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 [1]:

```
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 DataFrame
```

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 (https://en.wikipedia.org/wiki/Decimal_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 Bedtime.

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 (https://doi.org/10.1371/journal.pone.0125181)

In [19]:

```
# Import the data
df = pd.read_csv("nap_no_nap.csv")
```

In [20]:

```
# First, look at the DataFrame to get a sense of the data
df.head(20)
```

Out[20]:

	id	sex	age (months)	dlmo time	days napped	napping	nap lights outl time	nap sleep onset	nap midsleep	nap sleep offset	nap wake time	durat
0	1	female	33.7	19.24	0	0	NaN	NaN	NaN	NaN	NaN	N
1	2	female	31.5	18.27	0	0	NaN	NaN	NaN	NaN	NaN	N
2	3	male	31.9	19.14	0	0	NaN	NaN	NaN	NaN	NaN	N
3	4	female	31.6	19.69	0	0	NaN	NaN	NaN	NaN	NaN	N
4	5	female	33.0	19.52	0	0	NaN	NaN	NaN	NaN	NaN	N
5	6	female	36.2	18.22	4	1	14.00	14.22	15.00	15.78	16.28	9:
6	7	male	36.3	19.28	1	1	14.75	15.03	15.92	16.80	16.08	10
7	8	male	30.0	21.06	5	1	13.09	13.43	14.44	15.46	15.82	12
8	9	male	33.2	19.38	2	1	14.41	14.42	15.71	17.01	16.60	15
9	10	female	37.1	19.93	3	1	13.12	13.42	14.31	15.19	15.30	10
10	11	male	32.9	18.79	4	1	13.99	14.03	14.85	15.68	16.10	9
11	12	female	35.0	19.65	5	1	13.18	13.45	14.33	15.21	15.35	10
12	13	male	35.1	19.83	3	1	13.94	14.48	15.26	16.03	15.78	9:
13	14	female	35.6	19.88	4	1	12.68	13.08	13.92	14.76	15.00	10
14	15	female	36.6	19.94	4	1	12.71	12.88	13.80	14.72	14.88	11
15	16	male	36.5	20.25	3	1	13.74	14.68	15.66	16.64	16.45	11
16	17	female	33.7	20.33	5	1	13.15	13.87	14.49	15.11	15.40	7
17	18	male	36.4	20.16	5	1	12.47	12.56	13.30	14.05	14.25	8
18	19	female	33.6	19.68	3	1	14.71	14.85	15.46	16.07	16.20	7:
19	20	male	33.8	20.51	3	1	12.68	13.54	14.30	15.07	15.23	9

In [21]:

```
df.shape
```

Out[21]:

```
(20, 25)
```

Question: What value is used in the column 'napping' to indicate a toddler takes a nap? (see reference article)

Question: What is the overall sample size n? What is the sample size for toddlers who nap, n_1 , and toddlers who don't nap, n_2 ?

Average bedtime confidence interval for napping and non napping toddlers

Create two 95% confidence intervals for the average bedtime, one for toddler who nap and one for toddlers who don't.

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

In [25]:

```
bedtime_nap = df.loc[df['napping']==1, 'night bedtime']
bedtime_nap
```

Out[25]:

```
5     19.95
6     20.60
7     22.01
8     20.24
9     20.78
10    19.45
11    20.18
12    20.22
13    20.26
14    20.28
15    20.46
16    20.43
17    20.02
18    19.50
19    20.18
Name: night bedtime, dtype: float64
```

In [26]:

```
bedtime_no_nap = df.loc[df['napping']==0, 'night bedtime']
bedtime_no_nap
```

Out[26]:

```
0    20.45
1    19.23
2    19.60
3    19.46
4    19.21
Name: night bedtime, dtype: float64
```

Now find the sample mean bedtime for nap and no_nap.

In [48]:

```
nap_mean_bedtime = bedtime_nap.mean()
```

In [49]:

```
no_nap_mean_bedtime = bedtime_no_nap.mean()
```

Now find the sample standard deviation for X_{nap} and $X_{no\ nap}$.

In [36]:

```
# The np.std function can be used to find the standard deviation. The
# ddof parameter must be set to 1 to get the sample standard deviation.
# If it is not, you will be using the population standard deviation which
# is not the correct estimator
nap_s_bedtime = bedtime_nap.std(ddof=1)
```

In [37]:

```
no_nap_s_bedtime = bedtime_no_nap.std(ddof=1)
```

Now find the standard error for $\bar{X}_{nap}$ and $\bar{X}_{no\ nap}$.

In [38]:

```
nap_se_mean_bedtime = nap_s_bedtime/np.sqrt(len(bedtime_nap))
```

In [39]:

```
no_nap_se_mean_bedtime = no_nap_s_bedtime/np.sqrt(len(bedtime_no_nap))
```

Question: Given our sample sizes of n_1 and n_2 for napping and non napping toddlers respectively, how many degrees of freedom (df) are there for the associated t distributions?

To build a 95% confidence interval, what is the value of t*? You can find this value using the percent point function (PPF):

```
from scipy.stats import t

t.ppf(probability, df)
```

This will return the quantile value such that to the left of this value, the tail probabiliy is equal to the input probabiliy (for the specified degrees of freedom).

Example: to find the t^* for a 90% confidence interval, we want t^* such that 90% of the density of the t distribution lies between $-t^*$ and t^*.

Or in other words if $X \sim t(df)$:

P($-t^*$ < X < t^*) = .90

Which, because the t distribution is symmetric, is equivalent to finding t^* such that:

P(X < t^*) = .95

(0.95 = 1 - (1 - confidence) / 2 = 1 - 0.1 / 2 = 1 - 0.05)

So the t^* for a 90% confidence interval, and lets say df=10, will be:

t_star = t.ppf(.95, df=10)

In [40]:

```
def welch_dof(x,y):
        dof = (x.var()/x.size + y.var()/y.size)**2 / ((x.var()/x.size)**2 / (x.size-1) + (y
        print(f"Welch-Satterthwaite Degrees of Freedom= {dof:.4f}")

welch_dof(bedtime_nap, bedtime_no_nap)
```

```
Welch-Satterthwaite Degrees of Freedom= 7.9679
```

In [64]:

```
n1 = df.napping.value_counts()[1]
print(n1)
n2 = df.napping.value_counts()[0]
print(n2)
```

```
15
5
```

In [79]:

```
# Find the t_stars for the 95% confidence intervals
from scipy.stats import t
nap_t_star = t.ppf(.975, df=n1-1)
nap_t_star
```

Out[79]:

```
2.1447866879169273
```

In [78]:

```
no_nap_t_star = t.ppf(.975, df=n2-1)
no_nap_t_star
```

Out[78]:

```
2.7764451051977987
```

In [ ]:

```

```

Quesion: What is t^* for nap and no nap?

Now to create our confidence intervals. For the average bedtime for nap and no nap, find the upper and lower bounds for the respective 95% confidence intervals.

In [81]:

```
UB_nap = nap_mean_bedtime + nap_t_star * nap_se_mean_bedtime/np.sqrt(n1)
LB_nap = nap_mean_bedtime-(nap_t_star*(nap_se_mean_bedtime/np.sqrt(n1)))
(LB_nap,UB_nap)
```

Out[81]:

```
(20.219486539635366, 20.38851346036463)
```

In [82]:

```

print('UB no_nap-',round(no_nap_mean_bedtime+(no_nap_t_star * (no_nap_se_mean_bedtime/np.sc
print('LB no_nap-',round(no_nap_mean_bedtime - (no_nap_t_star *(no_nap_se_mean_bedtime/np.s
```

```
UB no_nap- 19.8719
LB no_nap- 19.3081
```

```
**Question**: What are the 95% confidence intervals for the average bedtime for toddlers
who nap and for toddlers who don't nap?

CI = $\bar{X} \pm \ t^* \cdot s.e.(\bar{X})$
```

In [ ]:

```

```