nhanes_confidence_intervals

July 25, 2024

1 Confidence intervals case study using NHANES data

[3]:
```
%matplotlib inline
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
import statsmodels.api as sm
```

[4]:
```
da = pd.read_csv("nhanes_2015_2016.csv")
```

The specific definition of "smoker" used here (SMQ020) identifies a person as being a smoker if they self-report as having smoked 100 or more cigarettes in their lifetime.

Attempt to calculate the proportions of smokers separately for females and for males.

First replacing the numeric codes in the variables of interest with text labels, and setting the rare answers other than "yes" and "no" to be missing.

[6]:
```
da["SMQ020x"] = da.SMQ020.replace({1: "Yes", 2: "No", 7: np.nan, 9: np.nan})
da["RIAGENDRx"] = da.RIAGENDR.replace({1: "Male", 2: "Female"})
```

[7]:
```
dx = da[["SMQ020x", "RIAGENDRx"]].dropna()
pd.crosstab(dx.SMQ020x, dx.RIAGENDRx)
```

[7]:
```
RIAGENDRx  Female  Male
SMQ020x
No           2066  1340
Yes           906  1413
```

The confidence interval (CI) is constructed using two inputs: the sample proportion of smokers, and the total sample size for smokers and non-smokers combined. We calculate these values next.

[9]:
```
dz = dx.groupby(dx.RIAGENDRx).agg({"SMQ020x": [lambda x: np.mean(x=="Yes"), np.
 ↪size]})
dz.columns = ["Proportion", "Total_n"] # The default column names are unclear,␣
 ↪so we replace them here
dz
```

```
[9]:           Proportion  Total_n
     RIAGENDRx
     Female      0.304845     2972
     Male        0.513258     2753
```

Calculating standard error:

```
[11]: # Estimate the standard error for the proportion of females who smoke
      p = dz.Proportion.Female
      n = dz.Total_n.Female
      se_female = np.sqrt(p * (1 - p) / n)
      print(se_female)

      # Estimate the standard error for the proportion of males who smoke
      p = dz.Proportion.Male
      n = dz["Total_n"].Male
      se_male = np.sqrt(p * (1 - p) / n)
      print(se_male)
```

```
0.008444152146214435
0.009526078653689868
```

The standard errors for the estimated proportions of females and males who smoke are similar, and are each around 1%.

The estimated male smoking proportion is closer to 1/2 than the estimated female smoking proportion, and the male sample size is smaller than the female sample size. Both of these factors lead to the male standard error being larger than the female standard error, although the difference between the female and male standard errors is small.

Calculating the 95% confidence intervals for the proportions of female and male smokers:

```
[10]: # 95% CI for the proportion of females who smoke (compare to value above)
      sm.stats.proportion_confint(906, 906+2066)
```

```
[10]: (0.2882949879861214, 0.32139545615923526)
```

The results above indicate that any population proportion (for female lifetime smokers) between 0.288 and 0.321 would be compatible with the data that we observed in NHANES.

```
[11]: # 95% CI for the proportion of males who smoke (compare to value above)
      sm.stats.proportion_confint(1413, 1413+1340)
```

```
[11]: (0.49458749263718593, 0.5319290347874418)
```

These results indicate that any population proportion (for male lifetime smokers) between 0.493 and 0.531 would be compatible with the NHANES data.

1.1 Confidence intervals comparing two independent proportions - Smoking Rates

The point estimate of the difference between female and male smoking rates is -0.208 (0.305 minus 0.513, taking numbers from above). That is, the smoking rate is about 20 percentage points higher in men than in women. This difference of around 20 percentage points is only a point estimate of the true value.

Calculating the standard error for the difference between the proportion of females who smoke and the proportion of males who smoke:

[15]:
```
se_diff = np.sqrt(se_female**2 + se_male**2)
se_diff
```

[15]:
```
0.012729881381407434
```

The standard error of around 0.013 indicates that the estimated difference statistic -0.208 is expected to fall around 0.013 units from the true value. We do not know in which direction the error lies, and we do not know that the error is exactly 0.013, only that it is around this size on average. For most purposes, a standard error of 0.013 relative to an observed difference of -0.21 would be considered very small. That is, we have a very accurate estimate of the difference between smoking rates in women and in men.

Now that we have the standard error, we can construct a 95% confidence interval for the difference in proportions by taking the estimate and subtracting and adding two (or 1.96) standard errors from it.

[16]:
```
d = dz.Proportion.Female - dz.Proportion.Male
lcb = d - 2*se_diff
ucb = d + 2*se_diff
print(lcb, ucb)
```

```
-0.2338728044024504 -0.18295327887682067
```

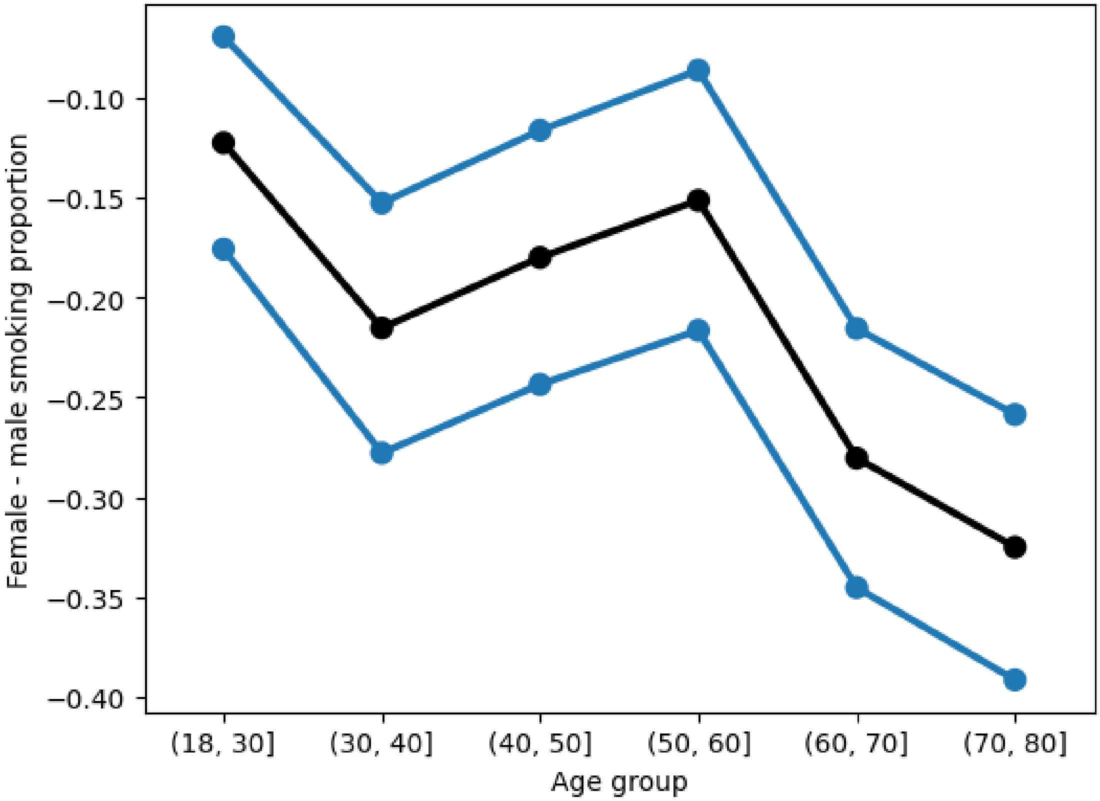
The 95% confidence interval above shows us that any value for the difference of population proportions (between females and males) lying between -0.233 and -0.183 is consistent with the observed data.

1.1.1 Confidence intervals for subpopulations - Confidence Band Viz

[17]:
```
# Calculate the smoking rates within age/gender groups
da["agegrp"] = pd.cut(da.RIDAGEYR, [18, 30, 40, 50, 60, 70, 80])
pr = da.groupby(["agegrp", "RIAGENDRx"]).agg({"SMQ020x": lambda x: np.
  ↪mean(x=="Yes")}).unstack()
pr.columns = ["Female", "Male"]

# The number of people for each calculated proportion
dn = da.groupby(["agegrp", "RIAGENDRx"]).agg({"SMQ020x": np.size}).unstack()
dn.columns = ["Female", "Male"]
```

```
# Standard errors for each proportion
se = np.sqrt(pr * (1 - pr) / dn)

# Standard error for the difference in female/male smoking rates in every age␣
 ↪band
se_diff = np.sqrt(se.Female**2 + se.Male**2)

# Standard errors for the difference in smoking rates between genders, within␣
 ↪age bands

# The difference in smoking rates between genders
pq = pr.Female - pr.Male

x = np.arange(pq.size)
pp = sns.pointplot(x=x, y=pq.values, color='black')
sns.pointplot(x=x, y=pq - 2*se_diff)
sns.pointplot(x=x, y=pq + 2*se_diff)
pp.set_xticklabels(pq.index)
pp.set_xlabel("Age group")
pp.set_ylabel("Female - male smoking proportion")
```

```
[17]: Text(0, 0.5, 'Female - male smoking proportion')
```



The plot above shows for each age band, the point estimate of the difference in smoking rates between genders (black dot), and the lower and upper end points of the 95% confidence interval (blue points). Based on this plot, we see that in the United States, smoking is more common in men than in women, not just overall, but also in every one of the age bands. The difference is largest for older people – for people older than 60, the smoking rate for males is around 30 percentage points greater than the smoking rate for females, while for people younger than 30, the smoking rate for males is only around 15 percentage points greater than the smoking rate for females.

Also note that the 95% confidence bands shown above are much wider than the 95% confidence intervals for the data that were not stratified by age. Stratifying by age leads to smaller sample sizes, which in turn results in wider confidence intervals.

1.2 Confidence intervals for the mean - Analysing mean BMI

Calculating the mean BMI for all women and for all men in the NHANES sample:

[19]:
```
da.groupby("RIAGENDRx").agg({"BMXBMI": np.mean})
```

[19]:
```
              BMXBMI
RIAGENDRx
Female     29.939946
Male       28.778072
```

[16]:
```
da.groupby("RIAGENDRx").agg({"BMXBMI": [np.mean, np.std, np.size]})
```

[16]:
```
              BMXBMI
                mean       std  size
RIAGENDRx
Female     29.939946  7.753319  2976
Male       28.778072  6.252568  2759
```

Calculating| the standard error of the mean BMI for women and for men:

[20]:
```
sem_female = 7.753 / np.sqrt(2976)
sem_male = 6.253 / np.sqrt(2759)
print(sem_female, sem_male)
```

```
0.14211938534506902 0.119045388988243
```

We see that the sample mean BMI for women is expected to be off by around 0.14 relative to the population mean BMI for women, and the sample mean BMI for men is expected to be off by around 0.12 relative to the population mean BMI for men.

The standard error of the mean for women is slightly larger for women than for men. The NHANES sample size for women is slightly larger than that for men, the data for women appears to be more dispersed.

The 95% confidence interval for female BMI:

[21]:

```
lcb_female = 29.94 - 1.96 * 7.753 / np.sqrt(2976)
ucb_female = 29.94 + 1.96 * 7.753 / np.sqrt(2976)
print(lcb_female, ucb_female)
```

```
29.661446004723665 30.218553995276338
```

Calculating the same using Statsmodels:

[22]:

```
female_bmi = da.loc[da.RIAGENDRx=="Female", "BMXBMI"].dropna()
sm.stats.DescrStatsW(female_bmi).zconfint_mean()
```

[22]:

```
(29.65987549809015, 30.220015806257674)
```

1.2.1 Confidence intervals for the difference between two means

[24]:

```
sem_diff = np.sqrt(sem_female**2 + sem_male**2)
sem_diff
```

[24]:

```
0.18539073420811059
```

The variance pooling rule gives us a value around 0.19 when comparing the female BMI to the male BMI.

Constructing a 95% confidence interval for the difference between the female and male mean BMI:

[25]:

```
bmi_diff = 29.94 - 28.78
lcb = bmi_diff - 2*sem_diff
ucb = bmi_diff + 2*sem_diff
(lcb, ucb)
```

[25]:

```
(0.789218531583779, 1.5307814684162213)
```

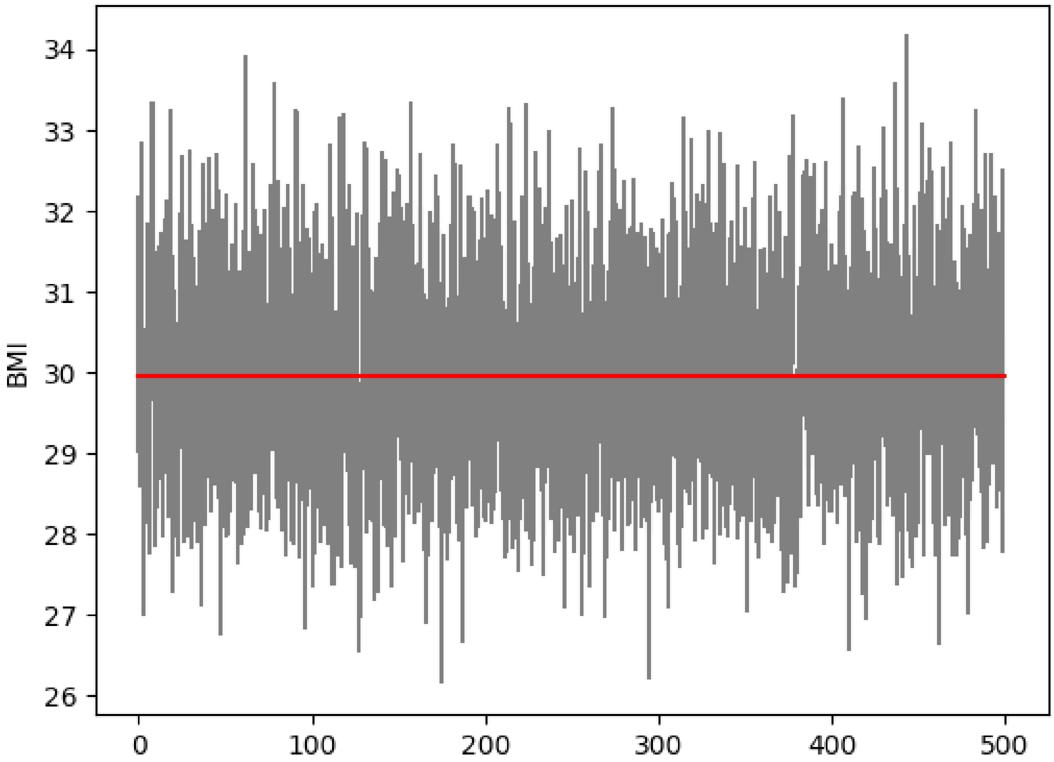
This finding indicates that while the point estimate shows that the women in our sample have around 1.1 unit greater BMI than the men in our sample, the true difference between the mean for all women in the population and for all men in the population could fall between 0.79 and 1.53, and still be consistent with the observed data.

1.2.2 Relationship between confidence intervals and sample size

[28]:

```
dx = da.loc[da.RIAGENDRx=="Female", ["RIAGENDRx", "BMXBMI"]].dropna()

all_cis = []
for n in 100, 200, 400, 800:
    cis = []
    for i in range(500):
        dz = dx.sample(n)
        ci = sm.stats.DescrStatsW(dz.BMXBMI).zconfint_mean()
        cis.append(ci)
    cis = np.asarray(cis)
    mean_width = cis[:, 1].mean() - cis[:, 0].mean()
    #print(n, mean_width)
    all_cis.append(cis)
ci = all_cis[0]
for j, x in enumerate(ci):
    plt.plot([j, j], x, color='grey')
    plt.gca().set_ylabel("BMI")
mn = dx.BMXBMI.mean()
plt.plot([0, 500], [mn, mn], color='red')
```

```
[28]: [<matplotlib.lines.Line2D at 0x7707191900d0>]
```



The vertical grey bars below each correspond to a confidence interval.While the individual intervals are quite different from each other, it appears that the vast majority of them cover the population value.