Statistical Inference with Confidence Intervals

Throughout week 2, we have explored the concept of confidence intervals, how to calculate them, interpret them, and what confidence really means.

In this tutorial, we're going to review how to calculate confidence intervals of population proportions and means.

To begin, let's go over some of the material from this week and why confidence intervals are useful tools when deriving insights from data.

Why Confidence Intervals?

Confidence intervals are a calculated range or boundary around a parameter or a statistic that is supported mathematically with a certain level of confidence. For example, in the lecture, we estimated, with 95% confidence, that the population proportion of parents with a toddler that use a car seat for all travel with their toddler was somewhere between 82.2% and 87.7%.

This is ***different*** than having a 95% probability that the true population proportion is within our confidence interval.

Essentially, if we were to repeat this process, 95% of our calculated confidence intervals would contain the true proportion.

How are Confidence Intervals Calculated?

Our equation for calculating confidence intervals is as follows:

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

Where the *Best Estimate* is the **observed population proportion or mean** and the *Margin of Error* is the **t-multiplier**.

The t-multiplier is calculated based on the degrees of freedom and desired confidence level. For samples with more than 30 observations and a confidence level of 95%, the t-multiplier is 1.96

The equation to create a 95% confidence interval can also be shown as:

$$Population\ Proportion\ or\ Mean\ \ \pm (t - multiplier *\ Standard\ Error)$$

Lastly, the Standard Error is calculated differenly for population proportion and mean:

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

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

Let's replicate the car seat example from lecture:

In [1]:

```
import numpy as np
```

In [2]:

```
tstar = 1.96
p = .85
n = 659

se = np.sqrt((p * (1 - p))/n)
se
```

Out[2]:

```
0.01390952774409444
```

In [3]:

```
lcb = p - tstar * se
ucb = p + tstar * se
(lcb, ucb)
```

Out[3]:

```
(0.8227373256215749, 0.8772626743784251)
```

In [4]:

```
import statsmodels.api as sm
```

In [5]:

```
sm.stats.proportion_confint(n * p, n)
```

Out[5]:

```
(0.8227378265796143, 0.8772621734203857)
```

Now, lets take our Cartwheel dataset introduced in lecture and calculate a confidence interval for our mean cartwheel distance:

In [6]:

```
import pandas as pd

df = pd.read_csv("Cartwheeldata.csv")
```

In [7]:

```
df.head()
```

Out[7]:

	ID	Age	Gender	GenderGroup	Glasses	GlassesGroup	Height	Wingspan	CWDistance	C
0	1	56	F	1	Y	1	62.0	61.0	79	
1	2	26	F	1	Y	1	62.0	60.0	70	
2	3	33	F	1	Y	1	66.0	64.0	85	
3	4	39	F	1	N	0	64.0	63.0	87	
4	5	27	M	2	N	0	73.0	75.0	72	

In [8]:

```
mean = df["CWDistance"].mean()
sd = df["CWDistance"].std()
n = len(df)

n
```

Out[8]:

```
25
```

In [9]:

```
tstar = 2.064

se = sd/np.sqrt(n)

se
```

Out[9]:

```
3.0117104774529704
```

In [10]:

```
lcb = mean - tstar * se
ucb = mean + tstar * se
(lcb, ucb)
```

Out[10]:

```
(76.26382957453707, 88.69617042546294)
```

In [11]:

```
sm.stats.DescrStatsW(df["CWDistance"]).zconfint_mean()
```

Out[11]:

```
(76.57715593233024, 88.38284406766977)
```

In [ ]:

```

```