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

**Lab5: Biological hypothesis testing using Python.**

**Name: Anushka Udara**

**Index number: s14234**

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)

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.

Ho: UN -98.6= 0

There is no significant difference between the normal human body temperature mean and 98.6F.

HA: UN -98.6≠ 0

There is a significant difference between the normal human body temperature mean and 98.6F.

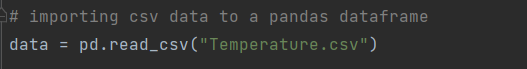
* 1. What are the assumptions when performing the above test?

Data population is normally distributed.

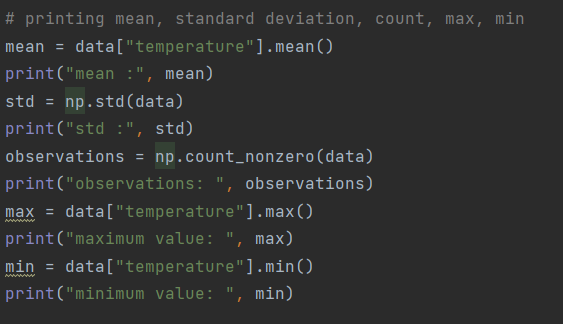
Samples are randomly selected.

Independent observations.

* 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.

****

**Mean: 98.524**

**Std: 0.664096**

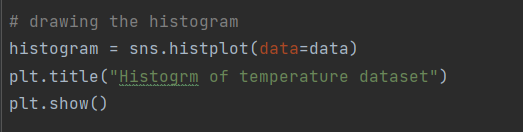
**Observations: 25**

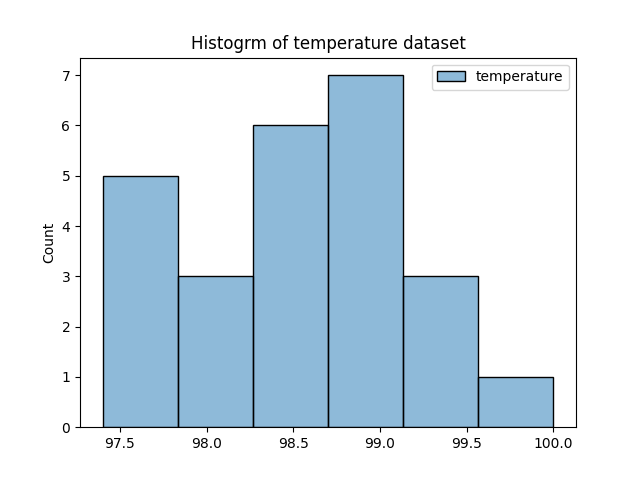
**Maximum value: 100**

**Minimum value: 97.4**

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

First, draw the histogram for the variable.



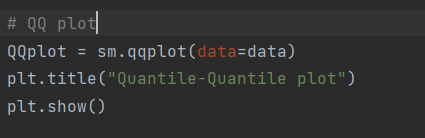


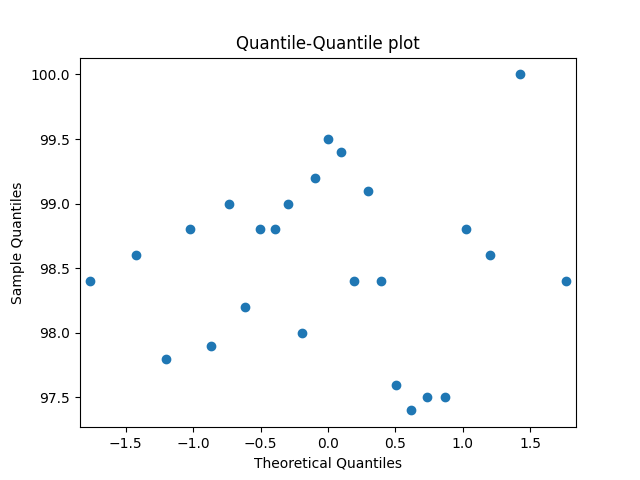
Histogram

There are five values in the range of 97.5 to 98, three values in the range of 97.7 to 98.3, six values in the range of 98.3 to 98.7, seven values in the range of 98.7 to 99.1.

There are three values in the range of 99.1 to 99.6 and one value in the range of 99.6 to 100.

Then, draw a Quantile-Quantile plot (QQ plot).





QQ plot

QQ plot gives an indication of univariate normality. No any sign of a straight line. Therefore, the data is not normally distributed.

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)?

Code

nt = ss.shapiro(x=data)  
print(nt)

Wilk test for normality

Ho: data is normally distributed.

HA: data is not normally distributed.

ShapiroResult(statistic=0.9721594452857971, pvalue=0.7001229524612427)

p-value >significance level(0.05)

null hypothesis fails to be rejected.

Therefore, data is normally distributed.

Since the dataset is normally distributed, a parametric test is recommended.

Dataset is less than 30 and population variance is not known.

Therefore, one sample t-test.

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

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

One sample t-test

Code

one\_sampt = ss.ttest\_1samp(a=data, popmean=98.6, alternative="two-sided")  
print(one\_sampt)

TtestResult(statistic=array([-0.56064519]), pvalue=array([0.58023628]), df=array([24]))

p-value > 0.05

null hypothesis fails to be rejected.

Conclusion

At the 0.05 significance level, there is no significant difference between the normal human body temperature mean and 98.6F.

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, *Laniusludovicianus*. 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.

Ho: US – UD=0

There is no significant difference between the horn lengths of horned lizards that survived and dead.

HA: US – UD≠0

There is a significant difference between the horn lengths of horned lizards that survived and dead.

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

Data population is normally distributed.

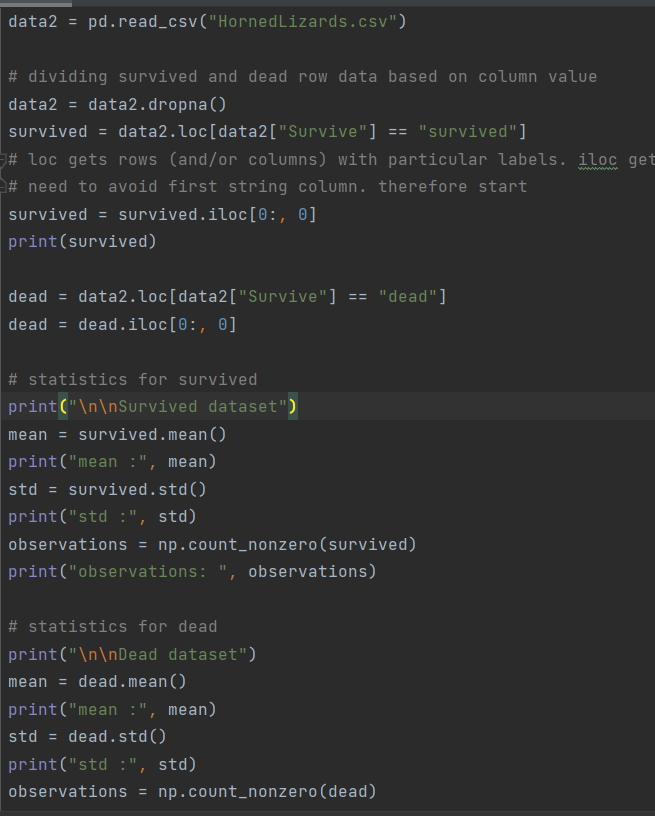
Samples are randomly selected.

Independent observations.

Equal variance in populations are assumed.

* 1. Import the data set into a Pandas Data Frame. 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.*



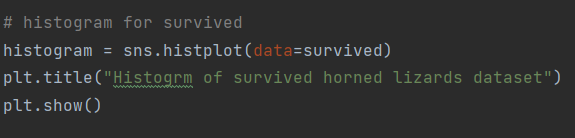
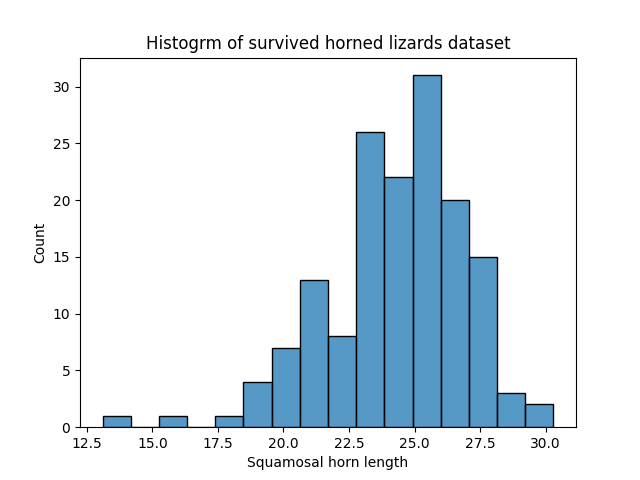
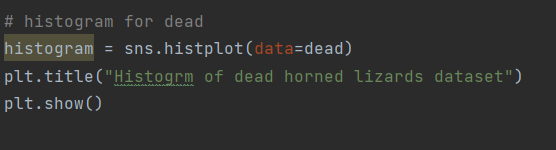
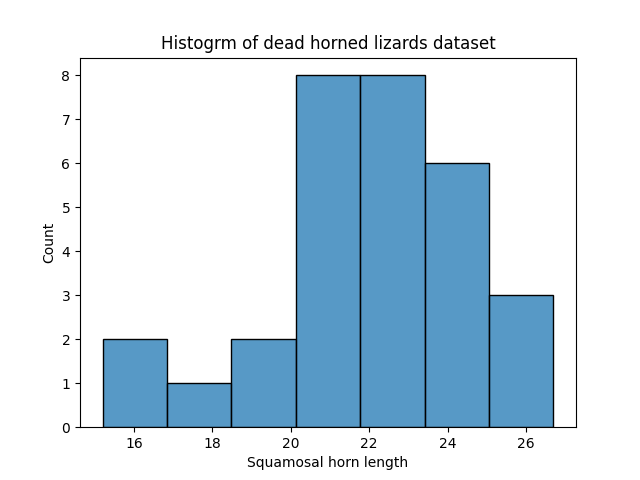
Survived dataset Dead dataset

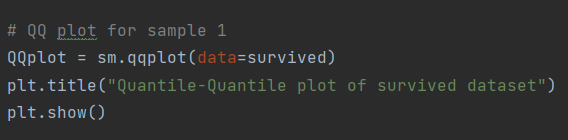
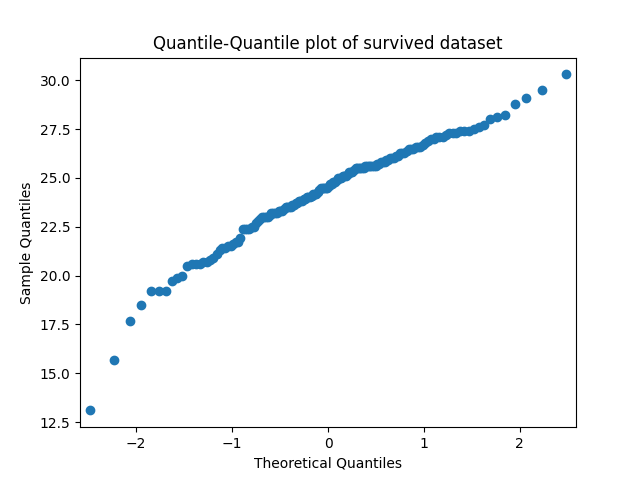
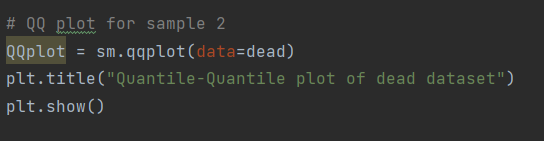
mean: 24.28116883116883 mean: 21.986666666666668

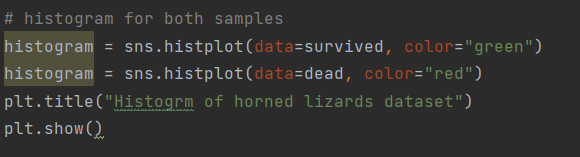
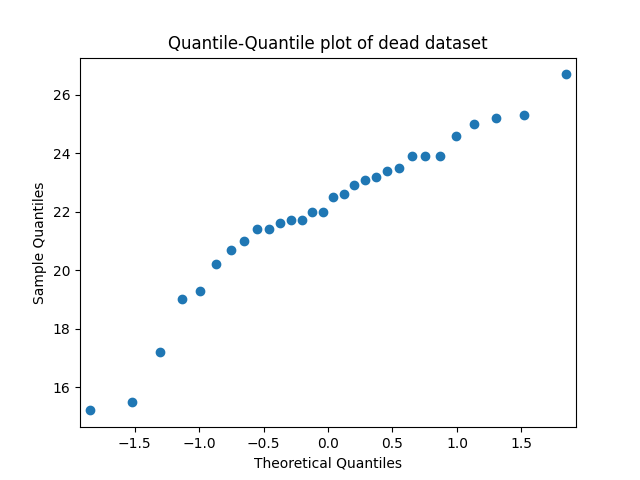
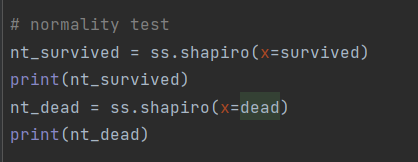
Std: 2.630782320597537 std: 2.7094640433670363 observations: 154 observations: 30

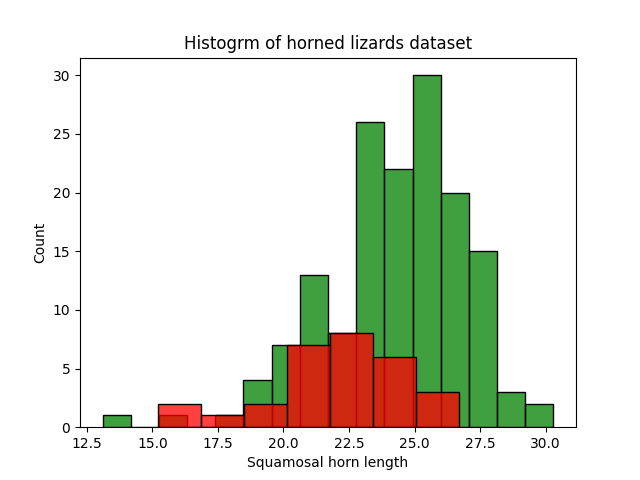
* 1. Testing the normality assumption for the two independent samples.

First, draw histograms and QQ plots for each sample.





Draw the two histograms in the same plot.



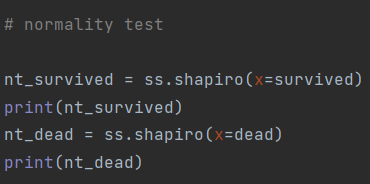
Then, interpret the results of the histograms and QQ plots.

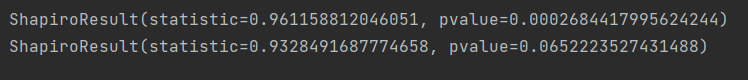
Histogram

Both show the frequency of data values in their relevant ranges.

QQplot

Both QQplots for survived and dead datasets show rough straight lines.

Then, perform the Shapiro-Wilk test for normality on each variable.



Write the test hypotheses, P-value, and the conclusion for each variable.

**Shapiro-Wilk test for normality**

**Ho: data is normally distributed.**

**HA: data is not normally distributed.**

**p-value for survived dataset <significance level(0.05)**

**null hypothesis is rejected.**

**Therefore, survived data is not normally distributed.**

**p-value for dead dataset >significance level(0.05)**

**null hypothesis is not rejected.**

**Therefore, dead data is normally distributed.**

What type of test would you recommend (parametric or non-parametric)?

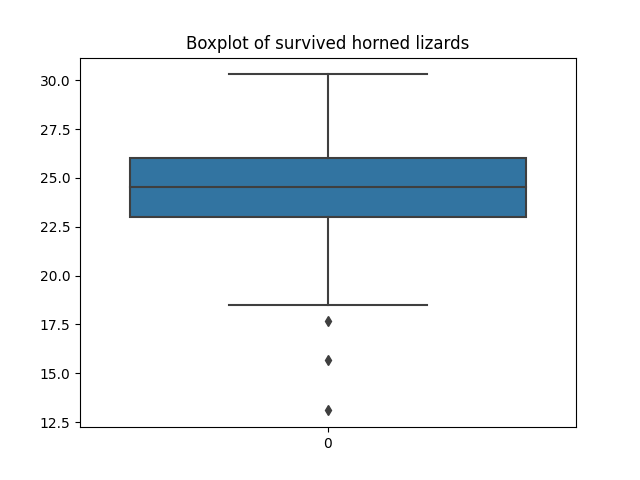
*Hint: use statmodels package for QQ plot generation.*

**Since both datasets are almost normally distributed, a parametric test is recommended.**

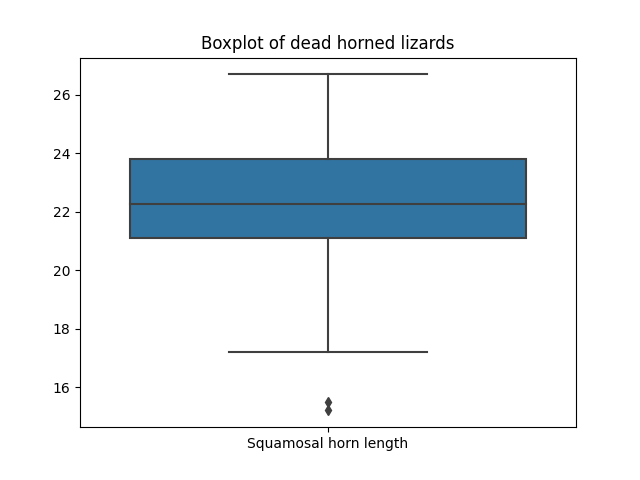
**One dataset is 30 and population variance is not known.**

**Therefore, two independent sample t-test.**

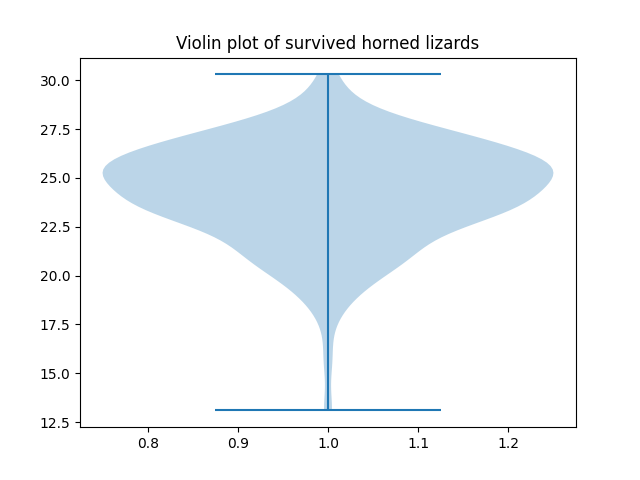
* 1. Comparison of means. To further visualize the two samples, draw boxplots and a violin plots to compare the distribution of the two samples. What are your interpretations for each plot?



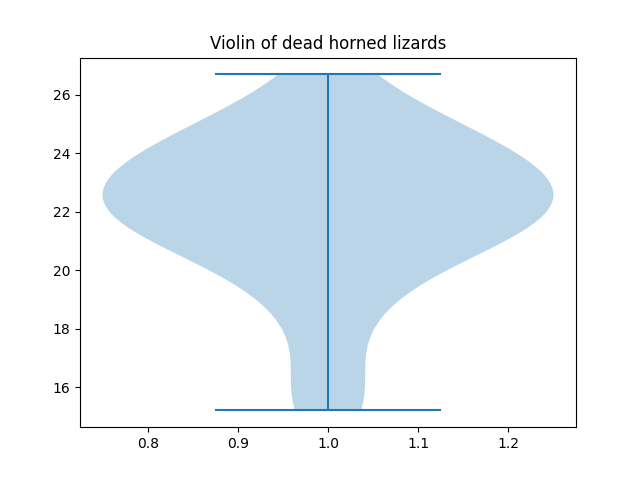
Median value is between 24 and 25. 50% percent of data falls between 22.5 and 26. Data range is between 18 and 31.



Median value is between 22 and 23. 50% percent of data falls between 21and 24. Data range is between 17 and 27.

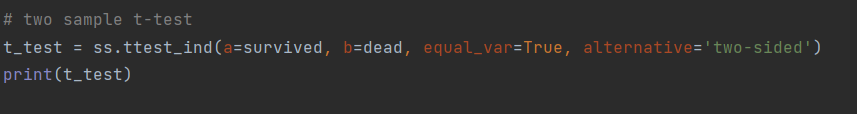


Maximum value near 30. Minimum value is near 13. Value with the highest frequency is 25.5.



Maximum value near 27. Minimum value is near 15. Value with the highest frequency is 25.6.

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





p-value > 0.05

null hypothesis is not rejected.

Conclusion

At the 0.05 significance level, there is no significant difference between the horn lengths of horned lizards that survived and dead.

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 malered-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 non-pathogenic 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.

Ho: UH – UL=>0

The immunocompetence of birds with higher testosterone is not less compared to birds with lower testosterone.

HA: UH – UL <0

The immunocompetence of birds with higher testosterone is less compared to birds with lower testosterone.

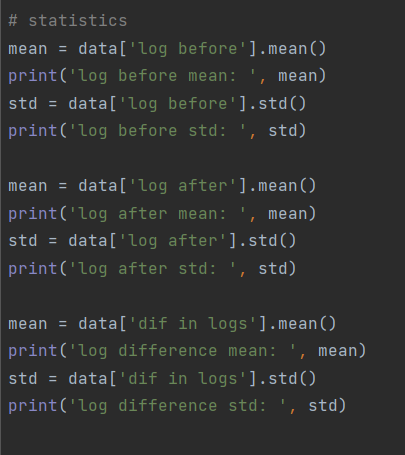
* 1. What are the assumptions when performing the above test?

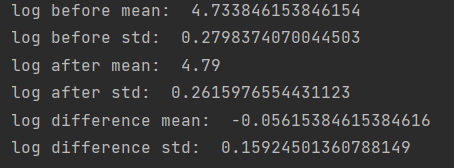
Data population is normally distributed.

Samples are randomly selected.

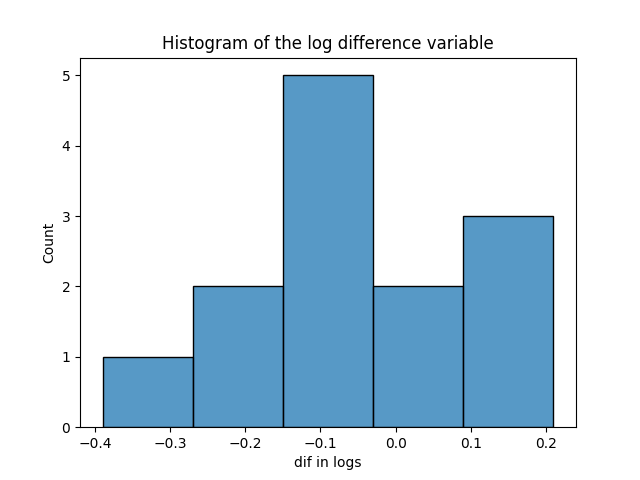
Independent observations.

Equal variance in populations are assumed.

* 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.

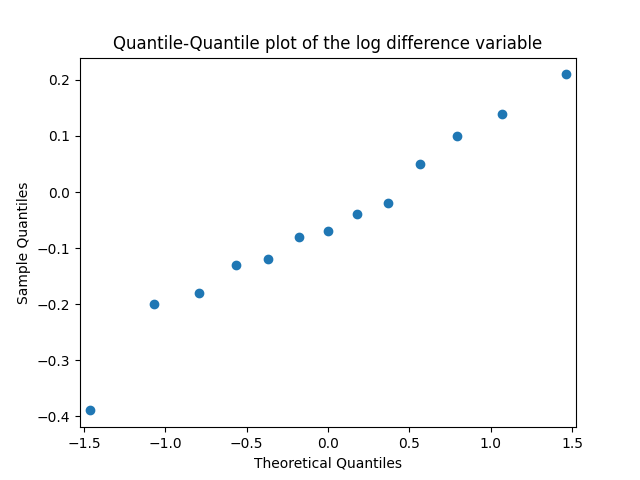


* 1. Testing the normality assumption for the log difference variable.

First, draw a histogram and a QQ plot for the variable.

There is one value in the range of -0.4to -0.26, two values in the range of -0.26 to -0.15, five values in the range of -0.15 to -0.3, two values in the range of -0.3 to 0.1

And three values in the range of 0.1 to 0.2.



This plot shows a straight line. Therefore, the data is normally distributed.

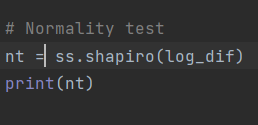
Interpret the results of the two plots.

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.

**Shapiro-Wilk test for normality**

**Ho: data is normally distributed.**

**HA: data is not normally distributed.**

****

**p-value for dataset >significance level(0.05)**

**null hypothesis is not rejected.**

**Therefore, dead is normally distributed.**

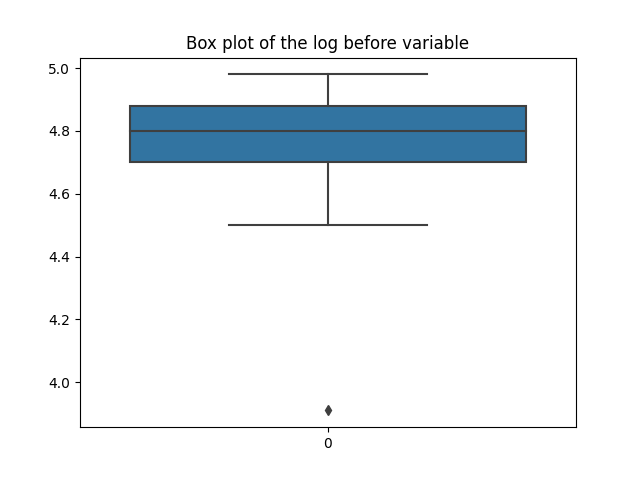
What type of test would you recommend (parametric or non-parametric)?

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

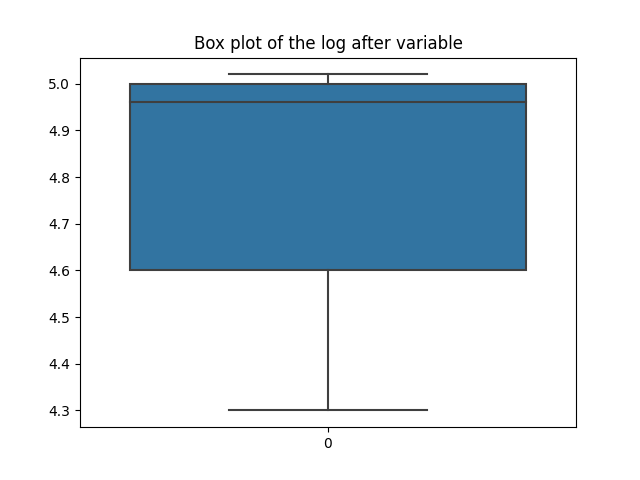
**Since the dataset is normally distributed, a parametric test is recommended.**

**Datasets are less (13) than 30 and population variances are not known.**

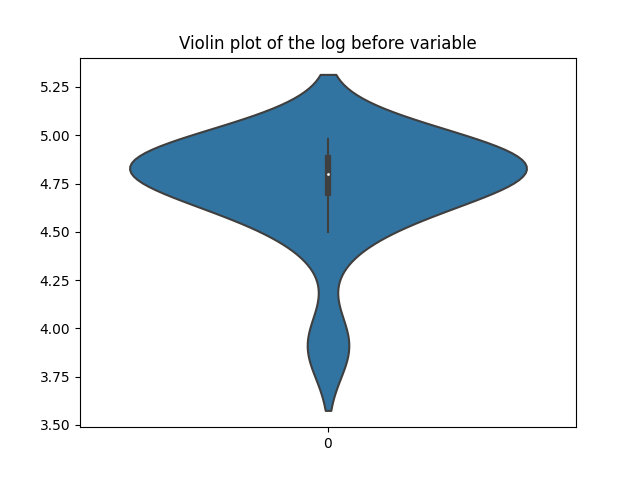
**Therefore, paired sample t-test.**

* 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.

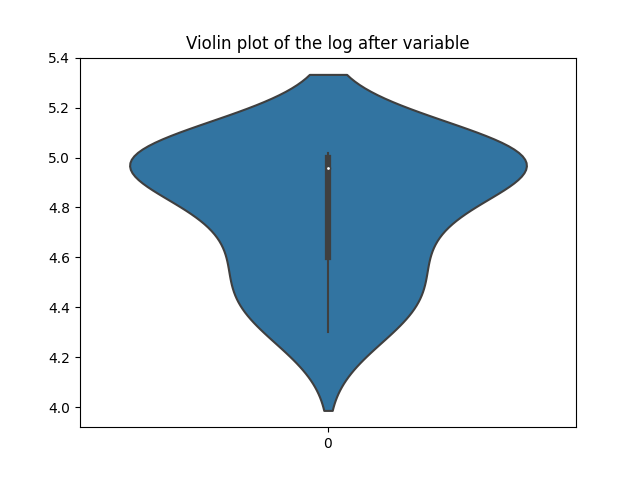
Median value is between 4.8. 50% percent of data falls between 4.9 and 4.7. Data range is between 4.5 and 5.0.



Median value is between 4.95. 50% percent of data falls between 4.6 and 5.0. Data range is between 4.3 and 5.1.

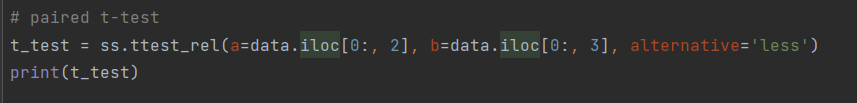


Maximum value near 5.30. Minimum value is near 3.50. Value with the highest frequency is 4.8.



Maximum value near 5.4. Minimum value is near 4.0. Value with the highest frequency is 5.0.

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





p-value > 0.05

null hypothesis is not rejected.

Conclusion

At the 0.05 significance level, the immunocompetence of birds with higher testosterone is not less compared to birds with lower testosterone.

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*

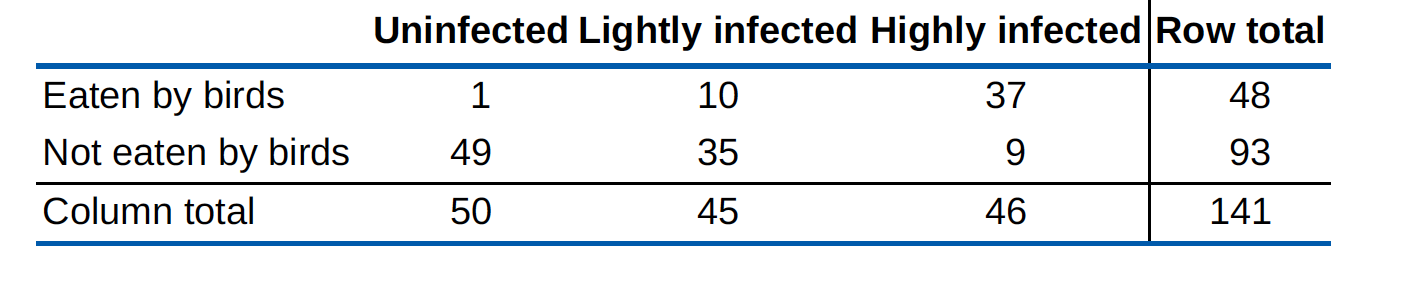
*californiensis*use three hosts during their life cycle. Worms mature in birds and lay eggs that pass out of the bird in its feces. The horn snail *Cerithidea californica* eats these eggs, which hatch and grow to another life stage in the snail, sterilizing the snail in the process. When an infected snail is eaten by the California killifish *Fundulusparvipinnis*, 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.

Ho: PU – PL – PH =0

The probability of being eaten by birds doesn’t depend on the infection status.

HA: PU – PL – PH ≠0

The probability of being eaten by birds, depends on the infection status.

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

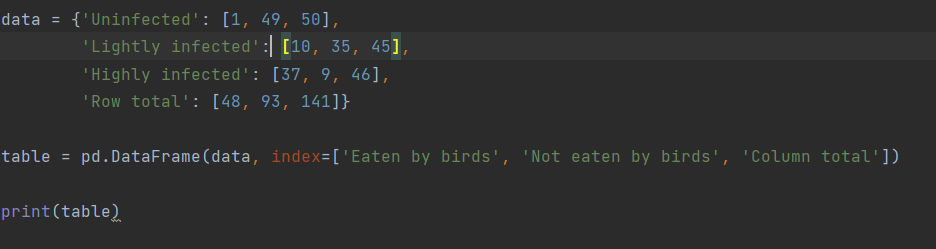
Data populations are normally distributed.

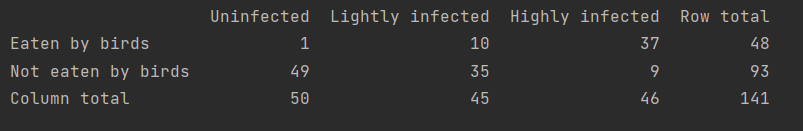
All variables are categorical.

Samples are randomly selected.

Independent observations.

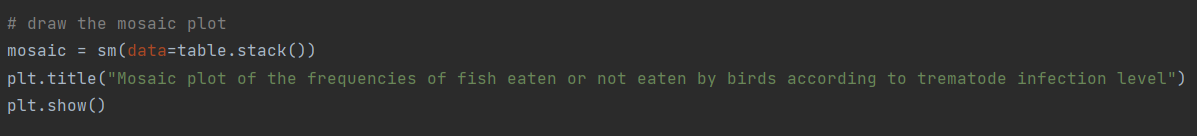
* 1. Create the above contingency table in a Pandas Data Frame and print it.

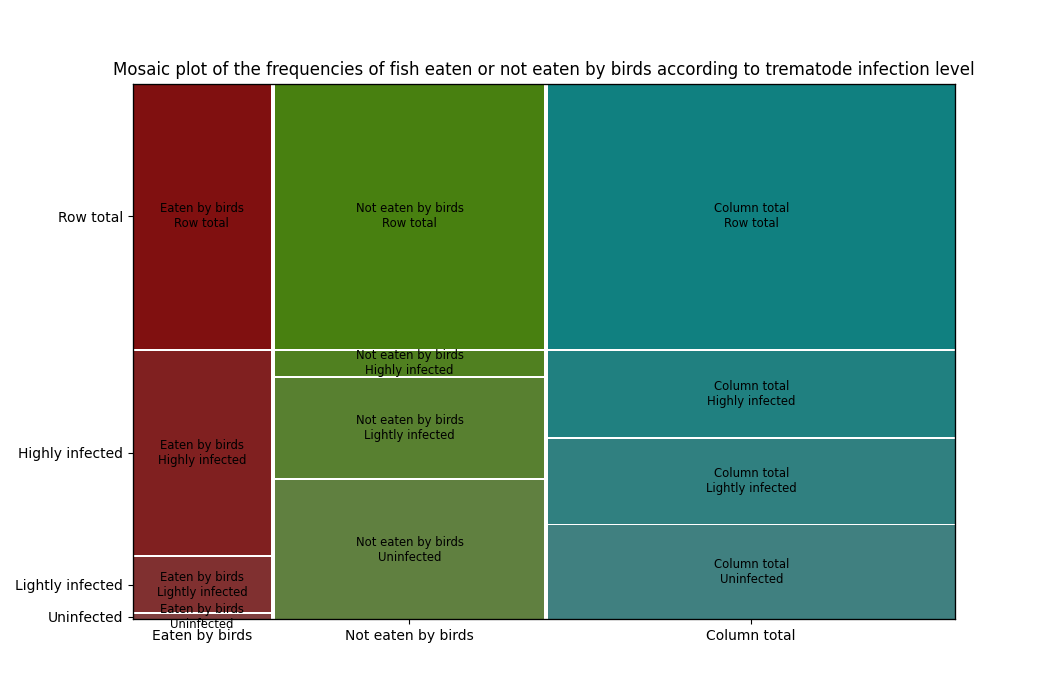




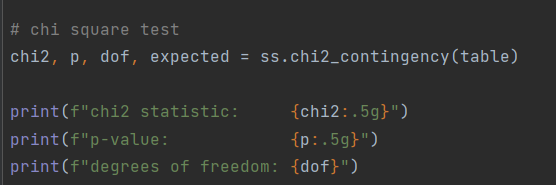
* 1. Draw a mosaic plot for the above table.

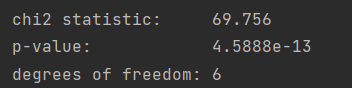
*Hint: Use the statmodels package to draw the plot. Also, look into the pandas DataFrame.stack() function. Maybe it will be useful to you.*



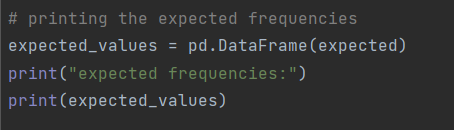


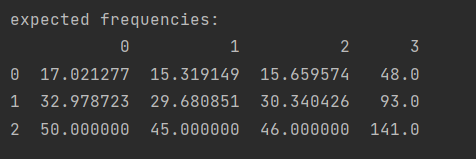
* 1. Perform a Chi-square contingency test on above data. Write the Chi-square statistic, the P-value and the degree of freedom value below.





* 1. Output the expected value table. You can create a new Data Frame to store the expected values.





* 1. Write your conclusion based on above results.

p-value <0.05

null hypothesis is rejected.

Conclusion

At 0.05 significance level, the probability of being eaten by birds, depends on the infection status.

**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.