Guide for exam 1

Melissa Pulley

First exam

This first exam will be an in-class, Canvas exam. I recommend you bring the following items to the exam:

- Basic 4 function calculator
- Pen or pencil
- Scratch paper

The exam **IS NOT** open book. You will **NOT** be asked to code, but you might need to interpret some code, or interpret code results.

The exam will have an access code.

What type of questions will there be?

The exam will have multiple choice, fill in the blank, short answers, very basic math questions, etc. While knowing definitions is important, the focus will be on critical thinking and *understanding concepts* and being able to apply that understanding when presented with scenarios.

You can email me if you have questions about this guide. If you need help with a specific topic or questions, email me, and we can meet on Wednesday 6th during class hours.

About this guide

This guide is meant to help you study by:

1. Telling you about definitions and concepts you should know
2. Specifying certain *skills* you should have (e.g., be able to estimate the probability of simple events)
3. It will tell you certain topics that you should understand
4. Some exercises that will help you study and will help you understand concepts

The guide lists the concepts and definitions you should understand, but it doesn't explain them. You should check the lecture material and the labs in order to find the explanations/definitions. You are also welcome to check any order books or online explanations (I recommend the ones in the syllabus!)

Guide for exam 1

Basic Definitions, types of data, and data visualization

Understand the basic concepts and definitions that we use in statistics. *Note: It's more important to understand the definitions that to memorize them!* Examples of concepts and definitions you need to understand are:

- Population
- Sample
- Study Design
- Inference
- Census
- Statistics
- Descriptive statistics
- Inferential Statistics
- Data
- Observation
- Variables
- Data
- Accuracy

- Bias
- Precision

You should be able to identify and describe the different types of data.

Q1. 2 pts. Give an example of each of the four types of data we discussed about in class during week 2

A1. We have four type of data
- Ordinal: This is ordered qualitative data. Data might include evalution items on a rubric to grade an asignment. An evaluation item for a particular category may be rated on Beginning, Developing, Acccomplished, or Exemplary performance.
- Nominal: This is unordered qualitative data.This code be location data. A survey of UTK students might ask students that are TN residents their home county.
- Discrete: This is numerical data with integers. This could be represented by daily temporal data within a calendar year. Data could be assigned to the day of the year using Julian Day.
- Continuous data: This code be temporal time-stamp data, where the exact time and day are recorded when data is sampled. These numbers can be decimals.

Summary function. Looking at the following code output (using the iris dataset in R), you should be able to understand the structure and some central tendency measures from your dataset:

```
summary(iris)
```

```
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :50
 versicolor:50
 virginica :50
```

Q2. 2 pts. Answer: How many variables and how many observations are in the iris dataset?

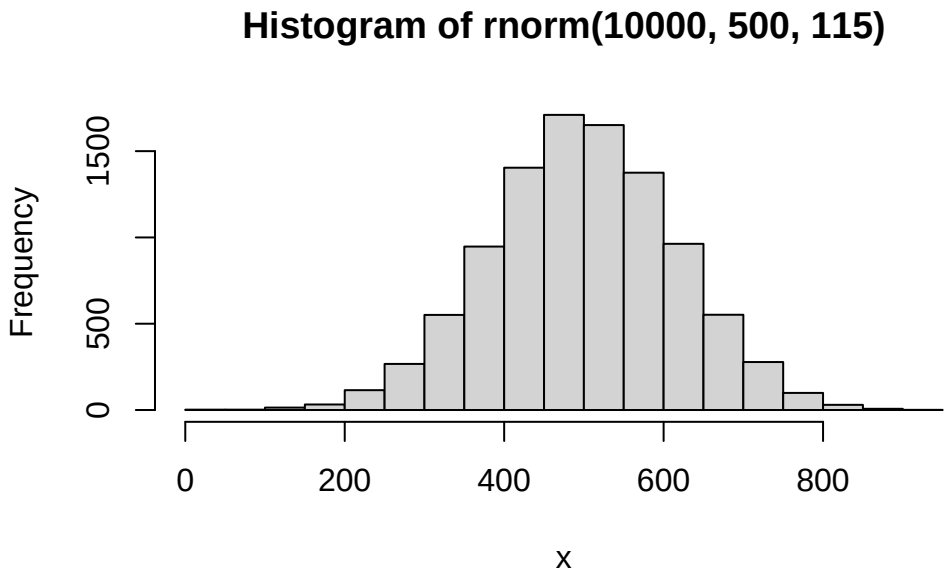
There are 5 variables in this set. There are 150 observations.

Measures of central tendency and dispersion

You should be able to identify and estimate mean and median. And understand how the symmetry of your data distribution affects them.

Q3. 2 pts. In the following histogram, are mean and median the same or different

```
hist(rnorm(10000,500,115),xlab="x")
```

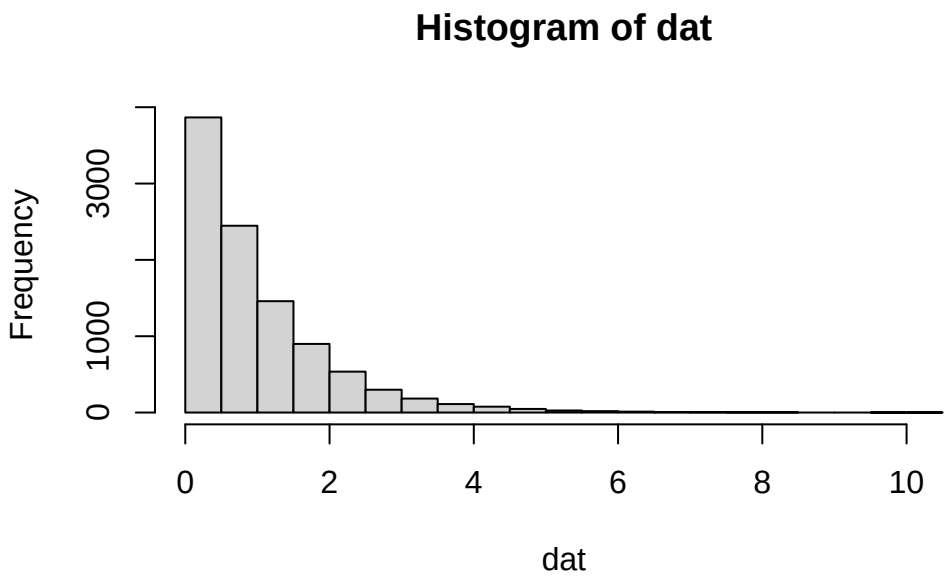


A3. Since the histogram is roughly symmetric, the mean and median should roughly equal.

Q4. 2 pts. Try creating a histogram in which the value of the median is less than the mean

The below histogram is right-skewed so the median is less than the mean

```
set.seed(5)
dat=rexp(10000,1)
hist(dat)
```



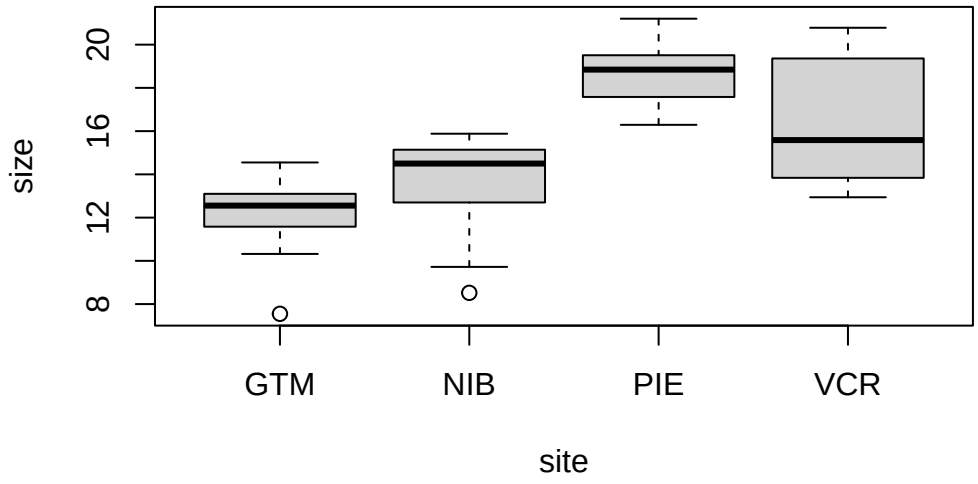

```
mean(dat)
```

```
[1] 1.004031
```

```
median(dat)
```

```
[1] 0.7041072
```

You should understand what a percentile is, as well as variance and standard deviation. Identify the parts of a boxplot. For example, this Fiddler Crab Data:



Q5. 2 pts. Mention one site that is "skewed left" in the Fiddler Crab boxplot. Mention what site has the largest IQR, and which value is higher: NIB's 50th percentile, or VCR's 25th percentile

Probability and distributions

Understand the following concepts:

- Sample Space
- Outcome
- Event

You should be able to estimate the probability of simple events (e.g., getting 2 Heads when tossing 3 coins).

Know the difference between a Bernoulli trial and a Binomial Distribution.

Be **VERY!** familiar with the following functions: *dbinom()*, *pbinom()*.

Q6. 2 pts. Remember when in an assignment (lab2) you were tasked with taking a sample of n=3 from an experiment of 100 coin tosses and you had answer whether you thoughtit was fair or not? Let's do something similar. This time, you tossed 100 coins two times. The first time you got 59 heads. Using the pbinom() function figure out the probability of getting 59 (or more heads) when tossing a fair coin. Would you PERSONALLY say it's a fair coin? (only yes or no, maybe's are not allowed for this answer). After that you throw 62, and 60 heads in the following toss. How about now?

You should be able to define a distribution (and it's OK if it's your own words).

More concepts you need to know:

- Random process and random event
- Random variable
- Understand the three steps to define a discrete probability distribution dbinom() is the way we usually do it in R.
- Probability mass function

You should be able to estimate the expected value. Try to estimate it for the following example:

X = number of heads out of 5 coins

Probability of heads = 0.2 (unfair coin)

```
x<-0:5
probs<-dbinom(0:5,5,0.2)
```

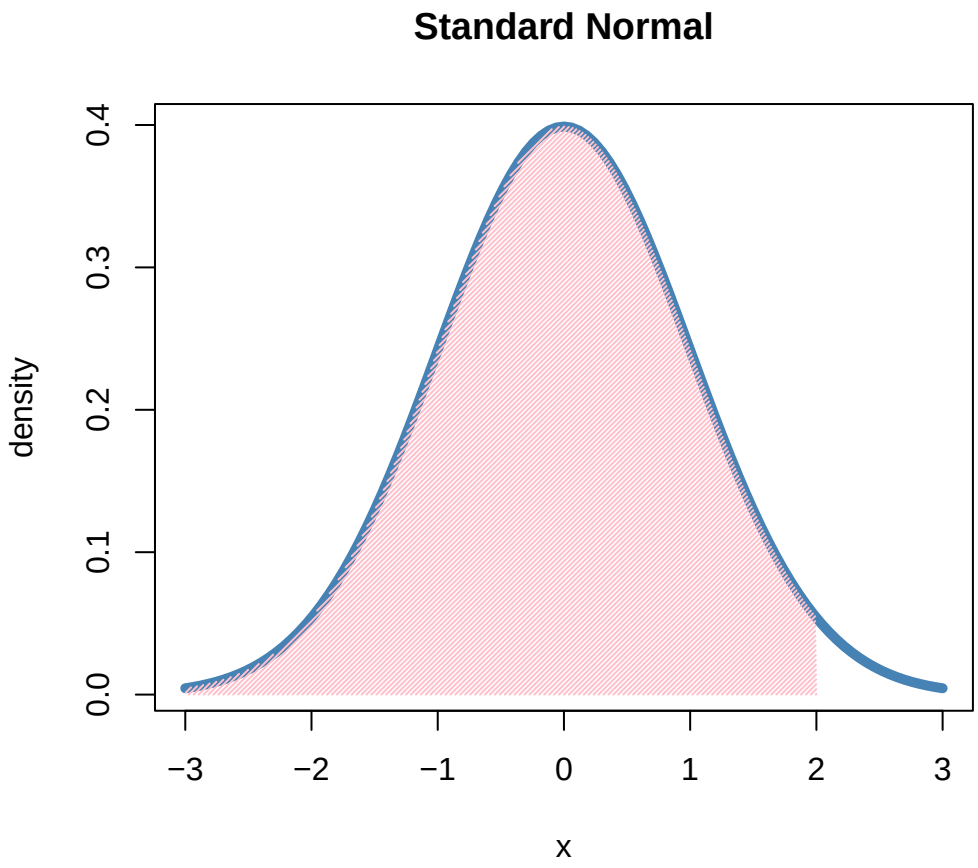
Q7. 2 pts. Either using R, or by hand, estimate the Expected Value, and tell me, what would the expected number (heads) be?

Continuous distributions

Understand the differences between discrete and continuous probability distributions.

Normal Distribution: Be able to identify and describe it

You should understand the standard normal distribution, and how probability is distributed. For example, what proportion (approximately) of the area under the curve is shaded in the following normal distribution?



Understanding Z-values and how you obtain them is important. You should know the following equation:

$$z_i = \frac{x_i - \mu}{\sigma} \tag{1}$$

Imagine the following are different lengths of the lizard *Scelloporus gramicus:*

```
length_Scelloporus<-c(49.79345, 48.27289, 47.01204,
                      50.04918, 33.23075, 36.95524,
                      45.49277, 44.33337, 40.92777,
                      40.02123)
```

Q8. 2 pts. Estimate the z values for each observation

Remember the exercise in which people sampled a population by "snorkeling" and getting mussel sizes? Remember that even though the population was the same, we obtained different values? You should be able to explain why that happened.

You should understand what is a standard error

Why is the Central Limit Theorem important when estimating confidence intervals?

What is a confidence interval? How is it estimated?

Null Hypothesis Statistical Testing

You should be able to define:

- P value
- Null Hypothesis
- Alternative hypothesis
- alpha (significance)
- Type I and type II errors
- One and two sample tests
- T-test
- paired test
- Parametric and non-parametric tests
- Covariance and correlation (and how they are different)
- One tailed and two tailed test
- R-squared

You should be able to look at an output like the following (from the teporingos data):

```
teporingospaired<-read.csv("teporingospaired.csv")
t.test(teporingospaired$diet, teporingospaired$original,
       paired=T, alternative='two.sided')
```

```
    Paired t-test

data:  teporingospaired$diet and teporingospaired$original
t = 5.8934, df = 29, p-value = 2.135e-06
alternative hypothesis: true mean difference is not equal to 0
95 percent confidence interval:
  9.82815 20.27515
sample estimates:
mean difference
       15.05165
```

and interpret it.

Q9. 2 pts. Interpret the output for the teporingos data. Define the null and alternative hypothesis, as well as the decision.

Linear models

Understand the structure of a linear model. Particularly, the following equation:

$$Y_i = \beta_0 + beta_1 X_{1,i} + \epsilon_i$$

where,

$$\epsilon_i \sim N(0, \sigma^2)$$

and know that as many terms as possible can be added to the linear model.

You should know the 4 assumptions of a linear model. Also, please remember that LINEARITY is the most important one!

You should know how to estimate a predicted value for a response variable, given certain values of the explanatory variables. What I mean with this. If you are given a model with 2

predictors, and given the value of the coefficients β_0, β_1, and β_2, as well as some values for X1 and X2, you should be able to obtain the predicted value.

You should understand what an ANOVA null and alternative hypothesis is, and what a pairwise comparison (post-hoc) test is.

Be able to interpret the output of a linear model.

Understand multicollinearity and why it can be undesirable.

Multiple linear regression

The lecture for this topic will be on Wednesday 28 and Friday 1st

You should be able to interpret an output of a multiple linear regression, for example. Try to describe the following model and it's output (from the iris dataset in R). You can explore that dataset if you need more help.

```
summary(lm(Petal.Length~Petal.Width*Species,data=iris))
```

```
Call:
lm(formula = Petal.Length ~ Petal.Width * Species, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-0.84099 -0.19343 -0.03686  0.16314  1.17065

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    1.3276     0.1309  10.139  < 2e-16 ***
Petal.Width                    0.5465     0.4900   1.115   0.2666
Speciesversicolor              0.4537     0.3737   1.214   0.2267
Speciesvirginica               2.9131     0.4060   7.175 3.53e-11 ***
Petal.Width:Speciesversicolor  1.3228     0.5552   2.382   0.0185 *
Petal.Width:Speciesvirginica   0.1008     0.5248   0.192   0.8480
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3615 on 144 degrees of freedom
Multiple R-squared:  0.9595,    Adjusted R-squared:  0.9581
F-statistic: 681.9 on 5 and 144 DF,  p-value: < 2.2e-16
```

Q10. 2 pts. What is the predicted petal length for an individual from the Setosa family with a petal width of 2

Model selection, AIC and Maximum Likelihood

Lecture Monday 4th

Be able to define Likelihood and Maximum Likelihood, understand the benefits of model selection over traditional hypothesis testing, and how to interpret AIC.

Q11. 2 pts. List the two things that you've found the most interesting regarding this course