Study guide for midterm:

Schedule: Feb 27 and March 1: lecture as normal

March 1 (lab time): lab to go through this study guide.

Spring break: March 6 & March 8

Monday March 13th: Mid-term (morning class).

Wed March 15th: AF in D.C. – In class exercise during lecture and lab time.

------------------------------------------------------------------------------------------------------------------------------

Questions on the mid-term will be a mix of questions on this study guide and questions you haven’t seen before, but no new topics will be introduced.

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

Mid-term is closed book.

(1) What is the mean of these numbers { -10,-5,0,5,10}.

(2) What is the median of these numbers : { -30,-20,4,10,15}

(3) True or false: The median calculation is more robust to outliers than the mean. Why?

(4) The variance of a set of numbers is 16. What is the standard deviation?

(5) Define the following data-types in R: lists, vectors, arrays, data frames and matrices.

(6) Draw the following probability density functions:

The uniform distribution;

The binomial distribution with p = .25 and N = 100;

The beta distribution with alpha = 1, beta =1

The beta distribution with alpha = 25, beta = 25

The beta distribution with alpha = 100, beta = 25

The chi-square distribution with 25 degrees of freedom

(7) Define mean and variance for a discrete distribution.

(8) In your own words, describe the law of large numbers.

(9) Define the mean and variance for continuous distributions.

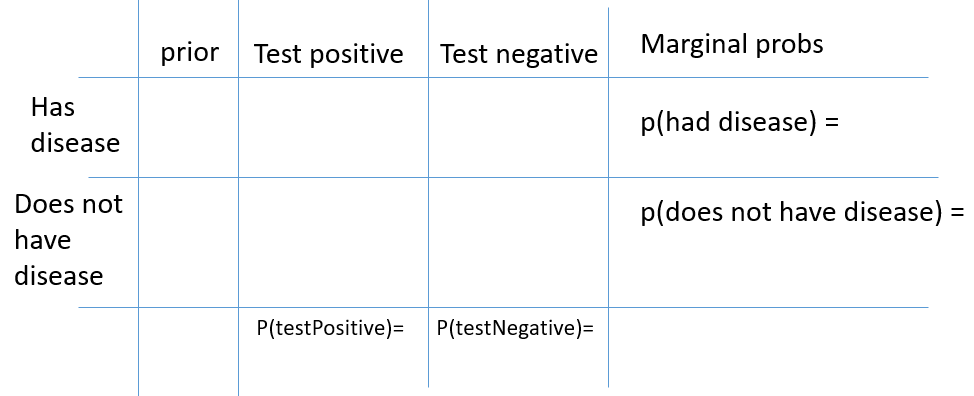
(10) Derive Bayes’ law

(11) Understand the Bayesian universe for a discrete two state test (i.e. the Cylon detector)

(12) Define the following terms: posterior probability, prior probability, likelihood probability, marginal probabilites,Bayesian update, joint probability, conditional probability.

(13) Our prior belief is that 20% of all people have a disease. We have a test where p(positiveTest|patient has disease) = 0.90 and p(positiveTest|patient does not have the disease) = 0.1

Fill out the Bayesian universe for this situation (you don’t have to work out the math). So you can write “(.2) \* (.9)” and not .18



We see a positive test in the above situation. What is our posterior probability that the person has the disease? ( Again, you don’t have to work out the math: so for example you could say

(0.2) \* (0.3) / (0.2 + 0.1)

rather than 0.2)

(14) What is “5 choose 2”

(15) What is the expected mean and variance of the binomial distribution? Of the Poisson distribution?

(16) What is the difference between pbinom, dbinom and rbinom in R?

(17) Consider three coins with a probability of getting a head of 0.3, 0.5 and 0.7. We pick up one of the coins (we don’t know which one). Starting with a uniform prior, fill out the Bayesian universe below… (again, you don’t need to work through all the math, so you can say “0.2\*0.3” rather than “0.06”.

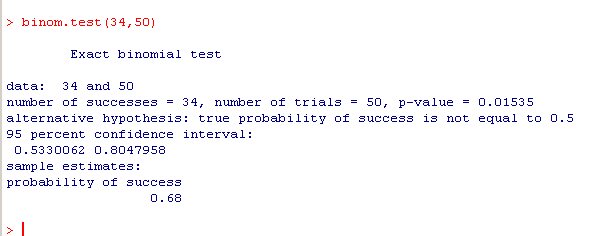
(18) In the above example, you see a head. What is the posterior probability that you’ve picked up the fair coin (the coin with p(head) =0.5). Show your work.

(19) How is the beta distribution defined (i.e. what is the PDF of the beta distribution)? What is the relationship between the beta distribution and the binomial distribution? Define the term conjugate prior. Why is the beta distribution the conjugate prior of the binomial distribution?

(20) The beta distribution is used as a prior with alpha=5 and beta=5. Then 26 heads and 24 tails are observed. What are the values for alpha and beta for the beta posterior?

(21) What values for alpha and beta should be used to use the beta distribution as a uniform distribution. Why is the beta distribution not uniform whenever alpha and beta are equal?

(22) You pick up a coin and flip it 50 times, observing 34 heads. To determine if the coin is fair, you type:

****

Is this is a two sided or one-sided test?

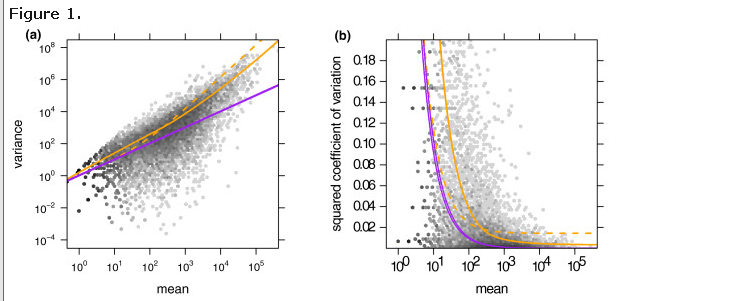
In your own words, explain how the p-value of 0.01535 is calculated.

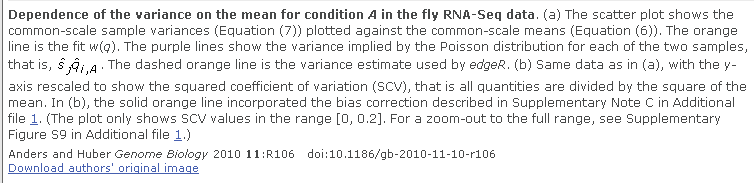
What is the probability of success (0.68). In your own words, how is that defined?

What is the relationship between the hypergeometric distribution and the binomial distribution? Under what circumstances do the hypergeometric and binomial distributions converge?

What is the relationship between the Poisson distribution and the binomial distribution? Under what circumstances do the Poisson distribution and the binomial distribution converge?

(23) In your own words, explain (a) in this figure:



****

How does this figure demonstrate that modeling RNA-seq data with the Poisson distribution might be problematic? How is the negative binomial distribution a more appropriate alternative? How is the term “over-dispersion” defined? How is the negative binomial distribution used to model cases where the variance > mean?

(24) What is the relationship between the negative binomial distribution and the binomial distribution?

(25) Choosing among the following distributions: binomial, beta, chi-square, uniform, Poisson, hypergeometric and negative binomial describe which distribution would be most appropriate for each of the following datasets. Where appropriate, indicate what the parameters would be when using the “d” series of functions in R to generate the probability density function.

(A) The results of rolling a fair 100 sided die a large number of times

(B) The results of flipping a fair coin 50 times.

(C) The results of flipping a coin 50 times in which the p(head) = 0.25.

(D) The expected p-values for each gene from an RNA-seq experiment comparing two technical replicates (i.e. where the null hypothesis is always true).

(E) Describe a prior probability capturing our belief about a coin after observing 10 heads and 5 tails.

(F) Describe how many marked cards that would be drawn when drawing 4 cards from a deck with 20 marked and 40 unmarked cards.

(G) Describe how many heads would be expected to be observed from a coin with p(head)=0.9 before observing 3 tails.

(H) For two genomes of 5 billion nucleotides each, the expected number of SNP differences if a SNP difference was expected to occur once every 200,000 nucleotides.

(I) A disease kills 40% of the people who have it. In a cohort of 75 people, the number of people we would expect to survive.

(J) The number of wins that would be expected in a tournament in which a player plays until they lose 5 games with a p(win) = 0.6

(K) The value



Where S is an estimate of the true variance  calculated from n samples.

(26) The prevalence of a disease within a population is 1 in 10,000. A test is developed with:

p(positive result|a person has the disease) = .99 and

p(positive result|a person does not have the disease) = 0.01

A person takes the test and gets a positive result. What is the posterior probability of our belief that the person has the disease.

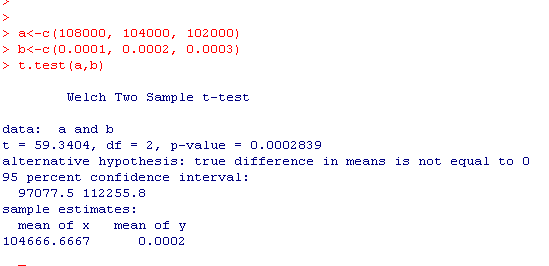
(27) What is the average (expected) value for a fair 5 sided die?

(28) What is the difference between the [] and [[]] operators in R?

(29) What is the expected (average) number of heads in 50 flips when the probability of a head = 0.0? When it is 0.5? When it is 1.0?

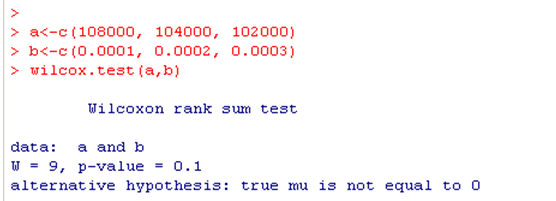
(30) In your own words, describe how the Metropolis algorithm works? How does grid approximation work? Given that grid approximation can produce a result that is not dependent on chance, why not always use grid approximation to sample a posterior distribution?

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

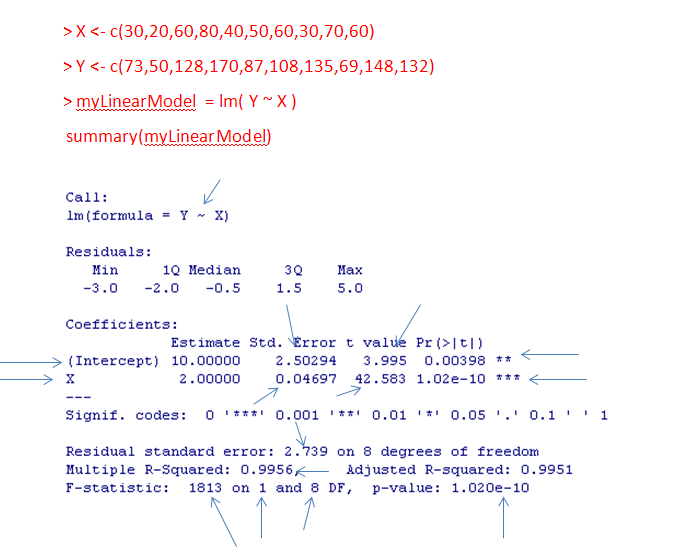


What is the null hypothesis being evaluated? What are the assumptions that are used to generate the p-value?

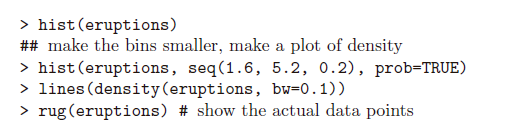
**(32)** In R the following is typed:

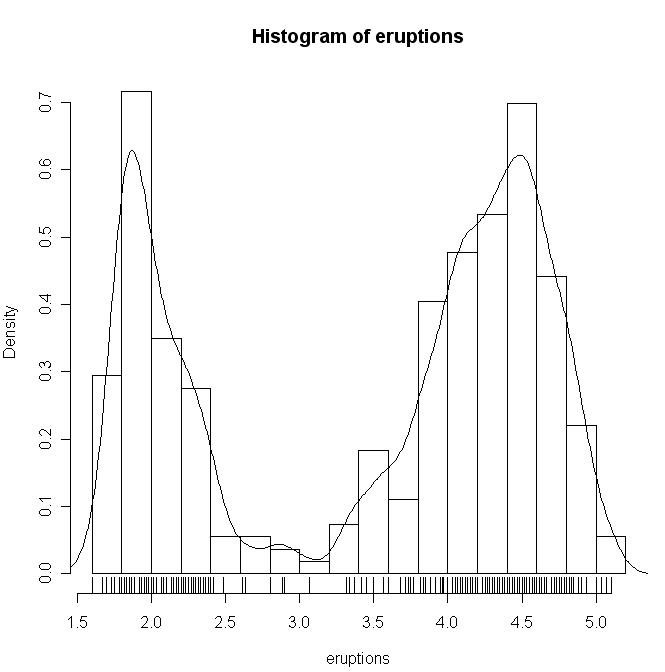
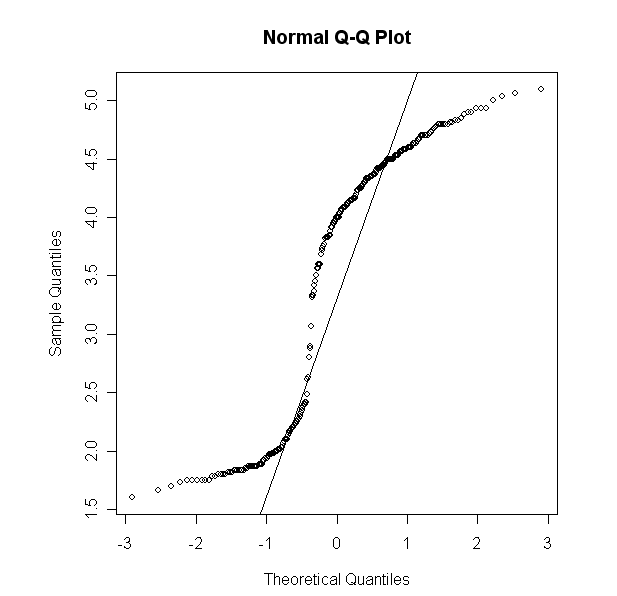
****

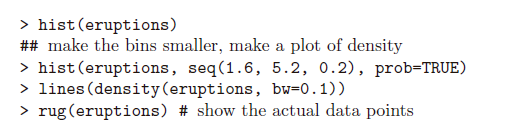
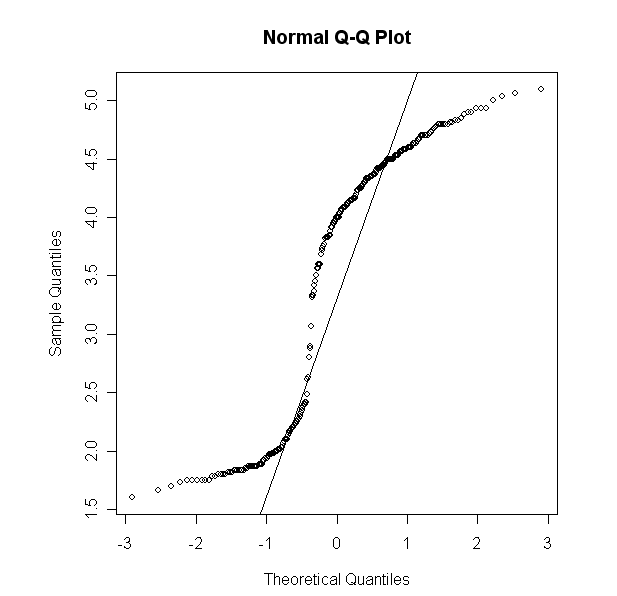
What is the null hypothesis being evaluated? What are the assumptions that are used to generate the p-value? Why is the p-value so much larger (closer to 1) than in question 31?

(33) 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.****

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

****

** **

****

(35) Define the F distribution.

(36) 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.

(37) You have an RNA-seq experiment with 10,000 genes. You perform a statistical test to evaluate a null hypothesis that the genes have the same distribution in two conditions ( say patients with and without cancer). The most significant gene has a p-value of 0.000001. The second most significant gene has a p-value of 0.000003. What is the Benjamini–Hochberg FDR corrected p-value for these two genes. Why do you need to adjust p-values in genomics experiments to correct for multiple hypothesis testing?

(38) 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?

(39) 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?

(40) 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)

{

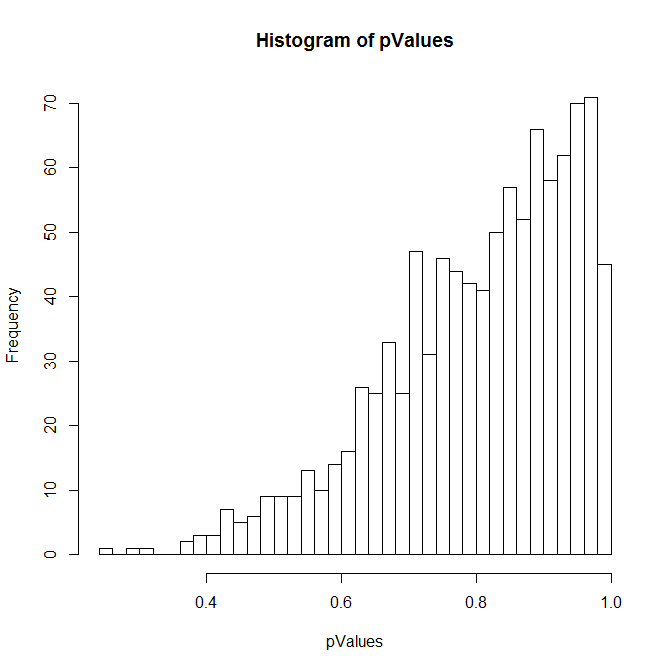
manyValues <- rnorm(10,mean=100,sd=12)

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

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

}

hist(pValues,breaks=50)



(41) 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

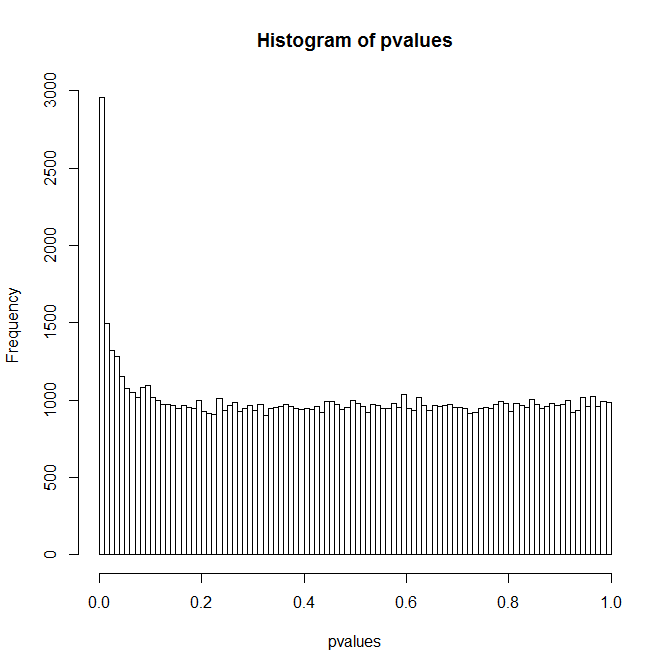
myVector <- rnorm(sampleSize,mean=100,sd=12)

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

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

}

hist(pvalues,breaks=100)



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

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