Population Mean with Confidence

Research Question

What is the **average** cartwheel distance (in inches) for adults?

- **Population**: All adults
- **Parameter of Interest**: population mean cartwheel distance μ

Construct a 95% confidence interval for the mean cartwheel distance for the population of all such adults.

Cartwheel Distance Summary

```
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
```

→

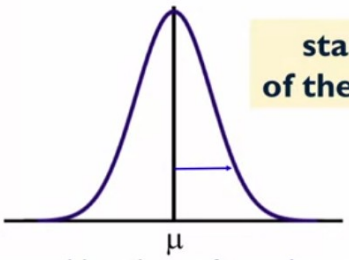
n = 25 observations

Minimum = 63 inches
Maximum = 115 inches

Mean = 82.48 inches
Standard Deviation = 15.06 inches

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 sample mean

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

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

O erro padrão pode ser calculado quando se conhece o desvio padrão populacional. Quando este é desconhecido, podemos estimar o erro padrão a partir do desvio padrão amostral.

Como nesse exemplo, não temos tantos dados a ponto da distribuição ser aproximadamente normal, precisaremos usar a distribuição T Student para obter o valor relacionado ao intervalo de confiança de acordo com a quantidade de dados que possuímos. Veja que quanto maior a quantidade de dados, mais esse valor se aproxima ao valor da distribuição normal (95% de confiança = 1,96 na distribuição normal).

95% Confidence Interval Calculations

Best Estimate ± Margin of Error

Sample mean ± "a few" · estimated standard error of sample mean

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

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

95% confidence

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

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

95% Confidence Interval Calculations

Mean = 82.48 inches
Standard Deviation = 15.06 inches
n = 25 observations → $t^* = 2.064$

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

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

$$82.48 \pm 2.064(3.012)$$

$$82.48 \pm 6.22$$

Interpreting the Confidence Interval

"range of reasonable values for our parameter"

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

What does "with 95% confidence" mean?

If this procedure were repeated over and over, each time producing a 95% confidence interval estimate,

we would **expect 95% of those resulting intervals to contain the population mean cartwheel distance.**

The coverage probability is defined in terms of (hypothetical) repeated sampling of multiple data sets from the population of interest. Over many such repeated samples, constructing one CI from each sample, there will be a fraction of the confidence intervals that cover the target. **This fraction is the coverage probability.**
We rarely have multiple independent samples from the same population, so we cannot usually verify that a confidence interval attains its intended coverage probability.
The confidence intervals we have seen so far are all constructed using two key quantities:

1. an unbiased estimate of a population parameter (if we are interested in estimating the population mean based on an independent and identically distributed sample of data, the unbiased estimate is the sample mean , and the standard error of this estimate is s/sqrt(n). where s is the standard deviation of the data, and n is the sample size). We need the "Z-score" (x̄ -

mu) / s to fall between -K and K with probability alpha. As long as this holds, then the interval $\bar{x}$ +/- K *σ/√n will have the intended coverage probability. The constant K plays a very important role in determining the properties of a CI.

There are two ways we can obtain values of K to use in constructing the CI. One approach is based on making the very strong assumption that the data are independent and identically distributed, and follow a normal (Gaussian) distribution. If this is the case, then the Z-score follows a Student-t distribution with n-1 degrees of freedom.

An alternative and much more broadly applicable basis for obtaining a value for K is to use the "central limit theorem" (CLT). The CLT states that the sample mean of independent and identically distributed values will be approximately normally distributed. The CLT also implies that the Z-score will be approximately normally distributed. Importantly, the CLT provides these guarantees even when the individual data values have distributions that are not normal, as long as the sample size is "sufficiently large." Unfortunately, there is no universal rule that defines how large the sample size should be to invoke the central limit theorem. In general, if the data distribution is close to being normal, then the Z-scores will be close to normally-distributed even when the sample size is quite small (e.g. around 10). If the individual data values are far from being normally distributed (e.g. they are strongly skewed or have heavy tails), then the CLT may not be relevant until the sample size is larger, say around 50.

Another common practice is to use K as calculated from the Student-t distribution, even when the data are not taken to be normal. The rationale for doing this is that even though the Z-scores do not follow a Student-t distribution in this setting, the values of K obtained using the t-distribution will always be slightly larger than 1.96. Thus, the coverage will be slightly higher when using the t-distribution to calculate K compared to when using the normal distribution. Using a slightly larger value of K helps compensate for several possible factors that could lead to the Z-scores being slightly heavier-tailed than predicted by a normal distribution. As the sample size grows, the values of K obtained from the normal and t-distributions will become very similar. The distinction between using these two approaches is therefore mainly relevant when the sample size is smaller than around 50.

When working with strongly skewed data, another practical technique for improving the coverage properties of intervals is to transform the data with a skew-reducing transformation, e.g. a log transformation, then calculate the interval in the usual way.

Two such factors that can cause major problems with CI coverage probabilities are clustering or other forms of dependence in the data, and overt or hidden pre-testing or multiplicity in the analysis.

2. the standard error of this estimate.

standard error: estimate of the standard deviation of the sampling distribution of estimates that would arise if we had drawn repeated samples of the same size and computed the same estimate for each random sample. In a simplified sense, the standard error gives us a sense of the uncertainty associated with our estimate. Estimates with smaller standard errors are thus consider more precise.

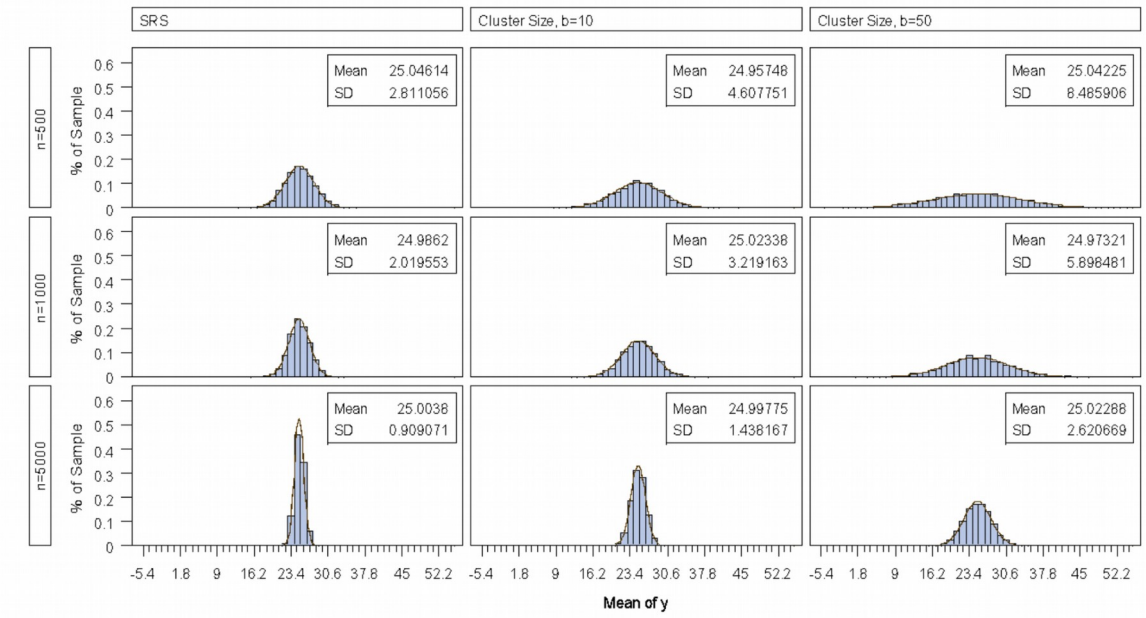
So what exactly impacts a standard error in terms of a study design?

- variance: the more variability that is associated with a given variable being measured, the more imprecise estimates based on that variable will be. So it is important to measure the variable of interest carefully

- Sample size: larger samples will tend to produce sampling distributions with less variability (or, in other words, estimates with smaller standard errors). Also, very unusual measures (outliers) could have strong influence on the variance of a given variable, and this requires careful descriptive assessment.
-
- The amount of dependence in the observations collected, possibly due to cluster sampling. In studies were clusters of units with similar characteristics are measured, the data collected will not be entirely independent with a given cluster. This is because units coming from the same cluster will generally have similar values on the variables of interest. This lack of independence in the observations collected reduces our effective sample size. We can account for this dependence within clusters by using specialized statistical procedures to estimate standard errors in a way that accounts for cluster sampling.

  The same problem arises in longitudinal studies, where we collect repeated measurements from the same individuals over time. While it may look like we have a large sample of observations, many of these observations will be strongly correlated with each other, and we need to account for this.

  In general, with these types of clustered data, standard errors will tend to be much larger. The larger the sample size selected from each cluster (and thus the smaller the sample of clusters), the larger the standard errors will tend to be.



- Stratification: if we select a stratified sample from a target population, we will tend to produce estimates with increased precision, because we are removing between-stratum variance from the variability of our estimates by design!

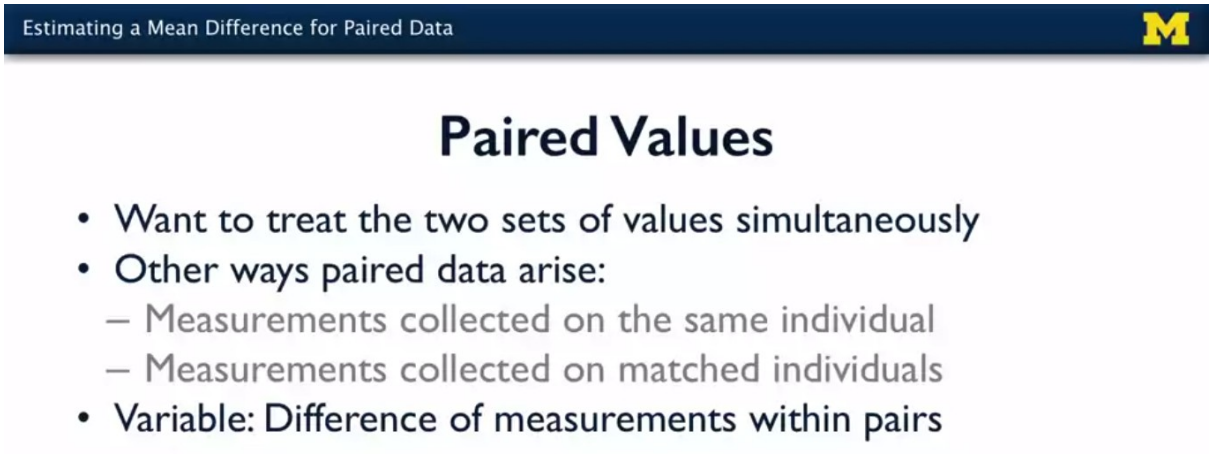
- Sampling weights: the use of weights in estimation can inflate the variance of our estimates. We can use specialized statistical procedures to make sure that our standard errors reflect the uncertainty in our estimates due to weighting. In general, the higher the variability in our weights, the more variable our estimates will be.

- Imputation: other design features may also ultimately affect standard errors (e.g., imputation of missing data).

Summary

- Confidence Intervals are used to give an *interval* estimate for our parameter of interest **~ a population mean**
- Center of the Confidence Interval is our best estimate **~ the sample mean**
- Margin of Error is "a few" (estimated) standard errors **~ for means we use t* multipliers**
- Assumptions for Confidence Interval for Population mean
  ~ data considered a random sample
  ~ population of responses is normal (else n large helps)

Difference in Means for Paired Data



- Want to treat the two sets of values simultaneously
- Other ways paired data arise:
  - Measurements collected on the same individual



pré e pós tratamento, medidas de um mesmo indivíduo

- Other ways paired data arise:
  - Measurements collected on the same individual
  - Measurements collected on matched individuals

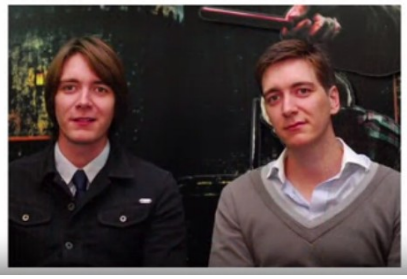

gêmeos, parentes próximos

Example:

Twin Education Levels

Twin Days in Twinsburg, Ohio annually since 1976

Variable: Education Level of Twins



Research Question

What is the *average* difference between the older twin's and younger twin's self-reported education?

Population - All identical twins
Parameter of Interest - Population mean difference of self-reported education level μ_d

Construct a 95% confidence interval for the mean difference of self-reported education for a set of identical twins.

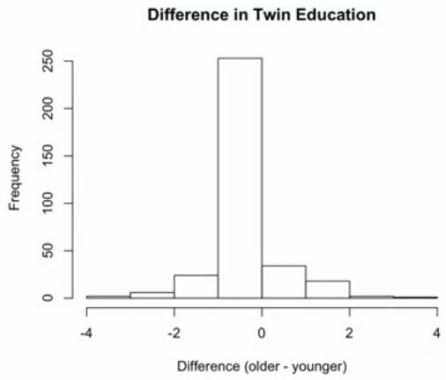
Difference Calculation

Difference = older twin - younger twin

Older twin education	Younger twin education	Difference (older - younger)
16	16	0
18	16	2
12	12	0
14	14	0
13	15	-2

Difference Summary

Difference = older twin - younger twin



n = 340 observations

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

Mean = 0.0838 years
Standard Deviation = 0.7627 years

95% Confidence Interval Calculations

Best Estimate ± Margin of Error

Sample mean difference ± "a few" · estimated standard error

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

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 = 0.084 years
Standard Deviation = 0.76 years

n = 340 observations ⟶ t* = 1.967

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

$0.084 \pm 1.967\ (0.76/\sqrt{340})$
$0.084 \pm 1.967\ (0.04)$
0.084 ± 0.0814
$(0.0025, 0.1652)$ years

"range of reasonable values for our parameter"

With 95% confidence, the population mean difference of the older twin's less the younger twin's self-reported education is estimated to be between 0.0025 years and 0.1652 years.

Há alguma diferença, mas não tanta, visto que todo o range é positivo e 0.0000 não faz parte dele.

Intervals for Differences

- Is there a mean difference between the education level of twins?
- If education levels are generally equal ⟶ mean difference is **0**
- If education levels are unequal ⟶ mean difference is not **0**
- Look for **0** in the range of reasonable values

Assumptions

We need to assume that we have a **random sample of identical twin sets**.

Population of differences is normal (or a **large enough sample size** can help to bypass this assumption).

Extension of the one mean confidence interval
~use difference variable now
Data need to be paired to calculate a difference variable
~two measurements on same individual
~two measurements on similar, matched individuals
0 not in the confidence interval
~implies the mean difference is not 0 → true difference

Difference in Means for Independent Groups

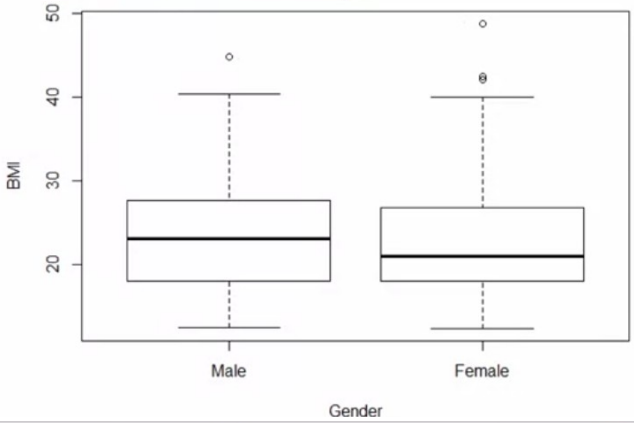
Research Question

Considering Mexican-American adults (ages 18 - 29) living in the United States, do males and females differ significantly in mean Body Mass Index (BMI)?

- **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)

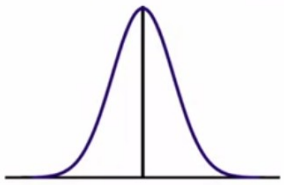
BMI Variable Summary

	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

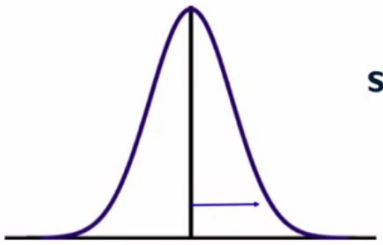
If models for both populations of responses are approximately normal (or sample sizes are both 'large' enough), distribution of the difference in sample means is (approximately) normal.



$\mu_1 - \mu_2$

All possible values of difference in sample means

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



$\mu_1 - \mu_2$

All possible values of difference in sample means

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

$$\text{Estimated Standard Error} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Confidence Interval Basics

Best Estimate ± Margin of Error

Best Estimate = Unbiased Point Estimate

Margin of Error = *"a few"* Estimated Standard Errors

"a few" = multiplier from appropriate distribution based on desired confidence level and sample design

95% Confidence Level ↔️ 0.05 Significance

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

Unpooled Confidence Interval Calculations

Best Estimate ± Margin of Error

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

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

The df for the t^* multiplier can be found using Welch's approximation.

If technology is not available, a conservative approach can be used by taking the smaller of $n_1 - 1$ and $n_2 - 1$ (i.e. $df = min(n_1 - 1, n_2 - 1)$)

Pooled Confidence Interval Calculations

Best Estimate ± Margin of Error

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

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

t^* multiplier comes from a t-distribution with $n_1 + n_2 - 2$ degrees of freedom

Again, this approach can be used if we assume the population variances are equal.

Normality Assumption: models for both populations of responses are approximately normal (or sample sizes are both 'large' enough)

Both distributions have a slight-to-moderate right skew, but the large sample sizes let us apply the CLT and continue.

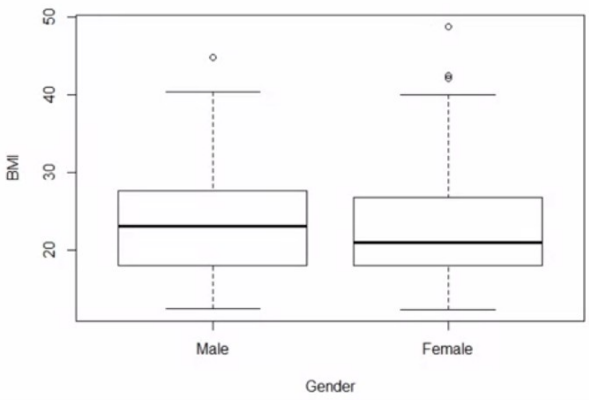
CLT: Teorema do Limite Central. Se o n for grande, podemos assumir que é aprox. normal.

Outra assumption para pooled variance:

Variance Assumption: if we have enough evidence to assume equal variances between the two populations, we can use the "pooled" approach

The IQR's and the standard deviations are similar enough to make this assumptions → the pooled approach will be used!

	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



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

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

Here, a t^* multiplier of 1.98 will be used

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

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

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

Interpreting the Confidence Interval

$$\left(-0.385 \ {}^{kg}/_{m^2}, 1.865 \ {}^{kg}/_{m^2}\right)$$

"range of reasonable values for our parameter"

With 95% confidence, the difference in mean body mass index between males and females for all Mexican-American adults (ages 18 - 29) in the U.S. is estimated to be between -0.385 kg/m^2 and 1.865 kg/m^2.

What does "with 95% confidence" mean?

If this procedure were repeated over and over, each time producing a 95% confidence interval estimate,

we would **expect 95% of those resulting intervals to contain the difference in population mean BMI.**

Summary

- Confidence Intervals are used to give an *interval* estimate for our parameter of interest **~ difference in population means**
- Center of the Confidence Interval is our best estimate **~ difference in sample means**
- Margin of Error is "a few" (estimated) standard errors **~ for two means we use t* multipliers (pooled vs. unpooled)**
- Assumptions for CI's for Difference in Population Means
  ~ data are two simple random samples, independent
  ~ both populations of responses are normal (else n large helps)
- Know how to interpret the **interval** and the **level**