BT-3172: Special Topics in Bioinformatics: Practical computing for bioinformatics
Lab 5: Biological hypothesis testing using Python.

W.M. Ayesha Sanahari | s13722 | 2017s16470

In this practical, you will learn how to import and use Numpy, Scipy, Pandas, Matplotlib, and Seaborn packages to analyze, describe, and visualize biological data and perform biological hypothesis testing.

After using PyCharm to write your scripts, **copy the codes to the appropriate space below the questions**. Also, submit the Python files separately so we can test them. Use the following format to name each script: YourIndexNo_PrimaryQuestion.py (submit four programs for the four questions)

1) One-sample t-test.
 Question description: Normal human body temperature, as kids are taught in North America, is 98.6°F. But how well is this supported by data? Researchers obtained body-temperature measurements on randomly chosen healthy people (Shoemaker 1996), which can be found in "Temperature.csv" file.

   I. Write the null and alternative hypotheses for the above research question.

      Null hypothesis (H_O): Population mean of normal human body temperature is 98.6°F.

      Alternative hypothesis (H_A): Population mean of normal human body temperature is not 98.6°F.

   II. What are the assumptions when performing the above test?

      Population is normally distributed.

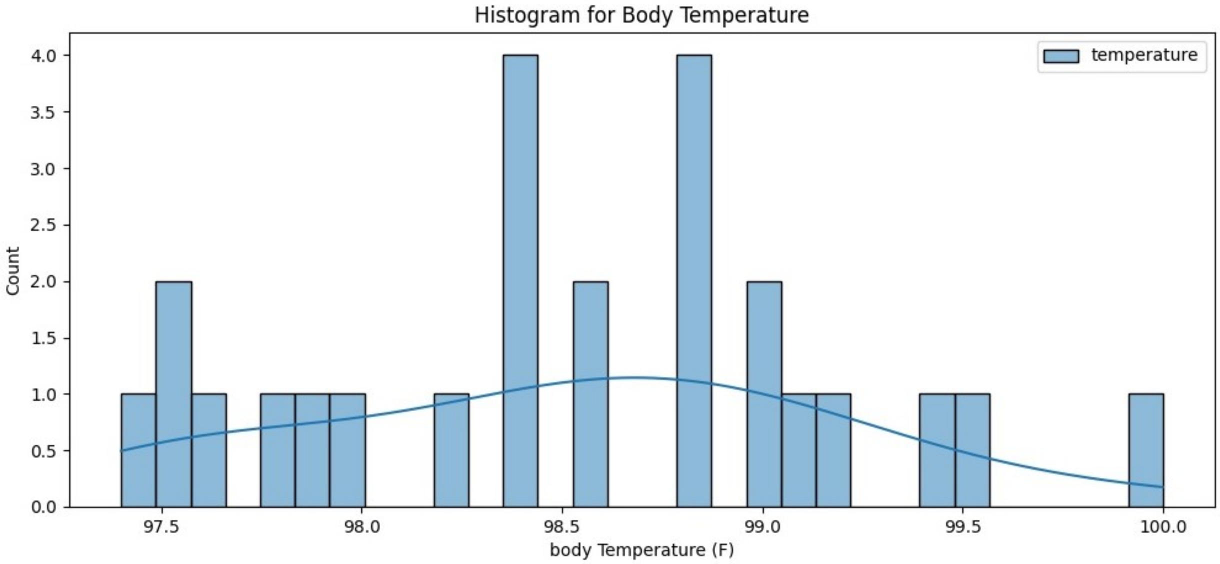
      Random sampling was done.

      Independent observations/ measurements

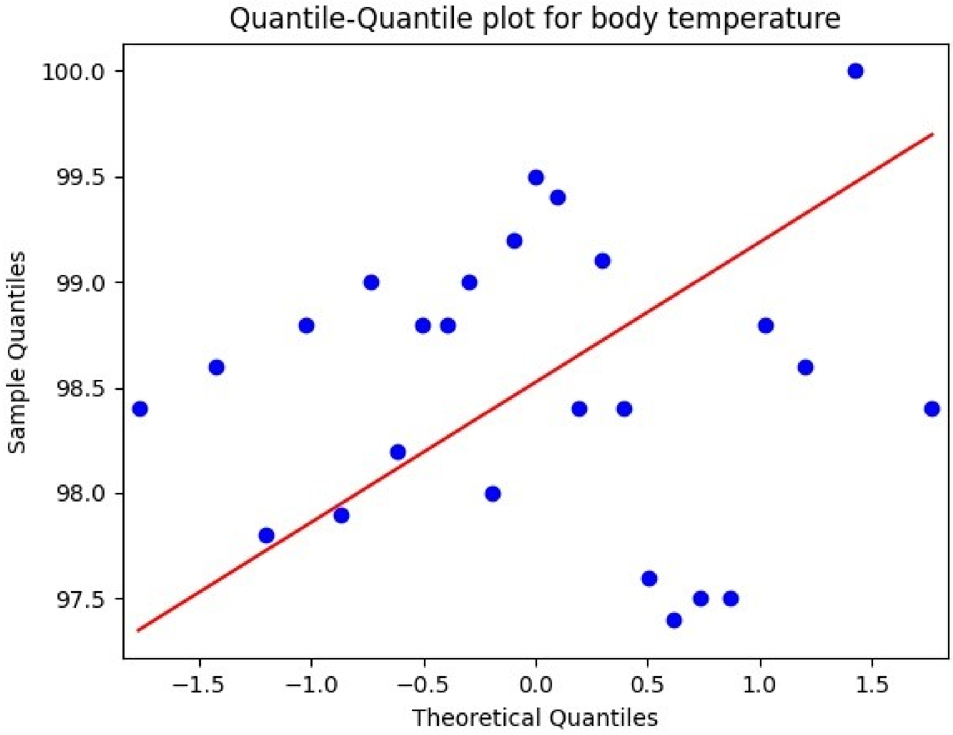
   III. Import the Scipy-stat, Matplotlib, and Seaborn packages/sub modules. Import the data set into a Pandas DataFrame. Write down the following statistics for the human temperature variable: mean, standard deviation, number of observations/count, minimum and maximum values.

```
      temperature
count   25.000000
mean    98.524000
std      0.677791
min     97.400000
max    100.000000
```

   IV. Testing the normality assumption for the temperature variable.
   First, draw the histogram for the variable. Then, draw a Quantile-Quantile plot (QQ plot). Interpret the results of the two plots. Then, perform the Shapiro-Wilk test for normality on the variable. Write the test hypotheses, P-value, and the conclusion for the Shapiro-Wilk test. What type of test would you recommend (parametric or non-parametric)?
   Hint: use statmodels package for the QQ plot generation.



It cannot say that the data is most correctly normal distributed. But up to a certain level, the data seems to be normally distributed.



As data points are scattered and not on the red line, the data is not normally distributed.

Shapiro-Wilk test

Null hypothesis (H_O): Data is Normally distributed.
Alternative hypothesis (H_A): Data is not Normally distributed.

p value: 0.7001275420188904
Conclusion: At 5% significant level, as p value (0.7001275420188904) > α (0.05), H_o is not rejected. So, the data set is Normally distributed.

Parametric test is recommended as the data set is normally distributed.

V. Perform the test you recommended in the previous question. Write the test statistic, P-value and the conclusions clearly.

Test statistics
n<30 (n=25),
σ unknown,
The dataset has only one sample.
Therefore, **one sample t-test** is the most suitable statistical test for this experiment.

p value: 0.58023628
Calculated t value: -0.56064519

Decision
At 5% significance level, as the p value/significance value (0.58023628) is greater than α (0.05), the null hypothesis H_0 is not rejected.

Conclusion
At 5% significance level (or 95% confidence level), there is enough evidences to say that the population mean of normal human body temperature is 98.6°F.
Therefore, normal human body temperature 98.6°F is correct as kids are taught in North America.

```
"""
Author : Ayesha Sanahari
Date : 20/Jan/2021
Perform one sample t test using python
Input: Temperature.csv data file
Output: Results of Statistical test performed
"""

import scipy.stats as sci
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
import statsmodels.api as sm

dataFrame1 = pd.read_csv("./La5_datasets/Temperature.csv")
print(dataFrame1)
print(dataFrame1.describe())

# Draw Histogram
fig,ax = plt.subplots(figsize = (12,5))
```

```
sns.histplot(dataFrame1, ax=ax, bins=30, kde=True)
plt.title("Histogram for Body Temperature")
plt.xlabel("body Temperature (F)")
#plt.show()
plt.savefig('Q1 histogram.jpg')

# qqplot
plot2 = sm.qqplot(dataFrame1, line="s")
plt.title("Quantile-Quantile plot for body temperature")
plt.savefig('Q1 qqplot.jpg')
#plt.show()

#Shapiro test for normality test
stat,p =sci.shapiro(dataFrame1)
print("p value:", p)

#t test
checkvalue= 98.6
t,p = sci.ttest_1samp(dataFrame1,checkvalue)
print("t value:",t)
print("p value:",p)
```

2) Independent two-sample t-test.

Question description: The horned lizard *Phrynosoma mcallii* has many unusual features, including the ability to squirt blood from its eyes. The species is named for the fringe of spikes surrounding the head. Herpetologists recently tested the idea that long spikes help protect horned lizards from being eaten, by taking advantage of the gruesome but convenient behavior of one of their main predators—the loggerhead shrike, *Lanius ludovicianus*. The loggerhead shrike is a small predatory bird that skewers its victims on thorns or barbed wire, to save for later eating. The researchers identified the remains of 30 horned lizards that had been killed by shrikes and measured the lengths of their horns (Young et al. 2004). As a comparison group, they measured the same trait on 154 horned lizards that were still alive and well. These data can be found in "HornedLizards.csv" file. Compare the mean horn lengths of the dead lizards with those of the living lizards.

I. Write the null and alternative hypotheses for the above research question.

Null hypothesis (H_O): The mean horn length of the dead lizards is greater than or equal to the those of living lizards.

Alternative hypothesis (H_A): The mean horn length of the dead lizards is less than the those of living lizards.

II. What are the assumptions when performing the above test?

Population is normally distributed.

Random sampling was done.

Independent observations/ measurements

Equal variance

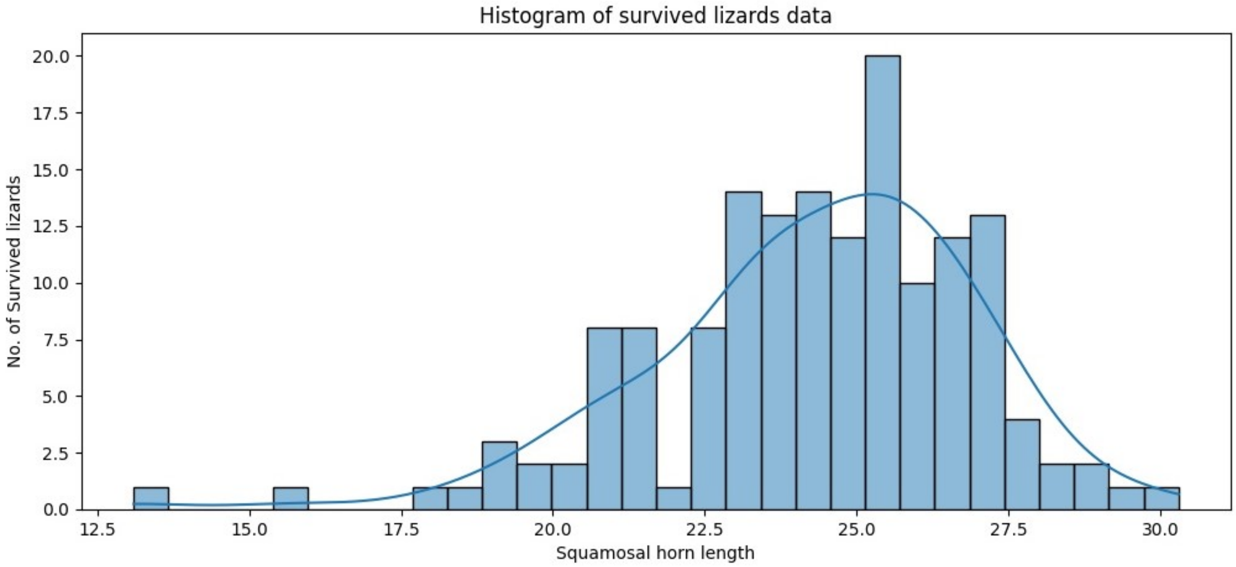
III. Import the data set into a Pandas DataFrame. Write down the following statistics for each variable: mean, standard deviation, number of observations/count.

Hint: sometimes there are missing values in data sets and they should be handled.

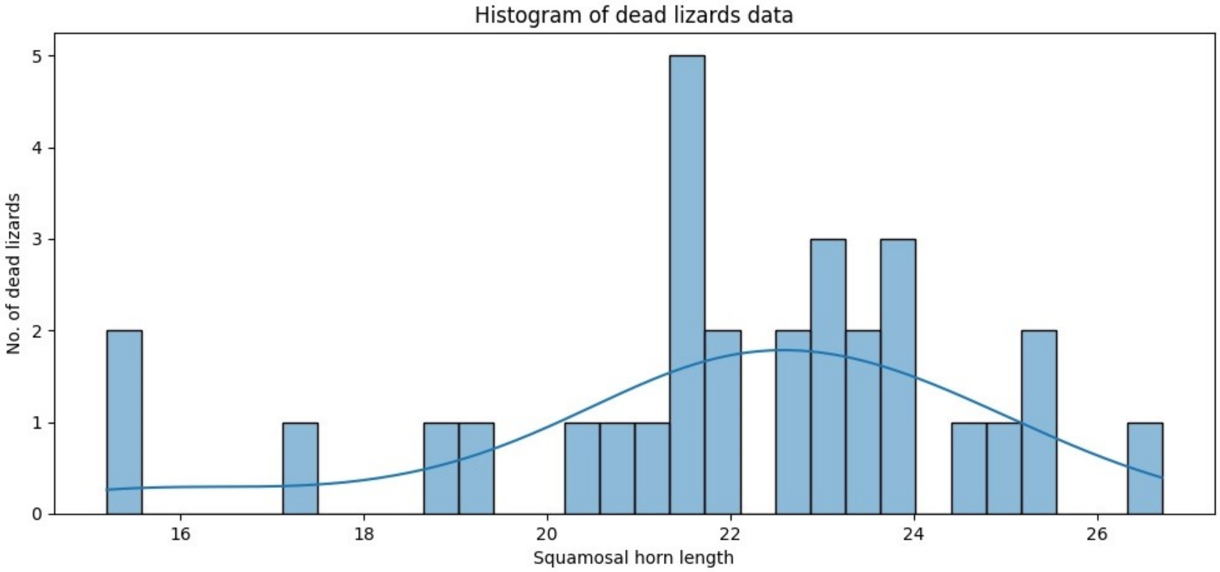
```
  Squamosal horn length                         ...
                         count       mean       std  ...    50%   75%   max
Survive                                              ...
dead                      30.0  21.986667  2.709464  ...  22.25  23.8  26.7
survived                 154.0  24.281169  2.630782  ...  24.55  26.0  30.3
```

IV. Testing the normality assumption for the two independent samples.
First, draw histograms and QQ plots for each sample. Draw the two histograms in the same plot. Then, interpret the results of the histograms and QQ plots. Then, perform the Shapiro-Wilk test for normality on each variable. Write the test hypotheses, P-value, and the conclusion for each variable. What type of test would you recommend (parametric or non-parametric)?

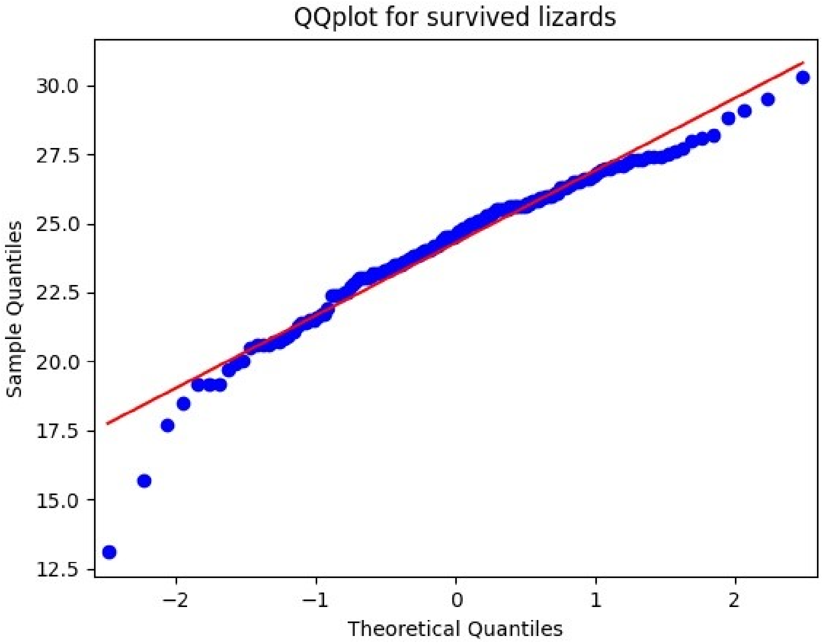
Hint: use statmodels package for QQ plot generation.



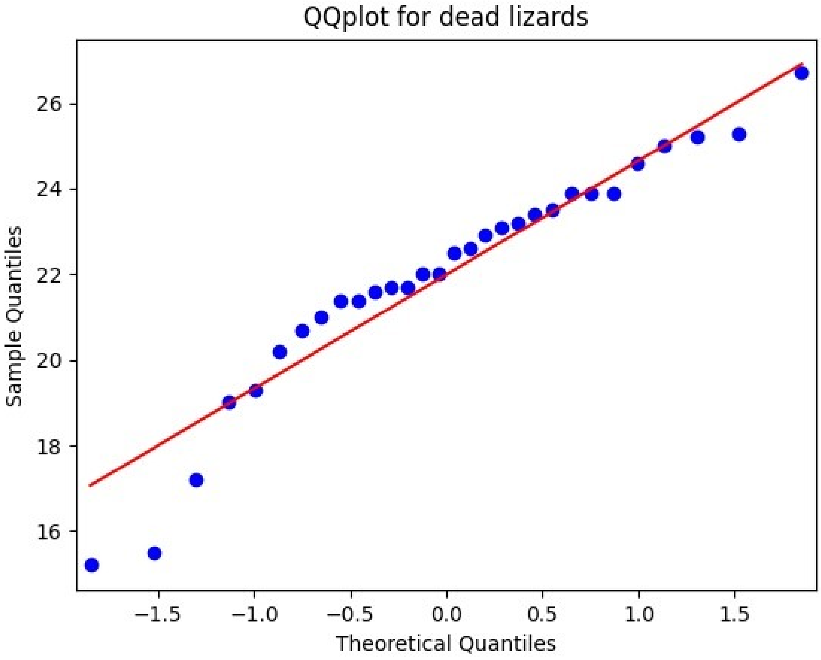
This shows approximately normal distribution of data in the suvived lizard horn length dataset.



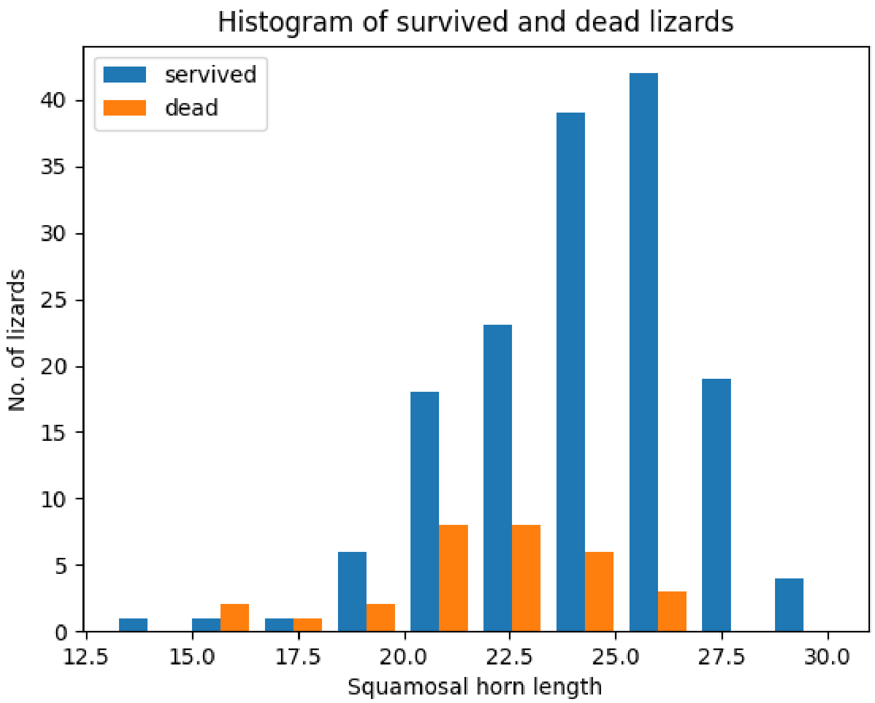
This do not show exact normal distribution of data and left skewed in the dead lizard horn length dataset.



This shows approximately normal distribution of data as most of data points are lies on the red line.



This do not show normal distribution of data as data points are not on the red line.



Shapiro-Wilk test

For survived lizards

Null hypothesis (H_O): Horn length of survived lizards data is Normally distributed.
Alternative hypothesis (H_A): Horn length of survived lizards data is not Normally distributed.

P value for survived: 0.00022339491988532245

Conclusion: At 5% significant level, as p value (0.00022339491988532245) < α (0.05), H_o is rejected. So, the data set of Horn length of survived lizards is not Normally distributed.

For dead lizards

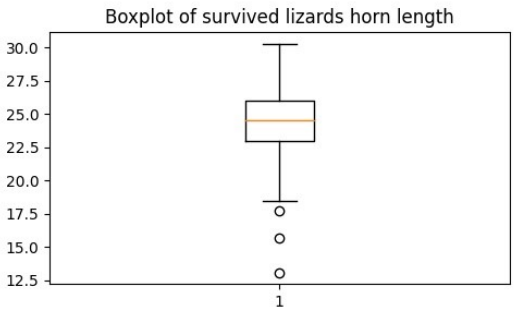
Null hypothesis (H_O): Horn length of dead lizards data is Normally distributed.
Alternative hypothesis (H_A): Horn length of dead lizards data is not Normally distributed.

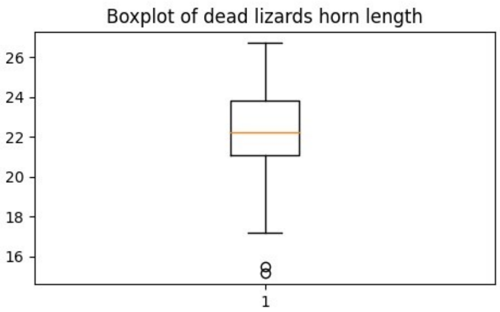
P value for dead: 0.06482069194316864

Conclusion: At 5% significant level, as p value (0.06482069194316864) > α (0.05), H_o is not rejected. So, the data set of Horn length of dead lizards is Normally distributed.

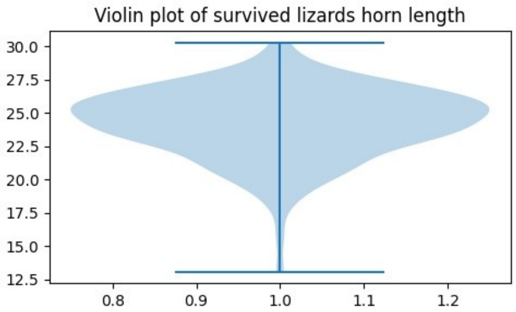
Non-Parametric test is recommended as one of the data set is not normally distributed.

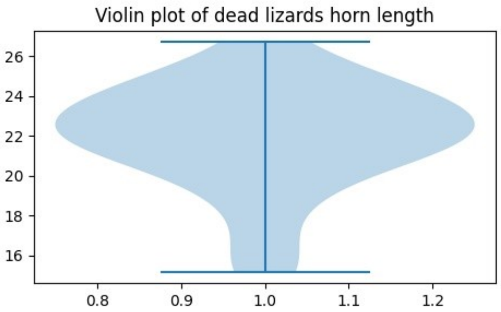
V. Comparison of means. To further visualize the two samples, draw boxplots and a violin plots to compare the distribution of the two samples. What are your interpretations for each plot?


Boxplot of survived lizards horn length


Boxplot of dead lizards horn length

The mean of survived lizards' horn length is greater than that of dead lizards.


Violin plot of survived lizards horn length


Violin plot of dead lizards horn length

The shape of the violin display frequencies of values. Thicker part means the values in that section of the violin has higher frequency, and the thinner part implies lower frequency. So violin plot of dead lizards horn length shows high frequency distribution than survived ones data.

VI. Perform the test you recommended in question IV. Write the test statistic, P-value and the conclusions clearly

Test statistic - Kruskal-Wallis H test
P-value: 2.3459268307095783e-05

Decision
At 5% level of significance, As p value < α (0.05), null hypothesis is rejected
Conclusion
At 5% level of significance, there is a significant difference in the horn length of the survived lizards and dead lizards.

```
"""
Author : Ayesha Sanahari
Date : 20/Jan/2021
Perform Independent two-sample t-test using python
Input: "HornedLizards.csv" data file
Output: Results of Statistical test performed
"""

import scipy.stats as sci
import matplotlib.pyplot as plt
```

```python
import seaborn as sns
import numpy as np
import pandas as pd
import statsmodels.api as sm
from statsmodels.graphics.gofplots import qqplot

dataFrame2 = pd.read_csv("./La5_datasets/HornedLizards.csv")

# Drop rows with missing values
dataFrame2.dropna(inplace=True)

# Group By survive
df_by_survive = dataFrame2.groupby("Survive")
print("\n\n")
print(df_by_survive.describe().head())


# seperate into two variables by grouping variables
servived = dataFrame2.loc[dataFrame2['Survive'] == "survived"]
survived_data=servived['Squamosal horn length']

dead = dataFrame2.loc[dataFrame2['Survive'] == "dead"]
dead_data=dead['Squamosal horn length']

# QQ plot for servived
qqplot(survived_data, line='s')
plt.title('QQplot for survived lizards')
plt.savefig('Q2 QQplot for survived.jpg')

# QQ plot for dead
qqplot(dead_data, line='s')
plt.title('QQplot for dead lizards')
plt.savefig('Q2 QQplot for dead.jpg')


# Histogram of survived lizards data
fig,ax = plt.subplots(figsize = (12,5))
sns.histplot(survived_data, ax=ax, bins=30, kde=True)
plt.title("Histogram of survived lizards data")
plt.xlabel("Squamosal horn length")
plt.ylabel("No. of Survived lizards")
plt.savefig('Q2 histogram survived.jpg')

# Histogram of dead lizards data
fig,ax = plt.subplots(figsize = (12,5))
sns.histplot(dead_data, ax=ax, bins=30, kde=True)
plt.title("Histogram of dead lizards data")
plt.xlabel("Squamosal horn length")
plt.ylabel("No. of dead lizards")
plt.savefig('Q2 histogram dead.jpg')
```

```
# two histograms on one plot
plt.hist([survived_data, dead_data], label=['servived', 'dead'])
plt.title('Histogram of survived and dead lizards')
plt.legend(loc='upper left')
plt.xlabel("Squamosal horn length")
plt.ylabel("No. of lizards")
plt.savefig('Q2 two histo.jpg')

#Shapiro test for servived
stat, p = sci.shapiro(survived_data)
print('P value for survived:',p)

#Shapiro test for dead
stat, p = sci.shapiro(dead_data)
print('P value for dead:', p)

# draw boxplots for two samples
fig,ax =plt.subplots(1,2, figsize=(12,3))
ax[0].boxplot(survived_data)
ax[0].set_title("Boxplot of survived lizards horn length")
ax[1].boxplot(dead_data)
ax[1].set_title("Boxplot of dead lizards horn length")
plt.savefig("q2 boxplot.jpg")

# draw violin plots for two samples
fig,ax =plt.subplots(1,2, figsize=(12,3))
ax[0].violinplot(survived_data)
ax[0].set_title("Violin plot of survived lizards horn length")
ax[1].violinplot(dead_data)
ax[1].set_title("Violin plot of dead lizards horn length")
plt.savefig("q2 violin plot.jpg")

#independed sample t-test
stat,p = sci.kruskal(survived_data,dead_data)
print('P value for independent two sample t-test',p)
```

3) Paired t-test.

In many species, males are more likely to attract females if the males have high testosterone levels. Are males with high testosterone paying a cost for this extra mating success in other ways? One hypothesis is that males with high testosterone might be less able to fight off disease—that is, their high levels of testosterone might reduce their immunocompetent. To test this idea, Hasselquist et al. (1999) experimentally increased the testosterone levels of 13 male red-winged blackbirds by surgically implanting a small permeable tube filled with testosterone. They measured immunocompetence as the rate of antibody production in response to a nonpathogenic antigen in each bird’s blood serum both before and after the implant. The antibody production rates were measured optically, in units of log 10−3 optical density per minute (ln[mOD/min]). The data is available in “BlackbirdTestosterone.csv”

I. Write the null and alternative hypotheses for the above research question

μ_{after} - mean rate of antibody production after treatment with high testosterone levels

μ_{before} - mean rate of antibody production before treatment with high testosterone levels

Null Hypothesis (H0) ; $\mu_{before} - \mu_{after} <= 0$

The mean rate of antibody production difference between before & after treatment with high testosterone levels is less than or equal to zero in male birds.

Alternative Hypothesis (HA) ; $\mu_{before} - \mu_{after} > 0$

The mean rate of antibody production difference between before & after treatment with high testosterone levels is greater than zero in male birds.

II. What are the assumptions when performing the above test?

Population is normally distributed.
Random sampling was done.
Independent observations/ measurements
Equal variance

III. Import the data set into a Pandas DataFrame. Write down the following statistics for each log before, log after, and log difference variable: mean, standard deviation.

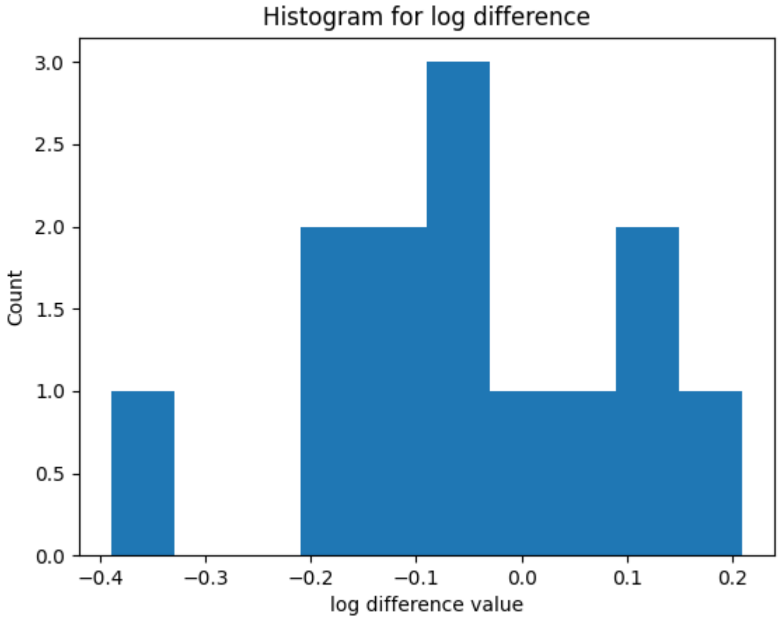
log before:
mean 4.733846
std 0.279837

log after:
mean 4.790000
std 0.261598

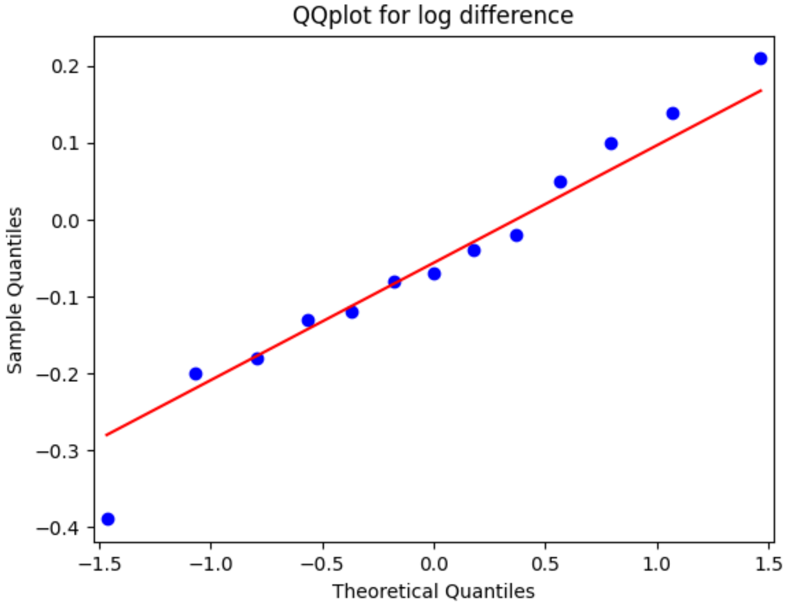
log diff:
mean -0.056154
std 0.159245

IV. Testing the normality assumption for the log difference variable.
First, draw a histogram and a QQ plot for the variable. Interpret the results of the two plots. Then, perform the Shapiro-Wilk test for normality on the variable. Write the test hypotheses, P-value, and the conclusion for the Shapiro-Wilk test. What type of test would you recommend (parametric or non-parametric)?

Hint: use statmodels package for QQ plot generation.



Do not show a normal distribution



Approximately shows a normal distribution of log difference data.

Shapiro-Wilk test

Null hypothesis (H_O): log difference data is Normally distributed.
Alternative hypothesis (H_A): log difference data is not Normally distributed.

P value of log difference: 0.9769595265388489

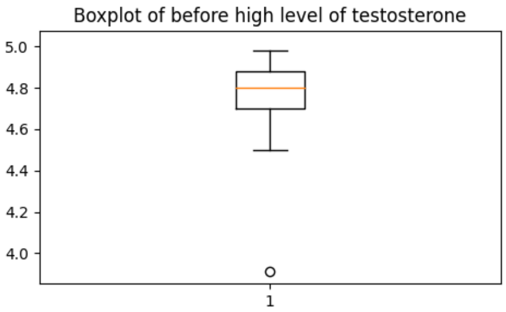
Conclusion: At 5% significant level, as p value (0.9769595265388489) > α (0.05), H_o is not rejected. So, the data set of log difference is Normally distributed.

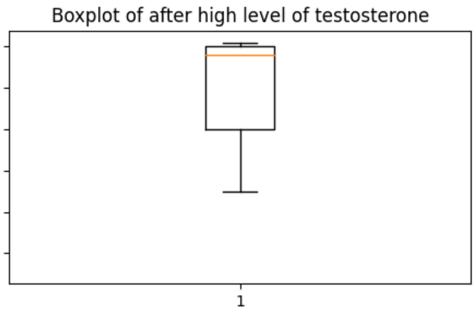
Parametric test is recommended as the log difference data set is normally distributed.

P value of log before: 0.0011588883353397
25

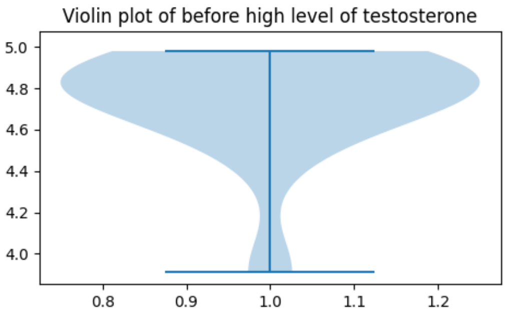
P value of log after: 0.011307741515338421

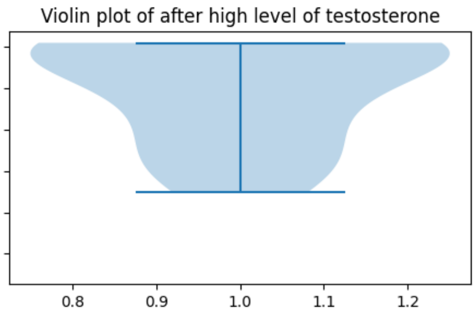
V. Comparison of means. To further visualize the two samples, draw boxplots and violin plots to compare the distribution of the two before and after samples. What are your interpretations for each plot?


Boxplot of before high level of testosterone


Boxplot of after high level of testosterone

The mean rate of antibody production after treatment with high testosterone levels is high than that of before treatment with high testosterone levels.


Violin plot of before high level of testosterone


Violin plot of after high level of testosterone

The shape of the violin display frequencies of values. Thicker part means the values in that section of the violin has higher frequency, and the thinner part implies lower frequency. So, violin plot of after testosterone treatment shows high frequency distribution than before treatment with testosterone.

VI. Perform the test you recommended in question IV. Write the test statistic, P-value and the conclusions clearly

Test statistic

n<30 (n=13),

σ unknown,

The dataset is **dependent** since the **same set of birds is used** to obtain data for before and after treatment with testosterone.

Therefore, **paired samples t-test** is the most suitable statistical test for this experiment.

P value for paired sample t-test 0.2276738727254288

Decision

At 5% level of significance, As p value (0.2276738727254288) > α (0.05), null hypothesis is not rejected.

Conclusion

At 5% level of significance, the mean rate of antibody production difference between before & after treatment with high testosterone levels is significantly less than or equal to zero in male birds.

So, males with high testosterone levels might not reduce their immunocompetent.

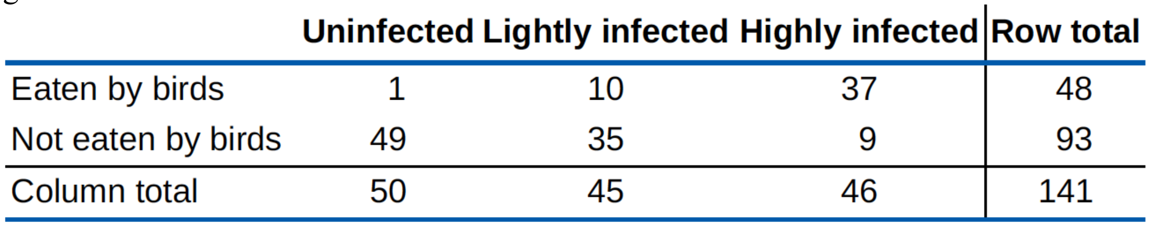
```
"""
 Author : Ayesha Sanahari
 Date : 23/Jan/2021
 Perform Independent two-sample t-test using python
 Input: “BlackbirdTestosterone.csv” data file
 Output: Results of Statistical test performed
"""

import scipy.stats as sci
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
import pandas as pd
from statsmodels.graphics.gofplots import qqplot

# read the datafile
dataFrame3 = pd.read_csv('./La5_datasets/BlackbirdTestosterone.csv')
print(dataFrame3)

# seperate columns
log_before=dataFrame3['log before']
log_after=dataFrame3['log after']
log_diff=dataFrame3['dif in logs']

print('log before:', log_before.describe())
print('log after:', log_after.describe())
print('log diff:', log_diff.describe())

# draw qq plot and histogram for log difference
qqplot(log_diff, line='s')
plt.title('QQplot for log difference')
plt.show()

plt.hist(log_diff)
plt.title("Histogram for log difference")
plt.xlabel('log difference value')
plt.ylabel('Count')
plt.show()

#Shapiro test for Normality
stat, p = sci.shapiro(log_diff)
print('P value of log difference:',p)

stat, p = sci.shapiro(log_before)
print('P value of log before:',p)
stat, p = sci.shapiro(log_after)
print('P value of log after:',p)

# Draw boxplots
fig,ax =plt.subplots(1,2, figsize=(12,3), sharey=True)
ax[0].boxplot(log_before)
ax[0].set_title("Boxplot of before high level of testosterone")
ax[1].boxplot(log_after)
ax[1].set_title("Boxplot of after high level of testosterone")
plt.show()

# Draw violin plots
fig,ax =plt.subplots(1,2, figsize=(12,3), sharey=True)
ax[0].violinplot(log_before)
ax[0].set_title("Violin plot of before high level of testosterone")
ax[1].violinplot(log_after)
ax[1].set_title("Violin plot of after high level of testosterone")
plt.show()

# Paired sample t-test
stat,p = sci.ttest_rel(log_before,log_after)
print('P value for paired sample t-test',p)
```

4) Chi-square contingency test.

Many parasites have more than one species of host, so the individual parasite must get from one host to another to complete its life cycle. Trematodes of the species *Euhaplorchis californiensis* use three hosts during their life cycle. Worms mature in birds and lay eggs that pass out of the bird in its feces. The horn snail *Cerithidea californica* eats these eggs, which hatch and grow to another life stage in the snail, sterilizing the snail in the process. When an infected snail is eaten by the California killifish *Fundulus parvipinnis*, the parasite develops to the next life stage and encysts in the fish's braincase. Finally, when the killifish is eaten by a bird, the worm becomes a mature adult and starts the cycle again.

Researchers have observed that infected fish spend excessive time near the water surface, where they may be more vulnerable to bird predation. This would certainly be to the worm's advantage, as it would increase its chances of being ingested by a bird, its next host. Lafferty and Morris (1996) tested the hypothesis that infection influences risk of predation by birds. A large outdoor tank was stocked with three kinds of killifish: unparasitized, lightly infected, and heavily infected. This tank was left open to foraging by birds, especially great egrets, great blue herons, and snowy egrets.

Observed frequencies of fish eaten or not eaten by birds according to trematode infection level is given below.



	Uninfected	Lightly infected	Highly infected	Row total
Eaten by birds	1	10	37	48
Not eaten by birds	49	35	9	93
Column total	50	45	46	141

It is essential to test whether the probability of being eaten by birds differs according to infection status.

I. Write the null and alternative hypotheses for the above research question.

Null Hypothesis (H0) ;

There is no significant difference in the mean number of fish eaten by birds between the different infection status.

Alternative Hypothesis (HA) ;

There is a significant difference in the mean number of fish eaten by birds between the different infection status.

II. What are the assumptions when performing the above test?

Random sampling was done

Data is normally distributed

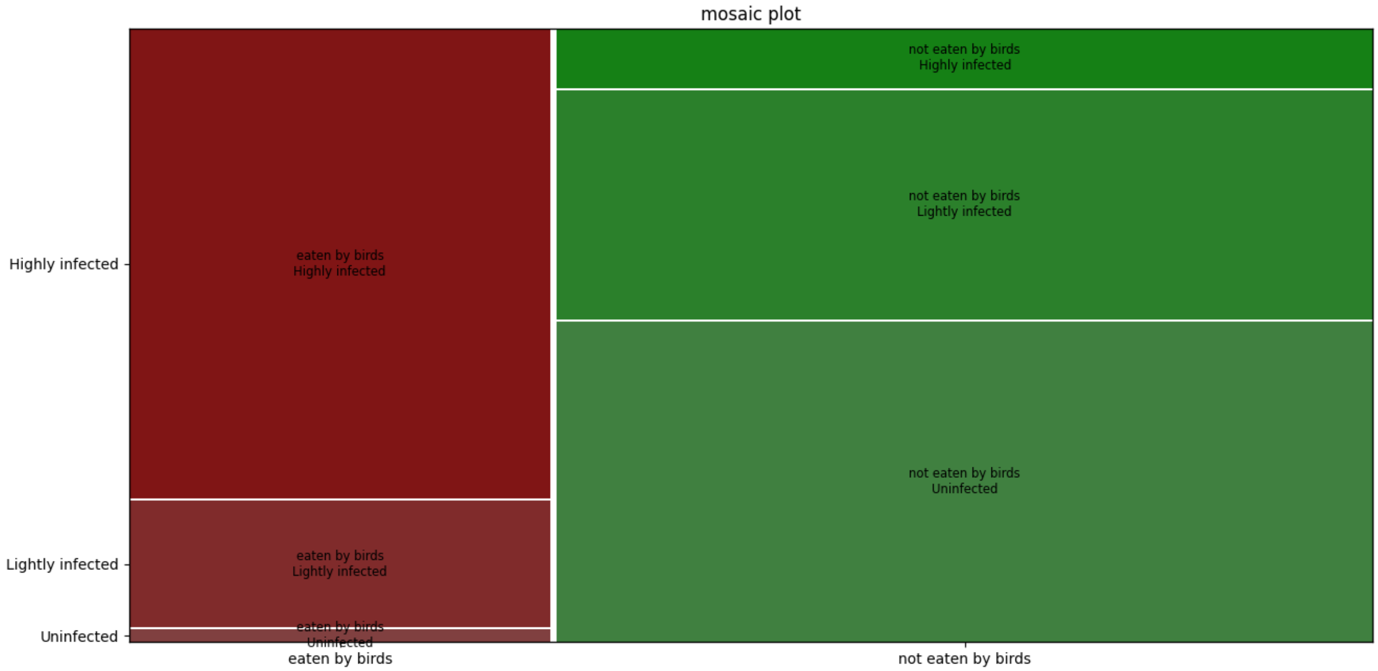
Independent measurement

Sample variances are equal

III. Create the above contingency table in a pandas DataFrame and print it.

IV. Draw a mosaic plot for the above table.

Hint: Use the statmodels package to draw the plot. Also, look into the pandas DataFrame.stack() function. Maybe it will be useful to you.



V. Perform a Chi-square contingency test on above data. Write the Chi-square statistic, the P-value and the degree of freedom value below.

Chi-square statistic: 69.75570515817361
P value : 7.124281712683879e-16
degree of freedom : 2

VI. Output the expected value table. You can create a new DataFrame to store the expected values.

```
          0          1          2
0  17.021277  15.319149  15.659574
1  32.978723  29.680851  30.340426
```

VII. Write your conclusion based on above results.

At a 5% significance level, as p value(7.124281712683879e-16) < α (0.05) H_o is rejected. There is a significant difference in the mean number of fish eaten by birds between the different infection status.
So infection influences risk of predation by birds.

```
"""
 Author : Ayesha Sanahari
 Date : 22/Jan/2021
 Perform Chi-square contingency test using python
"""
import scipy.stats as sci
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.graphics.mosaicplot as mosaicplot

# Creating a Dataframe from the data and printing it
data = {'Uninfected': [1, 49],'Lightly infected': [10, 35],'Highly infected':[37,9]}
dataFrame4 = pd.DataFrame(data, index=['eaten by birds','not eaten by birds'])
print(dataFrame4)

# plot a mosaicplot for the dataframe
mosaicplot.mosaic(dataFrame4.stack(),title='mosaic plot')
plt.show()

# do a chi-square contingency test
chi_value,p,dof,expected = sci.chi2_contingency(dataFrame4)
print('Chi values :',chi_value)
print('P value :',p)
print('degree of freedom :',dof)
```

```
expected = pd.DataFrame(expected)
print(expected)
```

References

- Whitlock, Michael C., and Dolphcoaut Schluter. The analysis of biological data. No. 574.015195 W5. 2009.
- Young, Kevin V., Edmund D. Brodie Jr, and Edmund D. Brodie III. "How the horned lizard got its horns." Science 304.5667 (2004): 65-65.
- Hasselquist, Dennis, et al. "Is avian humoral immunocompetence suppressed by testosterone?." Behavioral Ecology and Sociobiology 45.3-4 (1999): 167-175.
- Lafferty, Kevin D., and A. Kimo Morris. "Altered behavior of parasitized killifish increases susceptibility to predation by bird final hosts." Ecology 77.5 (1996): 1390-1397.