Confidence_Intervals_Differences_Population_Parameters

June 24, 2020

1 Confidence Intervals

This tutorial is going to demonstrate how to load data, clean/manipulate a dataset, and construct a confidence interval for the difference between two population proportions and means.

We will use the 2015-2016 wave of the NHANES data for our analysis.

*Note: We have provided a notebook that includes more analysis, with examples of confidence intervals for one population proportions and means, in addition to the analysis I will show you in this tutorial. I highly recommend checking it out!

For our population proportions, we will analyze the difference of proportion between female and male smokers. The column that specifies smoker and non-smoker is "SMQ020" in our dataset.

For our population means, we will analyze the difference of mean of body mass index within our female and male populations. The column that includes the body mass index value is "BMXBMI".

Additionally, the gender is specified in the column "RIAGENDR".

```
In [ ]: import pandas as pd
        import numpy as np
        import matplotlib
        matplotlib.use('Agg')
        import seaborn as sns
        %matplotlib inline
        import matplotlib.pyplot as plt
        import statsmodels.api as sm
```

```
In [ ]: url = "nhanes_2015_2016.csv"
        da = pd.read_csv(url)
```

1.0.1 Investigating and Cleaning Data

```
In [ ]: da["SMQ020x"] = da.SMQ020.replace({1: "Yes", 2: "No", 7: np.nan, 9: np.nan})
        da["SMQ020x"]
```

```
In [ ]: da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
        da["RIAGENDRx"]
```

```
In [ ]: dx = da[["SMQ020x", "RIAGENDRx"]].dropna()
        pd.crosstab(dx.SMQ020x, dx.RIAGENDRx)
```

```
In [ ]: dz = dx.groupby("RIAGENDRx").agg({"SMQ020x": [np.mean, np.size]})
        dz.column = ["Proportion", "Total n"]
        dz
```

1.0.2 Constructing Confidence Intervals

Now that we have the population proportions of male and female smokers, we can begin to calculate confidence intervals. From lecture, we know that the equation is as follows:

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

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

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

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

The Standard Error (SE) is calculated differenly for population proportion and mean:

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

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

Lastly, the standard error for difference of population proportions and means is:

$$Standard\ Error\ for\ Difference\ of\ Two\ Population\ Proportions\ Or\ Means = \sqrt{(SE_1)^2 + (SE_2)^2}$$

Difference of Two Population Proportions

```
In [ ]: p = .304845
        n = 2972
        se_female = np.sqrt(p * (1 - p)/n)
        se_female

In [ ]: p = .513258
        n = 2753
        se_male = np.sqrt(p * (1 - p)/ n)
        se_male

In [ ]: se_diff = np.sqrt(se_female**2 + se_male**2)
        se_diff

In [ ]: d = .304845 - .513258
        lcb = d - 1.96 * se_diff
        ucb = d + 1.96 * se_diff
        (lcb, ucb)
```

Difference of Two Population Means

```
In [ ]: da["BMXBMI"].head()

In [ ]: da.groupby("RIAGENDRx").agg({"BMXBMI": [np.mean, np.std, np.size]})

In [ ]: sem_female = 7.753319 / np.sqrt(2976)
        sem_male = 6.252568 / np.sqrt(2759)
        (sem_female, sem_male)

In [ ]: sem_diff = np.sqrt(sem_female**2 + sem_male**2)
        sem_diff

In [ ]: d = 29.939946 - 28.778072

In [ ]: lcb = d - 1.96 * sem_diff
        ucb = d + 1.96 * sem_diff
        (lcb, ucb)
```