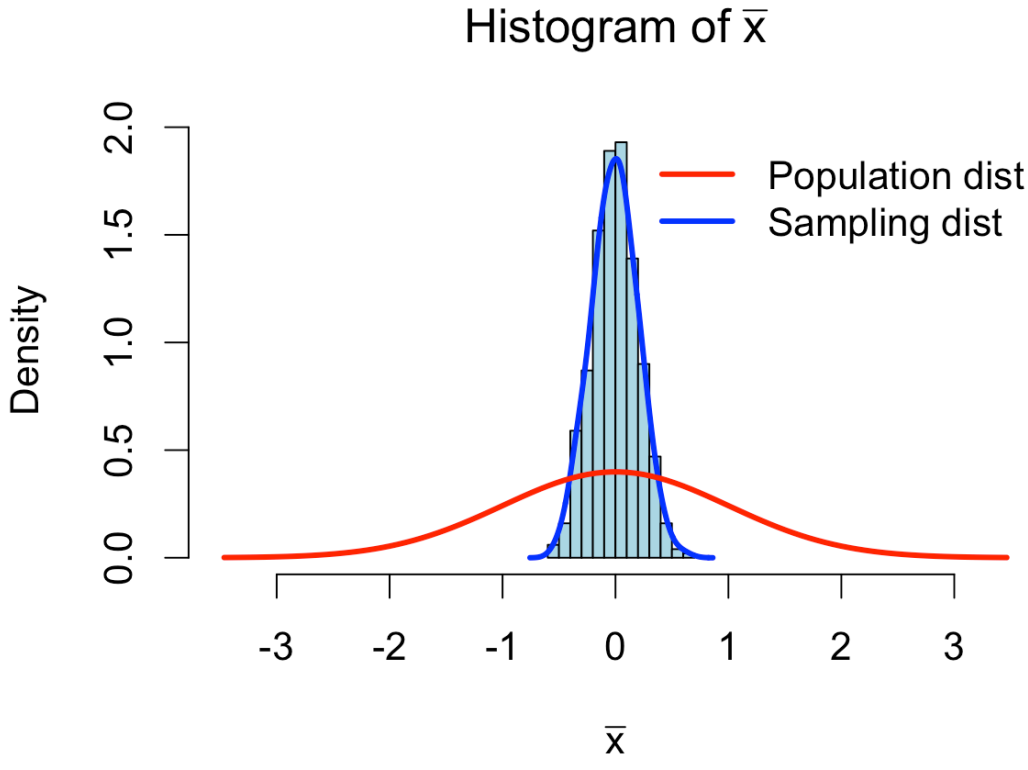
$$\mu_{\bar{x}} = \mu,$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

And $\bar{x}$ has the $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ distribution.

This says that if $X \sim N(\mu, \sigma)$, then

$$\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Sampling distribution of a sample mean




```
mean(mean_x)
```

```
## [1] -0.006801407
```

```
sd(mean_x)
```

```
## [1] 0.202523
```

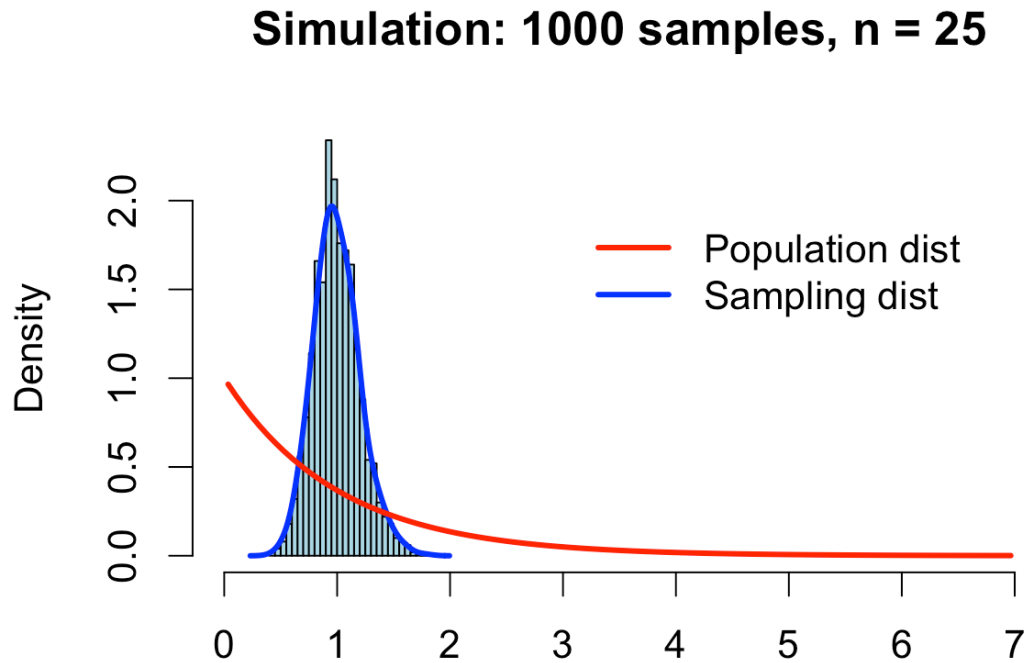
- Population distribution: $X \sim N(0, 1)$
- Sampling distribution of $\bar{x}$ (by **theory**): $\bar{x} \sim N\left(0, \frac{1}{\sqrt{25}}\right) = N(0, 0.2)$
- Sampling distribution of $\bar{x}$ (by **simulation**): $\bar{x} \sim N(-0.007, 0.203)$
- The results from simulation are very close to the theoretical results.

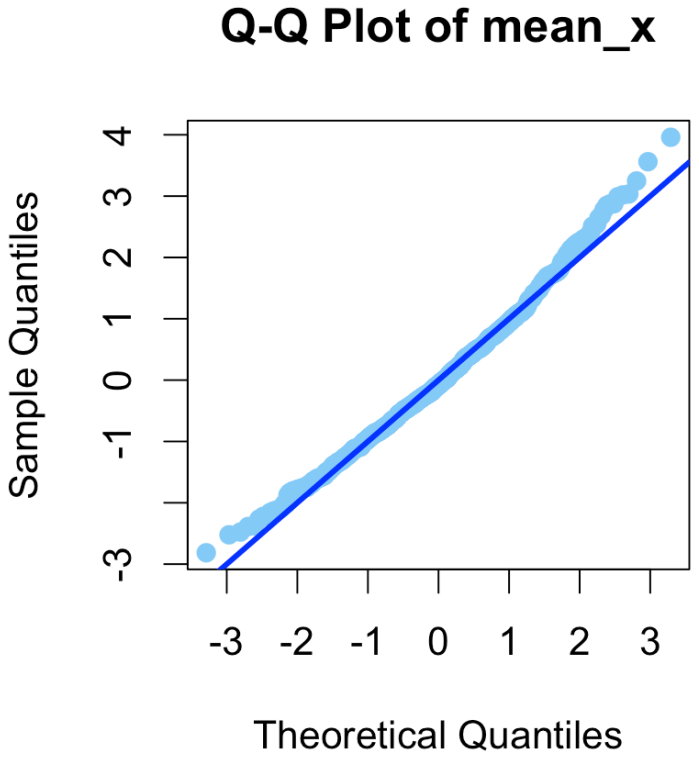
Sampling distribution of a sample mean

- When population distribution is **Normal** $X \sim N(\mu, \sigma)$, the distribution of mean $\bar{x}$ is also Normal.
  - The center of $\bar{x}$ is the same as the center of the population: $\boldsymbol{\mu_{\bar{x}} = \mu}$
    - Simple random sampling assures this.
  - The spread of $\bar{x}$ is related to both the spread of the population and the sample size: $\boldsymbol{\sigma_{\bar{x}} = \sigma/\sqrt{n}}$
    - If the population data has large variability, the sample mean also has large variability.
    - Larger sample size leads to smaller variability.
- What if the population distribution is **NOT Normal**?
- Will the shape of the sample mean $\bar{x}$ still the same as the shape of the population distribution?
- How about center and spread? Sampling distribution simulation

Sampling distribution of a sample mean

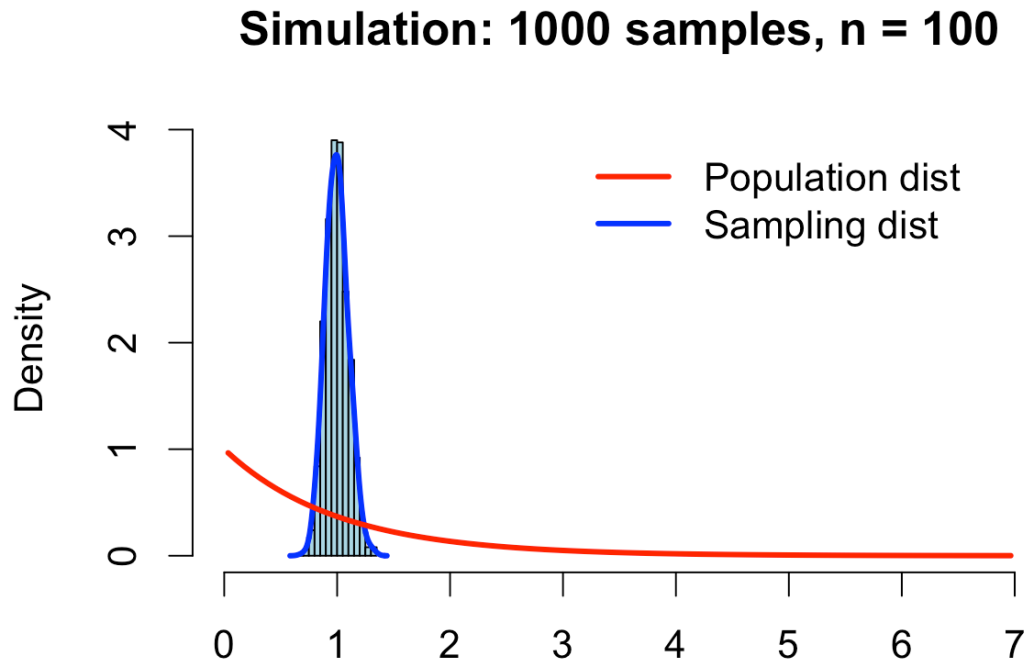
When **population distribution is NOT Normal** even highly skewed, the distribution of the sample mean $\bar{x}$ **is still approximately Normal**!

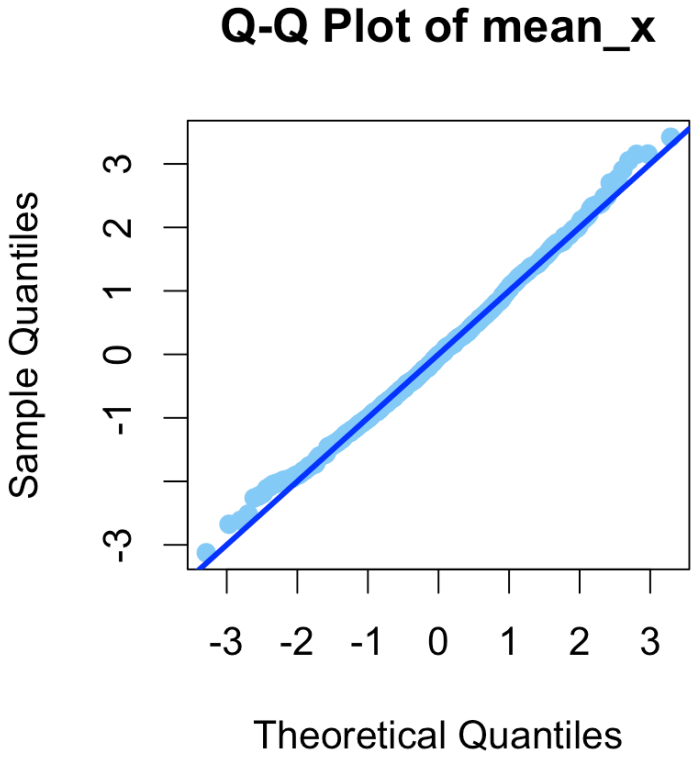




Sampling distribution of a sample mean

When **population distribution is NOT Normal** even highly skewed, the distribution of the sample mean $\bar{x}$ **is still approximately Normal**!





Central Limit Theorem

Draw an SRS of size n from **any population** with mean μ and finite standard deviation σ. When ***n* is large**, the sampling distribution of the sample mean $\bar{x}$ is approximately Normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

$$\bar{x} \overset{approx.}{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Central limit theorem holds for data with any population distribution.
- **Amazing Fact!**

Central Limit Theorem

Sample size for CLT to hold

Population distribution	Sample size n	Sampling distribution of $\bar{x}$
Exactly Normal	Any	Exactly Normal
Close to Normal	$n > 20$	Approximately Normal
Highly skewed	$n > 60$	Approximately Normal

- These suggestions for samples size are only rules of thumb.
- The appropriate sample size always depends on each specific problem.
- Generally, as sample size n increases, the sampling distribution of mean $\bar{x}$ will become **less variable** and **more Normal**.

Example 1 - Female height

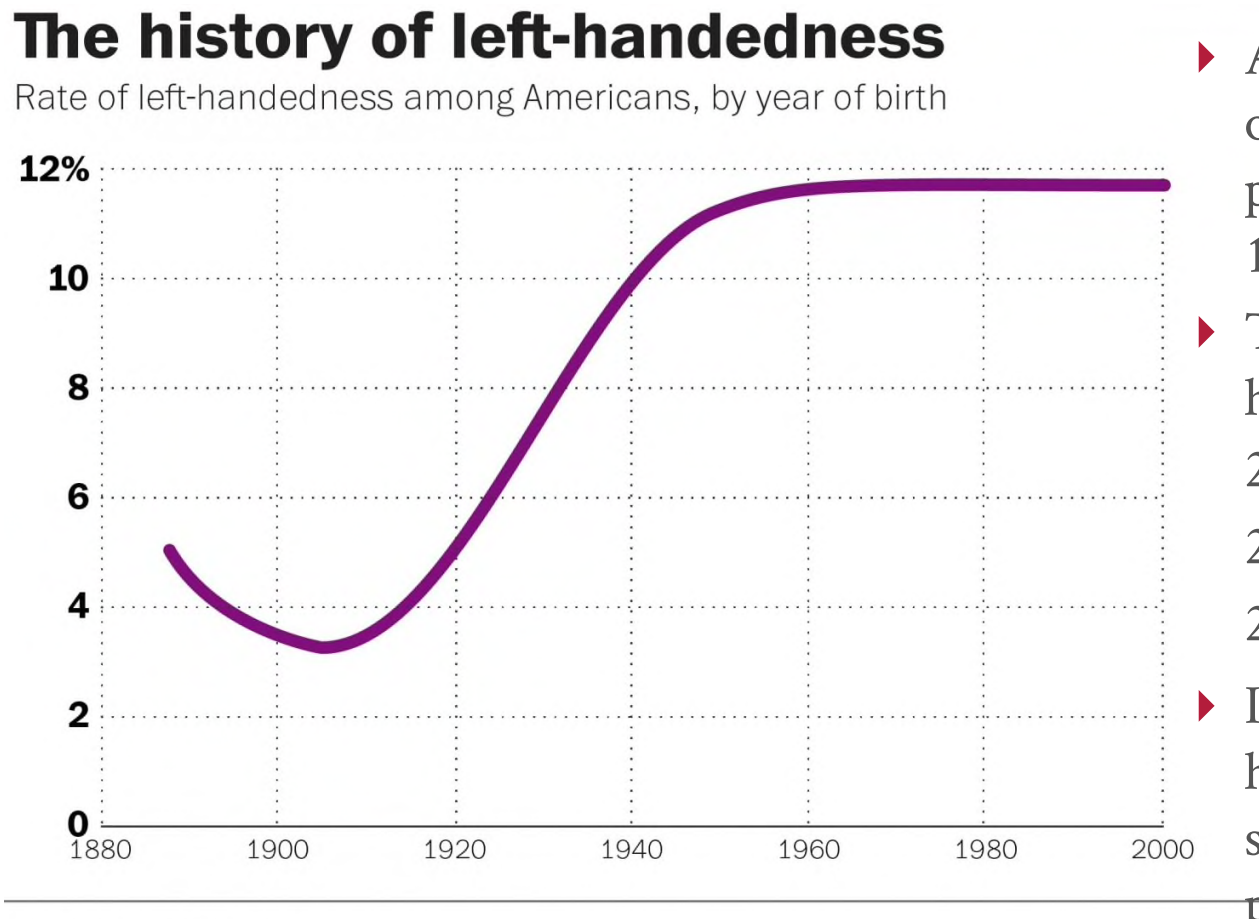
US population female height (not necessarily Normal): Mean 64 inches, SD 3 inches (assume these are true population parameter values).

STAT011 female height ($n = 58$): Mean 64.9 inches, SD 3.4 inches.

Is the STAT011 female height the same as the US population female height?

- $\mu = 64, \sigma = 3, n = 58, \bar{x} = 64.9$
- Applying CLT, $\bar{x} \overset{approx.}{\sim} N(64, \frac{3}{\sqrt{58}}) = N(64, 0.4)$
- 68% of the sample mean $\bar{x}$ should fall between 63.6 and 64.4 inches.
- 95% of the sample mean $\bar{x}$ should fall between 63.2 and 64.8 inches.
- The chance is quite low that we get a sample of size 58 from the US population and the mean is 64.9 inches.
- Therefore, STAT011 female height is quite different from (somewhat larger than) the US population female height.

Example 2 - Left-handedness



- After 1960's, the proportion of left-handedness in US population stays around 11.8%.
- The proportion of left-handedness in STAT 11

  2016 class: $10/122 = 0.082$

  2017 class: $7/94 = 0.074$

  2019 class: $4/112 = 0.036$
- Is the proportion of left-handedness in STAT 11 the same as the US population proportion?

Example 2 - Left-handedness

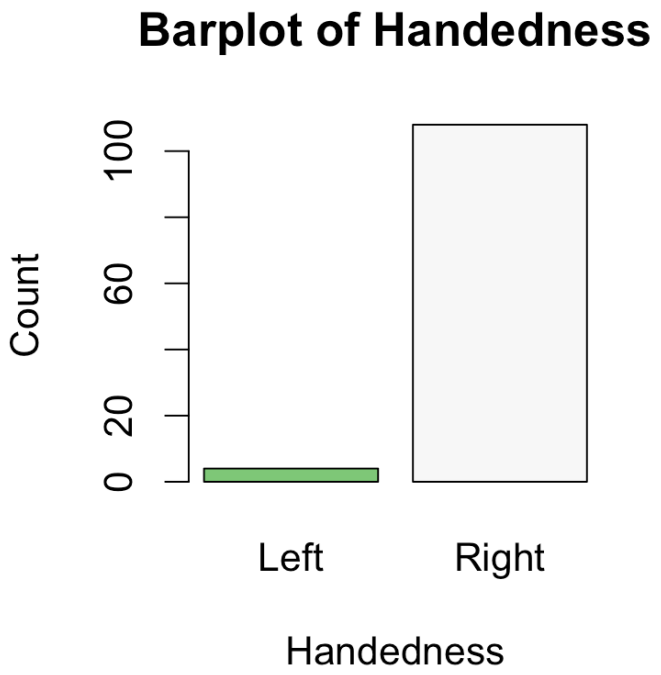
```
Survey$Handedness
```

```
##   [1] Right Right Right Right Right Right Right Right Right Right Right Right
##  [13] Right Right Right Right Right Right Right Right Right Right Right Right
##  [25] Right Right Right Right Right Right Right Right Right Right Right Right
##  [37] Right Right Right Right Right Right Right Right Right Right Right Right
##  [49] Right Right Right Right Right Right Right Right Right Right Left  Right
##  [61] Right Right Right Right Left  Right Right Right Right Right Right Right
##  [73] Right Right Right Right Right Right Right Right Right Right Right Right
##  [85] Right Right Right Right Right Right Right Right Right Right Right Right
##  [97] Left  Right Right Right Right Right Right Right Right Right Right Left
## [109] Right Right Right Right
## Levels: Left Right
```

- *Handedness* is a binary variable with two categories: *Left* and *Right*.
- It is not even a quantitative variable and thus far away from a Normal distribution or any skewed/bimodal distribution we have seen.
- Since the values are qualitative, we do not use mean and SD to describe the data.
- How to solve the problem?

Example 2 - Left-handedness

2019 Class	Left	Right	Total
Count	4	108	112
Proportion	3.6%	96.4%	100%



- Any binary variable can be transformed to a **dummy variable** with values 0 and 1.
- Transform *Handedness* to a dummy variable: Assign value 1 to *Left* and value 0 to *Right*.
- This new *Handedness* variable has 4 **1**'s and 108 **0**'s - all become numerical values.
- We can now calculate its mean

$$\bar{x} = \frac{0 + \cdots + 0 + 1 + 1 + 1 + 1}{112} = \frac{4}{112} = 0.036,$$

which is the proportion of left-handedness.

- The question becomes "is the sample mean/proportion 0.036 the same as the population mean/proportion 0.118?"

Bernoulli Distribution

- Denote population proportion as p and sample proportion as $\hat{p}$.

	Mean	Standard deviation	Proportion
Population Parameter	μ	σ	p
Sample Statistic	$\bar{x}$	s	$\hat{p}$

- A dummy variable X with values 0 and 1 follows a **Bernoulli distribution**

$$X \sim Bernoulli(p)$$

- The proportion of $X = 1$ (usually called success) is p.
- The proportion of $X = 0$ (usually called failure) is $1 - p$.
- The mean of X is p.
- The SD of X is $\sqrt{p(1-p)}$.

Normal distribution vs. Bernoulli distribution

	Normal distribution	Bernoulli distribution
Form	$X \sim N(\mu, \sigma)$	$X \sim Bernoulli(p)$
Values	Quantitative	1 or 0
Parameter(s)	μ, σ	p
Sample statistic(s)	$\bar{x}, s$	$\hat{p}$
Mean	μ	p
Standard deviation	σ	$\sqrt{p(1-p)}$

- According to CLT, the sample mean $\bar{x} \sim N(\mu, \sigma/\sqrt{n})$
- What is the distribution of $\hat{p}$?

Sampling distribution of a proportion

Normal approximation for proportions

Draw an SRS of size n from a large population having population proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. When n is large, the sampling distribution of $\hat{p}$ is approximately Normal with mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$:

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

As a rule of thumb, we will use this approximation for values of n and p that satisfy $np \geq 10$ and $n(1-p) \geq 10$.

Bernoulli distribution - Example

$$X \sim Bernoulli(p)$$

A classic example of Bernoulli distribution is **coin toss**. Suppose we have a fair coin. The probability of getting a head or tail is $p = 0.5$. Denote the outcome as X, where $X = 1$ for head and 0 for tail. The coin is tossed for 20 times and we get a total of 7 heads. Is this coin fair or not?

- $X \sim Bernoulli(0.5)$, population proportion $p = 0.5$, sample size $n = 20$, sample proportion $\hat{p} = 7/20 = 0.35$.
- Mean of X is $p = 0.5$; SD of X is $\sqrt{p(1-p)} = \sqrt{0.5 \times 0.5}$.
- Applying CLT, $\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N\left(0.5, \sqrt{\frac{0.5 \times 0.5}{20}}\right) = N(0.5, 0.11)$
- If the coin is fair, 68% of the proportions of head should fall within $[0.39, 0.61]$ and 95% fall within $[0.28, 0.72]$. The chance of getting $\hat{p} = 0.35$ is somewhat high. This is mostly likely a fair coin.

Bernoulli distribution - Example

What if we toss the coin for **100** times and get 35 heads?

- $X \sim Bernoulli(0.5), p = 0.5, n = \mathbf{100}, \hat{p} = 35/100 = 0.35$.
- When $n = 100$ instead of 20,

$$\hat{p} \overset{approx.}{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) = N\left(0.5, \sqrt{\frac{0.5 \times 0.5}{\mathbf{100}}}\right) = N(0.5, \mathbf{0.05})$$

- 68% of the proportions of head should fall within [0.45, 0.55] and 95% fall within [0.4, 0.6].
- The chance of getting $\hat{p} = 0.35$ is very low. This is mostly likely an unfair coin.

Example 2 - Left-handedness

- Proportion of left-handedness in the US population is 0.118.
- Proportion of left-handedness in STAT 11
  - **2016 class**: $10/122 = 0.082$
  - 2017 class: $7/94 = 0.074$
  - 2019 class: $4/112 = 0.036$
- Is the proportion of left-handedness in STAT 11 the same as the US population proportion?
- $X \sim Bernoulli(0.118), p = 0.118$
- 2016 class: $n = 122, \hat{p} \overset{approx.}{\sim} N(0.118, \sqrt{\frac{0.118(1-0.118)}{122}}) = N(0.118, 0.029)$
- The 95% interval $[0.060, 0.176]$ contains $\hat{p} = 0.082$.
- The proportion of left-handedness in STAT 11 2016 class is quite similar as the US population proportion.

Example 2 - Left-handedness

- Proportion of left-handedness in the US population is 0.118.
- Proportion of left-handedness in STAT 11
  - 2016 class: $10/122 = 0.082$
  - **2017 class**: $7/94 = 0.074$
  - 2019 class: $4/112 = 0.036$
- Is the proportion of left-handedness in STAT 11 the same as US population?
- $X \sim Bernoulli(0.118), p = 0.118$
- 2017 class: $n = 94, \hat{p} \overset{approx.}{\sim} N(0.118, \sqrt{\frac{0.118(1-0.118)}{94}}) = N(0.118, 0.033)$
- The 95% interval $[0.052, 0.184]$ contains $\hat{p} = 0.074$.
- The proportion of left-handedness in STAT 11 2017 class is quite similar as the US population proportion.

Example 2 - Left-handedness

- Proportion of left-handedness in the US population is 0.118.
- Proportion of left-handedness in STAT 11
  - 2016 class: $10/122 = 0.082$
  - 2017 class: $7/94 = 0.074$
  - **2019 class**: $4/112 = 0.036$
- Is the proportion of left-handedness in STAT 11 the same as US population?
- $X \sim Bernoulli(0.118), p = 0.118$
- 2019 class: $n = 112, \hat{p} \overset{approx.}{\sim} N(0.118, \sqrt{\frac{0.118(1-0.118)}{112}}) = N(0.118, 0.030)$
- The 95% interval $[0.058, 0.178]$ does NOT contain $\hat{p} = 0.036$.
- The proportion of left-handedness in STAT 11 2019 class is quite different from the US population proportion.

Summary

- Sampling distribution of a sample mean
  - When population distribution is Normal
  - When population distribution is not Normal
- Central Limit Theorem (CLT)
  - Definition
  - Examples
    - Sampling distribution of a proportion
      - Normal approximation
      - Examples