Final study guide:

Final Wed May 6th at 8:30 am

Questions on the final will be a mix of questions on this study guide and questions you haven't seen before.

(1) You are responsible for the material on the mid-term study guide:

http://afodor.github.io/classes/stats2015/MidtermStudyGuide.pdf

In general, you are responsible for the material in the lectures (on the PowerPoint slides).

Finally, you are responsible for the material in the labs. If you were asked to perform an operation in a lab exercise, you may be asked to perform a similar operation on the final.

---

(2) The following is typed into R. What is the null hypothesis being evaluated by the p-values? What assumptions went into producing that p-value? Write the equations for the full and reduced model. Draw graphs representing the full and reduced model.

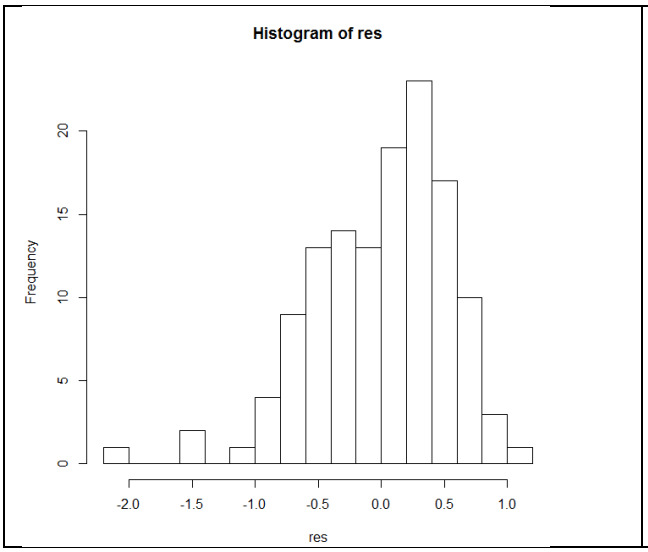
```
>
> AA <- c(4.3,2.3,4.5,5.6,4.2,3.9,2.8)
> Aa <- c(2.7,2.3,1.9,1.3,1.2,1.8,2.1)
> aa <- c(1.6,0.9,1.1,1.2,2.1,0.5,0.9)
> myData <- c(AA,Aa,aa)
> genotypes <- c( rep("AA",length(AA)), rep("Aa",length(Aa)), rep("aa",length(aa)))
> genotypes <- factor(genotypes)
> myLm <- lm( myData ~ genotypes, x=TRUE)
> anova(myLm)
Analysis of Variance Table

Response: myData
           Df Sum Sq Mean Sq F value    Pr(>F)
genotypes   2 28.666 14.3329  24.326 7.641e-06 ***
Residuals  18 10.606  0.5892
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

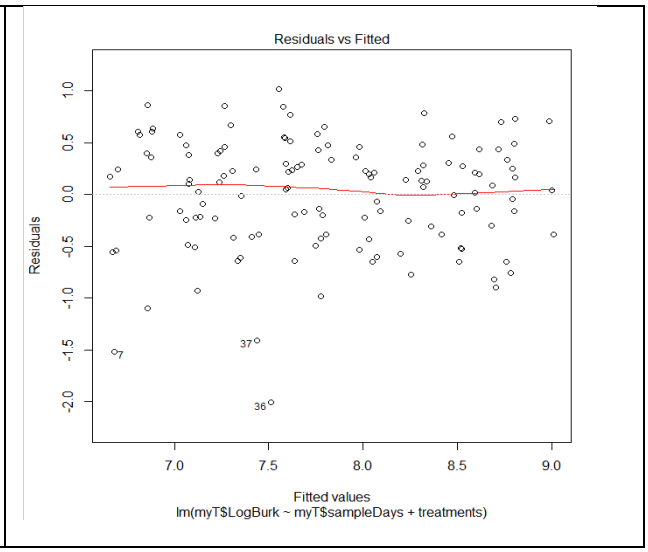
(3) The following is typed into R. What are the null hypotheses being evaluated by the p-value in the call to summary? What assumptions went into producing that p-value? Write the equations for the full and reduced models. Draw graphs representing the full and reduced model.

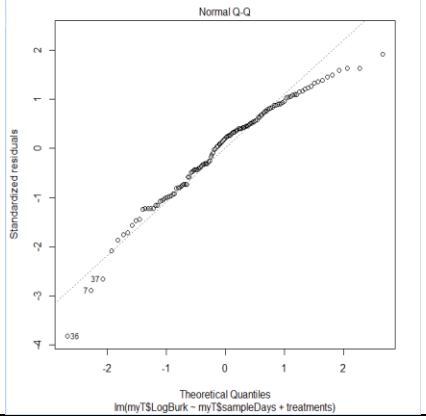
```
>
> rm(list=ls())
> AA <- c(4.3,2.3,4.5,5.6,4.2,3.9,2.8)
> Aa <- c(2.7,2.3,1.9,1.3,1.2,1.8,2.1)
> aa <- c(1.6,0.9,1.1,1.2,2.1,0.5,0.9)
> myData <- c(AA,Aa,aa)
> genotypes <- c( rep(0,length(AA)), rep(1,length(Aa)), rep(2,length(aa)))
> plot(genotypes, myData)
> myLm <- lm( myData ~ genotypes, x=TRUE)
> anova(myLm)
Analysis of Variance Table

Response: myData
          Df Sum Sq Mean Sq F value    Pr(>F)
genotypes  1 26.606 26.6064  39.915 4.603e-06 ***
Residuals 19 12.665  0.6666
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> summary(myLm)

Call:
lm(formula = myData ~ genotypes, x = TRUE)

Residuals:
     Min       1Q   Median       3Q      Max
-1.42143 -0.46429 -0.04286  0.47857  1.87857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.7214     0.2817  13.211 5.03e-11 ***
genotypes    -1.3786     0.2182  -6.318 4.60e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8164 on 19 degrees of freedom
Multiple R-squared: 0.6775,     Adjusted R-squared: 0.6605
F-statistic: 39.91 on 1 and 19 DF,  p-value: 4.603e-06
```

(4) Given the below results, does the assumption of normality seem appropriate for this linear model? Justify your answer.

```
>
> rm(list=ls())
>
> setwd("D:\\June_2012_Kannapolis16SRUN\\2012-06-18_KavanaghIlluminaSeq\\josh\\")
>
> myT <- read.table("qPCRWithSampleDays.txt", header=TRUE, sep="\t")
>
> treatments <- factor( myT$treatmentStatus)
> treatments <- relevel( treatments, ref ="Treatment")
> myLm <- lm( myT$LogBurk~ myT$sampleDay  + treatments, x=TRUE)
> res <- residuals(myLm)
> hist(res,breaks=15)
>
> ks.test(res,"pnorm", mean=mean(res),sd=sqrt(var(res)))

        One-sample Kolmogorov-Smirnov test

data:  res
D = 0.088, p-value = 0.2667
alternative hypothesis: two-sided

>
```

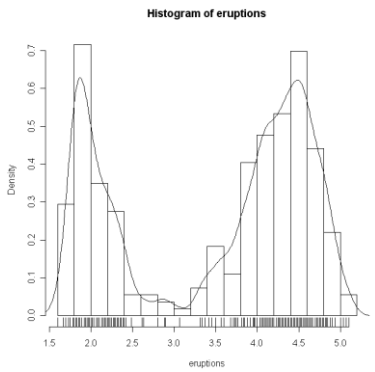


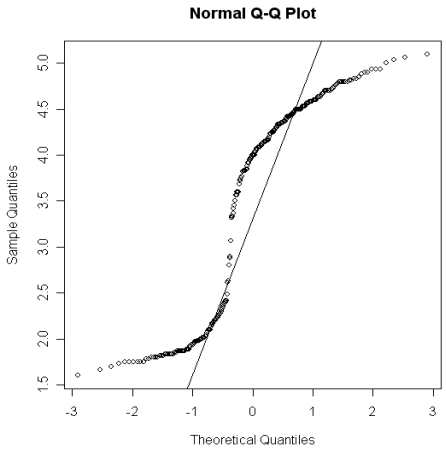


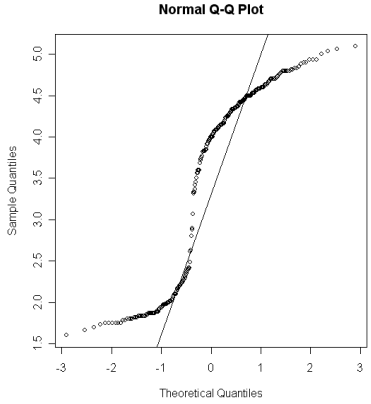


(5) The data for waiting time eruptions in Yellowstone is shown below. Do the assumptions of normality apply to these data? Support your answer.

```
> hist(eruptions)
## make the bins smaller, make a plot of density
> hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
> lines(density(eruptions, bw=0.1))
> rug(eruptions) # show the actual data points
```








```
> hist(eruptions)
## make the bins smaller, make a plot of density
> hist(eruptions, seq(1.6, 5.2, 0.2), prob=TRUE)
> lines(density(eruptions, bw=0.1))
> rug(eruptions) # show the actual data points
```

(6) In R the following is typed:

```
>
>
> a<-c(108000, 104000, 102000)
> b<-c(0.0001, 0.0002, 0.0003)
> t.test(a,b)

        Welch Two Sample t-test

data:  a and b
t = 59.3404, df = 2, p-value = 0.0002839
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  97077.5 112255.8
sample estimates:
  mean of x   mean of y
104666.6667      0.0002
```

What is the null hypothesis being evaluated. What are the assumptions that are used to generate the p-value?

(7) In R the following is typed:

```
>
> a<-c(108000, 104000, 102000)
> b<-c(0.0001, 0.0002, 0.0003)
> wilcox.test(a,b)

        Wilcoxon rank sum test

data:  a and b
W = 9, p-value = 0.1
alternative hypothesis: true mu is not equal to 0
```

What is the null hypothesis being evaluated. What are the assumptions that are used to generate the p-value? Why is the p-value so much larger (closer to 1) than in question 6?

(8)In R the following is typed. In your own words, explain how the numbers pointed to by the arrow are calculated and what they mean. For p-values, state what assumptions were used to generate those p-values.

```
> X <- c(30,20,60,80,40,50,60,30,70,60)
> Y <- c(73,50,128,170,87,108,135,69,148,132)
> myLinearModel = lm( Y ~ X )
summary(myLinearModel)


Call:
lm(formula = Y ~ X)

Residuals:
   Min     1Q Median     3Q    Max
  -3.0   -2.0   -0.5    1.5    5.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.00000    2.50294   3.995  0.00398 **
X            2.00000    0.04697  42.583 1.02e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.739 on 8 degrees of freedom
Multiple R-Squared: 0.9956,     Adjusted R-squared: 0.9951
F-statistic:  1813 on 1 and 8 DF,  p-value: 1.020e-10
```

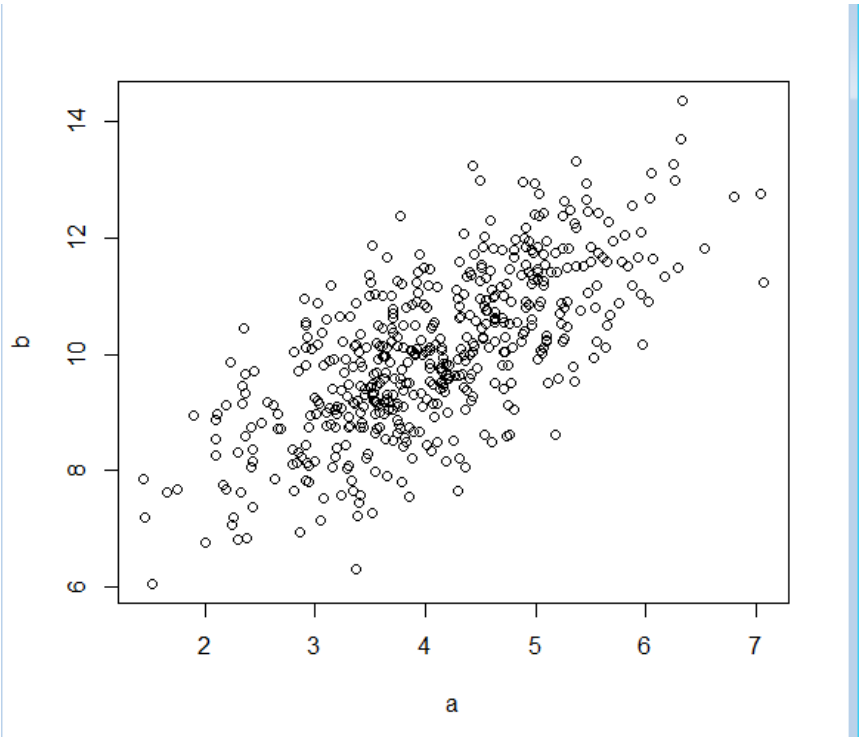
(9) **Define the F distribution.**

(10) An ANOVA test is used to discriminate two models. The full model has 5 degrees of freedom and an associated residual sum squared of 100. The reduced model has 10 degrees of freedom and an associated residual sum squared of 200. What is the value of the F statistic? Show your work.

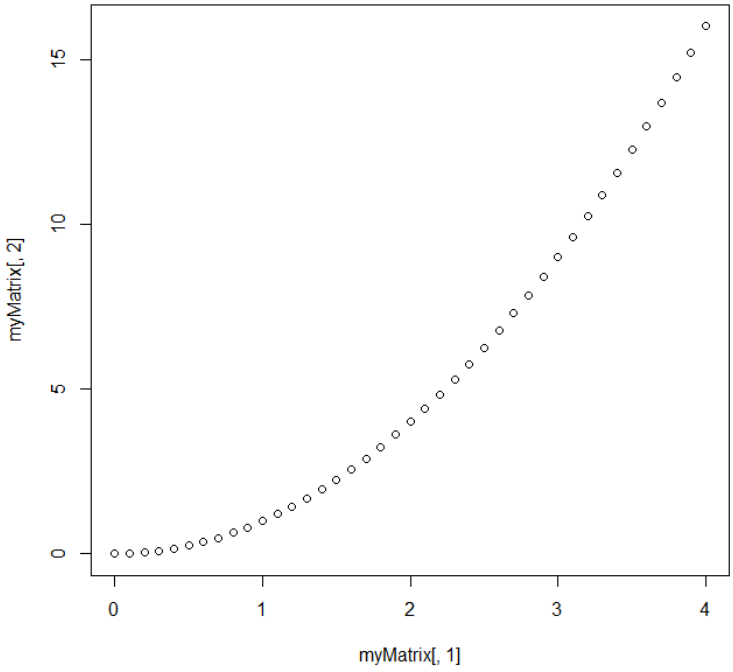
(11) The following is typed into R. What is the null hypothesis being tested. What are the assumptions being used to generate the p-values. Is this a one-sided or two-sided test?

```
>
> rm(list=ls())
> a <- c(2.1,3.2,2.0,1.9)
> b <- c(1.8,1.3,1.7,1.2,1.1)
> ( myData <- c(a,b) )
[1] 2.1 3.2 2.0 1.9 1.8 1.3 1.7 1.2 1.1
> ( categories <- c( rep('A', length(a)), rep('B', length(b)) ) )
[1] "A" "A" "A" "A" "B" "B" "B" "B" "B"
> ( categories <- factor(categories) )
[1] A A A A B B B B B
Levels: A B
> myLm <- lm( myData ~ categories, x=TRUE)
> anova(myLm)
Analysis of Variance Table

Response: myData
           Df Sum Sq Mean Sq F value  Pr(>F)
categories  1 1.7209 1.72089  8.0956 0.02486 *
Residuals   7 1.4880 0.21257
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

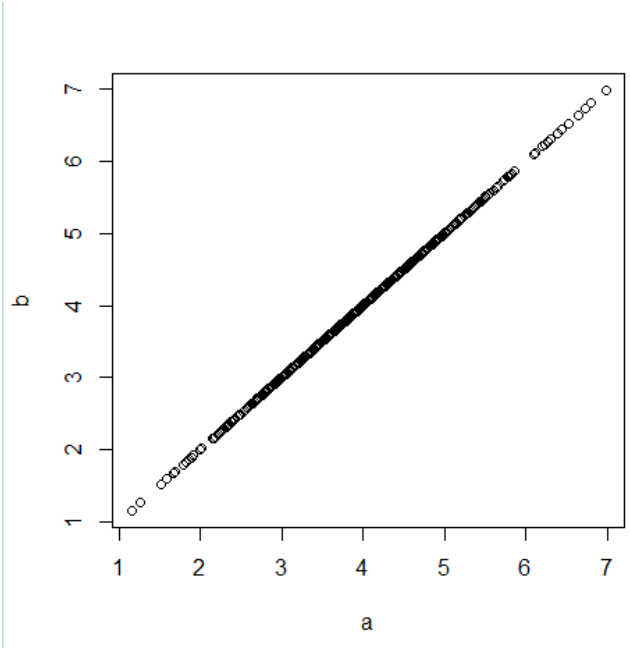
(12) An analyst tells you that "a" and "b" below are the first two principle components of a multi-dimensional dataset? Is this possible? Why or why not?



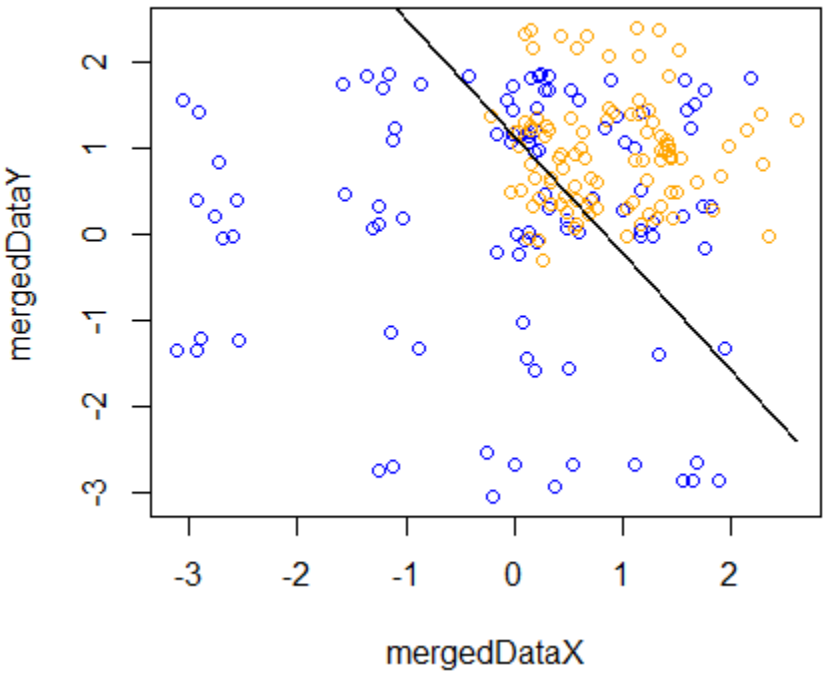
(13) Draw a graph representing the below data transformed into the first two principle components.



(14) In a PCA transformation of the below data, what % variance would be explained by the first principle component? What % variance would be explained by the 2nd principle component?



(15) In the below dataset, why would a k-means nearest neighbor classification model yield better results than the linear discrimination model below?

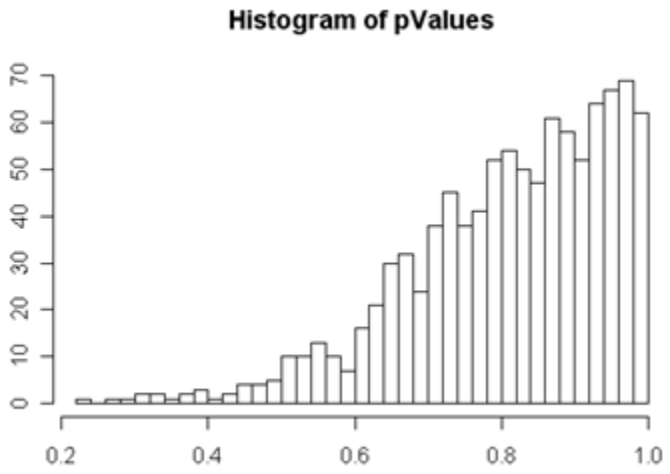


(16) In the above graph, a nearest neighbor model with k=1 could correctly classify every data point as blue or orange? Why not just always use that model as a classifier? What is the bias vs. variance trade-off?

(17) As the number of nearest neighbors goes up, does the complexity of the model go up or down? Is a complex model more likely to underfit or overfit a training data set?

(18) The following code simulates 10 students taking an exam with a mean of 100 and then tests a null hypothesis that the mean score of the 10 students was 100. Are the p-values it produces uniform? If not, why not? Fix the code so the p-values are uniform.

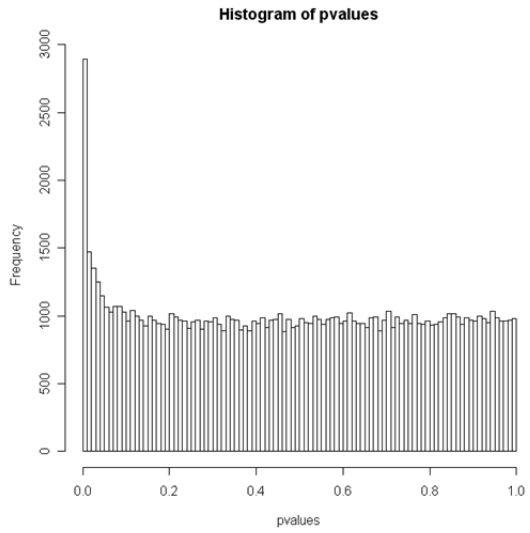
```
rm(list=ls())
pValues <- array()
numTrials <- 1000
for( i in 1:numTrials)
{
        manyValues <-  rnorm(10,mean=100,sd=12)
        manyValues = - abs(( mean(manyValues)  - 100 ) /12);
        pValues[i] = 2 * abs(pnorm(manyValues));
}

hist(pValues,breaks=50)
```



(19) What is the relationship between the normal distribution, the chi-square distribution, the t-distribution and the F-distribution?

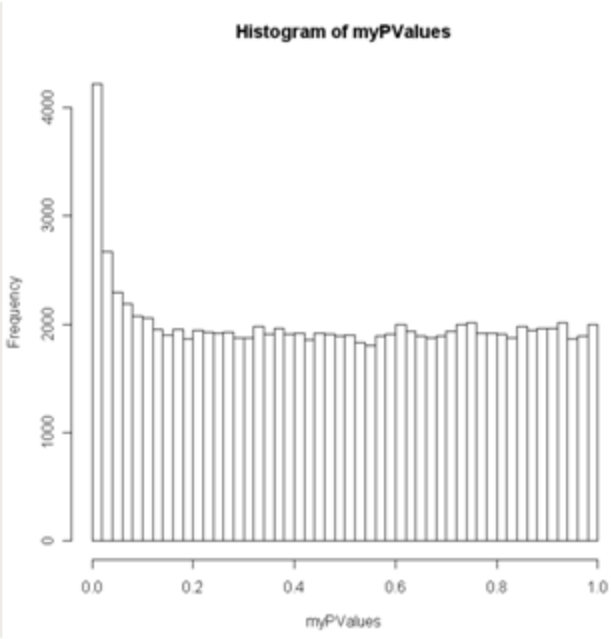
(20) The following code simulates 10 students taking an exam with a mean of 100 and then tests a null hypothesis that the mean score of the 10 students was 100. Are the p-values it produces uniform? If not, why not? Fix the code so the p-values are uniform.

```
rm(list=ls())
s <- seq(1,100000)
pvalues <- array();
for( i in s )
{
  sampleSize <-10
  myVector <- rnorm(sampleSize,mean=100,sd=12)
  stderr = sd(myVector) / sqrt(sampleSize)
  zVal = -abs( (mean(myVector)-100)/stderr)
  pvalues[i]=2  * pnorm(zVal)
}
hist(pvalues,breaks=100)
```



(21) The following code simulates 10 students taking an exam with a mean of 100 and then tests a null hypothesis that the mean score of the 10 students was 100. Are the p-values it produces uniform? If not, why not? Fix the code so the p-values are uniform.

```
>
> numTrials <- 100000
> sampleSize <- 5
>
> myPValues <- array()
>
> for( i in 1 : numTrials)
+ {
+ aSamples <- rnorm(sampleSize)
+ bSamples <- rnorm(sampleSize,sd=10)
+ myPValues[i] <- t.test(aSamples,bSamples,var.equal=TRUE)$p.value
+ }
>
> hist(myPValues,breaks=50)
```



(18) How is the AIC criteria defined? When is a model "better" under the AIC criteria? In a maximum likelihood fit, why not always just take the most likely model?

(19) You have an experimental design in which 10 hospitals use drug A and 10 other hospitals use drug B. You measure the weight of each patient in the study. Why can't you use a simple lineal model like:

lm(weight ~ drug + hospital)

to analyze your data? How can you analyze your data to test the null hypothesis that drug has no effect?

(20) What are the different assumptions of a variance-covariance matrix for residual errors in a linear model that looks like this:

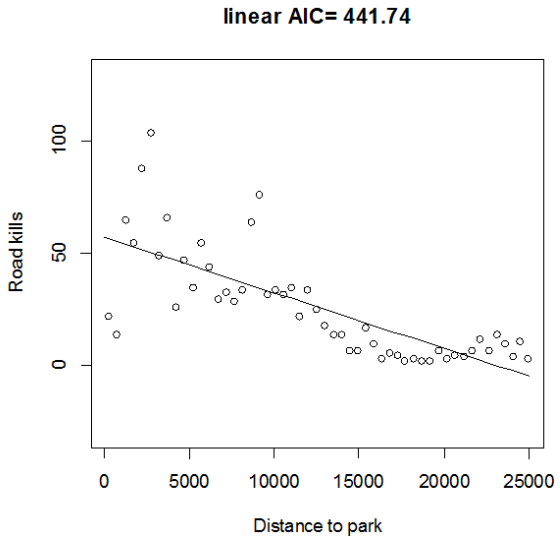
σ^2	**0**	**0**	**0**	**0**
0	σ^2	0	0	0
0	0	σ^2	0	0
0	0	0	σ^2	0

0 0 0 0 σ^2

Vs. a variance-covariance matrix for residuals that looks like this:

$$\mathbf{V}_i = \Sigma_i = \begin{pmatrix} \sigma^2 & \varphi & \varphi & \varphi & \varphi \\ \varphi & \sigma^2 & \varphi & \varphi & \varphi \\ \varphi & \varphi & \sigma^2 & \varphi & \varphi \\ \varphi & \varphi & \varphi & \sigma^2 & \varphi \\ \varphi & \varphi & \varphi & \varphi & \sigma^2 \end{pmatrix}$$

(21) What are two problems with the fit shown below? What are some alternative models that might solve these problems?




```
plot(RK$D.PARK, RK$TOT.N, xlab = "Distance to park",ylab = "Road kills",ylim=c(-30,130))

M0 <- lm( RK$TOT.N ~ RK$D.PARK )
plot(RK$D.PARK, RK$TOT.N, xlab = "Distance to park",ylab = "Road kills",ylim=c(-30,130),main=paste( "linear AIC=",
format(AIC(M0),digits=5)))

xRange<- seq(from = 0,to = 25000, by = 1000)
linearMeans <- coef(M0)[1] + coef(M0)[2] * xRange
lines( xRange, linearMeans)
```

(22) What is the relationship between a "standard" linear model and generalized linear models based on the Poisson and negative binomial distributions?

(23) What is the relationship between a logistic regression and the binomial distribution? When should you use a logistic regression?

(24) What is a zero-inflated generalized linear model?

(25) Understand how these equations define a generalized linear model based on the Negative Binomial distribution (and the equivalent equations for the Poisson distribution and for logistic regression)

$$Y_i \sim NB(\mu_i, k)$$

$$E(Y_i) = \mu_i \quad \text{and} \quad \text{var}(Y_i) = \mu_i + \frac{\mu_i^2}{k}$$

$$\log(\mu_i) = \eta(X_{i1}, \cdots, X_{iq}) \quad \text{or} \quad \mu_i = e^{\eta(X_{i1}, \cdots, X_{iq})}$$