Exercise1_SOLVED

September 13, 2023

1 Exercise 1: Statistical study of smokers

1.0.1 NBE-4070: Basics of Biomedical Data Analysis

Stephane Deny: stephane.deny@aalto.fi

Carlos Sevilla Salcedo: carlos.sevillasalcedo@aalto.fi

Hyunkyung Choo: hyunkyung.choo@aalto.fi

2 Problem description

You are a medical researcher studying the effects of smoking on heart health. Over the years, previous research has found that blood pressure is an important indicator of heart health. Your objective here is to assess the effect of smoking on two types of blood pressure which are equally important in monitoring heart health: systolic (pressure in the arteries when the heart beats) and diastolic (pressure in the arteries when the heart rests between beats).

You have access to a dataset of 100 subjects: 34 smokers and 66 non-smokers. For each subject, the 4 following features have been measured: - age, - weight, - systolic blood pressure, - diastolic blood pressure.

NOTE: This dataset has been artificially generated for the purpose of the exercise. It has not been collected from real subjects.

2.1 Data loading

The dataset is stored in the file `HospitalData.pickle`. The code below loads this file and stores the feature values for each subject in variable `X`, the identifier of whether a subjects is or not a smoker in variable `y` (0: non-smoker, 1: smoker), and the name of each feature in variable `feats`.

2.2 Note:

Run every code in order of comparison.

[1]:

```python
# loading all the relevant packages for the exercise
import numpy as np
import matplotlib.pyplot as plt
import scipy.io
from scipy import stats
import pickle
```

```
%matplotlib inline

# a function that loads file encoded in the "pickle" format
def load_pickle_file(name):
    with open(name + '.pickle','rb') as f:  # Python 3: open(..., 'wb')
        obj = pickle.load(f)
    return obj

# loading the data file
(X, y, feats) = load_pickle_file('HospitalData')

# printing the shapes of the different data fields
print(X.shape, y.shape, feats)
```

```
(100, 4) (100,) ['Age', 'Weight', 'BloodPressure_Systolic',
'BloodPressure_Diastolic']
```

2.3 1. Vizualizing the data (1 point)

First, vizualise the dataset with the help of bar plots. For this, follow these steps: - Compute the mean systolic and diastolic blood pressure for smokers and non-smokers. Print these values. - Compute the standard deviations around each of those means. Print these values. - In a bar plot, show the mean systolic blood pressure of smokers vs. non-smokers. Indicate the standard deviations around those means with an error bar. - Generate a similar bar plot for diastolic blood pressure.

NOTE 1: As a first step, you can store the data corresponding to subjects that don't smoke in a variable `X0`, and the data from subjects that smoke in a variable `X1`.

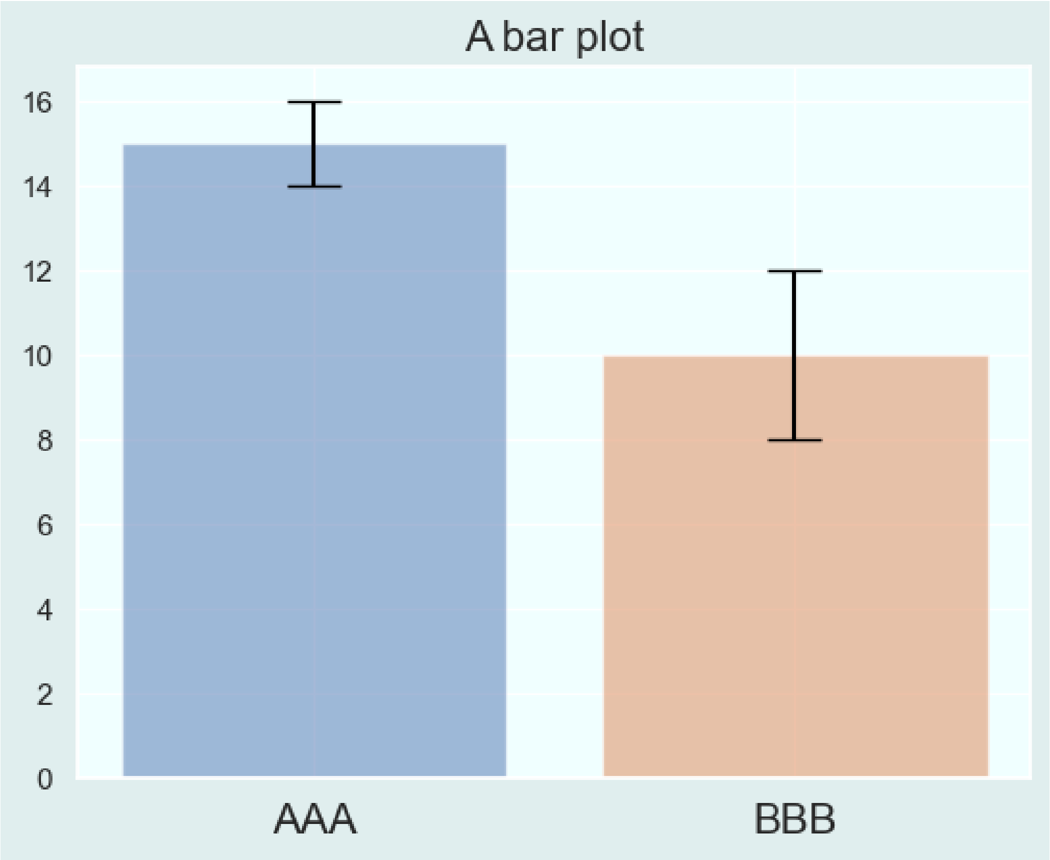
NOTE 2: Plots should always have a title. X and Y axes should always be labelled.

[2]:

```
#FIND SOME HELP HERE

###I'll change color's palette
import seaborn as sns
from matplotlib import colors

color0 = '#458B74'
color1 = '#4B0082'
color2 = '#8B5742'
color3 = '#EE7600'
sns.set(rc={"axes.facecolor": "#F0FFFF", "figure.facecolor": "#E0EEEE"})
palette = ["#2E86C1", "#E67E22",'#8B3626','#8B6914']
cluster_labels = [0, 1, 2, 3]
colors = [color0, color1, color2, color3]
###
```

```python
# This instruction selects only the lines of the matrix X corresponding to␣
 ↪non-smokers (boolean condition y == 0)
X0 = X[y==0]

# This instruction computes the mean of a list or array
m = np.mean([1,2,3])

# This instruction computes the standard deviation of a list or array
sd = np.std([1,2,3])

#Printing values in a nice format
print('The blood pressure of the non-smoking cohort has %.2f mean, %.2f␣
 ↪standard deviation' %(10.0, 15.0))

# This set of instructions draws a simple bar plot with random values
fig, ax = plt.subplots()
ax.bar(0, 15, yerr=1, align='center', alpha=0.5, ecolor='black', capsize=10,␣
 ↪label='Smokers')
ax.bar(1, 10, yerr=2, align='center', alpha=0.5, ecolor='black', capsize=10,␣
 ↪label='Non-smokers')
ax.set_xticks(range(2))
ax.set_xticklabels(['AAA', 'BBB'], fontsize = 16)
ax.set_title('A bar plot', fontsize = 16)
ax.yaxis.grid(True)

print(len(X0))
```

```
The blood pressure of the non-smoking cohort has 10.00 mean, 15.00 standard
deviation
66
```

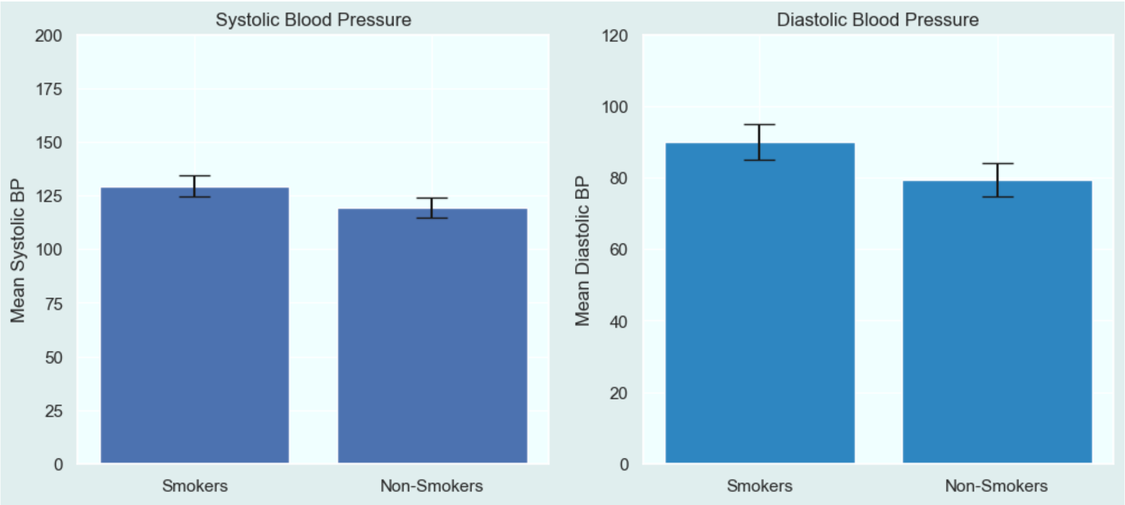


[3]:

```python
#CODE YOUR SOLUTION HERE

# Data separation for smokers and non-smokers
X1 = X[y == 1]  # Data for smokers
X0 = X[y == 0]  # Data for non-smokers

# Compute the mean and standard deviation values
def calculate_stats(data):
    mean = np.mean(data)
    std = np.std(data)
    return mean, std

# Calculate mean and standard deviation for systolic and diastolic blood pressure
mean_systolic_smokers, std_systolic_smokers = calculate_stats(X1[:, 2])  # Column 2 represents systolic blood pressure
mean_systolic_nonsmokers, std_systolic_nonsmokers = calculate_stats(X0[:, 2])
mean_diastolic_smokers, std_diastolic_smokers = calculate_stats(X1[:, 3])  # Column 3 represents diastolic blood pressure
mean_diastolic_nonsmokers, std_diastolic_nonsmokers = calculate_stats(X0[:, 3])
```

```python
# Print the mean and standard deviation values
print("Systolic Blood Pressure:")
print("Mean for Smokers:", mean_systolic_smokers)
print("Mean for Non-Smokers:", mean_systolic_nonsmokers)
print("Standard Deviation for Smokers:", std_systolic_smokers)
print("Standard Deviation for Non-Smokers:", std_systolic_nonsmokers)

print("\nDiastolic Blood Pressure:")
print("Mean for Smokers:", mean_diastolic_smokers)
print("Mean for Non-Smokers:", mean_diastolic_nonsmokers)
print("Standard Deviation for Smokers:", std_diastolic_smokers)
print("Standard Deviation for Non-Smokers:", std_diastolic_nonsmokers)

# Create subplots for bar plots
fig, axes = plt.subplots(1, 2, figsize=(12, 5))

# Plot data for each label (systolic and diastolic)
for i, label in enumerate(['Systolic', 'Diastolic']):
    smokers_mean = mean_systolic_smokers if i == 0 else mean_diastolic_smokers
    nonsmokers_mean = mean_systolic_nonsmokers if i == 0 else␣
 ↪mean_diastolic_nonsmokers
    smokers_std = std_systolic_smokers if i == 0 else std_diastolic_smokers
    nonsmokers_std = std_systolic_nonsmokers if i == 0 else␣
 ↪std_diastolic_nonsmokers
    colore = "#2E86C1" if i == 1 else None  # Set color only for 'Diastolic'␣
 ↪label

    axes[i].bar(['Smokers', 'Non-Smokers'], [smokers_mean, nonsmokers_mean],␣
 ↪yerr=[smokers_std, nonsmokers_std], capsize=10, color=colore)
    axes[i].set_ylabel(f'Mean {label} BP')
    axes[i].set_title(f'{label} Blood Pressure')
    axes[i].set_ylim(0, 200) if i == 0 else axes[i].set_ylim(0, 120)

plt.show()
```

```
Systolic Blood Pressure:
Mean for Smokers: 129.35294117647058
Mean for Non-Smokers: 119.39393939393939
Standard Deviation for Smokers: 4.910267471696284
Standard Deviation for Non-Smokers: 4.631508648374012

Diastolic Blood Pressure:
Mean for Smokers: 89.91176470588235
Mean for Non-Smokers: 79.37878787878788
Standard Deviation for Smokers: 4.978587367911785
Standard Deviation for Non-Smokers: 4.647466404968155
```



2.3.1 Questions (1 point):

- In general, what does the standard deviation represent?
- Do you notice a difference in the range of blood pressure between smokers and non-smokers for each type of blood pressure (systolic and diastolic)?
- Using the chart below, do you think that the smokers in this study are healthy?

source: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

Preliminary

```
 *Systolic Blood Pressure for Smokers*: Smokers in our dataset have an average systolic blo

*Systolic Blood Pressure for Non-Smokers*: Non-smokers in our dataset have an average systo

*Diastolic Blood Pressure for Smokers*: In the case of diastolic blood pressure, smokers ha

*Diastolic Blood Pressure for Non-Smokers*: Non-smokers also exhibit stable diastolic blood
```

2.3.2 Answer

```
1. Standard deviation measures the dispersion of data points around the mean. In simpler te

2. For both systolic and diastolic blood pressure: Smokers exhibit relatively consistent va

3. Small Dataset Caution: It's crucial to consider that our dataset is relatively small, an
```

2.4 2. Assessing statistical significance by vizualizing confidence intervals (1 point)

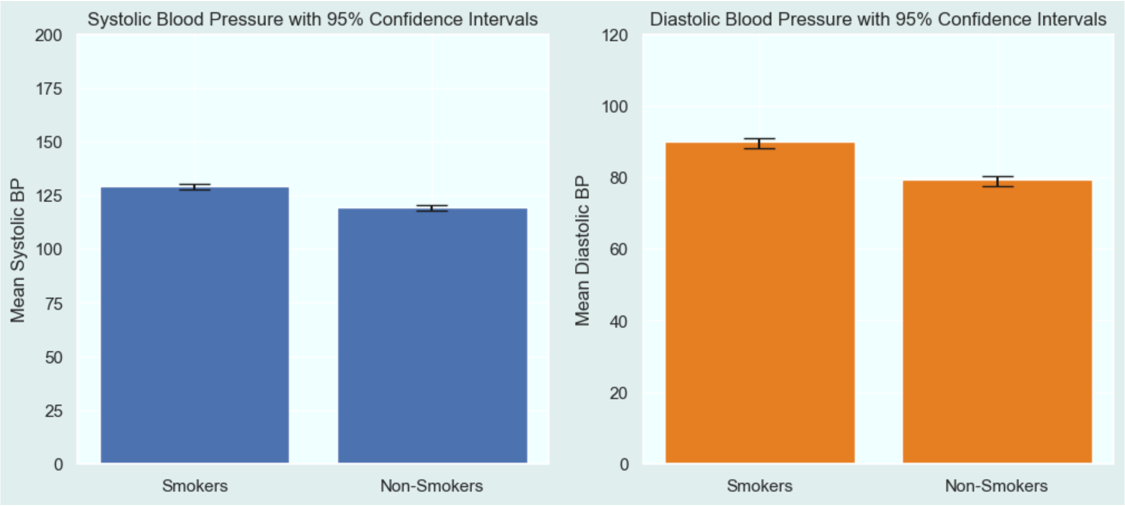
You would now like to know if the differences in blood pressure observed between smokers and non-smokers are statistically significant.

First, vizualise the 95% confidence intervals around the means: - Compute the 95% confidence interval around the mean for all conditions (smokers, non-smokers, systolic, diastolic). - Generate the same bar plots as previously, but where the error bars indicate 95% confidence intervals instead of standard deviations.

[4]:

```
# Define a function to calculate confidence intervals
def calculate_confidence_intervals(data):
    mean = np.mean(data)
    std = np.std(data)
    n = len(data)
    confidence_interval = stats.norm.interval(0.95, loc=mean, scale=std/np.
 ↪sqrt(n))
    return confidence_interval

# Define labels and data for systolic and diastolic
labels = ['Systolic', 'Diastolic']
data = [X1[:, 2], X1[:, 3], X0[:, 2], X0[:, 3]]

# Calculate confidence intervals for each condition
confidence_intervals = [calculate_confidence_intervals(d) for d in data]

# Extract lower and upper bounds of confidence intervals
lower_bounds = [ci[0] for ci in confidence_intervals]
upper_bounds = [ci[1] for ci in confidence_intervals]

# Create subplots
fig, axes = plt.subplots(1, 2, figsize=(12, 5))

# Plot data for each label (systolic and diastolic) with confidence intervals
for i in range(2):
    mean_smokers, mean_nonsmokers = np.mean(data[i]), np.mean(data[i + 2])

    color = "#E67E22" if i == 1 else None  # Set color only for 'Diastolic'␣
 ↪label
    axes[i].bar(['Smokers', 'Non-Smokers'],  [mean_smokers,␣
 ↪mean_nonsmokers],color=color, yerr=[(mean_smokers - lower_bounds[i],␣
 ↪upper_bounds[i] - mean_smokers), (mean_nonsmokers - lower_bounds[i + 2],␣
 ↪upper_bounds[i + 2] - mean_nonsmokers)], capsize=10)
    axes[i].set_ylabel(f'Mean {labels[i]} BP')
    axes[i].set_title(f'{labels[i]} Blood Pressure with 95% Confidence␣
 ↪Intervals')
    axes[i].set_ylim(0, 200) if i == 0 else axes[i].set_ylim(0, 120)

plt.show()

# Print confidence intervals
print("95% Confidence Intervals:")
```

```
for i, label in enumerate(labels):
    print(f"{label} - Smokers: ({lower_bounds[i]:.2f}, {upper_bounds[i]:.2f})")
    print(f"{label} - Non-Smokers: ({lower_bounds[i + 2]:.2f}, {upper_bounds[i␣
 ↪+ 2]:.2f})")
```



```
95% Confidence Intervals:
Systolic - Smokers: (127.70, 131.00)
Systolic - Non-Smokers: (118.28, 120.51)
Diastolic - Smokers: (88.24, 91.59)
Diastolic - Non-Smokers: (78.26, 80.50)
```

2.4.1 Questions (1 point):

- What does the 95% confidence interval represent?
- From these vizualisations, would you guess that there is a statistically significant difference in average blood pressure between smokers and non-smokers, for the different types of blood pressure?

```
1.The 95% confidence interval represents a range of values within which we can be 95% confiden

2.In both cases, the confidence intervals are relatively narrow when compared to the average va
```

2.5 3. Assessing statistical significance using t-tests (1 point)

You will now perform t-tests to verify that the average blood pressure observed in smokers and non-smokers are statistically different: - Decide which type of t-test is appropriate for this test: paired, unpaired, one-tailed, two-tailed? - Apply the selected t-test for the different blood pressure types (systolic, diastolic). - Print the p-values resulting from these tests.

[5]:
```
#CODE YOUR SOLUTION HERE
from scipy.stats import ttest_ind
```

```python

# Perform unpaired two-tailed t-test for systolic blood pressure
t_stat_systolic, p_value_systolic = ttest_ind(X1[:, 2], X0[:, 2])

# Perform unpaired two-tailed t-test for diastolic blood pressure
t_stat_diastolic, p_value_diastolic = ttest_ind(X1[:, 3], X0[:, 3])

# Print the p-values
print("Systolic Blood Pressure - p-value:", p_value_systolic)
print("Diastolic Blood Pressure - p-value:", p_value_diastolic)
```

```
Systolic Blood Pressure - p-value: 2.231437597803922e-16
Diastolic Blood Pressure - p-value: 1.8963069008457136e-17
```

[6]:

```python
#FIND SOME HELP HERE

# t-test documentation:
# https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.
 ↪html
# https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_rel.
 ↪html
```

2.5.1 Questions (1 point):

- What is a t-test?
- Which type of t-test did you apply and why?
- Do you find statistically significant differences in average blood pressure between smokers and non-smokers, for the different types of blood pressure?

```
1. A t-test is a statistical hypothesis test used to determine if there is a significant diffe

2. I applied the unpaired two-tailed t-test. This choice was made because we were comparing two

3. These p-values, which are both extremely close to zero, indicate that there is a highly stat
```

2.6 4. Testing for possible confounding factors (1 point)

In this dataset, we have found differences in blood pressure between smokers and non-smokers. However, other existing differences between the smokers and non-smokers selected for this study may explain the differences in blood pressure. Such other possibles causes for the effect observed are called confounding factors. For example, the smokers in this study could be older than the non-smokers: age could be a confouding factor.

In order to test for potential confounding factors: - Run a t-test on the age and weight of the two cohorts (smokers vs. non-smokers). - Print the p-values for these tests.

[7]:

```python
# CODE YOUR SOLUTION HERE

# Select the age and weight data for smokers and non-smokers
```

```python
age_smokers = X1[:, 0]  # Age data for smokers
age_nonsmokers = X0[:, 0]  # Age data for non-smokers

weight_smokers = X1[:, 1]  # Weight data for smokers
weight_nonsmokers = X0[:, 1]  # Weight data for non-smokers

# Perform t-tests
age_t_stat, age_p_value = ttest_ind(age_smokers, age_nonsmokers)
weight_t_stat, weight_p_value = ttest_ind(weight_smokers, weight_nonsmokers)

# Print the p-values for the t-tests
print("Age t-test p-value:", age_p_value)
print("Weight t-test p-value:", weight_p_value)
```

```
Age t-test p-value: 0.5517313283409775
Weight t-test p-value: 0.03122827941228747
```

2.6.1 Questions (1 point):

- Do you find statistically significant differences in age or weight between the two cohorts (smokers vs. non-smokers)?
- How do you interpret these results? Comment on the possible causality relations between weight, smoking status and blood pressure in this study.

```
1. There are no statistically significant differences in age between smokers and non-smokers, n

2. The results indicate that age is not a significant factor contributing to the differences in
On the other hand, the significant difference in weight between smokers and non-smokers suggest
However, it's important to note that these are observational findings, and establishing causali
```