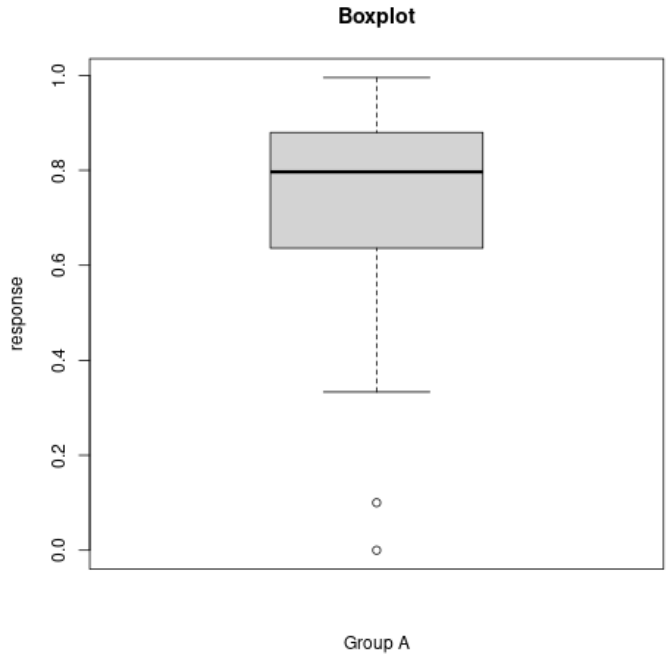
Midterm 1
100 points
Due April 13, 2021

This take-home exam is open notes, but solutions must represent your own efforts, without assistance from any other person (though you are welcome to email me if you have any questions). Include any R code that you used in your derivations.



Problem 1. (10 points) Consider the boxplot given above.

a. Is the data represented in this plot right skewed or left skewed? How can you tell?

b. Roughly, what is the interquartile range of this data (IQR)?

c. How many outliers exist in this dataset?

Problem 2. (10 points) You wish to assess whether or not there is a relationship between age and the width of a particular bony structure in a model organism. 140 subjects are measured age 8 to 18 months old. The data from the study are plotted below

a. Which type of correlation, Pearson correlation or the Spearman rank correlation, would you use in assessing this data? Why? What would you conclude based on the correlation statistic you chose?

b. Regardless of what you chose above, which correlation statistic would have a larger value? Why?

Problem 3. (20 points) You wish to assess whether variants in the melanocortin-1 receptor (MNC1R) are associated with hair color. The MC1R gene is highly polymorphic, but a subset of variants "most likely to decrease the function of the MC1R gene most dramatically" were identified.

You assemple a sample of 50 persons with natural red hair and a sample of 50 persons with natural dark hair (brown or black) and genotype them. Each subject is classified as to whether or not the subject carries at least one "decreased function" variant. The investigators found that 47 of the red-haired subjects carried at least one "decreased function" MC1R gene variant, and 18 of the dark-haired subjects were found to carry one such variant. Assess the magnitude of association, if any, between hair color and the presence of the MC1R gene variants using the odds ratio.

	Decreased Variant Present		
Hair Color	Yes	No	Total
Red	47	3	50
Dark	18	32	50
Total	65	35	100

a. State and interpret the value of the odds ratio.

b. What is the mean value and standard deviation of the log odds test statistic?

c. Construct a 95% confidence interval for the odds ratio. What are your conclusions?

Problem 4. (10 points) Consider the following sample data, measured in mg:

$$24, 26, 41, 50, 42, 42, 58, 64$$

What is the median _____ What is the mean _____ What is the sample size _____

What is the minimum _____ What is the maximum _____ What is the range _____

Problem 5. (20 points) In assessing the effectiveness of a new drug, you sample 10 individuals with high blood pressure and collect measurements of diastolic blood pressure (mmHg). Three months after administering treatment, you record measurements of diastolic blood pressure from the same individuals. The data collected are given below

Subject	Before Treatment	After Treatment
1	104.96	93.73
2	107.93	97.57
3	124.03	113.63
4	110.63	100.52
5	111.16	101.72
6	125.44	113.65
7	114.15	103.65
8	98.61	90.58
9	103.82	93.12
10	105.99	96.46

a. What type of test would you use to detect this difference?

b. What are the degrees of freedom associated with this test?

c. What is the mean diastolic blood pressure before treatment? After?

d. What is the value of the test statistic?

e. What is the p-value from this test statistic?

f. What are your conclusions?

Problem 6. (20 points) In professional basketball games during the 2009-2010 season, when Kobe Bryant of the LA Lakers shot a pair of free throws, 8 times he missed both, 152 times he made both, 33 times he made only the first, and 37 times he made only the second. We are interested in knowing whether or not there is an association between successive free throws

	Kobe Freethrow		
	Make 2nd	Miss 2nd	Total
Make 1st	152	33	185
Miss 1st	37	8	45
Total	189	41	230

a. Is this a chi-square test of independence or homogeneity?

b. What are the degrees of freedom for this test?

c. What is the smallest expected value for this table? Is the chi-square test justified?

d. What is the value of the chi-square statistic?

e. What is the p-value from this procedure?

f. What are your conclusions?

Problem 7. (10 points) In a study of hereditary breast cancer carried out by the National Institute of Health, samples of tumors were collected from patients with mutations in either BRCA1 or BRCA2 genes. 3,226 genes were measured on a microarray and 3,226 t-tests were conducted to determine if the expression level of any other genes differed between BRCA1 and BRCA2 tumors.

a. Using the Bonferroni correction to control the FWER at $\alpha = 0.05$, what would be the adjusted α^* to determine significance of the test statistics?

b. From the breast cancer study, 207 genes had p-values below a cutoff of $\alpha = 0.01$. Based on this, what is the estimate of the false discovery rate?

Extra Credit: (5 pts) Find a paper in genetics utilizing the use of some statistic for use in your final presentation. Plan for about 10-15 minutes, where you will introduce the problem, describe the methodology used, justify use of that method, and then state the conclusions. You may pair up and present with someone in class. Include here the name of your paper and if you will be presenting alone or with a partner. Feel free to email me if you need any help finding relevant material.