

| | |
|--|---|
| Python Scientific Libs | 1 |
| Confidence Interval | 2 |
| Conservative Standard Error and Sample Size | 7 |
| Sample Size | 8 |
| Difference of Population Proportions with Confidence | 9 |

Python Scientific Libs

Python scientific libraries

Foundation Libraries

- Numpy - arrays and linear algebra
- Scipy - large collection of numerical algorithms
- Matplotlib - powerful graphing and plotting library

Data Science Libraries

- **Seaborn** - higher-level plotting
- **Pandas** - data containers, data management tools
- **Statsmodels** - statistical analysis, data modeling
- Sklearn - machine learning

Confidence Interval

Estimating a Population Proportion with Confidence



MottsPoll Background

What proportion of parents report they use a car seat for all travel with their toddler?

Population - Parents with a toddler

Parameter of Interest - A Proportion

Construct a 95% Confidence Interval for the population proportion of parents reporting they use a car seat for all travel with their toddler

Estimating a Population Proportion with Confidence



MottsPoll Background

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

540 parents responded 'yes' to this question.

95% Confidence Interval Calculations

Best Estimate \pm Margin of Error

$\hat{p} \pm$ Margin of Error

$\hat{p} \pm$ "a few" \cdot estimated $se(\hat{p})$

se = standard error

Different Z Multipliers

| 90% | 95% | 98% | 99% |
|--------------|-------------|--------------|--------------|
| 1.645 | 1.96 | 2.326 | 2.576 |


Best Estimate \pm Margin of Error

Best Estimate \pm “a few” (estimated) standard errors

More confident \rightarrow Larger Multiplier \rightarrow Wider Interval

Best Estimate = \hat{p}

Sample Size  $n = 659$

**Number
Responded “yes”**  $x = 540$
 $\hat{p} = x/n$
 $= 540/659$
 $= \mathbf{0.85}$

Best Estimate \pm Margin of Error

$\hat{p} \pm$ Margin of Error

$\hat{p} \pm$ “a few” \cdot estimated se(\hat{p})

$$\hat{p} \pm \mathbf{1.96} \cdot \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

95% Confidence Interval Calculations

$$\begin{aligned} & \hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} & \hat{p}=0.85 \\ & 0.85 \pm 1.96 \sqrt{\frac{0.85(1-0.85)}{659}} & n=659 \\ & 0.85 \pm 0.0273 \\ & \boxed{(0.8227, 0.8773)} \end{aligned}$$

Interpreting the Confidence Interval

“range of reasonable values for our parameter”

With 95% confidence, the population proportion of parents with a toddler who use a car seat for all travel is **estimated** to be between 82.27% - 87.73%.

Wrong Understanding of Confidence Level

95% chance or probability
that the population proportion is in
this already computed interval of (0.8227, 0.8773)

We don't know what the proportion is.. we are trying to estimate it based on sample.

Essa confiança se refere a quantidade de amostras em que a proporção da população esteve dentro do intervalo estimado.

Understanding Confidence Level

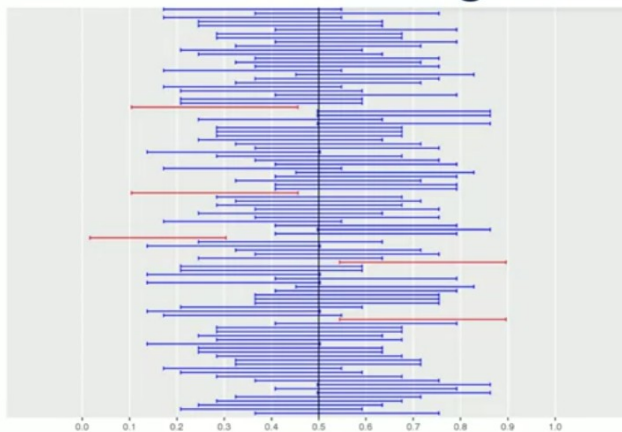
The screenshot shows a web-based simulation interface for creating confidence intervals. It includes three sliders: 'True Proportion' set to 0.5, 'What Confidence Level do you want to use?' set to 0.95, and 'How Many Confidence Interval Should We Make?' set to 100. A text box for 'Enter Your Sample Size (>=10)' contains the value 25. A red button at the bottom is labeled 'Create Confidence Intervals'.

Population Proportion = 0.50

Take 100 samples each of size 25

For each sample,
create a 95% confidence interval
for the population proportion

Understanding Confidence Level



95 of these **100**
generated intervals
did contain the true
proportion of **0.5**
while **5** did not.

Understanding Confidence Level



With a **95%** confidence level,
we would expect (in the long run)
about **95%** of the intervals
to contain the true proportion.

Car Seats for Toddlers Example

In a sample of 659 parents with a toddler, 540 (or **85%**) stated they **use a car seat** for all travel with their toddler.



90% CI:
 0.85 ± 0.0229
82.7% to 87.3%

95% CI:
 0.85 ± 0.0273
82.3% to 87.7%

99% CI:
 0.85 ± 0.0358
81.4% to 88.6%

What are the assumptions?

- **Best Estimate** – in order to get a *reliable* best estimate, we need a *SIMPLE RANDOM SAMPLE*
- **Simple Random Sample** – a representative subset of the population made up of observations (or subjects) that have an equal probability of being chosen

What are the assumptions?

- **Margin of error** – in order to use the critical z-value in our calculations, we need a *large enough sample size*
- **But what is ‘large enough’...?**

Se a amostra for grande, ela se aproximará da curva normal.

Checking Assumptions

- Simple Random Sample – analyze how the sample was collected, does it seem representative?
- Large Enough Sample Size – do we have at least 10 of each response outcome?

Conservative Standard Error and Sample Size

- Estimated standard error may be too small, or inaccurate based off sample so can employ conservative approach.
- Conservative approach → determine sample size needed based on a confidence level and desired margin of error.

Conservative Approach & Sample Size Consideration



Closer Look at estimated SE

$$\text{estimated standard error} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

What if **\hat{p}** is
not accurate?

which is maximized when **$\hat{p}=0.5$**

$$\text{conservative standard error} = \frac{1}{2\sqrt{n}}$$

Back to Motts Car Seat Example

95% Margin of Error is only dependent on sample size

$$\hat{p} \pm 2 \cdot \frac{1}{2\sqrt{n}}$$

$$\hat{p} \pm \frac{1}{\sqrt{n}}$$

$$\hat{p}=0.85$$

$$n=659$$

(0.81, 0.89)

Conservative 95% Confidence Interval, 4% Margin of Error

4:10 / 8:35

Sample Size

Sample Size Determination

Margin of Error (MoE) is only dependent on:

- 1) our confidence level (typically 95%) and
- 2) our sample size

What sample size would we need to have a 95% (conservative) confidence interval with a Margin of Error of only 3% (0.03)?

$$\text{MoE} = 0.03$$

$$\text{MoE} = \frac{1}{\sqrt{n}}$$

For 95% Confidence

$$n = (1/\text{MoE})^2$$

$$n = (1/0.03)^2$$

$$n = 1,111.11$$

$$n \geq 1,112$$

What if 3% MoE @ 99% Confidence?

$$\hat{p} \pm Z^* \cdot \frac{1}{2\sqrt{n}}$$

$$\text{MoE} = Z^* \cdot \frac{1}{2\sqrt{n}}$$

$$n = \left(\frac{Z^*}{2 \cdot \text{MoE}} \right)^2$$

What if 3% MoE @ 99% Confidence?

$$n = \left(\frac{Z^*}{2 \cdot \text{MoE}} \right)^2$$

$$Z^* = 2.576 \text{ (99\%)}$$

$$\text{MoE} = 0.03$$

$$n = 1843.27$$

$$n \geq 1844$$

Difference of Population Proportions with Confidence

Difference in Proportion Confidence Interval

Best Estimate \pm Margin of Error

$$\hat{p}_1 - \hat{p}_2 \pm \text{Margin of Error}$$

$$\hat{p}_1 - \hat{p}_2 \pm \text{"a few"} \cdot \text{se}(\hat{p}_1 - \hat{p}_2)$$

$$\hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Research Question

What is the difference in population proportions of parents reporting that their children age 6-18 have had some swimming lessons between white children and black children?

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

Parameter of Interest - Difference in population proportions ($p_1 - p_2$)

We'll let 1 = white and 2 = black

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 988 parents of white children age 6-18 was taken with 543 saying that their child has had some swimming lessons.

Best Estimate of the Parameter

$$\hat{p}_1 = 543/988 = 0.55$$

1 = white

$$\hat{p}_2 = 91/247 = 0.37$$

2 = black

$$\hat{p}_1 - \hat{p}_2 = 0.55 - 0.37 = 0.18$$

$$\begin{aligned} \hat{p}_1 - \hat{p}_2 \pm 1.96 \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \\ 0.18 \pm 1.96 \cdot 0.0345 \\ 0.18 \pm 0.0677 \\ (0.1123, 0.2477) \end{aligned}$$

Interpreting the Confidence Interval

“range of reasonable values for our parameter”

With 95% confidence, the population proportion of parents with white children who have taken swimming lessons is 11.23 to 24.77% higher than the population proportion of parents with black children who have taken 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 need $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, and $n_2(1-\hat{p}_2)$ In other words, we need at least 10 yes's and at least 10 no's for each sample.

Is there a difference between two parameters?

If parameters are equal \rightarrow difference is 0

If parameters are unequal \rightarrow difference is not 0

Look for 0 in the range of reasonable values