intro_confidenceintervals

February 9, 2022

0.1 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.

0.1.1 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.

0.1.2 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 [3]: # standard error for proportion p
        tstar = 1.96
        p = .85
        n = 659
        se = np.sqrt((p * (1 - p))/n)
        se

Out[3]: 0.01390952774409444

In [4]: # low and upper confidence level with 95% of confidence (1.96) for n=659 (big)
        lcb = p - tstar * se
        ucb = p + tstar * se
        (lcb, ucb)

Out[4]: (0.8227373256215749, 0.8772626743784251)

In [5]: # same calculation using the lib statsmodels
        import statsmodels.api as sm

In [6]: sm.stats.proportion_confint(n * p, n)

Out[6]: (0.8227378265796143, 0.8772621734203857)

In [7]: # another example
        # Lets suppose that on a certain website, out of 1126 visitors on a given day,
        # 310 clicked on an ad purchased by a sponsor. Lets construct a confidence interval fo
        # the population proportion of visitors who click the ad.
        from statsmodels.stats.proportion import proportion_confint

        proportion = 310 / 1126    # Sample proportion

        # Function for computing confidence intervals
        # Alpha, which is 1 minus the confidence level
        proportion_confint(count=310,    # Number of "successes"
                           nobs=1126,    # Number of trials
                           alpha=(1 - 0.95)) # default is 0,05

Out[7]: (0.24922129423231776, 0.30140037539468045)
```

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

```
In [10]: import pandas as pd

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

In [11]: df.head()
```

Out[11]:

	ID	Age	Gender	GenderGroup	Glasses	GlassesGroup	Height	Wingspan	\
0	1	56	F	1	Y	1	62.0	61.0	
1	2	26	F	1	Y	1	62.0	60.0	
2	3	33	F	1	Y	1	66.0	64.0	
3	4	39	F	1	N	0	64.0	63.0	
4	5	27	M	2	N	0	73.0	75.0	

	CWDistance	Complete	CompleteGroup	Score
0	79	Y	1	7
1	70	Y	1	8
2	85	Y	1	7
3	87	Y	1	10
4	72	N	0	4

```
In [16]: mean = df["CWDistance"].mean()
         sd = df["CWDistance"].std()
         n = len(df)
         mean, sd, n
```

Out[16]: (82.48, 15.058552387264852, 25)

```
In [17]: tstar = 2.064

         se = sd/np.sqrt(n)

         se
```

Out[17]: 3.0117104774529704

```
In [18]: lcb = mean - tstar * se
         ucb = mean + tstar * se
         (lcb, ucb)
```

Out[18]: (76.26382957453707, 88.69617042546294)

```
In [19]: sm.stats.DescrStatsW(df["CWDistance"]).zconfint_mean()
```

Out[19]: (76.57715593233024, 88.38284406766977)