Please enter your answers to this exam into ecampus by Monday May 4th at 5PM. This test is open note/internet. Do not get assistance from other students or PIs. If you have questions come to office hours (9-10 Friday or Monday)

Questions 1-10 are worth 2.5 points each

Questions 11-15 are worth 15 points each

**1. What is a p-value?**

A p-value describes the probability of observing our data assuming the null hypothesis is true.

A p-value tells us based on our data whether our alternative hypothesis is true.

A p-value describes the probability of generating a statistic as or more extreme under the null hypothesis.

A p-value describes the probability of a false negative given the null and our data.

**2. Why do we need to correct for multiple comparisons?**

To avoid an increased frequency of false positives.

To insure that our power is not reduced.

To know whether we have a large enough sample size for the experiment.

Because doing a number of tests with a single dataset increases our power to detect correlations.

**3. Choose one approach to dealing with the problem of using data that come from different species.**

Only use data from one species at a time. Data from multiple species will lead to false positives.

A standard ANOVA can be used when data comes from multiple species.

Use a Monte Carlo that incorporates the phylogeny into the simulation step.

Use a parametric method these are robust to phylogeny.

**4. An advantage of Bayesian approaches is that they \_\_\_\_\_\_\_\_\_\_\_.**

There are none they are flawed at heart

Bayesian methods are quicker and mathematically simpler than frequentist approaches.

Bayesian methods give clearer black and white answers than do frequentist approaches

Bayesian methods allow us to incorporate our biological knowledge via the prior.

**5. PCA is a method that is helpful in cases \_\_\_\_\_\_\_\_\_\_\_\_\_\_.**

Where we have just a few uncorrelated predictor variables.

Where we have many different response variables.

Where we have many possible variables and we aren’t sure which ones are meaningful.

Where we have only a discrete predictor variable.

**6. The starting portion of an MCMC will have low likelihood \_\_\_\_\_\_\_\_\_\_\_\_\_.**

Because the data is noisy and was not properly cleaned prior to analysis.

Because we have chosen a poor prior for one of our parameters.

Because our model is not appropriate for our data.

Because we start with random starting values that are a poor fit to our data.

**7. Why don’t we sample from the first portion of an MCMC run?**

It is the burnin and may have converged too quickly.

It is the burnin and has too little variation making it look like we have less uncertainty than we do.

It is the burnin and has not yet converged on the posterior distribution.

It is the burnin and only has the values that we started the MCMC with.

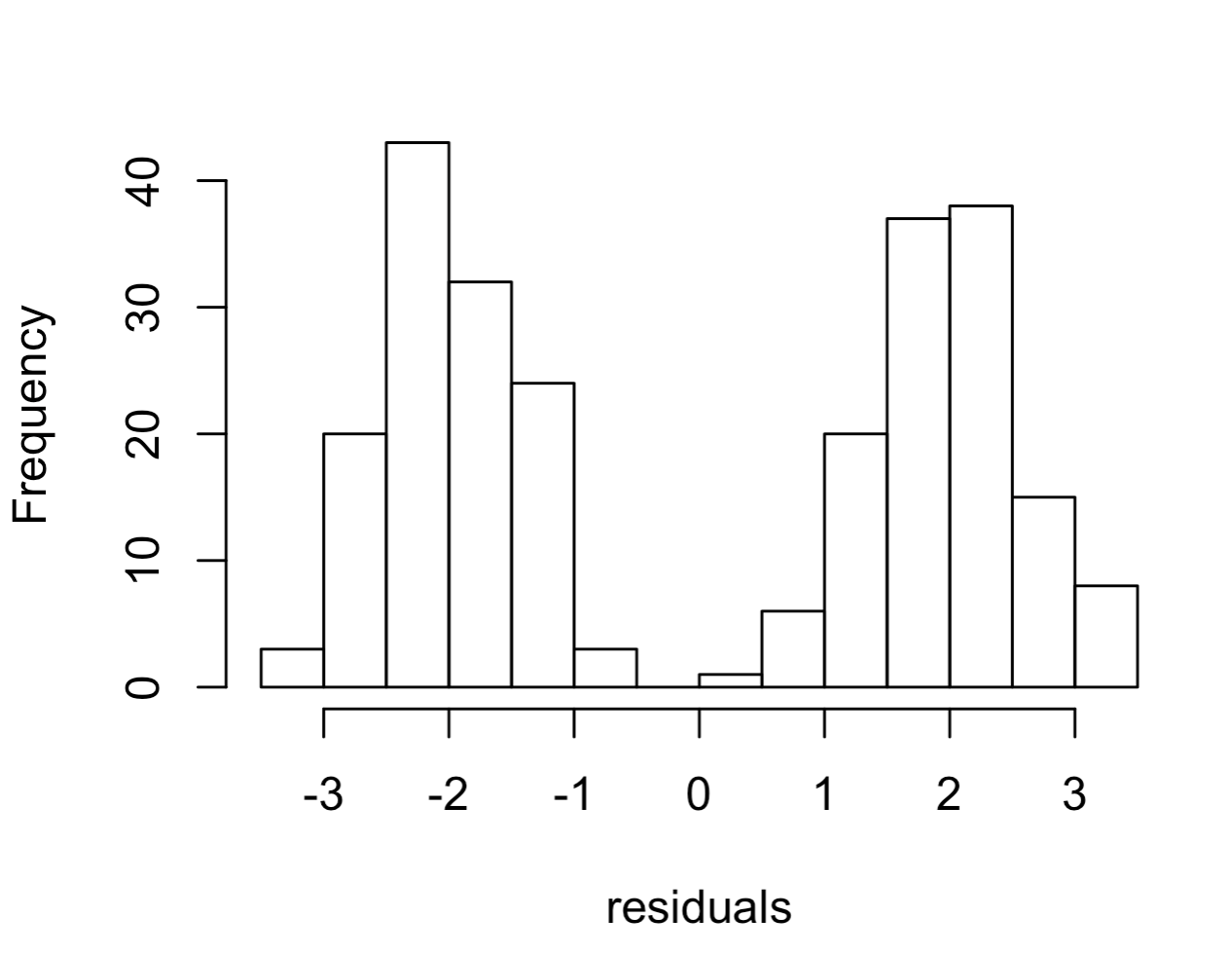
**8. The Beta coefficient in a linear model with a continuous predictor variable describes \_\_\_\_\_\_\_\_\_\_\_\_\_.**

The significance of the predictor variable.

The expected change in the response variable between levels of the predictor variable.

The expected change in the response variable for one unit of change in the predictor variable.

The expected change in the predictor variable for one unit of change in the response variable.

**9. The histogram below is of residuals for a linear model what potential problem do you see in them?**

They are too large

They are too far from zero

Nothing they look like we would expect

The are not normal

**10. Often in an experiment we will have a group that receives a sham treatment. For instance, if infecting an organism with a bacteria via injection we might have a second group that receives an injection of sterile saline. What is this called and what is its purpose?**

It is a treatment group and allows us to assess the significance of our predictor variable.

It is a control group and allows us to know how various confounding variables may change the baseline of our response variable.

It is a control that improves the skill of the experimenter in the experimental protocol.

It can increase our sample size so we get a significant result.

**These questions do not use the data from the review ensure that you are using the files specified for the final exam.**

**11. Download the test.frogs.csv morphology dataset from the course website this data includes measures for 3 species for 10 different traits as well as 10 unidentified species. To what species do these 10 unidentified samples belong (there could be unidentified samples that belong to different species).**

**12. Download the the mcmc log file test.log.csv file from the course website. Select the most appropriate estimate of the mean for the parameter “q01”.**

XXXX

XXXX

XXXX

**13. Download the test.dispersal.csv file from the course website this contains the proportion of individuals that attempt to disperse from 3 populations of beetles. Looking at the recorded variables what do you believe the most important predictor of dispersal is.**

Temperature

Sex

Strain

Media

**14. Use a full model with all possible predictor variables and select the answer that most closely matches the expected proportion of dispersing beetles for Marksville strain males at 27 degrees in conditioned media.**

XXXX

XXXX

XXXX

**15. Download the XXXXXXX file from the course website this data is a simplified version of what we get when we do Hi-C sequencing of genomes. It tells us about the spatial arrangement of chromosomes in the nucleus. The numbers in the table tell us how often two chromosomes were in close contact in the cells that were sequenced. Our expectation (null hypothesis) is that each chromosome should have lots of contact with itself and relatively few with all others. The size of a chromosome should predict the number of contacts it has with itself. In contrast the number of contacts between any two chromosomes should be a function of their sizes. However, some preliminary data suggests that micro chromosomes (very small chromosomes present in birds and some other reptiles) may be tightly associated in the nucleus because they contain genes that are in the same genetic pathways and can be co-regulated. In the table normal chromosomes (macro-chromosomes) are labeled with the prefix “MA” while micro-chromosomes are label with the prefix “MI”. Using a Monte Carlo approach does this data support a close association between any of the micro chromosomes in this dataset?**

Yes Mi1 and Mi2 appear to be closely associated in the nucleus

Yes Mi2 and Mi3 appear to be closely associated in the nucleus

Yes Mi1 and Mi3 appear to be closely associated in the nucleus

Yes more than one pair of micro chromosomes appear to be closely associated in the nucleus

No micro chromosomes appear to be closely associated