Lecture 8
Inference Statistics

Aug 06

Content

- Introduction to Inference Statistics
- Distributions:
  - Normal
  - Binomial
- Central limit theorem
- Confidence Intervals
- Hypothesis Testing

Distributions - Normal Distribution

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

- Normal Distribution
  - Sigma = SD
  - Mu = mean
- Probability of getting a score between 4.5 & 5.5 is the integration

$$\int_{4.5}^{5.5} P(x)$$

- The integration can be done numerically or there are look-up tables.
- Example: You have collected a sample of heights of lots of people.
  - The distribution curve will look like this.
  - Research Question: How many people are 5 inches taller that the average?
  - Solution: Take the area between the curve for heights between 4’9’’ and 5’11’’.

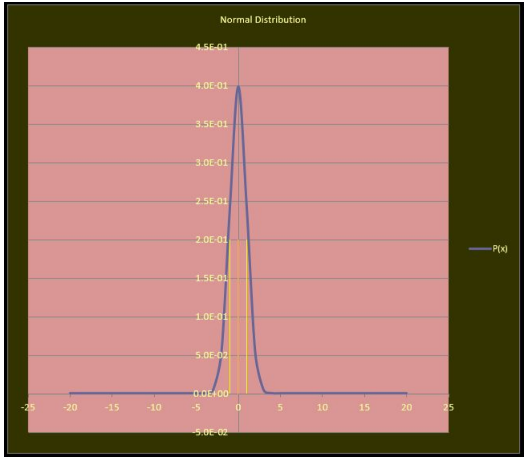
Distributions - Normal Distribution -2

- Can also be written as : $\frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{-1}{2}\left(\frac{x-\mu}{\sigma}\right)^2$
- Z-score: $\frac{x-\mu}{\sigma}$
    - How far from mean does ‘x’ lie.
- Re-writing: $\frac{1}{\sqrt{2\pi\sigma^2}}\left[exp\left(\frac{x-\mu}{\sigma}\right)^2\right]^{\frac{-1}{2}}$

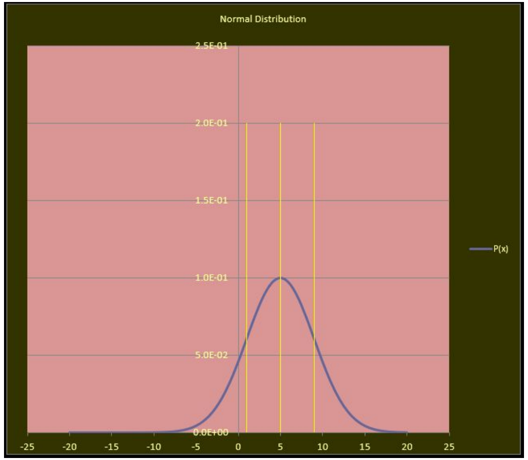
- Then $\frac{1}{\sqrt{2\pi\sigma^2 * exp(Z)}}$

Distributions - Normal Distribution -3

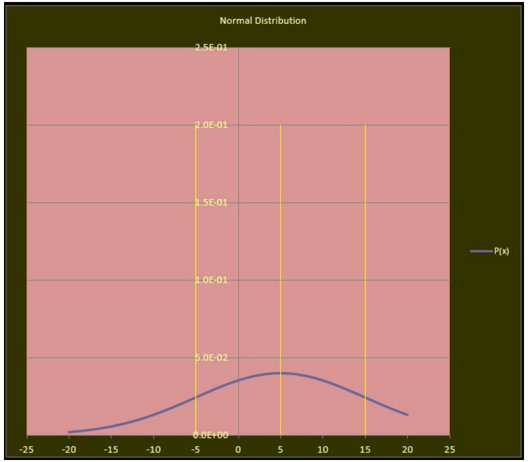
μ = 0
σ = 1



μ = 5
σ = 4



μ = 5
σ = 10



Distributions - Normal Distribution -4

- Standard Deviation
  - 1 SD: 68.3%
  - 2 SD: 95%
  - 3 SD: 99%

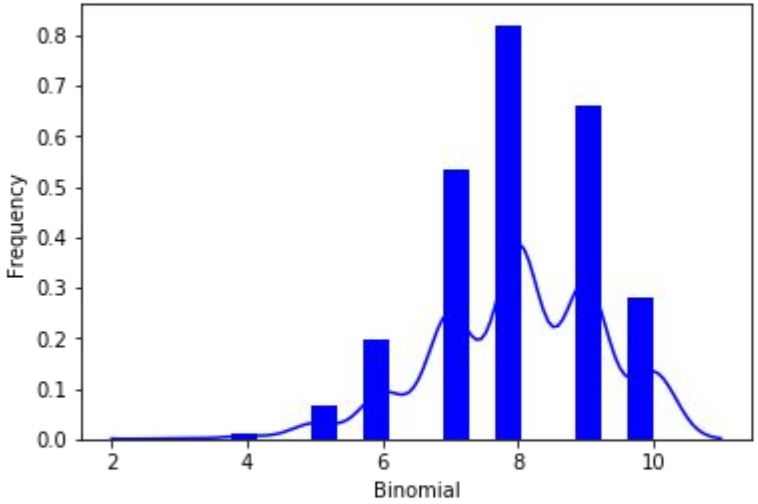
Distributions - Binomial -1

- X = NUmber of heads from flipping a coin 5 times
  - $\binom{N}{k}$ OR C_k^N
- $P(X=0) = \binom{5}{0} = \frac{5!}{0!(5-0)!} = \frac{5!}{5!}$
- $P(X=1) = \binom{5}{1} = \frac{5!}{1!(5-1)!} = \frac{5!}{4!}$

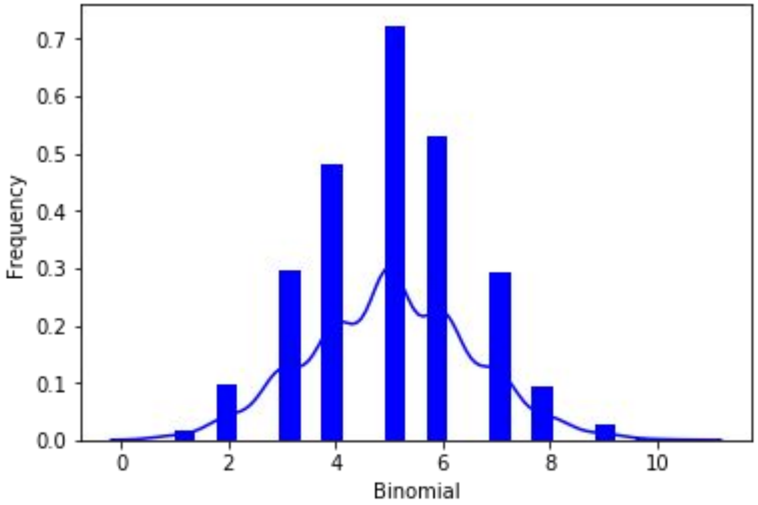
Distributions - Binomial -2

Spot the difference between these 2 plots.

Trials = 10, p=0.8



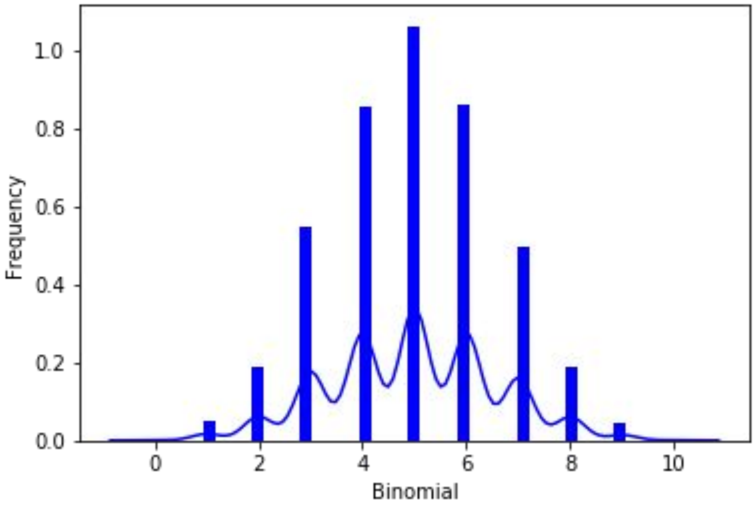
Trials = 10, p=0.5



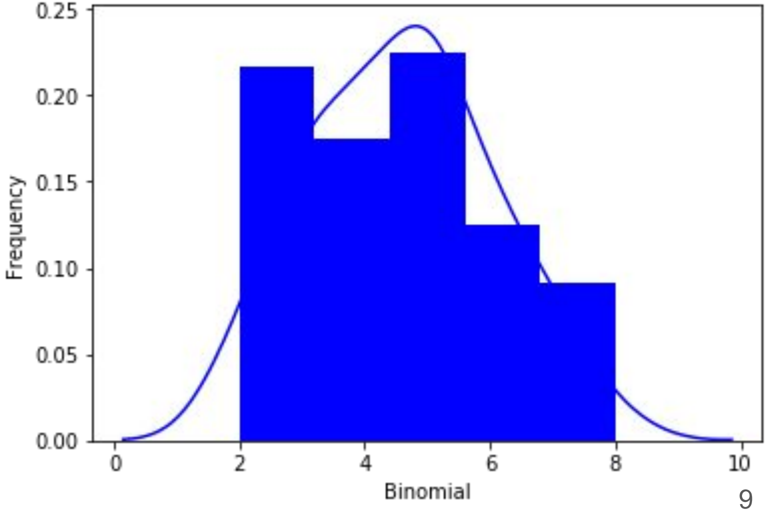
Distributions - Binomial -3

Spot the difference between these 2 plots.

Trials = 10, p=0.8, Sample size= 5000



Trials = 10, p=0.5, Sample size= 100



Central Limit Theorem

- **Observation**: Result from one trial of an experiment.
- **Sample**: Group of results gathered from separate independent trials.
- **Population**: Space of all possible observations that could be seen from a trial.

- Requirements on each **observation**:
  - is independent and obtained through a same process.
  - Draw from the same population distribution
  - In short, independent and identically distributed, i.i.d.

CLT -2

- Vs Law of Large numbers.
    - states that as the **size of a sample** is increased, the more accurate of an estimate the sample mean will be of the population mean.

    - CLT states about size and shape of the **sample means.**

- Example of Dice
    - Mean value of the sample:
        - (1 + 2 + 3 + 4 + 5 + 6) / 6 = 3.5
    - generate a specific number of random dice rolls (e.g. 50) between 1 and 6. (using randint)

Dataset

- National Health and Nutrition Examination Survey (NHANES).
- Assess the health and nutritional status of adults and children in the United States.

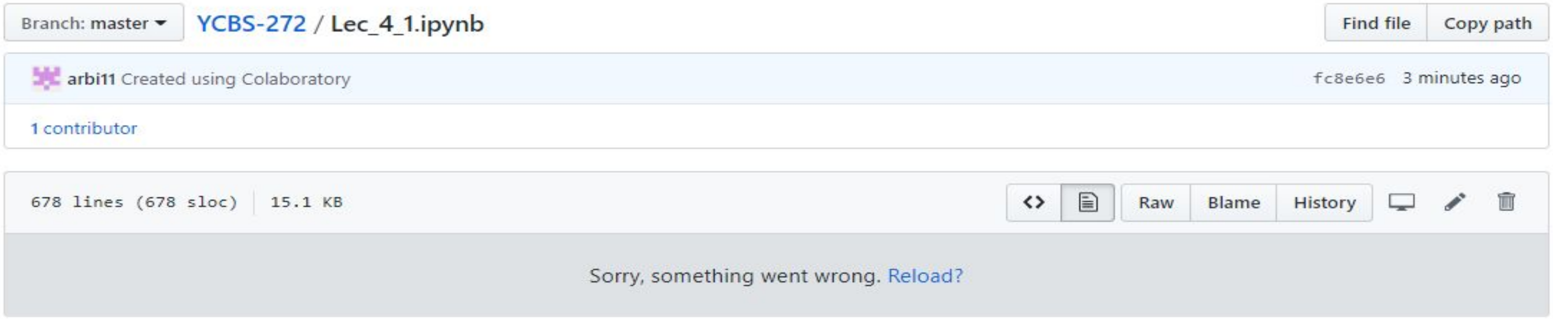
- The Cartwheel Dataset
- Contains: age, gender, glasses-wearing or not, height, weight, wingspan (arm length), completion, cartwheel distance, and overall cartwheel score.

- Nap or No Nap Dataset
- based on on a sleep study in toddlers, with basic confidence intervals and two independent means hypothesis testing.

Link for the notebook

https://github.com/arbi11/YCBS-272/blob/master/Lec_8_1.ipynb

If you see this error on github



Copy the link (of github page) and paste here:

https://nbviewer.jupyter.org/

Introduction to Inference methods

- Estimate some parameters of interest with confidence
    - test some theories about those parameters.
-
- We will be looking at estimating a population proportion with confidence.
- Estimating a population mean difference with confidence.

Convention Followed

Population mean	Mu (μ)	Sample proportion	$\hat{p}$
Sample mean	$\overline{x}$	normal distribution	N (μ, σ)
Population standard deviation	σ	Multiplier for forming 95% MoE	z* = 1.96
Sample standard deviation	s	Sampling distribution of the sample mean $\overline{x}$	Approx Normal
Population proportion	p	standard error of the sample mean*	$\sigma_{\overline{x}}$ or $se(\overline{x})$
standard error of the sample proportion	$\sigma_{\hat{p}}$ or $se(\hat{p})$	estimated standard error of the sample mean	$\hat{\sigma_{\overline{x}}}$ or $estimated\, se(\overline{x})$
estimated standard error of the sample proportion	$\hat{\sigma}_{\hat{p}}$ or $estimated\, se(\hat{p})$	Multiplier for forming 95% margin of error	z* = 1.96

Confidence Interval

- Why?
- How to calculate the confidence interval?
  - Best estimate plus or minus a margin of error.
    - $Best\ Estimate \pm Margin\ of\ Error$
  - Margin of Error (MoE) is defined as a few estimated standard errors.
    - Set the confidence level: 95%.
    - Implies, a significance level of 5%.
  - To create a 95% confidence interval can also be shown as:
    - $Population\ Proportion\ or\ Mean \pm (t - multiplier * \ Standard\ Error)$
  - The Standard Error is calculated differently for population proportion and mean:

$$Standard\ Error\ for\ Population\ Proportion = \sqrt{\frac{Population\ Proportion * (1 - Population\ Proportion)}{Number\ Of\ Observations}}$$

$$Standard\ Error\ for\ Mean = \frac{Standard\ Deviation}{\sqrt{Number\ Of\ Observations}}$$

Example

Research Question: "What proportion of parents reported they use a car seat for all travel with their toddler?"

Population: All parents with a toddler.

A sample of 659 parents with toddler was taken and asked if they used a toddler car seat for all their travel?

549 responded with ‘YES’.

$$Best\ Estimate \pm Margin\ of\ Error$$

Example: 95% Confidence Interval Calculations

$$Best\ Estimate \pm Margin\ of\ Error$$
$$\hat{p} \pm MoE$$

How to calculate sample mean?

n= 659 # Sample size

x= 540 # Responded ‘Yes’

$$\hat{p} = x/n = 540/649 = 0.85$$

Example: 95% Confidence Interval Calculations

$$Best\ Estimate \pm Margin\ of\ Error$$
$$\hat{p} \pm MoE$$

How to calculate MoE?

$$\hat{p} \pm 'few' * estimated\ \ se(\hat{p})$$

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \implies \hat{p} \pm 1.96 * \sqrt{\frac{0.85(1-0.85)}{659}}$$

$$0.85 \pm 0.0273 \implies (0.8227, 0.8773)$$

Confidence Interval - Two proportions

- Now we will look at estimating the difference in two population proportions with confidence.
- We are asking this question:
  - What is the difference of population proportions of parents that have their children (6-18) have had some swimming lesson.

- Population of interest:
  - Two; All parents of white/black children age 6-18.
- Parameter:
  - Difference in population proportions

$$\hat{p}_1 - \hat{p}_2$$

Confidence Interval - Two proportions - Example

- Significance level: 5%
- Survey:
  - 247 Black parents surveyed
    - 91 saying ‘YES’
  - 988 White parents surveyed
    - 543 said ‘YES’
- Confidence Interval

$$Best\ Estimate \pm Margin\ of\ Error$$

- Our Parameter $\hat{p}_1 - \hat{p}_2$

$$\hat{p}_1 - \hat{p}_2 \pm MoE \quad \Longrightarrow \quad \hat{p}_1 - \hat{p}_2 \pm 'few' * se(\hat{p}_1 - \hat{p}_2)$$

Confidence Interval - Two proportions - Example -2

$$\hat{p_1} - \hat{p_2} \pm 'few' * se(\hat{p_1} - \hat{p_2})$$

$$Best\ Estimate \pm 1.96 * \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} * \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$$

$$\hat{p_1} = \frac{543}{988} = 0.55 \qquad \hat{p_2} = \frac{91}{247} = 0.37$$

$$(0.55 - 0.37) \pm 1.96 * \sqrt{\frac{0.55(1-0.55)}{988} * \frac{0.37(1-0.37)}{247}}$$

$$0.18 \pm 0.0677 \Rightarrow (0.1123, 0.2477)$$

Confidence Interval - Two proportions - Conclusion

With 95% confidence, the population proportion of parents with white children, who have taken swimming lessons is 11.23 to 24.77% higher than that of population proportion of parents with black children.

Confidence Interval - One Mean

- Inference on population mean.
- Moving from Categorical to Quantitative.

- We will be using the Cart-wheel data.

```
df.describe()["CWDistance"]
```

```
count     25.000000
mean      82.480000
std       15.058552
min       63.000000
25%       70.000000
50%       81.000000
75%       92.000000
max      115.000000
Name: CWDistance, dtype: float64
```

→

n = 25 observations

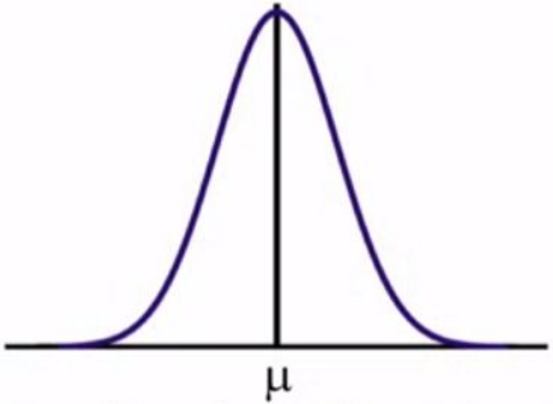
Minimum = 63 inches
Maximum = 115 inches

Mean = 82.48 inches
Standard Deviation = 15.06 inches

Confidence Interval - One Mean

Sampling Distribution of Sample Mean

If model for population of responses is approximately normal (or sample size is 'large' enough), distribution of sample mean is (approx.) normal.



All possible values of <u>sample mean</u>

standard error of the sample mean $= \dfrac{\sigma}{\sqrt{n}}$

Confidence Interval - One mean - Example

$$BestEstimate \pm MoE$$

$$BestEstimate = UnbiasedPointEstimate$$

$$Margin\ \ of\ \ Error =' few' * Estimated\ \ Standard\ \ Error$$

$$Best\ \ Estimate = \overline{x} \quad \Rightarrow \quad \overline{x} \pm ? * \left(\frac{s}{\sqrt{n}}\right)$$

t^* multiplier comes from a t-distribution with $n - 1$ degrees of freedom

95% confidence

$n = 25 \rightarrow t^* = 2.064$

$n = 1000 \rightarrow t^* = 1.962$

Confidence Interval - One mean - Example -2

Mean = 82.48 inches

Standard Deviation = 15.06 inches

n = 25 observations → t^* = 2.064

$$\overline{x} \pm t^*(\frac{s}{\sqrt{n}})$$

$$82.48 \pm 2.064 * (\frac{15.06}{\sqrt{25}})$$

Confidence Interval $(76.26 \ \ inches, 88.70 \ \ inches)$

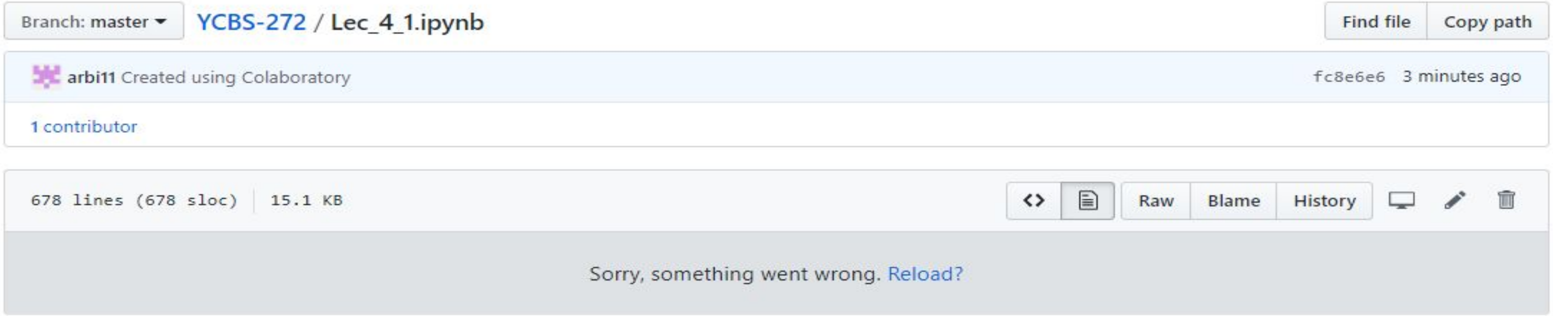
Confidence Interval - One mean - Interpretation

With 95% confidence, the population mean cartwheel distance for all adults is estimated to be between 76.26 inches and 88.70 inches.

Link for the notebook

https://github.com/arbi11/YCBS-272/blob/master/Lec_8_2_1_Confidence_Intervals.ipynb

If you see this error on github



Copy the link (of github page) and paste here:

https://nbviewer.jupyter.org/

Mean difference - Paired Data

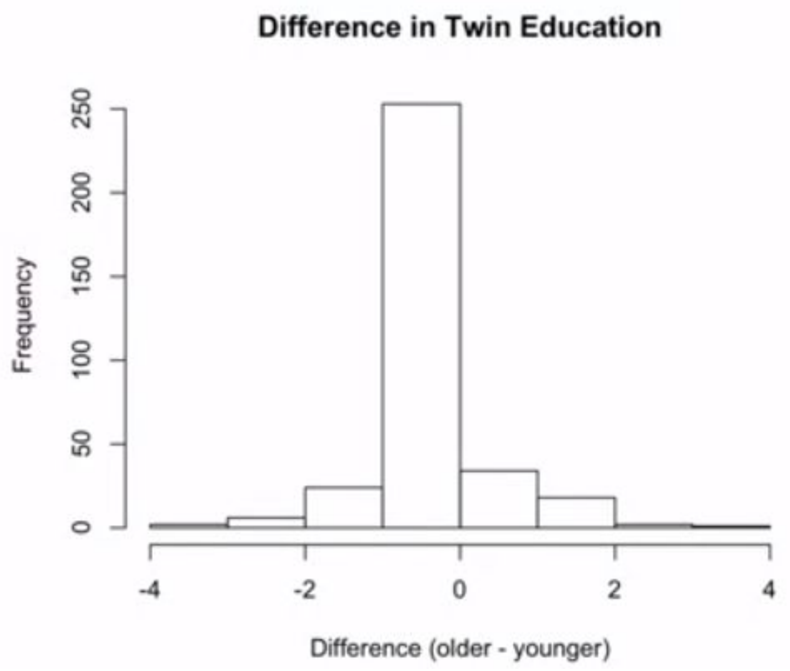
- **Parameter of interest :**
  - Population mean difference of self reported education level, μ_d

- **Difference** = Older - Younger

Research Objective:

Construct a 95% confidence interval for the **mean difference education** for that set of identical twins.

Mean difference - Paired Data -2



n = 340 observations

Minimum = -3.5 years
Maximum = 4 years
72.1% had a difference of 0 years

95% Confidence Interval

$$BestEstimate \pm MoE$$

$$BestEstimate = UnbiasedPointEstimate$$

$$Margin\ \ of\ \ Error =' few' * Estimated\ \ Standard\ \ Error$$

$$Best\ \ Estimate = \overline{x_d} \Rightarrow \overline{x_d} \pm ? * \left(\frac{\overline{s_d}}{\sqrt{n}}\right)$$

t^* multiplier comes from a t-distribution with $n - 1$ degrees of freedom

95% confidence

$n = 25 \rightarrow t^* = 2.064$

$n = 1000 \rightarrow t^* = 1.962$

Mean Difference Confidence Interval

Mean = 0.084 years
Standard Deviation = 0.76 years

n = 340 observations $\longrightarrow$ t* = 1.967

$$\bar{x}_{\text{d}} \pm t^{*}\left(\frac{s_{\text{d}}}{\sqrt{n}}\right)$$

Assumptions

- Random Sample of identical twins was collected.

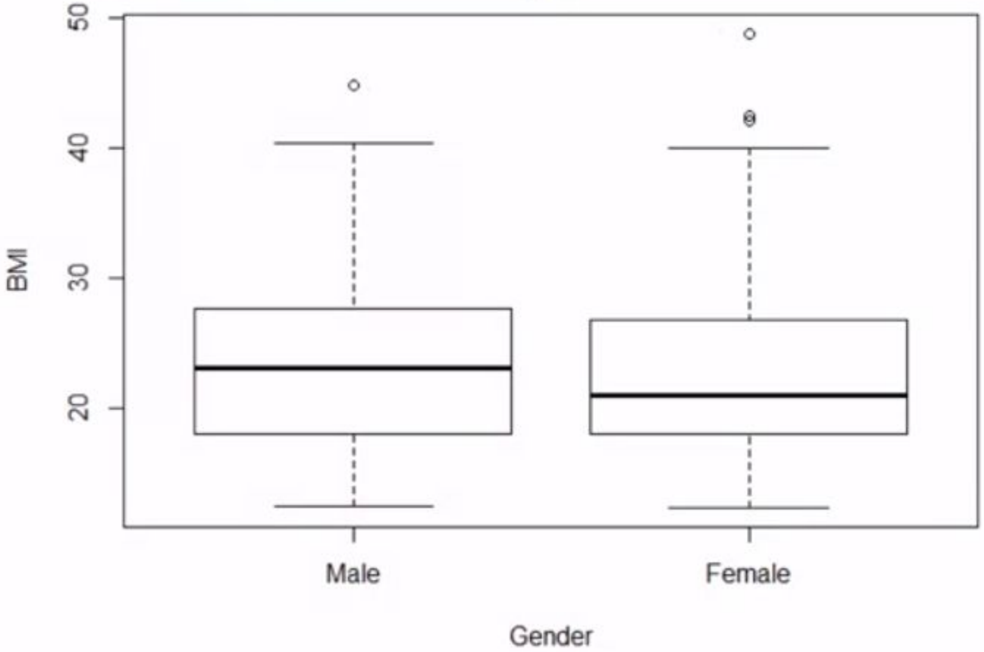
- Population of difference is normal.
  - Or a large sample size is needed to bypass this assumption.

CI - Difference in population mean

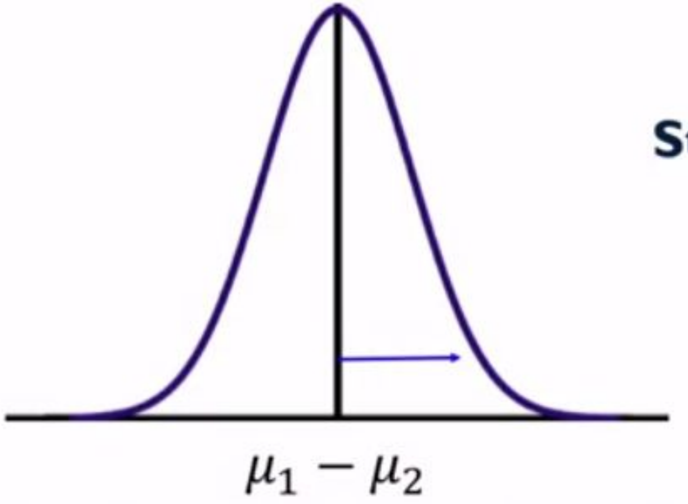
- Two independent groups.
- Research Question:
  - Do Male and Female in Mexican-American adults (18-29) have significantly different Body Mass Index?
- Parameter of Interest:
  - Body Mass Index (BMI)
  - $$\mu_1 - \mu_2$$

BMI Variable Summary

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
Min	12.5	12.4
Max	44.9	48.8
n	258	239



Sampling Distribution of the Difference in Two (Independent) Sample Means



Standard Error $= \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$

All possible values of <u>difference in sample means</u>

Confidence Interval Basics

Best Estimate ± Margin of Error

CI Approaches

- Pooled Approach
  - If the variances of both of our populations are close enough.

$$\sigma_1^2 = \sigma_2^2$$

- Unpooled Approach
  - If the variances of both of our populations are not close enough.

Unpooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\overline{x}_1 - \overline{x}_2) \quad \pm \quad t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Pooled Confidence Interval Calculations

Best Estimate ± Margin of Error

Difference in sample means ± "a few" · estimated standard error

$$(\bar{x}_1 - \bar{x}_2) \quad \pm \quad t^* \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

95% Confidence Interval - Example

	Male	Female
Mean	23.57	22.83
St. Dev.	6.24	6.43
n	258	239

$$(23.57 - 22.83) \quad \pm \quad 1.98\sqrt{\frac{(258-1)6.24^2+(239-1)6.43^2}{258+239-2}}\sqrt{\frac{1}{258}+\frac{1}{239}}$$

$$0.74 \quad \pm \quad 1.98 \; (6.33) \; (0.0898)$$

$$0.74 \quad \pm \quad 1.125 \longrightarrow \left(-0.385 \; {}^{kg}/_{m^2}, 1.865 \; {}^{kg}/_{m^2}\right)$$