Final study guide:

Dates: Last day of class. **May 3rd.** This lab (#7) is due.

Final exam. **May 12th;** Thursday May 12th from 11:00 am to 1:30 pm . In the 1st floor classroom. Bioinformatics 104.

Final project also due **May 12th.**

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

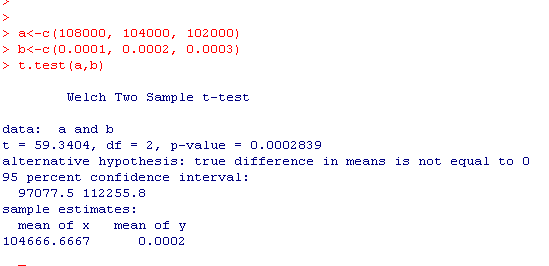
You are responsible for the material on the mid-term study guide:

https://fodorclasses.github.io/classes/stats2022/MidtermStudyGuide.docx

In general, you are responsible for the material in the lectures (on the PowerPoint slides) except **not lectures 22 and lectures 26.** You are not (obviously) ever responsible for details of Java implementations of alogorithms.

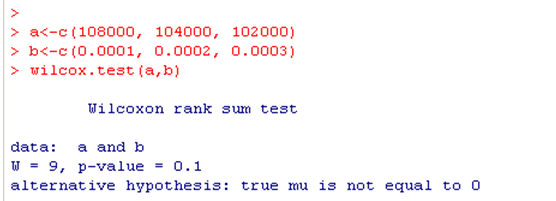
Finally, you are responsible for the material in the labs. If you were asked to perform an operation in a lab exercise, you may be asked to perform a similar operation on the final.

**(1)** In R the following is typed:

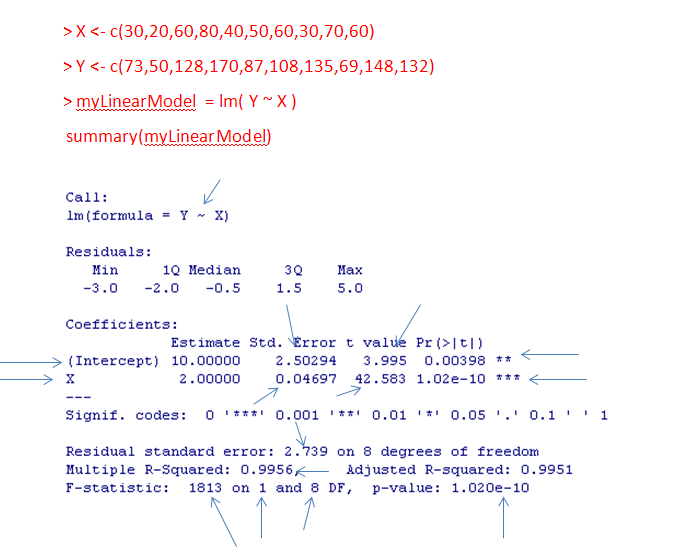


What is the null hypothesis being evaluated? What are the assumptions that are used to generate the p-value?

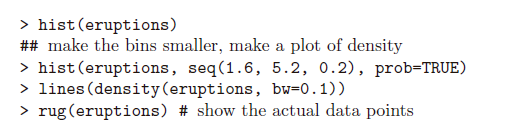
**(2)** In R the following is typed:

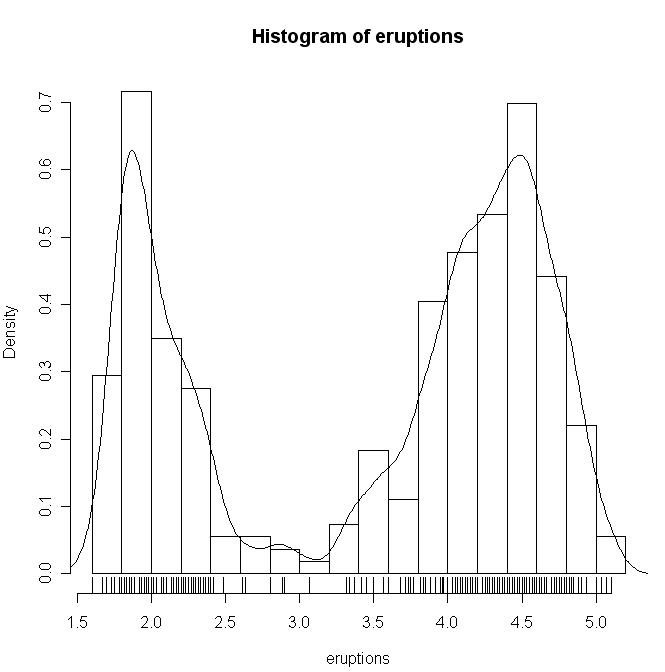
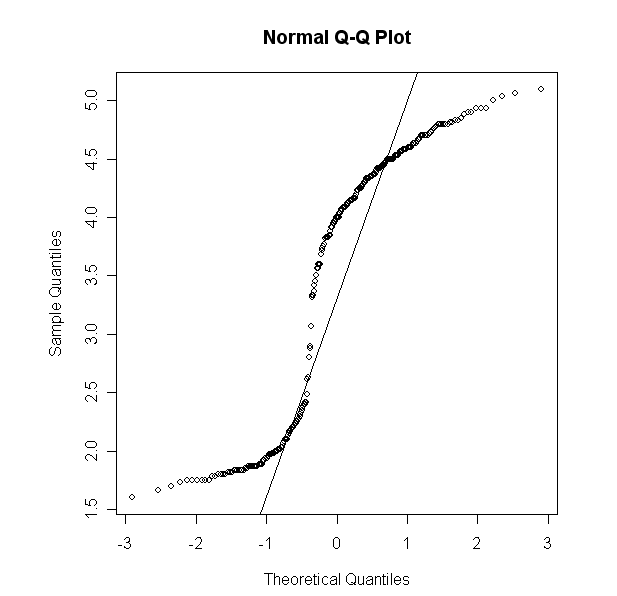
****

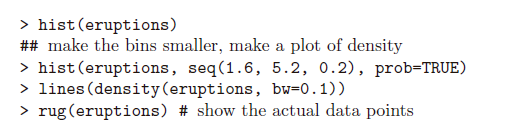
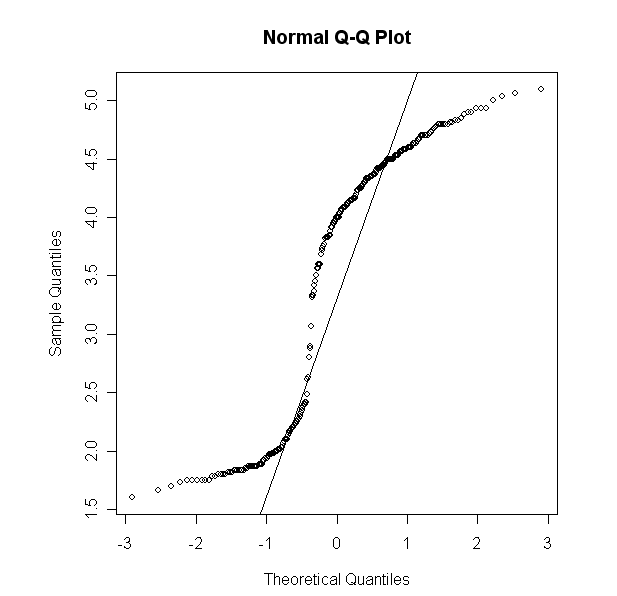
What is the null hypothesis being evaluated? What are the assumptions that are used to generate the p-value? Why is the p-value so much larger (closer to 1) than in question 31?

(3) 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.****

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

****

** **

****

(5) Define the F distribution.

(6) 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.

(7) In your own words, what does the central limit theorem state. Does the central limit theorem guarantee that every data set with sufficient sample size is normally distributed? Why or why not?

(8) How does the mean and variance of a distribution change when you add a constant to each datapoint in the distribution? When you multiply each datapoint in the distribution by a constant?

(9) The distribution of students who have taken an exam is 100 with a SD of 12 (and we know these values ahead of time). The following code takes the average scores of 10 students and then attempts to uses the normal distribution to compare the average to the expected average of 100. But the p-values that are generated are not uniform? Why? Fix the code so that a uniform distribution is produced.

rm(list=ls())

pValues <- vector()

numTrials <- 1000

for( i in 1:numTrials)

{

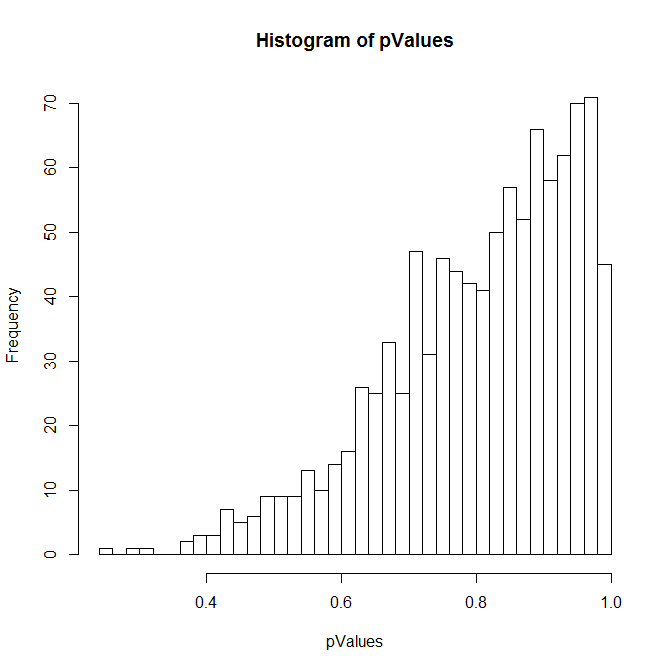
manyValues <- rnorm(10,mean=100,sd=12)

manyValues = - abs(( mean(manyValues) - 100 ) /12);

pValues[i] = 2 \* abs(pnorm(manyValues));

}

hist(pValues,breaks=50)



(10) The distribution of students who have taken an exam is 100 with a SD of 12. The following code takes the average scores of 10 students and then attempts to uses the z-test to compare the average to the expected average of 100 with the SD estimated from the 10 students. But the p-values that are generated are not uniform? Why? Fix the code so that a uniform distribution is produced.

rm(list=ls())

s <- seq(1,100000)

pvalues <- vector(length=length(s));

zVals <- vector(length=length(s));

for( i in s )

{

sampleSize <-10

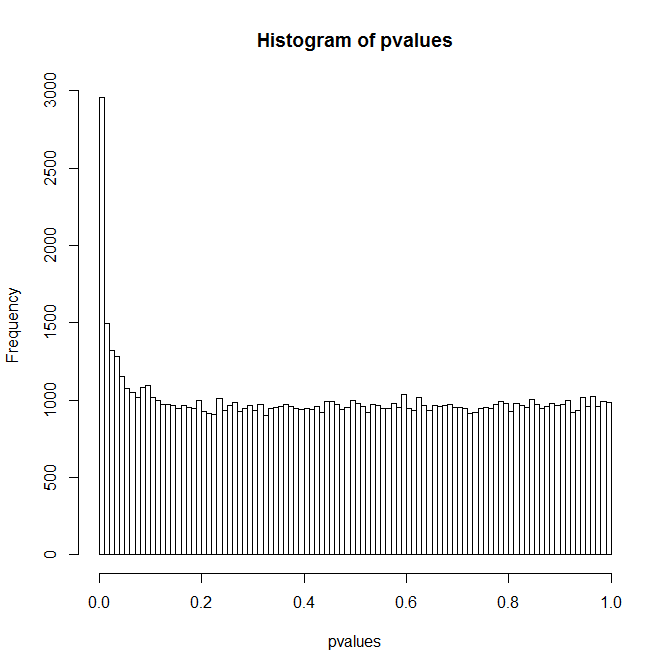
myVector <- rnorm(sampleSize,mean=100,sd=12)

zVals[i] = (mean(myVector)-100) \* sqrt(sampleSize)/sd(myVector)

pvalues[i]=2 \* pnorm(-abs(zVals[i]))

}

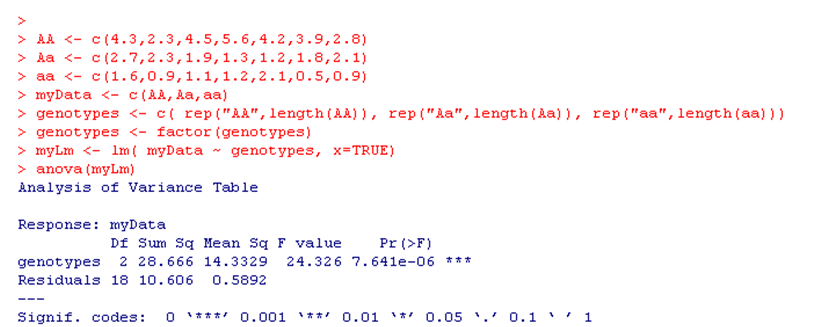
hist(pvalues,breaks=100)



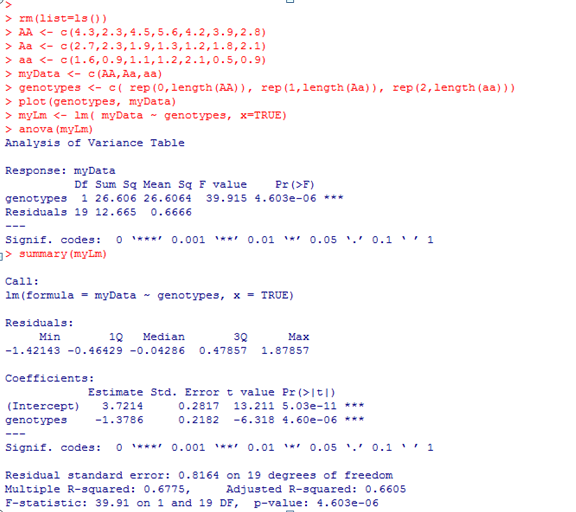
(11) How is a standard normal distribution defined? How is the standard error of the mean defined?

(12) How is the t-distribution defined? When does the t-distribution approach the standard normal distribution?

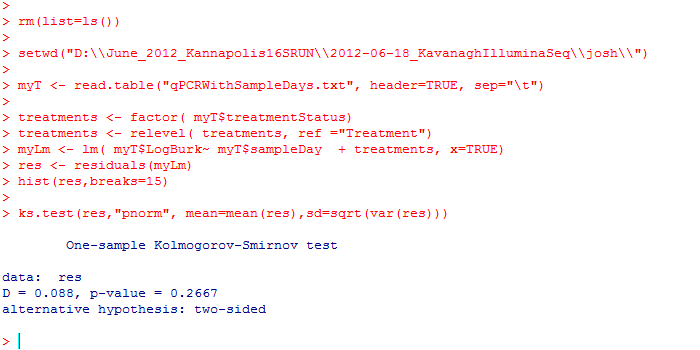
(13) The following is typed into R. What is the null hypothesis being evaluated by the p-values? What assumptions went into producing that p-value? Write the equations for the full and reduced model. Draw graphs representing the full and reduced model.



(14) 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.

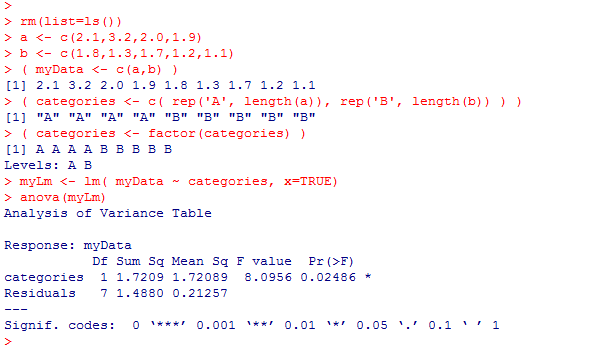


(15) Given the below results, does the assumption of normality seem appropriate for this linear model? Justify your answer.

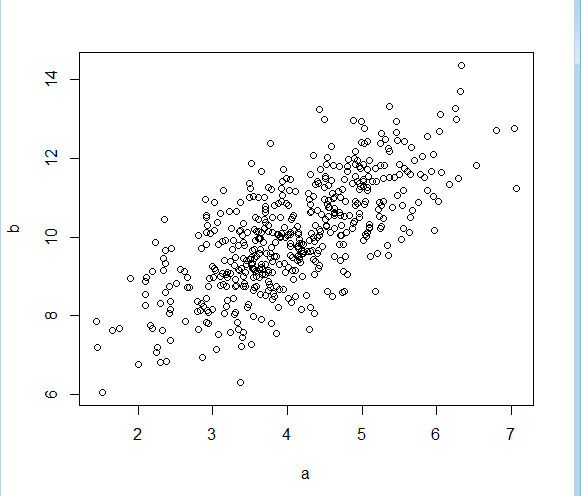


|  |  |
| --- | --- |
|  |  |
|  |  |

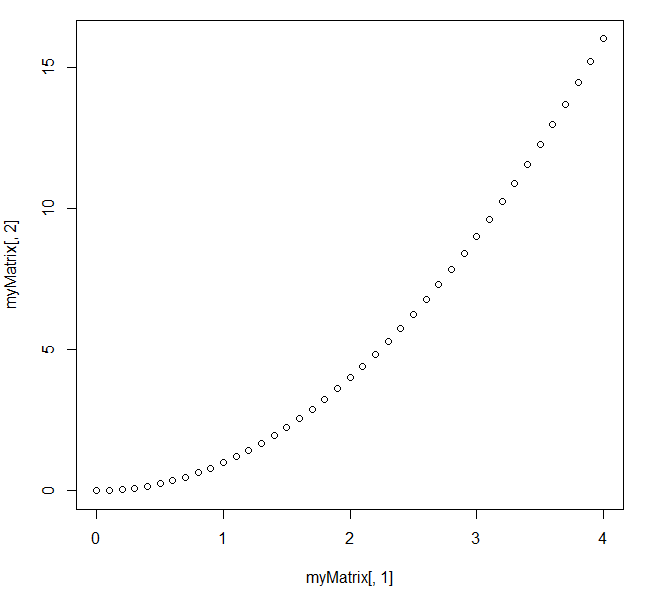
(16) The following is typed into R. What is the null hypothesis being tested. What are the assumptions being used to generate the p-values. Is this a one-sided or two-sided test? What is an alternative test that could relax the assumption of equal variance?

****

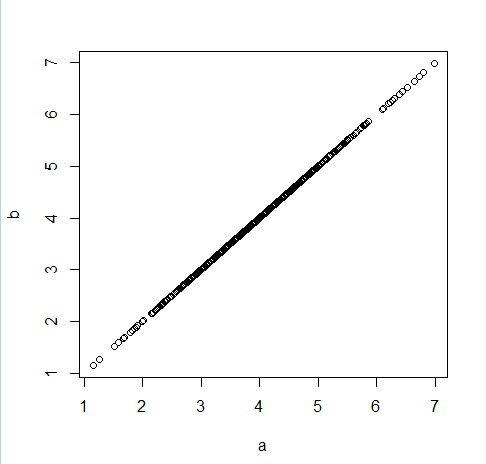
(17) An analyst tells you that “a” and “b” below are the first two principle components of a multi-dimensional dataset? Is this possible? Why or why not?



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



(19) 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?



(20) 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?

(21) 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. What are some of the limitations of a simple lineal model like:

lm( weight ~ drug + hospital)

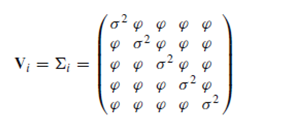
to analyze your data? How can you analyze your data to test the null hypothesis that drug has no effect?

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

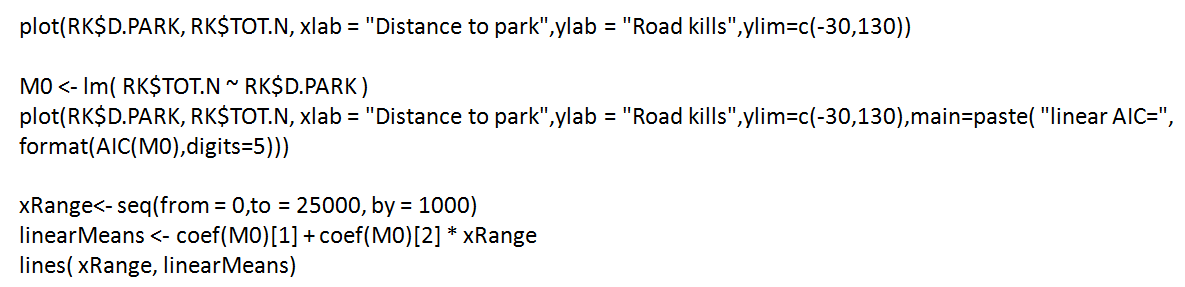
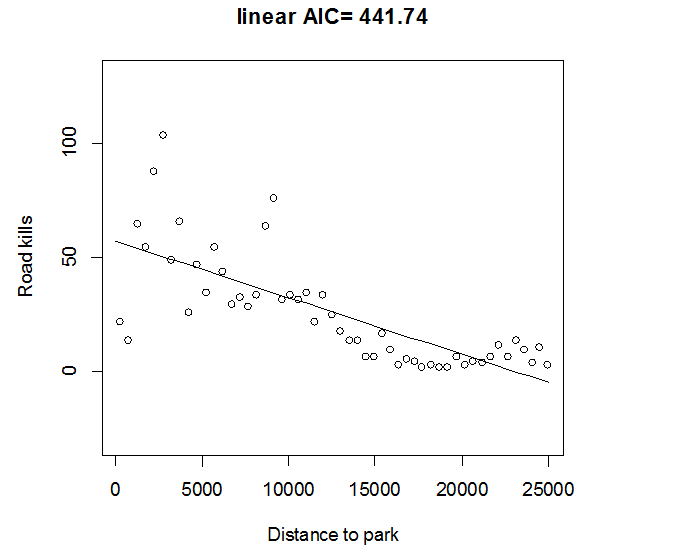
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **σ2** | **0** | **0** | **0** | **0** |
| 0 | σ2 | 0 | 0 | 0 |
| 0 | 0 | σ2 | 0 | 0 |
| 0 | 0 | 0 | σ2 | 0 |
| 0 | 0 | 0 | 0 | σ2 |

Vs. a variance-covariance matrix for residuals that looks like this (where the off-diagonal is

Zero for samples not in the same group) :



(23) What are two potential problems with the model shown below? What are some alternative models that might solve these problems?

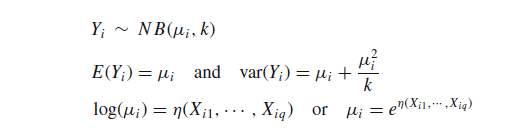


(24) What is the relationship between a “standard” linear model and generalized linear models based on the Poisson and negative binomial distributions? When is it appropriate to use one over the other?

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

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

(27) 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)



(28) What is the difference between the forward and backwards algorithm.

(29) What is the difference between the Viterbi algorithm and the forwards algorithm.

(30) What does the sentence: “First order Markov chains have no memory” mean?

(31) What is the difference between emission and transmission probabilities in a Markov chain?

(32) When working with posterior decoding on long strings, why do we need to work in log(probabilities)?