Hypothesis Tests in Python

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

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

In [2]:

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

In [3]:

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

In [4]:

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

Out[4]:

	id	sex	age (months)	dlmo time	days napped	napping	nap lights outl time	nap sleep onset	nap midsleep	nap sleep offset	nap wake time	na duratic
0	1	female	33.7	19.24	0	0	NaN	NaN	NaN	NaN	NaN	Na
1	2	female	31.5	18.27	0	0	NaN	NaN	NaN	NaN	NaN	Na
2	3	male	31.9	19.14	0	0	NaN	NaN	NaN	NaN	NaN	Na
3	4	female	31.6	19.69	0	0	NaN	NaN	NaN	NaN	NaN	Na
4	5	female	33.0	19.52	0	0	NaN	NaN	NaN	NaN	NaN	Na

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

Questions: What is the overall sample size n? What are the sample sizes of napping and non-napping toddlers?

Hypothesis tests

We will look at two hypothesis test, each with $\alpha = .05$:

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 equivalently:

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

First isolate `night bedtime` into two variables - one for toddlers who nap and one for toddlers who do not nap.

In [5]:

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

In [6]:

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

Now find the sample mean bedtime for nap and no_nap.

In [7]:

```
nap_mean_bedtime = nap_bedtime.mean()
```

In [8]:

```
no_nap_mean_bedtime = no_nap_bedtime.mean()
```

Question: What is the sample difference of mean bedtime for nappers minus no nappers?

In [9]:

```
mean_bedtime_diff = nap_mean_bedtime - no_nap_mean_bedtime
mean_bedtime_diff
```

Out[9]:

```
0.7139999999999951
```

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

In [10]:

```
# 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 = nap_bedtime.std(ddof=1)
```

In [11]:

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

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 [12]:

```
n1 = len(nap_bedtime)
n2 = len(no_nap_bedtime)
```

In [13]:

```
num1=(n1-1)*(nap_s_bedtime*nap_s_bedtime) + (n2-1)*(no_nap_s_bedtime*no_nap_s_bedtime)
den1=n1+n2-2
num2= (1/n1 + 1/n2)
pooled_se = np.sqrt((num1 * num2)/ den1)
pooled_se
```

Out[13]:

```
0.2961871280370147
```

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

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 [14]:

```
tstat = (nap_mean_bedtime-no_nap_mean_bedtime)/pooled_se
tstat
```

Out[14]:

```
2.4106381824626966
```

Question: What is the p-value for the first hypothesis test?

For a discussion of probability density functions (PDF) and cumulative distribution functions (CDF) see:

https://integratedmlai.com/normal-distribution-an-introductory-guide-to-pdf-and-cdf/ (https://integratedmlai.com/normal-distribution-an-introductory-guide-to-pdf-and-cdf/)

To find the p-value, we can use the CDF for the t-distribution:

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

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

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

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

returns $P(X > tstat)$

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

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

In [ ]:

```

```

In [17]:

```
deg_fre=18
p_val=1-t.cdf(np.abs(tstat), deg_fre)
p_val
```

Out[17]:

```
0.013417041438843036
```

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 [18]:

```
from scipy import stats

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

Out[18]:

```
Ttest_indResult(statistic=2.4106381824626966, pvalue=0.026834082877686044)
```

Question: For the $\alpha = .05$, do you reject or fail to reject the first hypothesis?

Question: For the $\alpha = .05$, do you reject or fail to reject the second hypothesis?