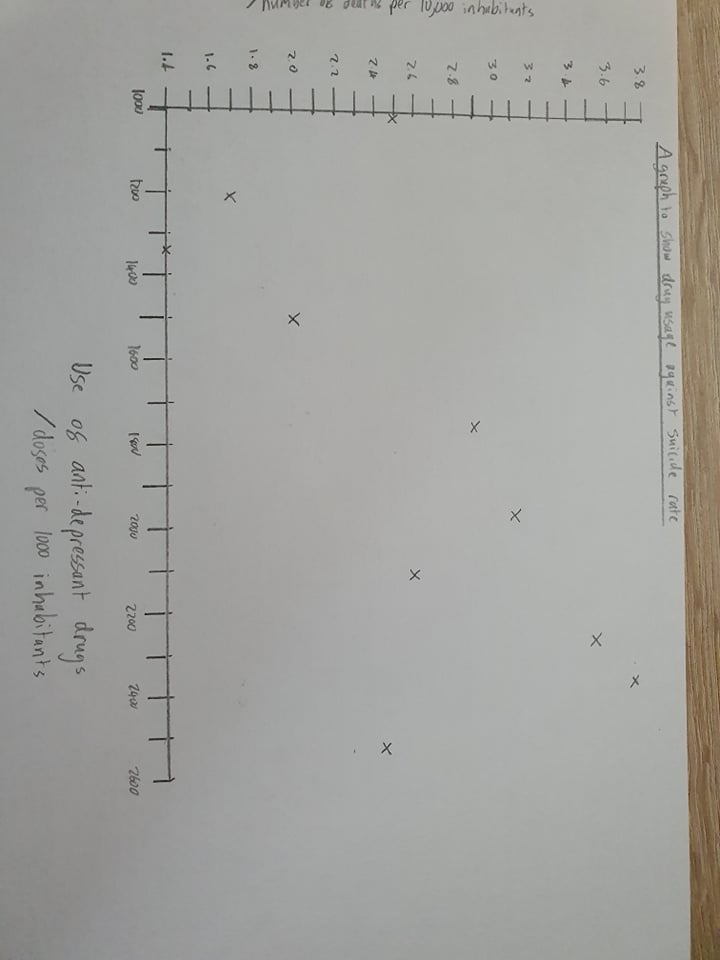
Statistics for Biology Coursework Element 010

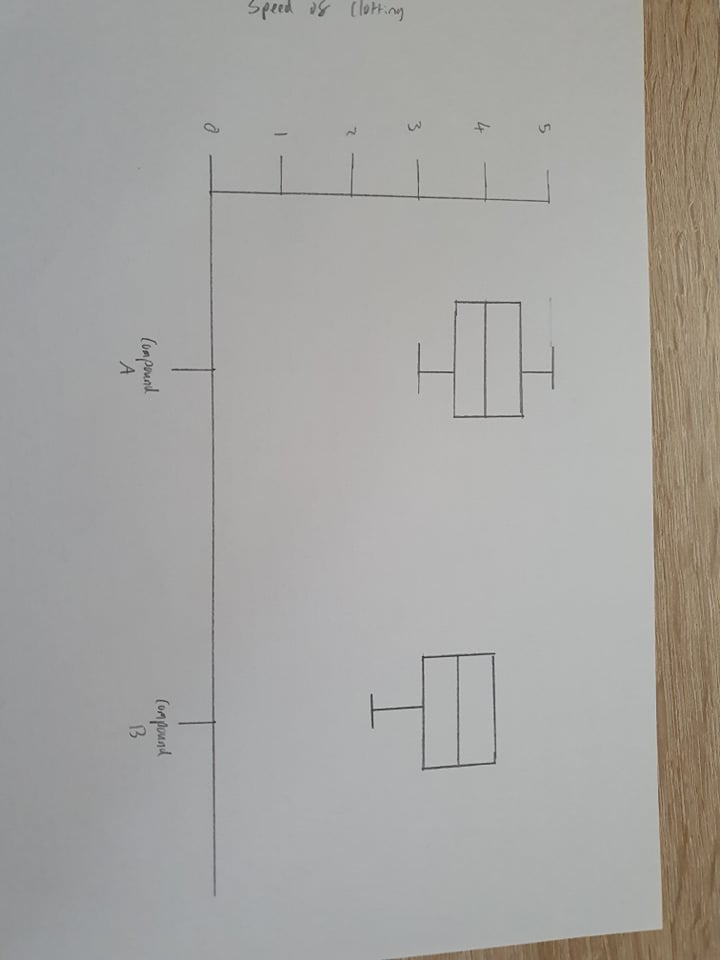
SID: 1838745

Part A:

Dataset A1:

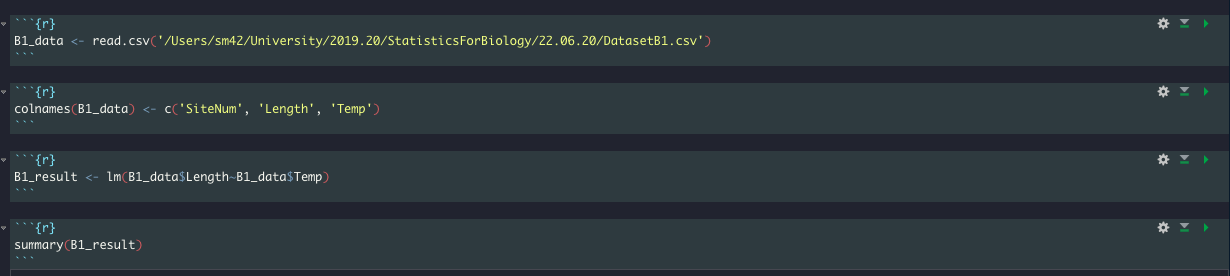
There are two variables involved in dataset A1. The first measures use of anti-depressant drug (doses per 1000 inhabitants) which is a scale level variable. The second measures suicide rate (number of deaths per 10,000 inhabitants) which is also a scale level variable. The variation in this dataset is natural and has not been manipulated. Both variables are measured in raw form. Each variable is measured for 10 different cities, one record of each per city. The data for use of anti-depressant drugs and the data for suicide rates are related datasets. This dataset is testing a relationship and the variables are interdependent (there is no dependent variable and independent variable). All this information tells me to choose either the Pearson correlation or the Spearman correlation test. The criteria for the parametric test (Pearson correlation) have been met, as there is no reason for me to assume data is not normally distributed, so I will apply a Pearson correlation. One reason for this is parametric test has more statistical power.

Dataset A2:

There are 2 variables involved in dataset A2. The first is speed of clotting which is an ordinal level variable. The second is compound A or compound B which is a nominal level variable. The variation in this dataset is natural and has not been manipulated. Both variables are measured in raw form. Each variable is measured for 8 subjects, one record of clotting speed for each compound per subject, which makes the data related. The data collected for this dataset is for testing a difference in clotting speed between the two compounds. All this information tells me to choose a paired t-test or Wilcoxon signed-rank test. The criteria for the parametric test (paired t-test) have not been met, as there is one variable of ordinal level, so I will apply a Wilcoxon signed-rank test.

Part B:

Dataset B1:

The following screenshot shows the code I used to import the B1 dataset to R, manipulate the data and perform linear regression on the data.

The following screenshot shows the output from the linear regression on dataset B1.

This is the value of the coefficient of determination (R2)



These are the values for the degrees of freedom

This is the P value

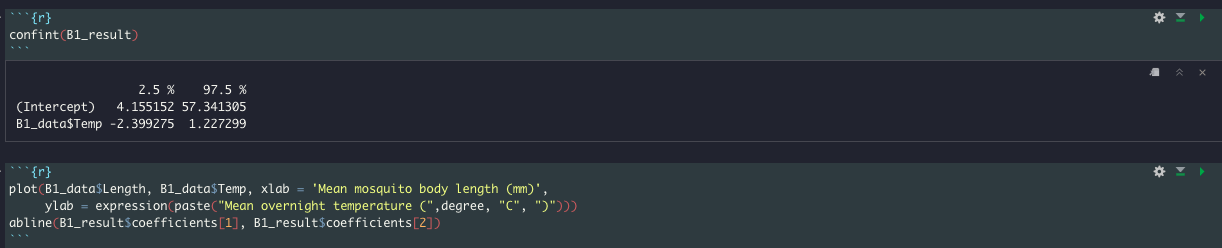
This is the value of the effect size, also known as the gradient (m)

This is the value of the y-intercept (c)

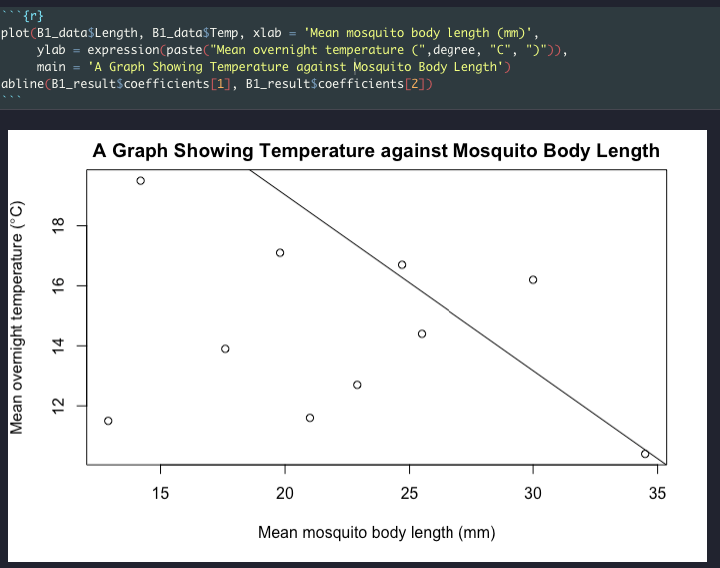
This is the value of the statistic produced by linear regression (F)

The key information is: y = -0.586x + 30.7, F1,8 = 0.555, P = 0.478, R2 = 0.0649

The following screenshot shows the code used to produce the confidence intervals and their values.



95% confidence intervals (left value is lower CI; right value is upper CI) of effect size

The following screenshot shows the code used to produce the graph for dataset B1 and the graph produced.

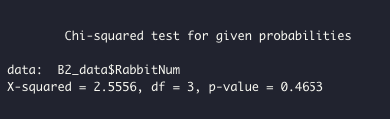
My data shows there is not a particular statistical pattern as the regression line does not fit the data very well, and the F statistic value is very low. So, body length cannot be predicted by overnight temperature.

Dataset B2:

The following screenshot shows the code I used to import the B2 dataset to R, manipulate the data and perform linear regression on the data.



The following screenshot shows the output from the One-way chi-square test on dataset B2.



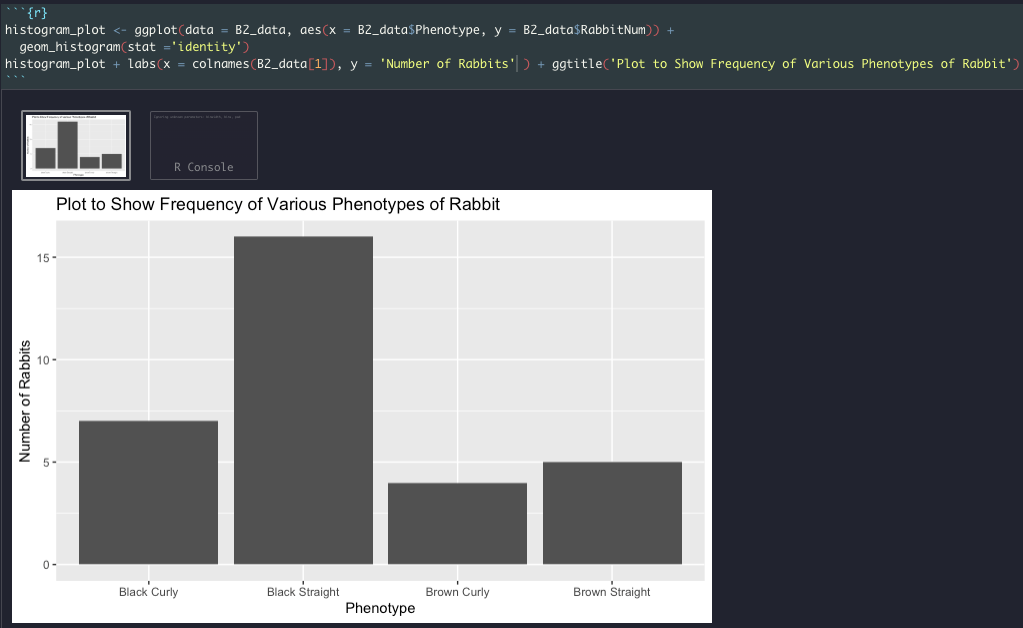
This is the P value

The value for the degrees of freedom

The value of the statistic produced from One-way chi-square test

The key information is: one-way chi-square: X2= 2.56, N = 32 (from dataset), Degrees of freedom = 3, P = 0.465

The following screenshot shows the code used to produce the graph for dataset B2 and the graph produced.



This data shows that although some rabbit phenotypes are more common than others there is not a dihybrid inheritance ration of 9:3:3:1.

Part C:

Scenario C1:

The word equation for this scenario is as follows:

Death = time of first dose (minutes) + patient age (years)

The assumptions that I made for this scenario are as follows:

* The data on whether a patient died or recovered in a 30-day period (known as death in the word equation) is normally distributed
* The data on how long after diagnosis the patient received their first dose of antibiotics (known as time of first dose in the word equation) is a continuous variable
* The data on the age of the patient (known as patient age in the word equation) could be categorical data (as years are discrete values) or continuous data but due to large range of values age could take, it will be treated as a continuous data for this model

The subframework model I would suggest is the general linear model, the reasons for this choice are:

* There are two continuous variables which are data types accepted when using the general linear model
* The data can be assumed to be normally distributed, as there is no reason to suggest otherwise, and normal distribution is accepted by the general linear model
* There is no binary data so the binary logistic model would not be an apt fit
* There is no data in count form and so loglinear model would not be a good fit

Scenario C2:

The word equation for this scenario is as follows:

Number who suffered cold = ethnic group (white, BAME, mixed or other) + total annual income (pounds)

The assumptions I made for this scenario are as follows:

* The data on number of people in each household who self-reported suffering a cold during 2019 (known as number who suffered cold in the word equation) is normally distributed
* The data on ethnic group each member of each household identifies with (known as ethnic group in the word equation) is a categorical variable
* The data on the total annual income of each household (known as total annual income in the word equation) is a continuous variable

The subframework model I would suggest using is the general linear model, the reasons for this choice are:

* There is one categorical variable and 1 continuous variable which are data types accepted when using the general linear model
* The data can be assumed to be normally distributed, as there is no reason to suggest otherwise, and normal distribution is accepted by the general linear model
* There is no binary data so the binary logistic model would not be an apt fit
* There is no data in count from and so the loglinear model would not be a good fit

Part D:

This part is document in the document ‘CourseworkCodeLibrary.Rmd’ which is also within this zip file

Please note that I read data in using read.csv rather than read.table as I do not like using the GUI file finder. If you want to run the code yourself then please change the pathway specified in read.csv or replace read.csv with the read.table function.

All data used in this library was is provided as .csv files in the folder. The names of the files correspond to the file names in the pathways of the read.csv function when they are used.

Part E:

This part is documented in the progress checkers and their reports which I do not have access to