Khan Inan
Transcriptomics 7653
Excercise #4

M4_1.1) In what scenarios is it better to use ANOVA compared to t-test? More specifically when can you NOT use a t-test?

It is better to use ANOVA test when there are large population counts, and you cannot use a t-test when the population count is larger than 30

M4_1.2) How is the ANOVA statistics calculated?

The formula is F=MSB/MSE where F is the test statistic

M4_1.3) What assumptions does ANOVA make about the data?

There are 3 assumptions which are that each factor level should be normally distributed for their population, the distributions have equal variance, and there is independent data

M4_2.1) What are some differences between pearson and spearman correlation?

The main difference is that a pearson coefficient works with linear relationships while the spearman coefficient can work with monotonic relationships in addition to linear relationships

M4_2.2) What is the range of values one can obtain from a correlation analysis?

The range is from -1 to 1

M4_3.1) What are the two variables that are needed to draw a straight line?

The two variables seem to be a and b which are slope and intercept

M4_3.2) What command can we use to create a linear model?

The lm() function can be used to make a linear model

M4_3.3) What command do we use to predict the values using the linear model?

From what I found there is a command called predict() which can be used to do that

M4_4.1) What is a limitation of the CHI-Square test?

The main limitation is the strict requirements in regards to sample size

M4_4.2) What is the null hypothesis of a fisher's exact test?

Its the proportion or way to represent how independent one variable is to the other variable

M4_4.3) Give an example of Hypergeometric tests can be used with biological data.

When doing gene sequencing it is necessary to compare an obtained sequence with a reference genome and that can be done using a hypergeometric test

M4_5.1) Why is it important to correct for multiple hypothesis testing?

If there is no corrections made then there would be a certain amount of null-hypothesis that would be falsely rejected

M4_5.2) Of the two methods discussed (Bonferroni and FDR) which method is more appropriate for life science (when we have few replicates).

FDR would be more appropriate compared to Bonferroni mainly because it increases in reliability and accuracy with the more tests that you do

M4_5.3) What is the difference between a TypeI and TypeII error?

Type I is when a null hypothesis is rejected that is actually true, and type II is the opposite where a false null hypothesis fails to be rejected

M4_6.1) Install the reshape2 package in R. What does the melt function do?

It pretty much re-formats the data. Data that is widely spread is converted to be long-wise spread

M4_6.2) Check out the help documentation for rnorm using help("rnorm"). What are the first three arguments of the function?

The frst three arguments are:

x,q : vector of quantities

p : vector of probabilities

n : number of observations. If length(n) > 1, the length is taken to be the number required

M4_6.3) How is the slope of the linear model and the difference in the means related in this example?

The difference in means makes it so the one with the higher mean has a higher positive slope while the one with the lower mean results in a negative slope

M4_7.1) What is the difference between the lm() function and aov() function? Do they give you the same result when looking at just one factor with two levels?

AOV is for typeI or for a sequential sum while lm() is typeII which is adjusted sum. They do seem to provide the same result for just one factor with two levels

M4_8.1) What is the argument for TukeyHSD and what analysis does it perform?

TukeyHSD essentially analyzes, amongst multiple variables, if their interaction with each other is statistically significant