**S13674**

**D.P.Bhagya Hendalage**

**BT-3172: Special Topics in Bioinformatics: Practical computing for bioinformatics**

**Lab 5: Biological hypothesis testing using Python.**

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.

* 1. Write the null and alternative hypotheses for the above research question

Null hypothesis(H0): Normal human body temperature is equal 98.6°F(µ=98.6°F)

Alternative hypothesis(HA): Normal human body temperature is different from 98.6°F(µ≠98.6°F)

* 1. What are the assumptions when performing the above test?

Data is normally distributed

Random sampling is done

Independent observations/measurements

Equal variance

* 1. 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.

**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
**import** pandas **as** pd  
**from** statsmodels.graphics.gofplots **import** qqplot  
**from** scipy **import** stats  
  
*# read the temperarure data file and analyze the descriptive statistics*tem = pd.read\_csv(**"Temperature.csv"**)  
print(tem.describe())

temperature

count 25.000000

mean 98.524000

std 0.677791

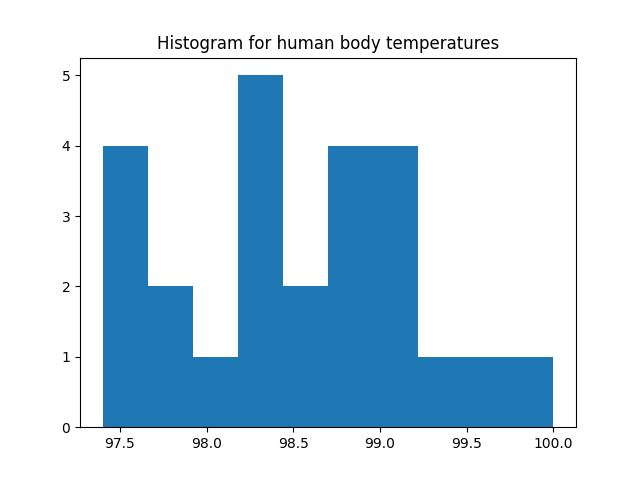
max 100.000000

* 1. Testing the normality assumption for the temperature variable.

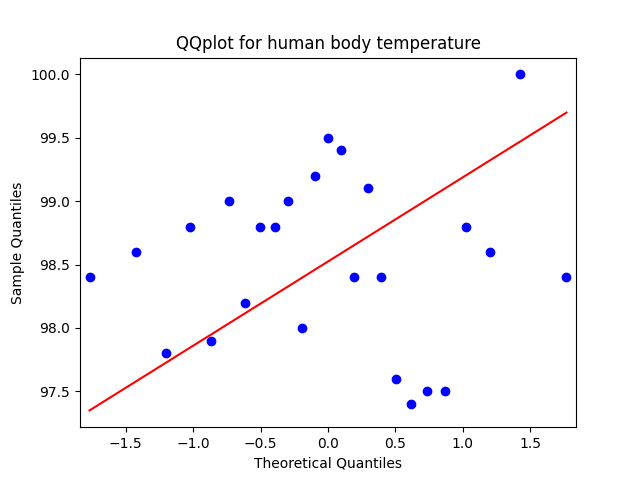
First, draw the histogram for the variable. Then, draw a Quantile-Quantile plot (QQ plot). Interpret the results of the two plots. Then, perform the Shapiro-Wilk test for normality on the variable. Write the test hypotheses, P-value, and the conclusion for the Shapiro-Wilk test. What type of test would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

*# histogram*plt.hist(tem)  
plt.title(**"Histogram for human body temperatures"**)  
plt.show()  
*# QQ plot*qqplot(tem, line=**'s'**)  
plt.title(**'QQplot for human body temperature'**)  
plt.show()

*Hint: use statmodels package for the QQ plot generation.*



When looking at the graph, there is not showing the normal distribution. It show some positive skewness.



When looking at the QQ plot, it show the data is not normally distributed.

Null hypothesis(H0): Data is normally distributed

Alternative hypothesis (HA): Data is not normally distributed

Assumptions: Random sampling is done

Independent observations/measurements

Equal variance

P value: 0.7001275420188904🡪0.7000>0.05🡪fails to reject the null hypothesis

The accept the null hypothesis

Conclusion: Data is normally distributed

The data set is normally distributed then parametric test can be used

One sample is given and n=25 variance is unknown, one sample t-test can be used

*#Shapiro*stat, p = stats.shapiro(tem)  
print(**'P value:'**,p)

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

the test statistic: one sample t-test is suitable for the analysis

P value: 0.58023628🡪0.580>0.05🡪fails to reject the null hypothesis🡪Accept the null hypothesis

Conclusion: Normal human body temperature is equal 98.6°F(µ=98.6°F) at 5% significance level.

*#t-test*checkvalue = 98.6  
t,p =stats.ttest\_1samp(tem, checkvalue)  
print(**'t value'**,t)  
print(**'P value'**,p)

1. 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.

* 1. Write the null and alternative hypotheses for the above research question.

µs: mean horn length of the dead lizards

µa: mean horn length of the alive lizards

Null hypothesis(H0): µd- µa =0

There is no difference in the mean horn length of the dead lizards and the alive lizards

Alternative hypothesis(HA): µd- µa ≠0

There is a difference in the mean horn length of the dead lizards and the alive lizards

* 1. What are the assumptions when performing the above test?

Data is normally distributed

Random sampling is done

Independent observations/measurements

Equal variance

* 1. 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 184.000000

mean 23.907065

std 2.769862

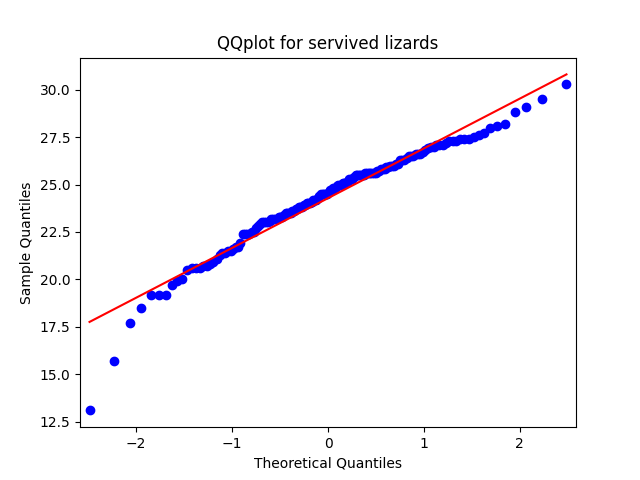
min 13.100000

max 30.300000

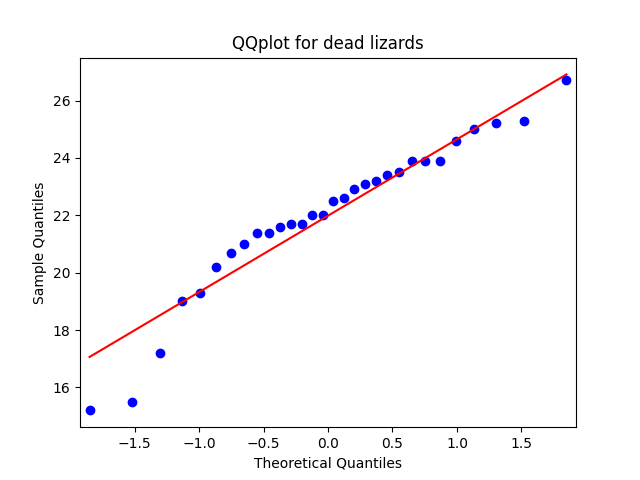
* 1. Testing the normality assumption for the two independent samples.

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)?

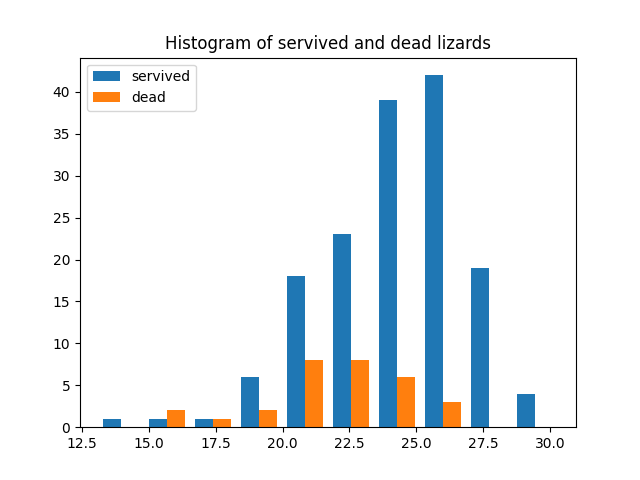
*Hint: use statmodels package for QQ plot generation.*



QQ plot for the survived lizards’ show not normal distribution but can not distinguish correctly.



QQ plot for dead lizards show normal distribution but cannot distinguish correctly.



When consider the histograms, survived lizards’ distribution is not show a normal distribution and it shows left skewed distribution. But in the dead lizards’ data is show a normal distribution.

**For survived lizards**

Null hypothesis(H0): Data is normally distributed

Alternative hypothesis (HA): Data is not normally distributed

Assumptions: Random sampling is done

Independent observations/measurements

Equal variance

P value for survived: 0.00022339491988532245🡪0.0002<0.05🡪does not fail to reject the null hypothesis

The reject the null hypothesis

Conclusion: Data is not normally distributed

**For dead lizards**

Null hypothesis(H0): Data is normally distributed

Alternative hypothesis (HA): Data is not normally distributed

Assumptions: Random sampling is done

Independent observations/measurements

Equal variance

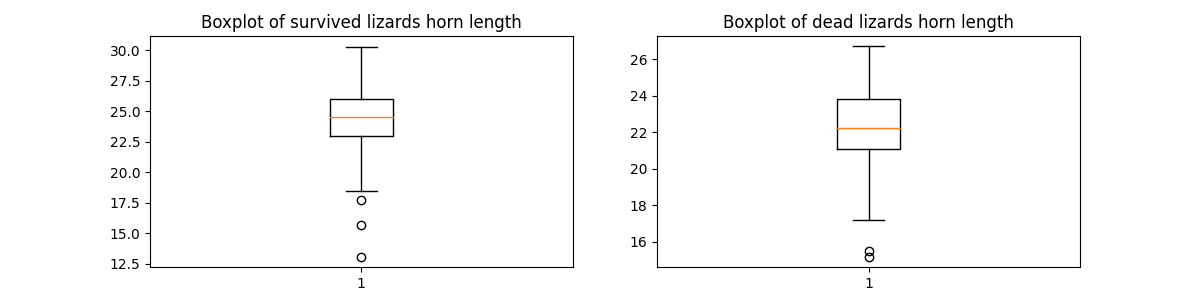
P value for dead: 0.06482069194316864🡪0.0648>0.05🡪fails to reject the null hypothesis

The accept the null hypothesis

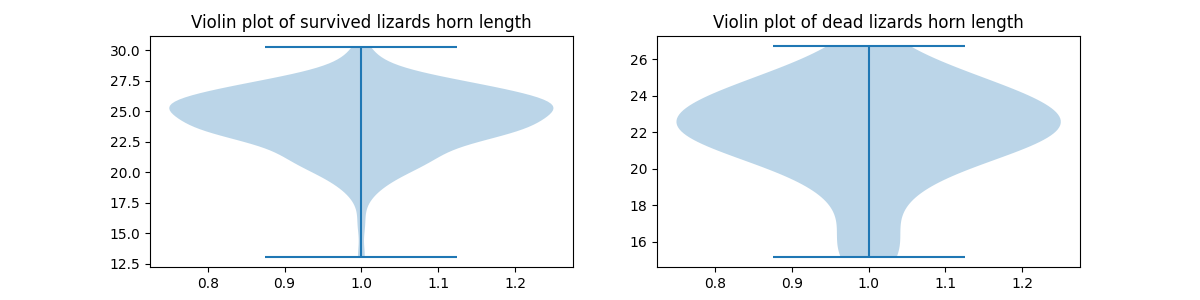
Conclusion: Data is normally distributed

One sample in those two sample data set is not normal distributed then it is suitable doing non-parametric test.

* 1. 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?



Mean of the sample for the survived lizards is much higher than the dead ones



In survived lizards’ violin plot shows shewed upper data density. But in dead lizards show much normally distributed data.

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

Test statistics: there are two sample have different distributions. Sample size is more than 5 and different sample size then it is suitable Kruskal-Wallis H-test for analysis

P value for independent two sample t-test 2.3459268307095783e-05 🡪2.3459e-05<0.05🡪Reject the null hypothesis

Conclusion: There are significant difference between mean horn length of dead lizards and survived lizards at 5% of significance level.

**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
**import** pandas **as** pd  
**from** statsmodels.graphics.gofplots **import** qqplot  
**from** scipy **import** stats  
*#read the datafile*lizards = pd.read\_csv(**'HornedLizards.csv'**)  
*#print(lizards)*lizards[**'Squamosal horn length'**].replace(**''**, np.nan, inplace=**True**)  
*#print(lizards)*lizards.dropna(subset=[**'Squamosal horn length'**], inplace=**True**)  
*#print(lizards)  
  
# Printing descriptive statistics*print(lizards.describe())  
*#print(lizards.rename(columns={'Squamosal horn length':'length'}))  
#to get rid from spaces of the colomn header*lizards=lizards.rename(columns={**'Squamosal horn length'**:**'length'**})  
  
servived = lizards.loc[lizards[**'Survive'**] == **"survived"**]  
servived\_data=servived[**'length'**]  
print(servived\_data)  
  
dead = lizards.loc[lizards[**'Survive'**] == **"dead"**]  
dead\_data=dead[**'length'**]  
print(dead\_data)  
*# QQ plot for servived*qqplot(servived\_data, line=**'s'**)  
plt.title(**'QQplot for servived lizards'**)  
plt.show()  
*# QQ plot for dead*qqplot(dead\_data, line=**'s'**)  
plt.title(**'QQplot for dead lizards'**)  
plt.show()  
  
*# two histograms on one plot*plt.hist([servived\_data, dead\_data], label=[**'servived'**, **'dead'**])  
plt.title(**'Histogram of servived and dead lizards'**)  
plt.legend(loc=**'upper left'**)  
plt.show()  
  
*#Shapiro for servived*stat, p = stats.shapiro(servived\_data)  
print(**'P value for servived:'**,p)  
  
*#Shapiro for dead*stat, p = stats.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(servived\_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.show()  
  
*# draw violin plots for two samples*fig,ax =plt.subplots(1,2, figsize=(12,3))  
ax[0].violinplot(servived\_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.show()  
  
*#independed sample t-test*stat,p = stats.kruskal(servived\_data,dead\_data)  
print(**'P value for independent two sample t-test'**,p)

1. 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”

* 1. Write the null and alternative hypotheses for the above research question.

Null hypothesis(H0)**:** .µhigh - µlow ≤ 0 The difference of mean rate of antibody production between males with high testosterone and low testosterone is equal or less than to zero.

Alternative hypothesis(HA)**:** µhigh - µlow > 0

The difference of mean rate of antibody production between males with high testosterone and low testosterone is greater than zero.

* 1. What are the assumptions when performing the above test?

Data is normally distributed

Random sampling is done

Independent observations/measurements

Equal variance

* 1. 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.

**before data :**

**mean 4.733846**

**std 0.279837**

**after data :**

**mean 4.790000**

**std 0.261598**

**difference data :**

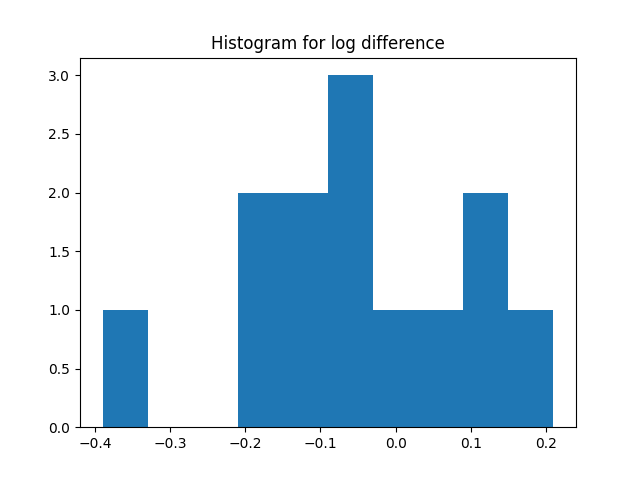
**mean -0.056154**

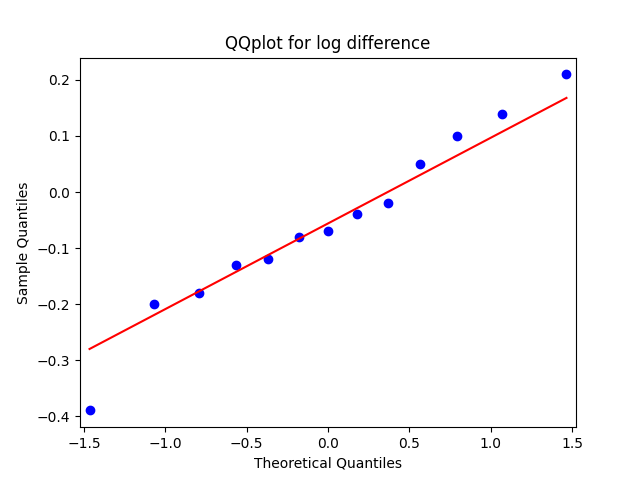
**std 0.159245**

* 1. 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.*





When consider two plots, they are showing normal distribution but can not recognize correctly.

Null hypothesis(H0): Data is normally distributed

Alternative hypothesis (HA): Data is not normally distributed

Assumptions: Random sampling is done

Independent observations/measurements

Equal variance

**P value of log difference:** 0.9769595265388489🡪0.976>0.05🡪fails to reject the null hypothesis

Accept the null hypothesis

Conclusion: Data is normally distributed

But P values for the dataset of before and after

**P value of log before: 0.001158888335339725🡪0.0011<0.05**

reject the null hypothesis

Conclusion: Data is not normally distributed

**P value of log after: 0.011307741515338421🡪0.011<0.05**

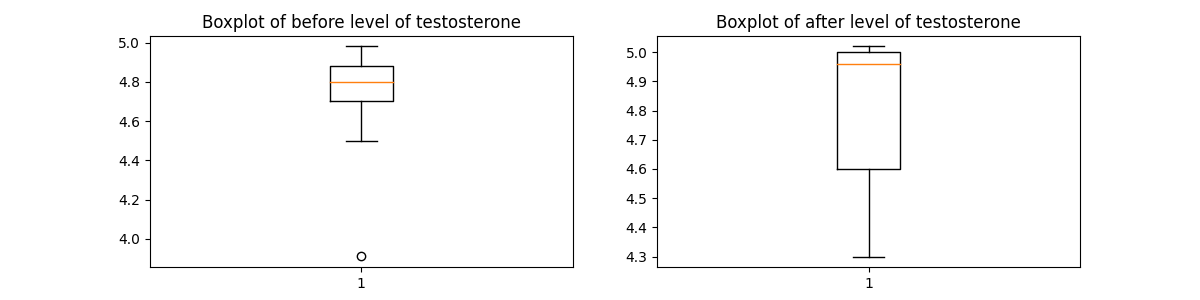
reject the null hypothesis

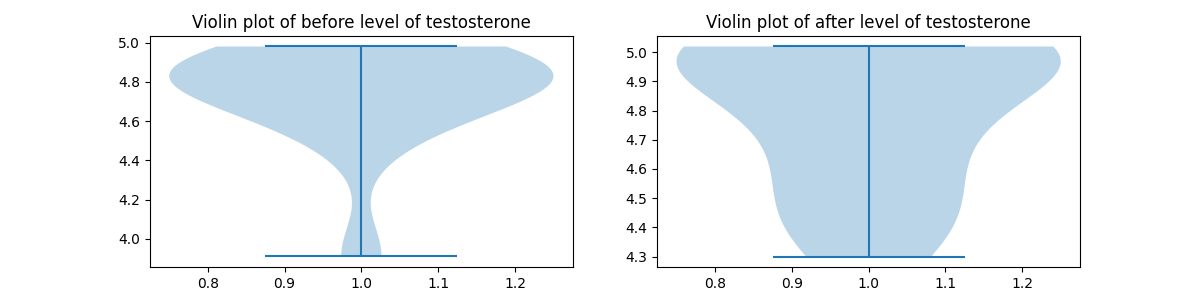
Conclusion: Data is not normally distributed

Then the datasets are not normally distributed for log after and log before values

When analysis before and after data , it is suitable for use non parametric test.

* 1. 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?





When looking at the boxplots there are significant difference in the means. Antibody level after the treatment is much higher than the before antibody level.

When looking at the violin plots there are not any plot shows normal distribution. They are show different data densities.

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

Test statistics: there are two sample have different distributions. Sample size is more than 5 and different sample size then it is suitable Kruskal-Wallis H-test for analysis

P value for paired sample t-test 0.3295474192963141 🡪0.3295>0.05🡪fails to reject the null hypothesis

Conclusion:The difference of mean rate of antibody production between males with high testosterone and low testosterone is equal or less than to zero at 5% of significance level.

**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**import** seaborn **as** sns  
**import** pandas **as** pd  
**from** statsmodels.graphics.gofplots **import** qqplot  
**from** scipy **import** stats  
*#read the datafile*blackbird = pd.read\_csv(**'BlackbirdTestosterone.csv'**)  
*#to get rid from spaces of the colomn header*blackbird=blackbird.rename(columns={**'log before'**:**'before'**,**'log after'**:**'after'**,**'dif in logs'**:**'diff'**})  
print(blackbird)  
*# seperate columns*before\_data=blackbird[**'before'**]  
print(before\_data)  
*# seperate columns*after\_data=blackbird[**'after'**]  
print(after\_data)  
*# seperate columns*diff\_data=blackbird[**'diff'**]  
print(diff\_data)  
*# discriptive statisctics*print(**'before data :'**,before\_data.describe())  
print(**'after data :'**,after\_data.describe())  
print(**'difference data :'**,diff\_data.describe())  
  
*# draw qq plot and histogram for log difference*qqplot(diff\_data, line=**'s'**)  
plt.title(**'QQplot for log difference'**)  
plt.show()  
plt.hist(diff\_data)  
plt.title(**"Histogram for log difference"**)  
plt.show()  
*#Shapiro*stat, p = stats.shapiro(diff\_data)  
print(**'P value of log difference:'**,p)  
stat, p = stats.shapiro(before\_data)  
print(**'P value of log before:'**,p)  
stat, p = stats.shapiro(after\_data)  
print(**'P value of log after:'**,p)  
*# draw boxplots for two samples*fig,ax =plt.subplots(1,2, figsize=(12,3))  
ax[0].boxplot(before\_data)  
ax[0].set\_title(**"Boxplot of before level of testosterone"**)  
ax[1].boxplot(after\_data)  
ax[1].set\_title(**"Boxplot of after level of testosterone"**)  
plt.show()  
  
*# draw violin plots for two samples*fig,ax =plt.subplots(1,2, figsize=(12,3))  
ax[0].violinplot(before\_data)  
ax[0].set\_title(**"Violin plot of before level of testosterone"**)  
ax[1].violinplot(after\_data)  
ax[1].set\_title(**"Violin plot of after level of testosterone"**)  
plt.show()  
  
*#independed sample t-test*stat,p = stats.kruskal(before\_data,after\_data)  
print(**'P value for paired sample t-test'**,p)

1. Chi-square contingency test.

Many parasites have more than one species of host, so the individual parasite must get from

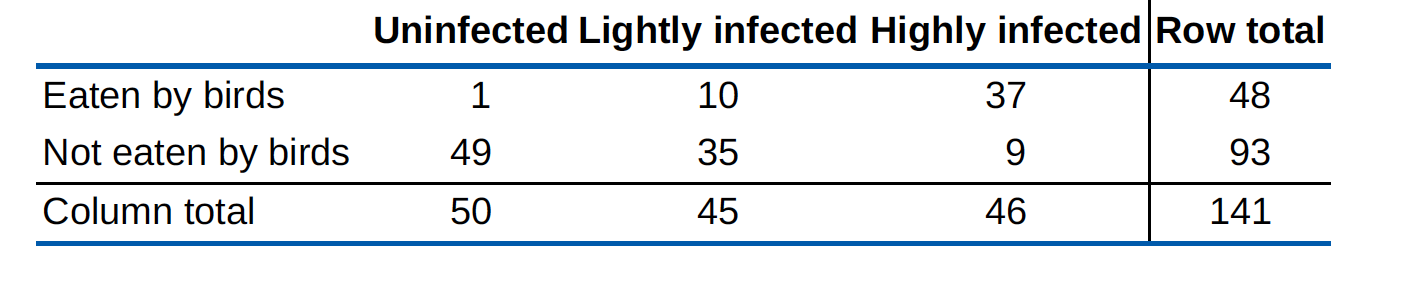
one host to another to complete its life cycle. Trematodes of the species *Euhaplorchis*

*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.



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

* 1. Write the null and alternative hypotheses for the above research question.

Null hypothesis(H0) : µu=µl=µh;

There is no significant difference of the mean number of fish eaten between the different infection levels

Alternative hypothesis (HA): µu≠µl≠µh;

There is a significant difference of the mean number of fish eaten between the different infection levels

* 1. What are the assumptions when performing the above test?

Data is normally distributed

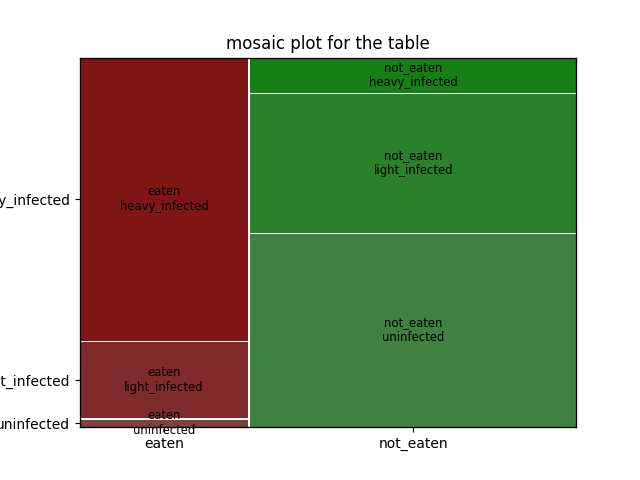
Random sampling is done

Independent observations/measurements

Equal variance

* 1. Create the above contingency table in a pandas DataFrame and print it.
  2. 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.*



* 1. 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 values : 69.75570515817361

P value : 7.124281712683879e-16

degree of freedom : 2

* 1. 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

* 1. Write your conclusion based on above results.

P value : 7.124281712683879e-16🡪7.12e-16<0.05🡪 reject the null hypothesis

Conclusion:

There is a significant difference of the mean number of fish eaten between the different infection levels at 5% significance level.

**import** numpy **as** np  
**import** matplotlib.pyplot **as** plt  
**import** pandas **as** pd  
**from** scipy **import** stats  
**import** statsmodels.graphics.mosaicplot **as** mosaicplot  
*# Creating a Dataframe from the data and printing it  
#fish = pd.DataFrame([[1,10,37],[49,35,9]],index=['eaten','not\_eaten'],columns=['uninfected','light\_infected','heavy\_infected'])  
#print(fish)  
  
# initialise data of lists.*data = {**'uninfected'**: [1, 49],**'light\_infected'**: [10, 35],**'heavy\_infected'**:[37,9]}  
*# Creates pandas DataFrame.*fish = pd.DataFrame(data, index=[**'eaten'**,**'not\_eaten'**])  
*# print the data*print(fish)  
*# plot a mosaicplot for the dataframe*mosaicplot.mosaic(fish.stack(),title=**'mosaic plot for the table'**)  
plt.show()  
  
*# do a chi-square contingency test*chi\_value,p,dof,expected = stats.chi2\_contingency(fish)  
print(**'Chi values :'**,chi\_value)  
print(**'P value :'**,p)  
print(**'degree of freedom :'**,dof)  
*#use the expected variable from it make a dataframe*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.