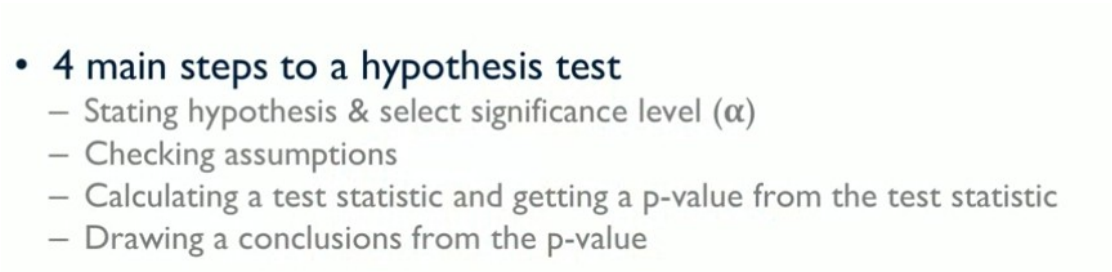
Teste de Hipóteses



One proportion

Proportions

One-Sample Z Proportion Test for $\hat{p} \Rightarrow$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Two-Sample Z Proportion Test

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, where \ldots \hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$$

Research Question

In previous years 52% of parents believed that electronics and social media was the cause of their teenager's lack of sleep. Do more parents today believe that their teenager's lack of sleep is caused due to electronics and social media?

Population - Parents with a teenager (age 13-18)

Parameter of Interest - p

Test for a significant increase in the proportion of parents with a teenager who believe that electronics and social media is the cause for lack of sleep.

Hypotheses

$H_0 : p = 0.52$

$H_a : p > 0.52$

Best Estimate of p is $\hat{p} = 0.56$

Where p is the population proportion of parents with a teenager who believe that electronics and social media is the cause of their teenager's lack of sleep

$\alpha = 0.05$

Survey Results

A random sample of **1018** parents with a teenager was taken and **56%** said they believe electronics and social media was the cause of their teenager's lack of sleep.

Assumptions

We need a **random sample** of parents
We also need a **large enough sample size** to ensure our distribution of sample proportions is normal

That is: **$n \cdot p$** be at least 10 ➡ **$n \cdot p_o$**
$n \cdot (1-p)$ be at least 10 ➡ **$n \cdot (1-p_o)$**

Random Sample ✓
$n \cdot p_o = 1018 \cdot (0.52) = 529$ ✓
$n \cdot (1-p_o) = 1018 \cdot (1-0.52) = 489$ ✓

Test Statistic

Best estimate - Hypothesized estimate
Standard error of estimate

$$\frac{\hat{p} - p_o}{s.e.}$$

$$s.e.(\hat{p}) = \sqrt{\frac{p \cdot (1-p)}{n}} \quad \Rightarrow \quad \boxed{s.e.(\hat{p}) = \sqrt{\frac{p_0 \cdot (1-p_o)}{n}}}$$

$$\frac{\hat{p} - p_o}{s.e.} \qquad \text{Null } s.e.(\hat{p}) = \sqrt{\frac{p_0 \cdot (1-p_o)}{n}}$$

$$Z = \frac{0.56 - 0.52}{0.0157}$$

Z = 2.555

Test Statistic Interpretation

Z = 2.555

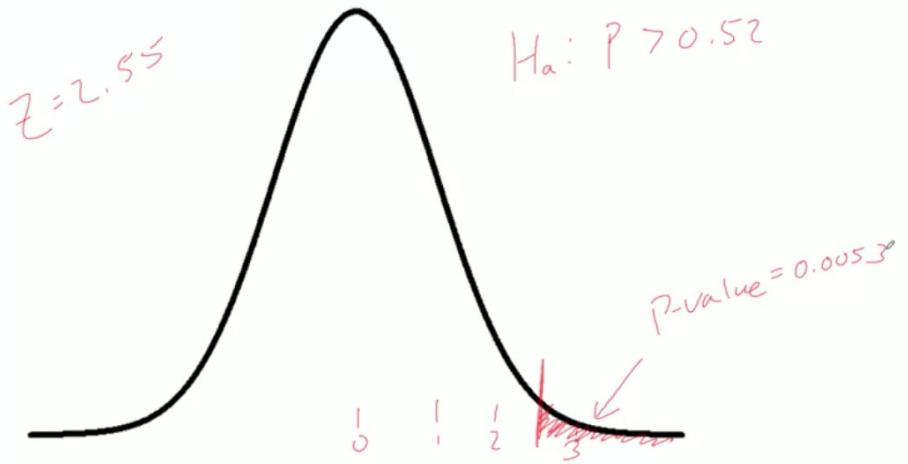
That means that our observed sample proportion is 2.555 null standard errors above our hypothesized population proportion

Test Statistic Distribution

- A Z test statistic is another random variable! It has a distribution.
- The Z test statistic will always follow a N(0,1)
- This is due to us centering and scaling our original data

Scales Data ← $\frac{\hat{p}-p_o}{s.e.(\hat{p})}$ → Centers Data

The P-Value



Conclusions

p-value = 0.0053 < α = 0.05

Reject the null hypothesis (H_0: p = 0.52)

There is sufficient evidence to conclude that the population proportion of parents with a teenager who believe that electronics and social media is the cause for lack of sleep is greater than 52%.

Two Proportions

Proportions

One-Sample Z Proportion Test for $\hat{p} \Rightarrow$

$$z = \frac{\hat{p} - \pi}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

Two-Sample Z Proportion Test

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, where \ldots \hat{p}_c = \frac{x_1 + x_2}{n_1 + n_2}$$

Research Question

Is there a significant difference between the population proportions of parents of black children and parents of Hispanic children who report that their child has had some swimming lessons?

Populations - All parents of black children age 6-18 and all parents of Hispanic children age 6-18

Parameter of Interest - $p_1 - p_2$

Test for a significant difference in the population proportions of parents reporting that their child has had swimming lessons at the 10% significance level.

Hypotheses

$H_0 : p_1 - p_2 = 0$
$H_a : p_1 - p_2 \neq 0$

$\alpha = 0.10$

Survey Results

- A sample of 247 parents of black children age 6 -18 was taken with 91 saying that their child has had some swimming lessons.
- A sample of 308 parents of Hispanic children age 6 -18 was taken with 120 saying that their child has had some swimming lessons.

Assumptions

We need to assume that we have **two independent random samples**.

We also need **large enough sample sizes** to assume that the distribution of our estimate is normal. That is, we need $n_1\hat{p}$, $n_1(1-\hat{p})$, $n_2\hat{p}$, and $n_2(1-\hat{p})$ to all be at least 10.

Checking Assumptions

$\hat{p} = (91+120)/(247+308) = 211/555 = 0.38$

$247(0.38) = 94$; $247(1-0.38) = 153$;
$308(0.38) = 117$; $308(1-0.38) = 191$

If this assumption is not met, we can perform different tests that bypass this assumption.

Best Estimate of the Parameter

$\hat{p}_1 = 91/247 = 0.37$ **1 = black**

$\hat{p}_2 = 120/308 = 0.39$ **2 = Hispanic**

$\hat{p}_1 - \hat{p}_2 = 0.37 - 0.39 = -0.02$

Test Statistic

Best estimate - Hypothesized estimate

Standard error of estimate

$$\frac{\hat{p}_1 - \hat{p}_2 - 0}{se(\hat{p})}$$

where $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

Test Statistic

$$\frac{\hat{p}_1 - \hat{p}_2 - 0}{se(\hat{p})}$$

where $se(\hat{p}) = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

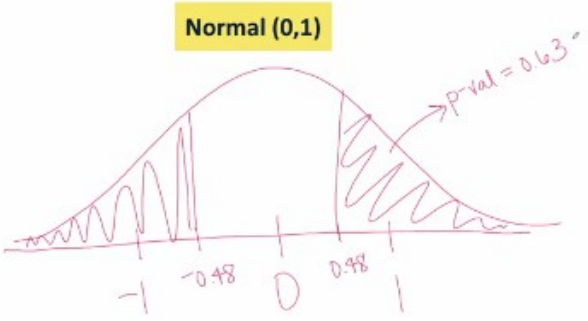
$$z = -0.02/0.041 = -0.48$$

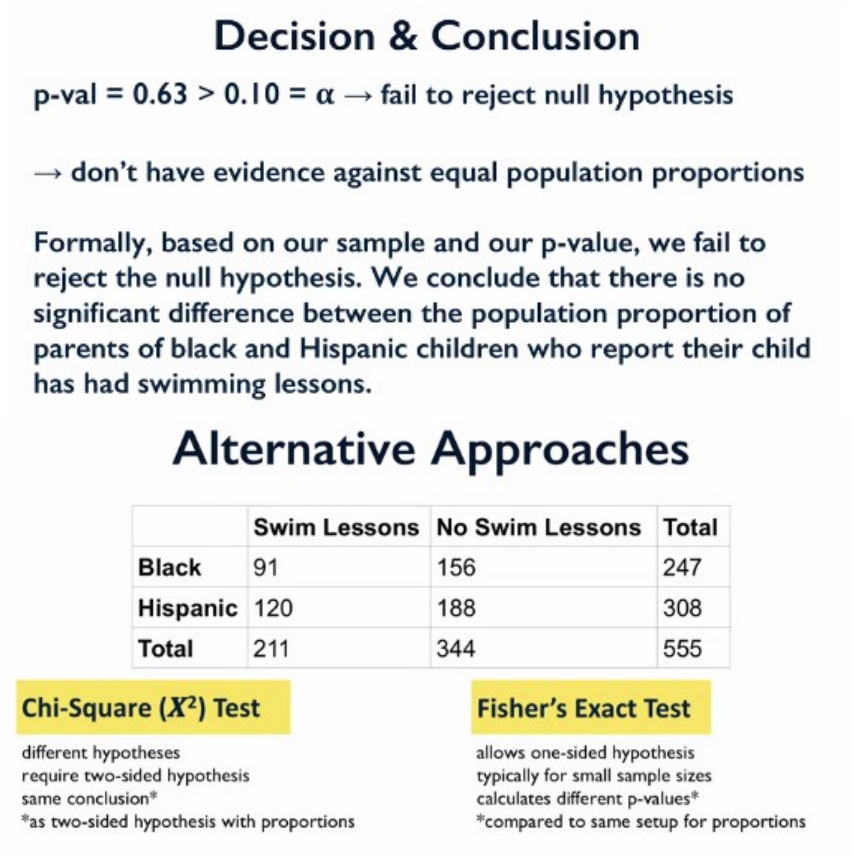
Test Statistic Interpretation

$$z = -0.48$$

That means that our observed difference in sample proportions is 0.48 estimated standard errors below our hypothesized mean of equal population proportions.

Test Statistic Distribution & P-value





Exemplo

Example of Two Sample Z Proportion Test (pooled)

Example: A car manufacturer aims to improve the quality of the products by reducing the defects and also increase the customer satisfaction. Therefore, he monitors the efficiency of two assembly lines in the shop floor. In line A there are 18 defects reported out of 200 samples. While the line B shows 25 defects out of 600 cars. At α 5%, is the differences between two assembly procedures are significant?

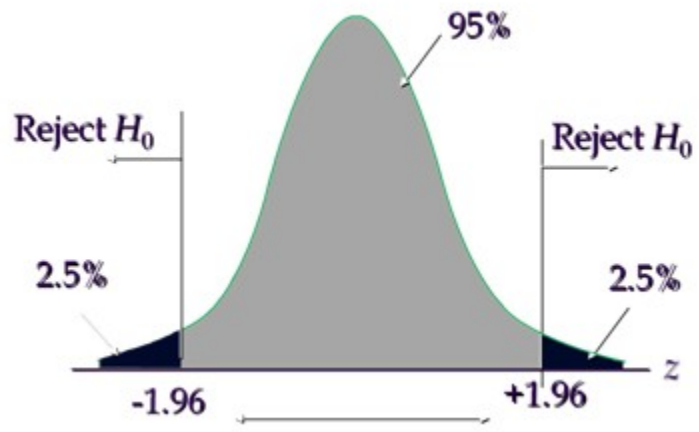
Define Null and Alternative hypothesis

- Null Hypothesis: Two proportions are the same
- Alternative Hypothesis: Two proportions are not the same

α=0.05

State decision rule

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756



Critical value is ±1.96, hence reject the null hypothesis if the calculated value is less than -1.96 or greater than +1.96

Calculate Test Statistic

- Line A= $\hat{p}_1$=18/200= 0.09 = 9%
- Line B= $\hat{p}_2$= 25/600 = 0.0416 = 4.16%

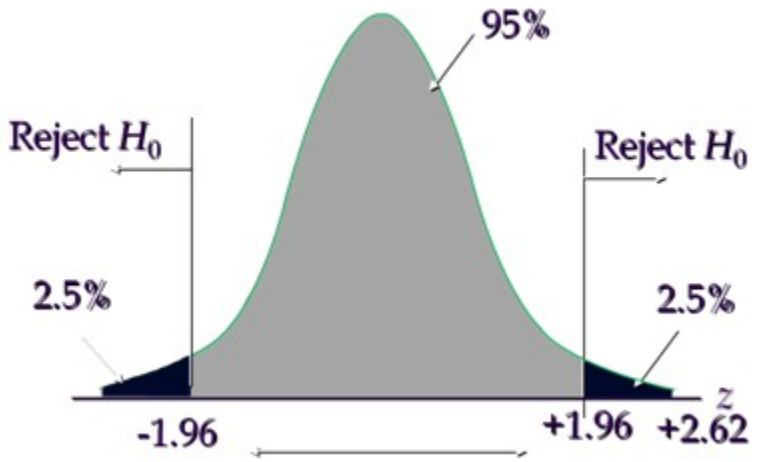
$$p_0 = \frac{X_1+X_2}{n_1+n_2}$$

- First compute p_0 = (18+25)/(200+600)=43/800=0.0537 =5.37%
- Now $\hat{p}_1 - \hat{p}_2$ = 0.09 – 0.0416 = 0.0484
- $p_0 * (1 - p_0)$ = 0.0537 * (1-0.0537) = 0.0537 * 0.9463 = 0.0508
- And (1/n1) + (1/n2) = 0.006667

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{p_0(1-p_0)(\frac{1}{n_1}+\frac{1}{n_2})}}$$

$$z = \frac{0.09 - 0.0416}{\sqrt{0.0537(1-0.0537)(\frac{1}{200}+\frac{1}{600})}}$$

- So, Z = (0.0484)/ SQRT ((0.0508) * (0.006667))
- Z = (0.0484)/ SQRT (0.000339)) = (0.0484)/ (0.018406) = 2.62



Interpret the results: Compare Z_{calc} to $Z_{critical}$. In hypothesis testing, a critical value is a point on the test distribution compares to the test statistic to determine whether to reject the null hypothesis. Calculated test statistic value 2.62 and it is in critical region, hence reject the null hypothesis, so, there is a significant difference in two line assembly procedures.

https://sixsigmastudyguide.com/two-sample-test-of-proportions/

Fisher Exact Test

	Men	**Women**	*Row total*
Studying	**1**	**9**	*10*
Not-studying	**11**	**3**	*14*
Column total	*12*	*12*	*24*

	Men	**Women**	*Row Total*
Studying	***a***	***b***	*a + b*
Non-studying	***c***	***d***	*c + d*
Column Total	*a + c*	*b + d*	*a + b + c + d (=n)*

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\,(c+d)!\,(a+c)!\,(b+d)!}{a!\;b!\;c!\;d!\;n!}$$

$$p = \binom{10}{1}\binom{14}{11}/\binom{24}{12} = \frac{10!\;14!\;12!\;12!}{1!\;9!\;11!\;3!\;24!} \approx 0.001346076$$

The formula above gives the exact hypergeometric probability of observing this particular arrangement of the data, assuming the given marginal totals, on the null hypothesis that men and women are equally likely to be studiers. To put it another way, if we assume that the probability that a man is a studier is p, the probability that a woman is a studier is also p.

Fisher showed that to generate a significance level, we need consider only the cases where the marginal totals are the same as in the observed table, and among those, only the cases where the arrangement is as extreme as the observed arrangement, or more so.

	Men	**Women**	*Row Total*
Studying	**0**	**10**	*10*
Non-studying	**12**	**2**	*14*
Column Total	*12*	*12*	*24*

For this table (with extremely unequal studying proportions) the probability is $p = \binom{10}{0}\binom{14}{12}/\binom{24}{12} \approx 0.000033652$.

In order to calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the null hypothesis is true, we have to calculate the values of p for both these tables, and add them together.
Em geral, você vai criando essas tabelas até uma das células atingir zero.
This gives a one-tailed test, with p approximately 0.001346076 + 0.000033652 = 0.001379728.
The smaller the value of p, the greater the evidence for rejecting the null hypothesis; so here the evidence is strong that men and women are not equally likely to be studiers.

https://en.wikipedia.org/wiki/Fisher%27s_exact_test

Outro exemplo:

Underweight	Normal	Total
8	11	19
3	11	14
11	22	33

KB	N	
9	10	19
2	12	14
11	22	33

KB	N	
10	9	19
1	13	14
11	22	33

KB	N	
11	8	19
0	14	14
11	22	33

- $\frac{(a+b)!\,(a+c)!\,(b+d)!\,(c+d)!}{N!\,a!\,b!\,c!\,d!}$

- $p1 = \frac{19!14!11!22!}{33!8!11!3!11!}$

$= \frac{4.758 \times 10^{56}}{3.3471 \times 10^{57}} = 0.142$

- Create 3 more extreme tables by deducting 1 from the smallest value. Continue to do so till the cell becomes zero;

KB	N	
9	10	19
2	12	14
11	22	33

KB	N	
10	9	19
1	13	14
11	22	33

KB	N	
11	8	19
0	14	14
11	22	33

- p2 = 0.0434
  p3 = 0.00668
  p4 = 0.00039

- Total p = 0.142+0.0434+0.00668+0.00039
  = 0.19247
- This is the p value for single-tailed test. To make it the p value for 2 tailed, times the value with 2; p = 0.385.
- p is larger than 0.05, therefore the null hypothesis is not rejected.
- There is no association between occupation and UW ;-)

https://www.youtube.com/watch?v=6xZHF8Y7w_8

P-values and P-Hackings

P-value é uma medida de “surpresa”. Maior p-value, menor a surpresa.

P-hacking é um termo amplo usado em pesquisa científica para descrever vários tipos de manipulação comumente empregados na análise de dados que levam a resultados estatisticamente significativos.

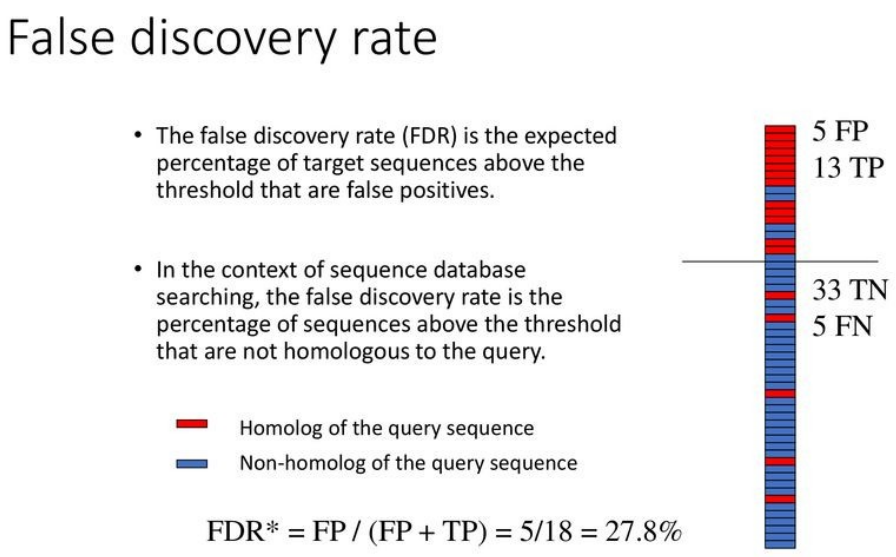
Bonferroni Correction

Essentially involves multiplying all p-values by the number of tests that were performed. For example, if we conduct 5 hypothesis tests, and one of them yields a p-value of 0.02, then we should adjust this p-value to 0.1 (= 0.02 * 5). Thus, this test which would have been deemed to be “statistically significant” if conducted in isolation will be not be seen as such following the Bonferroni adjustment.

FDR (False Discovery Rate)

The type I error rate is controlled by the researcher (say at 5%), but this only represents the risk of drawing a false conclusion from a single test (False Positive). In recent years, the notion of the “false discovery rate” (FDR) has been advanced to understand how often the conclusions drawn in a research process involving multiple formal inferential procedures are mistaken. Focusing on hypothesis tests, we can consider a situation where, for example, five tests are to be conducted. If two of the underlying null hypotheses are false, but the power to reject them (Power = 1-False Negative => FN = 1-0,2=0,8) is low (say it is only 20%), then around 25% of all the rejected null hypotheses were incorrectly rejected. This shows how the FDR is different than the type I error rate, which remains controlled here at 5%.

FDR = FP / (FP + TP)



Power

Power is often defined in a narrow sense as the probability of rejecting the null hypothesis when the null hypothesis is false. Loosely speaking, this is the probability of not making a type II error.

One Mean

Summary

- Hypothesis Tests are used to put theories about a parameter of interest to the test ~ parameter = population mean
- Basic Steps:
  - State hypotheses (and select significance level)
  - Examine results, check assumptions, summarize via test statistic
  - Convert test statistic to P-value
  - Compare P-value to significance level to make decision
- Assumptions for One-sample (t) Test for Population mean
  - Data considered a random sample
  - Population of responses is normal (else n large helps)

Research Question

Is the ***average*** cartwheel distance for adults more than 80 inches?

Population: All adults

Parameter of Interest: population mean cartwheel distance μ

Perform a one-sample test regarding the value for the mean cartwheel distance for the population of all such adults.

Step 1: Define the Null and Alternative

- Null: Population mean CW distance (μ) is 80 inches
- Alternative: Population mean is **_greater than (>)_** 80 inches

More compact notation:

- H_0: $\mu = 80$
- H_a: $\mu > 80$

where μ represents the population mean cartwheel distance (inches) for all adults

Significance Level = 5%

Step 2: Examine Results, Check Assumptions, Summarize Data via Test Statistic

```
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
```

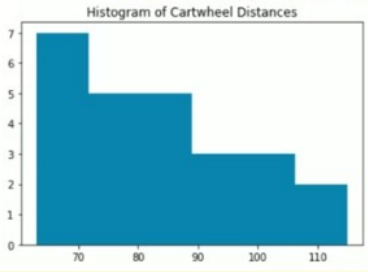
n = 25 observations

Minimum = 63 inches
Maximum = 115 inches

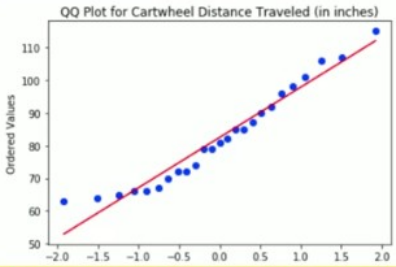
Mean = 82.48 inches
Standard Deviation = 15.06 inches

Assumptions:

- Sample of **CW Distance measurements** considered a **simple random sample**
- **Normal** distribution for CW Distances in population (not as critical given large sample size, but still graph our data!)



Histogram and Normal Q-Q Plot suggest modest deviations from **normality**



Note: reasonable sample size + CLT … normality assumption not so crucial

- Is sample mean of 82.48 inches *significantly greater than* hypothesized mean of 80 inches?

$H_0: \mu = 80$

$H_a: \mu > 80$

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

estimated **standard error of the sample mean** $= \frac{s}{\sqrt{n}}$

Test Statistic: Assuming sampling distribution of sample mean is normal,

$$t = \frac{best\ estimate - null\ value}{estimated\ std\ error} = \frac{\bar{x} - 80}{\frac{s}{\sqrt{n}}}$$

$$= \frac{82.48 - 80}{\frac{15.06}{\sqrt{25}}} = \frac{2.48}{3.012} = \mathbf{0.82}$$

Test Statistic Interpretation

$$t = \frac{best\ estimate - null\ value}{estimated\ std\ error} = \frac{\bar{x} - 80}{\frac{s}{\sqrt{n}}} = \frac{82.48 - 80}{\frac{15.06}{\sqrt{25}}} = \mathbf{0.82}$$

Our **sample** mean is **only 0.82** (*estimated*) **standard errors above null value of 80 inches**

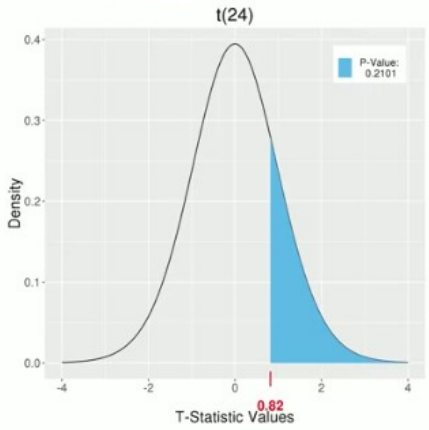
Step 3: Determine P-Value

- If null hypothesis was true, would a test statistic value of only t = 0.82 be unusual enough to reject the null?
- **P-value** = Probability of seeing test statistic of 0.82 or *more extreme* assuming the null hypothesis is true.
- If null hypothesis was true, t test statistic follows a Student t distribution with degrees of freedom n – 1 = 25 – 1 = 24.
- Since we have a one tailed test to the right
  → More extreme measured to the right (upper tail).

Step 3: Determine P-Value

P-value = 0.21

If population mean CW distance was really was 80 inches, then observing a sample mean of 82.48 inches (i.e. a t statistic of 0.82) or larger is **quite likely**.



Step 4: Make a Decision about the Null

Since our P-value is much bigger than 0.05 significance level, weak evidence against the null
→ we **fail to reject the null!**

90% Confidence Interval Estimate

Mean = 82.48 inches
Standard Deviation = 15.08 inches
n = 25 observations → $t^* = 1.711$

$$\bar{x} \pm t^* \left(\frac{s}{\sqrt{n}}\right)$$

$$82.48 \pm 1.711\left(\frac{15.06}{\sqrt{25}}\right)$$

$$82.48 \pm 1.711(3.012)$$

$$82.48 \pm 5.15$$

(77.33 inches, 87.63 inches)

Note: 80 inches is IN confidence interval of reasonable values for population mean CW distance

What if Normality Doesn't Hold?

- Not convinced that CW Distance follows a normal distribution in the population?
  → **non-parametric test** that does not assume normality

- Non-parametric analog of the one sample t-test
  = **Wilcoxon Signed Rank Test**
  ~ uses median to examine location of distribution of measurements

What if Normality Doesn't Hold?

Wilcoxon Signed Rank Test Result: p-value >> 0.05

Fail to reject the null that population median CW distance 80 inches

Conclusion is robust to potential violations of normality!

For the population of interest (all adults)
~ regardless of assumptions made and inference approach used
→ There is **not** sufficient evidence to support that the population mean CW distance is more than 80 inches

Difference in Means for Paired Data

20 homes remodeling their kitchens, requesting cabinet quotes from 2 suppliers

Is there an average difference in cabinet quotes from these two suppliers?

Research Question

Is there an average difference between the cabinet quotes from the suppliers?

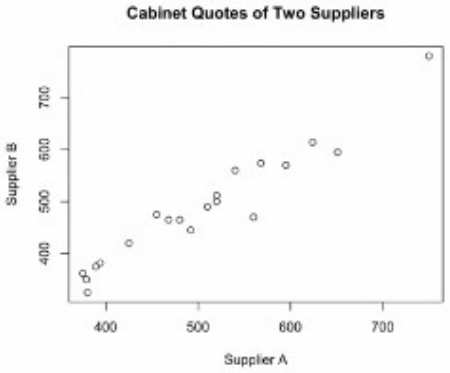
Populations - All houses

Parameter of Interest - Population mean difference of cabinet quotes μ_d (Supplier A - Supplier B)

Test for a significant mean difference in cabinet quotes at the 5% significance level.

Cabinet Data

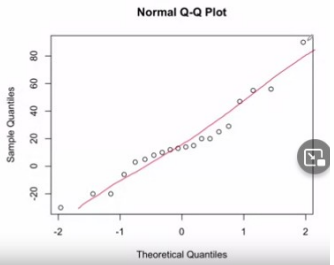
Supplier A	Supplier B	Difference
$380	$325	$55
$560	$470	$90
$425	$420	$5
$389	$375	$14
$568	$574	-$6
$651	$595	$56



Assumptions

We need to assume that we have a **random sample of differences**, i.e. a random sample of houses.

We also need the **population of differences to be normally distributed**. We can get around this assumption if we have a large sample size (about 25+).



n = 20 observations

Minimum = -$30
Maximum = $90

Median = $13.50
Mean = $17.30
Standard Deviation = $28.49

Test Statistic

Assuming the sampling distribution of the sample mean difference is normal,

$$t = \frac{\text{best estimate - hypothesized estimate}}{\text{estimated standard error of estimate}}$$

Best estimate - Hypothesized estimate

Estimated standard error of estimate

$$t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}} = \frac{17.30 - 0}{28.49 / \sqrt{20}}$$

n = 20 observations

Mean = \$17.30
SD = \$28.49

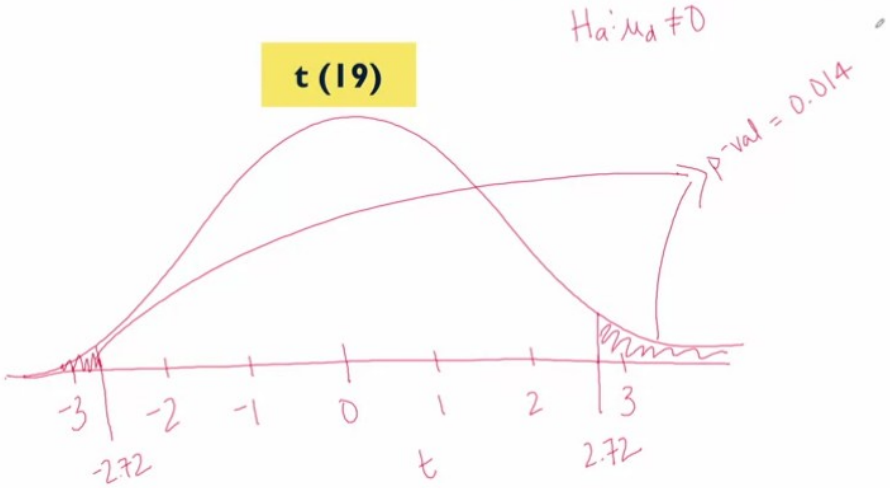
$$t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}} = \frac{17.30 - 0}{28.49 / \sqrt{20}} = \frac{17.30}{6.37}$$

$$= 2.72$$

$$t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}} = 2.72$$

Our observed mean difference is 2.72 (estimated) standard errors above our null value of 0.

Test Statistic Distribution & P-value



Decision & Conclusion

p-val = 0.014 < 0.05 = α → reject null hypothesis

→ have evidence against mean difference in cabinet quotes is 0

Formally, based on our sample and our p-value, we reject the null hypothesis. We conclude that the mean difference of cabinet quote prices for Suppliers A less B is **significantly different** from 0.

95% Confidence Interval

Mean = \$17.30
Standard deviation = \$28.49
$n = 20 \rightarrow t^* = 2.093$

$$\bar{x}_d \pm t^* \left(\frac{s_d}{\sqrt{n}}\right)$$

\$17.30 ± 2.093 (\$28.49/√20)
\$17.30 ± 2.093 (\$6.37)
\$17.30 ± \$13.33
(\$3.97, \$30.63)

Note 0 is NOT in our range of reasonable values for mean difference in cabinet prices.

Wilcoxon Signed Rank Test

If *normality* doesn't hold, we can use the Wilcoxon Signed Rank Test to test for the median.

p-val = 0.020

Again, we reject H_0 and conclude that the median difference in the cabinet quotes, Supplier A less B, is different from 0.

Difference in Means for Independent Groups (t-test)

Two Sample t Test for Independent Samples

Unpooled (assuming distributions have different variances)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Two-Sample Means t Test (unpooled/separate) with df

Pooled (assuming both distributions have same variance)

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Two-Sample Means t Test (pooled) with $(n_1 + n_2 - 2)$ df

Interpreting the Output: Pooled vs. Unpooled Variance

Before you can interpret your test statistic and reach a conclusion, you must determine whether to use the **pooled** or **unpooled** variances test statistic. If we can assume the two samples have *equal variances*, then we use the *pooled* t*. If, on the other hand, we determine that the two samples have *unequal variances*, then we must use the *unpooled* t*.

Research Question

Considering Mexican-American adults (ages 18 - 29) living in the United States, do males have a significantly higher mean Body Mass Index than females?

- **Population:** Mexican-American adults (ages 18 - 29) in the U.S.
- **Parameter of Interest ($\mu_1 - \mu_2$):** Body Mass Index or BMI (kg/m^2)

Task: Perform an independent samples t-test regarding the value for the difference in mean BMI between males and females.

Steps to Perform a Hypothesis Test

1. Define null and alternative hypotheses
2. Examine data, check assumptions, and calculate test statistic
3. Determine corresponding p-value
4. Make a decision about null hypothesis

Step 1: Define Hypotheses

Null: There is no difference in mean BMI

Alternative: There is a significant difference in mean BMI

(Both statements are for the specified populations)

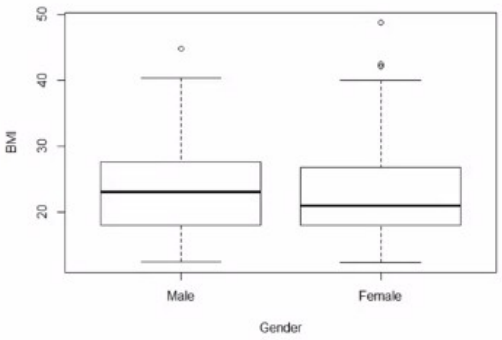
$H_0: \mu_1 = \mu_2$ (or $\mu_1 - \mu_2 = 0$)

$H_a: \mu_1 \neq \mu_2$ (or $\mu_1 - \mu_2 \neq 0$)

Significance Level = 5%

Step 2: Examine Data

	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



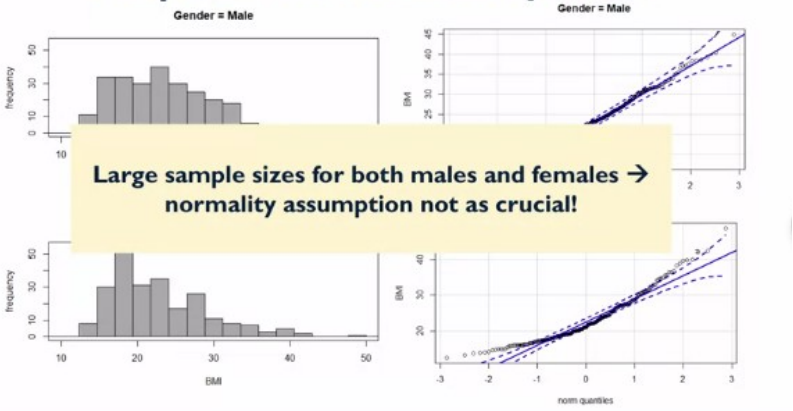
Step 2: Check Assumptions

Samples are considered simple random samples

Samples are independent from one another

Both populations of responses are approximately normal (or sample sizes are both 'large' enough)

Step 2: Check Assumptions



Large sample sizes for both males and females → normality assumption not as crucial!

Step 2: Calculate Test Statistic

$H_0: \mu_1 - \mu_2 = 0$ vs. $H_a: \mu_1 - \mu_2 \neq 0$

Best Estimate: $\bar{x}_1 - \bar{x}_2 = 23.57 - 22.83 = 0.74$

Is our sample mean difference of 0.74 kg/m^2 significantly different than 0?

Test Statistic

A measure of how far our sample statistic is from our hypothesized population parameter, in terms of estimated standard errors

The further away our sample statistic is, the less confident we'll be in our null hypothesized value

Step 2: Calculate Test Statistic

$$t = \frac{best\ estimate\ - null\ value}{estimated\ standard\ error}$$

Pooled Approach

The variance of the two populations are assumed to be equal $(\sigma_1^2 = \sigma_2^2)$

Unpooled Approach

The assumption of equal variances is dropped

Pooled Approach:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Pooled Approach:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

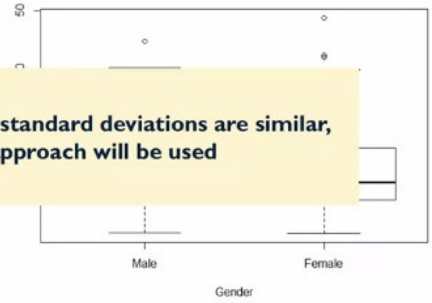
Unpooled Approach:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Pooled or Unpooled?

	Male	Female
Mean		
St. De		
Min		
Max		
n	258	239

Because the IQR's and standard deviations are similar, the pooled approach will be used



Step 2: Calculate Test Statistic

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(23.57 - 22.83)}{\sqrt{\frac{(257)6.24^2 + (238)6.43^2}{495}}\sqrt{\frac{1}{258} + \frac{1}{239}}}$$

$$t = \frac{0.74}{0.0898 * 6.332} = \mathbf{1.30}$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{(n_1}} \qquad \frac{(23.57 - 22.83)}{\frac{1}{58} + \frac{1}{239}}}$$

Our difference in sample means is only 1.30 (*estimated*) standard errors above the null difference of 0 kg/m²

$$t = \frac{}{0.0898 * 6.332} = \mathbf{1.30}$$

Degrees of Freedom for Independent Samples t-Test

$$df = (n_1 - 1) + (n_2 - 1)$$

$$df = (25 - 1) + (20 - 1) = 43$$

Step 3: Determine p-value

$$t = 1.30$$

If the null hypothesis ($\mu_1 - \mu_2 = 0$) were true, would a test statistic value of 1.30 be unusual enough to reject the null?

p-value: assuming the null hypothesis is true, it is the probability of observing a test statistic of 1.30 or more extreme

Using a $t(df)$ distribution where $df = n_1 + n_2 - 2$

Our alternative hypothesis is two-sided ($\mu_1 - \mu_2 \neq 0$) so we will check both the upper and lower tail

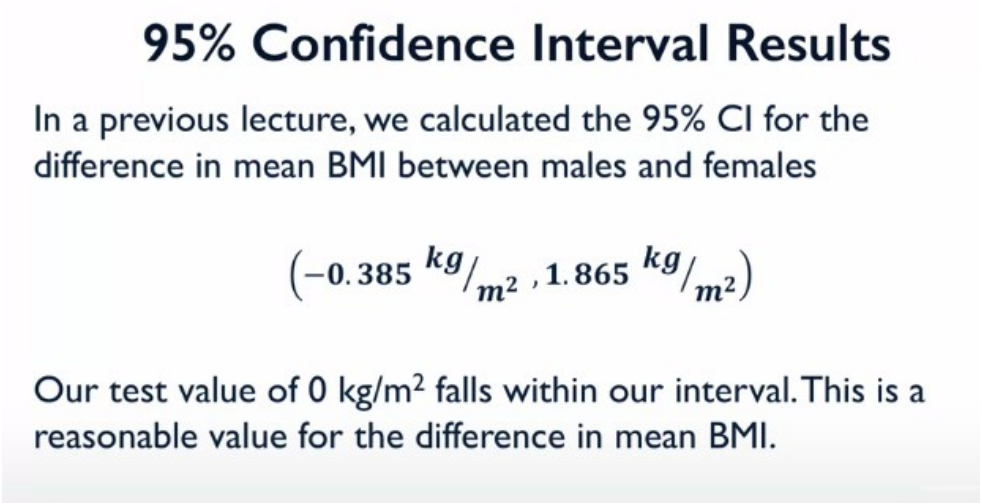
p-value = 0.19

If the difference in population mean BMI between males and females was really 0 kg/m², then observing a difference in sample means of 0.74 kg/m² (i.e. a t-statistic of 1.30) or more extreme is **fairly likely**.

Step 4: Make a Decision

Our p-value is larger than the 0.05 significance level, which means there is weak evidence against the null.

Thus, we **fail to reject the null!**



Considerations

One Population Proportion

Sample can be considered a simple random sample

Large enough sample size ()

- Confidence Interval: At least 10 of each outcome ()
- Hypothesis Test: At least 10 of each outcome ()

Two Population Proportions

Samples can be considered two simple random samples

Samples can be considered independent of one another

Large enough sample sizes ()

- Confidence Interval: At least 10 of each outcome ()
- Hypothesis Test: At least 10 of each outcome () - Where (the common population proportion estimate)

One Population Mean

Sample can be considered a simple random sample

Sample comes from a normally distributed population

- This assumption is less critical with a large enough sample size (application of the C.L.T.)

One Population Mean Difference

Sample of differences can be considered a simple random sample

Sample of differences comes from a normally distributed population of differences

- This assumption is less critical with a large enough sample size (application of the C.L.T.)

Two Population Means

Samples can be considered a simple random samples

Samples can be considered independent of one another

Samples each come from normally distributed populations

- This assumption is less critical with a large enough sample size (application of the C.L.T.)

Populations have equal variances – pooled procedure used

- If this assumption cannot be made, unpooled procedure used