Math 322: Biostatistics
Final Examination

NAME:_____

Due 4:00pm Friday, May 1st, 2020

9 Problems on 4 Pages

You may use the textbook, your class notes, a calculator or computer, and any other previously written reference, but you may not receive assistance from any other person.

Please use CrowdMark. Identify the commands, input, and output values used in your solutions.

Let MYSID be your student ID number. For all the problems below, set the random number seed to MYSID where indicated.

1. With the following `R` commands, generate a simulated joint probability table for the nucleotide pairs found at adjacent positions $(1, 2)$ in a DNA sequence:

```
MYSID <-
set.seed(MYSID); f<-runif(16,1,11)
jpdf<-matrix(f/sum(f),nrow=4)
aa1<-c("A1","C1","G1","T1")
aa2<-c("A2","C2","G2","T2")
dimnames(jpdf)<-list(aa1,aa2); jpdf
```

   Use it to answer the following questions:

   (a) What is P(C1)?

   (b) What is P(A2)?

   (c) What is P(T1 and G2)?

   (d) What is P(T1 | G2)?

2. Let $f(x, y)$ be a discrete joint pdf given by the joint probability table `jpdf` from Problem 1. Here x is index 1 (row number) and y is index 2 (column number).

   (a) Compute the marginal pdfs $f_X(x)$ and $f_Y(y)$.

   (b) Compute the complete conditional pdfs $f_{X|Y}(x, y)$ and $f_{Y|X}(x, y)$.

   (c) Implement a Gibbs sampler for this joint pdf and simulate taking 10 000 samples. Print the resulting matrix of counts as well as the normalized matrix that approximates the joint pdf.

3. Generate samples X1 and X2 from two normal populations as follows:

```
MYSID <-
set.seed(MYSID); X1 <- rnorm(12,0.5,1.5); X2 <- rnorm(13,0.9,1.1)
```

   Using the significance level $\alpha = 0.05$,

   (a) Test the hypotheses $H_0 : \ \sigma_1 = \sigma_2$ versus $H_A : \ \sigma_1 \neq \sigma_2$.

   (a') Test the hypotheses $H_0 : \ \sigma_1 \leq \sigma_2$ versus $H_A : \ \sigma_1 > \sigma_2$.

   Regardless of these results, assume homoscedasticity and then:

   (b) Test the hypotheses $H_0 : \ \mu_1 = \mu_2$ versus $H_A : \ \mu_1 \neq \mu_2$.

   (b') Test the hypotheses $H_0 : \ \mu_1 \geq \mu_2$ versus $H_A : \ \mu_1 < \mu_2$.

4. Generate a 200 sample data set as follows:

```
MYSID <-
set.seed(MYSID); data<- c(rnorm(101,mean=8,sd=0.7), rexp(99,rate=0.6))
```

   (a) Plot the histogram of `data`.

   (b) Find the mean and standard deviation of `data`.

   (b') Estimate the "standard error" of a 200-sample mean by $s/\sqrt{200}$ using the standard deviation from part b.

   (c) Find the median and the 1st and 3rd quartile values of `data`.

   Now apply the bootstrap method: using the same seed (MYSID), generate 200 replications of 200 samples of `data`, with replacement, and calculate their means and medians.

   (d) Calculate the mean and standard deviation of the 200 bootstrap means.

   (d') Calculate the mean and standard deviation of the 200 bootstrap medians.

   (e) Calculate the median and the 1st and 3rd quartile values of the 200 bootstrap means.

   (e') Calculate the median and the 1st and 3rd quartile values of the 200 bootstrap medians.

5. The following data are frequencies of smallmouth bass (a fish) found with and without flukes (a parasite) in two noncommunicating lakes:

Lake	With flukes	Without flukes
Nice	18	51
Sweet	35	44

(a) Using the Yates-corrected χ^2 test at the $\alpha = 0.05$ significance level, test H_0: the proportion of bass with flukes is the same in both lakes.

(b) Use the Fisher exact test at the 0.05 level to test H_A: the Lake Nice population bass are less likely to have flukes than those in the Lake Sweet population.

6. A follow-on study was performed on the same smallmouth bass data of Problem 5 but with the additional tabulation of gender:

	With flukes		Without flukes	
Lake	Male	Female	Male	Female
Nice	13	5	29	22
Sweet	16	19	21	23

(a) Test for mutual independence at the $\alpha = 0.05$ significance level.

(b) Test for partial independence at the $\alpha = 0.05$ significance level.

7. Generate a table of multivariate data as follows:

```
MYSID <-
set.seed(MYSID)
raw0<-cbind(rnorm(40),rnorm(40),rnorm(40),rnorm(40))
mat<-matrix(c(1,4,9,2, 1,-7,-7,6, 2,0,-1,9, 0,1,2,3), 4,4 )
cen<-matrix(c(1,-2,3,-4),1,4); c1<-matrix(rep(1,40),40,1)
raw<-raw0 %*% mat + c1 %*% cen
X<-raw[,1]; Y<-raw[,2]; Z<-raw[,3]; W<-raw[,4]
Species <- c(rep("Klingon",20),rep("Vulcan",20))
mvdata <- data.frame(X,Y,Z,W,Species); mvdata
```

Pretend that these are 4-variate responses of two imaginary alien species: *Klingon* (first 20 rows) and *Vulcan* (last 20 rows).

(a) Compute the correlation matrix for the four variables with both species' data combined.

(b) Compute the mean and the covariance matrix of the data for each species individually.

(c) Compute the multiple correlation coefficient R^2 for each variable in terms of the other three, and test $H_0 : R = 0$ at the 0.05 level in each case.

8. For this problem, use the Klingon and Vulcan data from Problem 7.

   (a) Plot all pairs of variables on a 4×4 grid of graphs using the `R` command `pairs()`. Identify the plotted points by species using "x" for *Klingon* and "o" for *Vulcan*.

   (b) Find the principal components of the data and determine how many components are needed to capture 95% of the variance.

   (c) An individual has the following values of the four variables:

Variable	Value
X	1.5
Y	-1.0
Z	6.0
W	-2.0

   Use linear discriminant analysis to judge whether it is likelier to be *Klingon* or *Vulcan*.

   (d) Use Mahalanobis distance to judge whether the individual in part (c) is likelier to be *Klingon* or *Vulcan*.

9. For this problem, use the Klingon and Vulcan data from Problem 7.

   (a) Plot a classification tree for the data including cutpoints and variable names.

   (b) Find the misclassification rate for the tree in part a.