week3_assessment

June 30, 2020

0.1 Creating confidence intervals in python

In this assessment, you will look at data from a study on toddler sleep habits.

The confidence intervals you create and the questions you answer in this Jupyter notebook will be used to answer questions in the following graded assignment.

```
In [59]: import numpy as np
         import pandas as pd
         from scipy.stats import t
         pd.set_option('display.max_columns', 30) # set so can see all columns of the DataFram
```

Your goal is to analyse data which is the result of a study that examined differences in a number of sleep variables between napping and non-napping toddlers. Some of these sleep variables included: Bedtime (lights-off time in decimalized time), Night Sleep Onset Time (in decimalized time), Wake Time (sleep end time in decimalized time), Night Sleep Duration (interval between sleep onset and sleep end in minutes), and Total 24-Hour Sleep Duration (in minutes). Note: Decimalized time is the representation of the time of day using units which are decimally related.

The 20 study participants were healthy, normally developing toddlers with no sleep or behavioral problems. These children were categorized as napping or non-napping based upon parental report of children's habitual sleep patterns. Researchers then verified napping status with data from actigraphy (a non-invasive method of monitoring human rest/activity cycles by wearing of a sensor on the wrist) and sleep diaries during the 5 days before the study assessments were made.

You are specifically interested in the results for the Bedtime, Night Sleep Duration, and Total 24- Hour Sleep Duration.

Reference: Akacem LD, Simpkin CT, Carskadon MA, Wright KP Jr, Jenni OG, Achermann P, et al. (2015) The Timing of the Circadian Clock and Sleep Differ between Napping and Non-Napping Toddlers. PLoS ONE 10(4): e0125181. https://doi.org/10.1371/journal.pone.0125181

```
In [34]: # Import the data
         df = pd.read_csv("nap_no_nap.csv")

In [35]: # First, look at the DataFrame to get a sense of the data
         df.head()
```

Out[35]:

	id	sex	age (months)	dlmo time	days napped	napping	\
0	1	female	33.7	19.24	0	0	
1	2	female	31.5	18.27	0	0	
2	3	male	31.9	19.14	0	0	
3	4	female	31.6	19.69	0	0	

```
4   5  female          33.0      19.52            0        0

   nap lights outl time  nap sleep onset  nap midsleep  nap sleep offset  \
0                   NaN              NaN           NaN               NaN
1                   NaN              NaN           NaN               NaN
2                   NaN              NaN           NaN               NaN
3                   NaN              NaN           NaN               NaN
4                   NaN              NaN           NaN               NaN

   nap wake time  nap duration  nap time in bed  night bedtime  \
0            NaN           NaN              NaN          20.45
1            NaN           NaN              NaN          19.23
2            NaN           NaN              NaN          19.60
3            NaN           NaN              NaN          19.46
4            NaN           NaN              NaN          19.21

   night sleep onset  sleep onset latency  night midsleep time  \
0              20.68                 0.23                 1.92
1              19.48                 0.25                 1.09
2              20.05                 0.45                 1.29
3              19.50                 0.05                 1.89
4              19.65                 0.45                 1.30

   night wake time  night sleep duration  night time in bed  \
0             7.17                 629.4              643.0
1             6.69                 672.4              700.4
2             6.53                 628.8              682.6
3             8.28                 766.6              784.0
4             6.95                 678.0              718.0

   24 h sleep duration  bedtime phase difference  \
0                629.4                     -1.21
1                672.4                     -0.96
2                628.8                     -0.46
3                766.6                      0.23
4                678.0                      0.31

   sleep onset phase difference  midsleep phase difference  \
0                         -1.44                       6.68
1                         -1.21                       6.82
2                         -0.91                       6.15
3                          0.19                       6.20
4                         -0.13                       5.78

   wake time phase difference
0                       11.93
1                       12.42
2                       11.39
```

```
3                    12.59
4                    11.43
```

Question: What variable is used in the column 'napping' to indicate a toddler takes a nap? napping

Question: What is the sample size n? 20

0.2 Hypothesis testing

We will look at two hypothesis test, each with $\alpha = .025$:

1. Is the average bedtime for toddlers who nap later than the average bedtime for toddlers who don't nap?

$$H_0 : \mu_{nap} = \mu_{no\ nap},\ H_a : \mu_{nap} > \mu_{no\ nap}$$

Or equivalently:

$$H_0 : \mu_{nap} - \mu_{no\ nap} = 0,\ H_a : \mu_{nap} - \mu_{no\ nap} > 0$$

2. The average 24 h sleep duration (in minutes) for napping toddlers is different from toddlers who don't nap.

$$H_0 : \mu_{nap} = \mu_{no\ nap},\ H_a : \mu_{nap} \neq \mu_{no\ nap}$$

Or equivalenty:

$$H_0 : \mu_{nap} - \mu_{no\ nap} = 0,\ H_a : \mu_{nap} - \mu_{no\ nap} \neq 0$$

Aside: This α level is equivalent to $\alpha = .05$ and then applying the Bonferonni correction.

Before any analysis, we will convert 'night bedtime' into decimalized time.

```
In [36]: # Convert 'night bedtime' into decimalized time
         df.loc[:,'night bedtime'] = np.floor(df['night bedtime'])*60 + np.round(df['night bedt
```

Now, isolate the column 'night bedtime' for those who nap into a new variable, and those who didn't nap into another new variable.

```
In [37]: nap_bedtime = df[df['napping'] == 1]['night bedtime']
```

```
In [38]: no_nap_bedtime = df[df['napping'] == 0]['night bedtime']
```

Now find the sample mean bedtime for nap and no_nap.

```
In [39]: nap_mean_bedtime = nap_bedtime.mean()
         nap_mean_bedtime
```

```
Out[39]: 1233.0666666666666
```

```
In [40]: no_nap_mean_bedtime = no_nap_bedtime.mean()
         no_nap_mean_bedtime
```

```
Out[40]: 1191.0

In [42]: sample_difference = nap_mean_bedtime - no_nap_mean_bedtime
         sample_difference

Out[42]: 42.066666666666606
```

Question: What is the sample difference of mean bedtime for nappers minus no nappers?
Now find the sample standard deviation for X_{nap} and $X_{no\ nap}$.

```
In [47]: nap_s_bedtime = np.std(nap_bedtime)
         nap_s_bedtime

Out[47]: 33.277553328866524

In [46]: no_nap_s_bedtime = np.std(no_nap_bedtime)
         no_nap_s_bedtime

Out[46]: 30.678983033992505
```

Question: What is the s.e.$(\bar{X}_{nap} - \bar{X}_{no\ nap})$?
We expect the variance in sleep time for toddlers who nap and toddlers who don't nap to be the same. So we use a pooled standard error.
Calculate the pooled standard error of $\bar{X}_{nap} - \bar{X}_{no\ nap}$ using the formula below.
$s.e.(\bar{X}_{nap} - \bar{X}_{no\ nap}) = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}(\frac{1}{n_1}+\frac{1}{n_2})}$

```
In [50]: pooled_se = np.sqrt(((((len(nap_bedtime)-1)*(nap_s_bedtime**2))+((len(no_nap_bedtime)
         pooled_se

Out[50]: 16.895468503923713
```

Question: Given our sample size of n, how many degrees of freedom (df) are there for the associated t distribution? 14, 4

```
In [53]: deg_fre=(len(nap_bedtime)-1)+(len(no_nap_bedtime)-1)
         deg_fre

Out[53]: 18
```

Now calculate the t-test statistic for our first hypothesis test using
* pooled s.e.$(\bar{X}_{nap} - \bar{X}_{no\ nap})$
* $\bar{X}_{nap} - \bar{X}_{no\ nap}$
* $\mu_{0,\ nap} - \mu_{0,\ no\ nap} = 0$, the population difference in means under the null hypothesis

```
In [51]: test_stats = (nap_mean_bedtime - no_nap_mean_bedtime)/pooled_se
         test_stats

Out[51]: 2.4898194836619814
```

Question: What is the p-value for the first hypothesis test?
To find the p-value, we can use the function:

```
t.cdf(y, df)
```

Which for $X \sim t(df)$ returns $P(X \leq y)$.

Because of the symmetry of the t distrubution, we have that

```
1-t.cdf(y, df)
```

returns $P(X > y)$

The function t.cdf(y, df) will give you the same value as finding the one-tailed probability of y on a t-table with the specified degrees of freedom.

Use the function t.cdf(y, df) to find the p-value for the first hypothesis test.

```
In [56]: p_val=1-t.cdf(test_stats, deg_fre)
         p_val

Out[56]: 0.011392413734599205
```

Question: What are the t-statistic and p-value for the second hypothesis test?

Calculate the t test statistics and corresponding p-value using the scipy function scipy.stats.ttest_ind(a, b, equal_var=True) and check with your answer.

Question: Does scipy.stats.ttest_ind return values for a one-sided or two-sided test?

Question: Can you think of a way to recover the results you got using 1-t.cdf from the p-value given by scipy.stats.ttest_ind?

Use the scipy function scipy.stats.ttest_ind(a, b, equal_var=True) to find the t test statistic and corresponding p-value for the second hypothesis test.

```
In [63]: from scipy import stats

         stats.ttest_ind(nap_bedtime, no_nap_bedtime, equal_var=True)

Out[63]: Ttest_indResult(statistic=2.367160052079275, pvalue=0.029334902861805394)
```

Question: For the $\alpha = .025$, do you reject or fail to reject the first hypothesis?

Question: For the $\alpha = .025$, do you reject or fail to reject the second hypothesis?