**Week 9 – CPaT Stats: The R Project for Statistical Computing**

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

PLEASE READ THE DIRECTIONS TO THIS LAB CAREFULLY!

Help Sessions prior to when this lab (and the final) are due:

1. Kara Karboski, who has been working with us during the Tuesday lab sessions will run a special 2 hour help session this Friday (week 9). Think about what time you would most like that to be. I suggest 1-3 or 3-5.
2. Robyn Andruszyn will conduct her regular help session on Monday (Week 10), in Lib 2617, 3-5pm

For information about R (and how to download it), see <http://www.r-project.org/>. R is installed in the Computer Center, but you might want a copy on your own computer. For more information you might also want to check out R Commander (<http://socserv.mcmaster.ca/jfox/Misc/Rcmdr/>) and <http://staff.washington.edu/tlumley/b514/R-fundamentals.pdf>

Note: you have the choice of using either the R command line interface or R Studio for this lab. R Studio has a command line window (Console), as well as 3 other useful windows, e.g., one with all data variables.

**You can choose to complete Part I or Part II of this lab.**

The lab report this week (for Part I) will consist of answering questions about your experience in using R – to be answered after you have completed the lab activities. These questions will be sent to you via email (as well as posted on the fileshare today by noon).

The lab report (for Part II) will consist of answering the questions embedded in the lab text. Please answer questions deleting all extraneous text; with just the question and answer.

Bring hardcopy of the lab report to next week’s lab – put it in your Statistics Portfolio.

Also, please upload any files you created for this lab:

Workspace/\_StatsLabReports/week\_9-R /

Wk7\_stats\_data\_cushingSkomraWeiss.???

where you insert your last names in place of “cushingSkomraWeiss”

**Part I. R Command Line Interface.**

Complete the R Tutorial[[1]](#footnote-1) at <http://www.cyclismo.org/tutorial/R/> . You might find it easier to use R Studio (rather than the “raw” command line interface).

Try to finish 1-10, and one of the two Case Studies. This will provide an excellent review of many of the concepts we covered this quarter, as well as an overview of R: inputting data to R, R data types, basic operations on data, basic probability, basic plots, basic linear least squares regression, confidence intervals, p-values, calculating the power of a test, working with tables and proportions.

Case Study I, taken from a statistics textbook (Moore & McCabe), is relatively simple, and looks at emissions of three pollutants from 46 different engines, in particular carbon monoxide data. You will transform the data so they are closer to being normally distributed, find the confidence interval for the mean and perform a significance test to evaluate whether or not the data are “close enough” to a fixed standard.

Case Study II is more complicated, and explores how to use R to confirm the results from a study reported in [Trends in Serum Lipids and Lipoproteins of Adults, 1960-2002](http://jama.ama-assn.org/content/vol294/issue14/index.dtl) , a paper from the Journal of the American Medical Association (JAMA). The paper examined trends of several studies of cholesterol levels of Americans conducted in 1960-1962, 1988-1994, 1976-1980, 1988-1994, and 1999-2002. Studies prior to 1999 indicated that overall cholesterol levels were declining. The authors of this paper focused on the changes between the two latest studies, 1988-1994 and 1999-2002 and concluded that for certain populations cholesterol levels decreased over this time. They speculated that “*the increase in the proportion of adults using lipid-lowering medication, particularly in older age groups, likely contributed to the decreases in total and LDL cholesterol levels observed.*” You will confirm the p-values given in the paperandthen calculate the power of the test to detect a prescribed difference in cholesterol levels.

**Part II. RStudio.**

For this part of the lab, you will use the R Studio interface to R to redo most of the first lab of the quarter. It should help you review the basics of probability theory and descriptive statistics.

# *R*eCap of Statistics Lab 1 - Probability and Descriptive Analysis

The purpose of this assignment is to review the basics of probability theory and descriptive statistics. You don't need to memorize this material, but the concept of probability is central to the study of statistics and you should be aware of the fundamentals.

The code shown in this lab is for the R statistical programming language. Although the function names may differ between the statistical packages you have been using (Excel and JMP), the concepts are the same.

# Probability

Probability is the formal study of the laws of chance. The laws of probability are used widely in statistics, but for most of human history the focus on probability theory came out of an interest in gambling.

## The gambler Chevalier DeMere

Games of chance were very popular during the middle ages, and a Renaissance period gambler named Chevalier DeMere was trying to figure out his odds with dice. Specifically, he wondered which was likelier, the odds of rolling at least one six in four throws of a single die or the odds of rolling at least one double six in 24 throws of a pair of dice? He reasoned that likelihood of both rolls was the same (see below), but empirical results did not conform to his expectations. So, he approached the mathematician Blaise Pascal with the question of why.

Pascal reasoned:

* Chance of one six =
* Chance of at least one six in four rolls = = (*wrong*)
* Chance of one double six in one roll =
* Chance of at least one double six in 24 rolls = = (*wrong*)

## 

## The sample space and elementary outcomes

The mathematician Blaise Pascal defined the following:

* A **random experiment** is the process of observing the outcome of a chance event.
* Individual chance events are referred to as a **trials**.
* The **elementary outcomes** are all the possible results of the random experiment.
* The **sample space** is the set of collection of all elementary outcomes.

If a *random experiemnt* consists of flipping a coin many times, the *elementary outcomes* would be our observations from the *sample space* of {heads, tails}. The sample space for a single dice is a little larger: {1, 2, 3, 4, 5, 6}, and the sample space for two die is a larger still: {1, 2, 3, ..., 36}

In a *random experiment* involving elementary outcomes that are equally likely, the probability of each outcome is the same and can be determined as 1/'sample space'. We don't always deal with equally likely outcomes, however. For example, if you had a loaded dice that was weighted to come of up one 25% of the time but all other rolls were equally likely, the probabilities would be: P(1) = 0.25 and P(2 or 3 or 4 or 5 or 6) = 0.75 so that P(2) = 0.75/5 = 0.15 (sincethe sum of the probabilities over the sample space always equals one.

## The binomial distribution

The binomial distribution is a discrete probability distribution that describes the outcomes of n independent trials in a random experiment in which there are only two possible *elementary outcomes* (i.e., sample space = 2). If the probability of a successful trial is p, then the probability of having k successful trials in an experiment of n independent trials is calculated as:

The notation is read n choose x and it defines the all the possible **combinations** of trials that would meet the criteria. For example, if I wanted to write all the possible outcomes of getting one head in three flips, I could start by putting the heads in any one of the n trials (e.g, {h}{t}{t}) but on the second round I'd have n-1 into which I could place the head (e.g., {X}{ }{ }); and on the third round there would be n-2 paces to put it (sounds like a factorial). In fact, we can calculate all the combinations of n choose k using what is known as the *binomial coefficient*:

Let’s use the binomial distribution to calculate the probability of getting four heads in six tosses of a fair coin. Start by putting what we know into the equation:

Now, let's do the math using R.

Start *RStudio* on your computer; you will see a console,

factorial(6)/(factorial(4) \* factorial(6 - 4)) \* (0.5^4) \* (1 - 0.5)^(6 - 4)

## [1] 0.2344

Q1: Interpret this result (i.e., what does it mean?):

Above, you used the bionomial distribution (as defined above using factorials to calculate the probability of getting 4 heads in 6 tosses, give a “fair coin”. In other words, you have calculated the probability of 4 specific events out of 6 trials, where probability of success is .5.

In R, we can write a function to do this calculation given three inputs:   
1) events – x , 2) trials – n, and 3) probability of success – p.

my.p.binom <- function(x, n, p) factorial(n)/(factorial(x) \* factorial(n - x)) \* (p^x) \* (1 - p)^(n - x)

Once you have taught R to do this calculation, you can type “my.p.binom”, R to get the definition of your new R function.

my.p.binom(4, 6, 0.5)

## [1] 0.2344

Luckily for us, statistical programming languages (and even Excel functions) have already been written for this type of calculation.

If in R you type *?dbinom*, you will see the help page for the binomial distribution function. The help page says that *dbinom* gives the probability density, while binomial gives the distribution function, binomial gives the quantile function and rbinom generates random deviates (we'll discuss the later three later).

Let's calculate the probability of getting four heads out of six flips of the fair coin using dbinom.

dbinom(4, 6, 0.5)

## [1] 0.2344

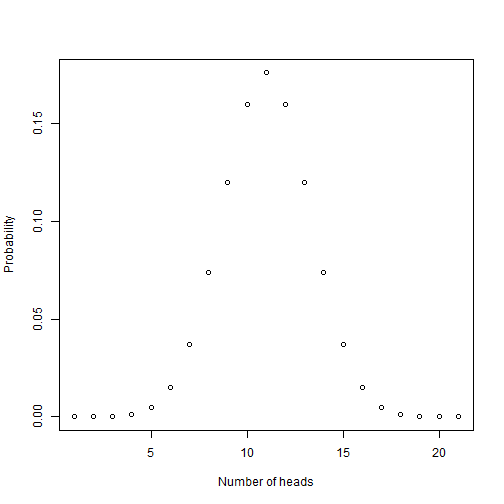
Another nice thing about an object-oriented statistical programming language like R or S[[2]](#footnote-2), is that it lets you combine functions. Let's create a histogram that describes the frequency of every possible outcome (i.e., heads = 0, 1, 2, 3, ..., 20) out of 20 flips).

In R, the primary data type is a vector. We can create a vector containing all the numbers from 0 to 20 by typing 0:20. We can then use this vector as input to the dbinom function. Let's try it.

outcomes <- 0:20 #create a vector of possible trialstrials <- 20 #Define the number of trials (i.e., 20)p.success <- 0.5 # Define the probability of success (head)results <- dbinom(outcomes, trials, p.success) #Record the resultsresults # List the probabilities

## [1] 9.537e-07 1.907e-05 1.812e-04 1.087e-03 4.621e-03 1.479e-02 3.696e-02## [8] 7.393e-02 1.201e-01 1.602e-01 1.762e-01 1.602e-01 1.201e-01 7.393e-02## [15] 3.696e-02 1.479e-02 4.621e-03 1.087e-03 1.812e-04 1.907e-05 9.537e-07

Let's try plotting the results using the standard plot command (see results on the next page).

plot(results, xlab = "Number of heads", ylab = "Probability")

Standard R plot of the probability density.

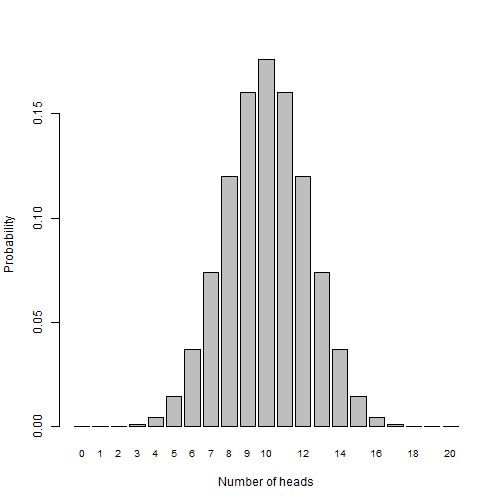
Notice that it is unlikely that you'll get 0 or 20 heads in 20 flips, and that the most probable outcome would be 10. An important note about probability densities is that they must sum to one. Let's see if ours does.

sum(results) #should equal one.

## [1] 1

Just for kicks, let's try plotting the data using a barplot (see results on the next page).

barplot(dbinom(0:20, 20, 0.5), xlab = "Number of heads", ylab = "Probability", names.arg = c(0:20), cex.names = 0.8)



Boxplot of the probability density.

*Note: One problem with command line languages is that you need to know the commands. The first part of the barplot might be self-explanatory, but the second part might not be. The parameter in the function call* names.arg = c(0:20) *gives the name for each of the bars and the parameter* cex.names = 0.8 *makes the bars 80% of their normal size so they all fit.*

## Basic probability operations

In math symbology (actually set theory – see the text), the logical constraint OR is written as and the logical constraint AND as .

### OR (addition)

If we want to know the probability of A OR B, we use the following formula:

This reads: “the probability of (A OR B) occurring is the sum of the probability of A and the probability of B minus the probability of (A and B). For mutually exclusive outcomes (i.e., the probability of two die simultaneously adding up to 3 AND 6) this formula reduces to:

where p is the individual probabilities of A and B.

If you have two die, what is the probability of either the white or black dice showing a 1? Based on the probability rule above, the probability of **either** showing 1 is:

# P(white 1)=6/36, P(black 1)=6/36, P(black and white are 1)=1/36

6/36 + 6/36 + 1/36

## [1] 0.3611

### 

### AND (multiplication)

If we want to know the probability of A AND B occurring, we use:

We'll get to below, but there is a special version of this function when A and B are independent (i.e., B doesn't depend on A), and that is:

What is the probability of rolling snake-eyes. The probability of a one is 1/6 and each roll is independent of the next roll, so the odds of rolling two ones is:

(1/6) \* (1/6)

## [1] 0.02778

### 

### NOT (subtraction)

Since we know that probabilities must range from 0 to 1, we know that:

What is the probability of getting just one six in four rolls of a single dice. This is one where it is helpful to work with the negative. We know that if the probability of getting a single six in one roll of the dice is 1/6, so the probability of NOT getting a six in one roll is:

1 - 1/6

## [1] 0.8333

### 

### Back to DeMures original problem

**What is the probaiblity of getting at least one six in four rolls?**

The 'at least' suggests that we should start with the the probability of not getting a six.

P(NOT\_6) AND P(NOT\_6) AND P(NOT\_6) AND P(NOT\_6) = 5/6^4 based on the AND rule and multiplication, so...

1 - (5/6)^4 #Probability of getting at least one six in four rolls

## [1] 0.5177

**What is the probabilty of getting at least one double six in 24 throws?**

Again, the ‘at least’ suggests that we should focus on the negative (likelihood of not getting a double six in 24 rolls).

1 - (35/36)^24 #Probability of getting at least one double dix in 24 rolls

## [1] 0.4914

### 

### Conditional probability

Above we demonstrated the case where we were interested in the probability of two independent events (A AND B) occurring, but many times we're interested in dependent events. For example, it is one thing to calculate the odds of drawing $\heartsuit 10$ from a standard deck of cards () or but the odds of drawing a flush in a game of poker is clearly not

In the cases where one event affects the probability of the next, we need to calculate conditional probabilities:

For example, the probability of the two die adding to three is zero if you roll a six with the first dice. What if you roll a one with the first dice?

If you roll a one with the first dice, the second dice must be two in order to add to three.

We'll discuss conditional probability in more detail later (e.g, Bayes’ theorem)

The key concepts in this lesson are that probability theory describes the expected outcomes from a large set of random trials, and we can use this theory to determine the likelihood of observing the data we observe given some hypothesis (e.g., we're flipping a fair coin). We'll discuss this in greater detail later as well.

## 

## Exercises (answers included)

Your turn.

Construct the appropriate command in R (using either equations [+,-, etc.] or functions [dbinom, etc.]), paste your command and results to each question

#### Can you use dbinom command in R to calculate the probability of getting 1 head in one flip of a fair coin?

dbinom(1, 1, 0.5) #R function for binomial distribution

## [1] 0.5

*The answer should be obvious, but it should make you feel better about the function you are using.*

*Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

#### 

#### If you had a biased coin that was weighted to come up heads 30% of the time, what would be the probability of getting 4 heads in 6 tosses?

dbinom(4, 6, 0.3)

## [1] 0.05954

#### *Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

#### If you gave a lady two cups of tea, what are the odds of hear picking the English tea 9 out of 10 times by random chance alone?

*Hint, assume that she has a 50:50 chance of picking either cup (P(English)=0.5)*

dbinom(9, 10, 0.5)

## [1] 0.009766

#### *Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

#### If a pitcher plant on average captures 1 fly out out every 10 that lands, what are the odds of you observing it catch 6 flies in a row?

*Note... 6 in a row is out of 6?*

dbinom(6, 6, 0.1)

## [1] 1e-06

#### *Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

#### If you roll a single dice one time, what are the odds that you'll get an even number?

*show your work*

1/6 + 1/6 + 1/6

## [1] 0.5

#### *Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

#### If you roll a single dice 3 times, what is the probability of getting 1 or a 2 on one of the rolls?

*Remember the rules about compounding probabilities. If you can determine the probability getting a 1 in 3 rolls and the probability of getting a 2 in three rolls, you can sum those proabilities to get the probability of getting a 1 OR 2 in three rolls.*

r sum(dbinom(1:2, 3, 1/6))

## [1] 0.4167

*Q: Interpret the result; to do this you will have to say explicitly what you gave the function dbinom as parameters:*

# Descriptive Analysis

After data are collected, they are typically analyzed in some way. Analyzing data helps us make sense of it and can be used to help us convey information to others.

For example, if I went into the Evergreen forest and laid out 100 20x20m plots and I counted the number of unique plant species, I might come back some counts that looked like this.

set.seed(50)

tree.count <- rpois(100, 10)

tree.count

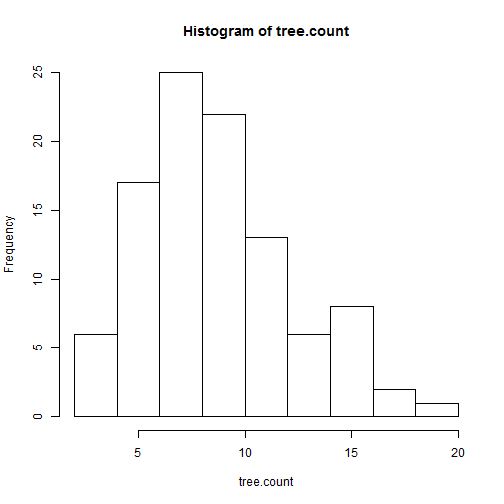
Q: In your own words, say what the commands above do.

## [1] 11 7 4 4 8 5 13 5 7 8 10 8 8 9 7 11 9 13 5 15 11 18 11## [24] 8 11 10 10 7 16 15 16 12 8 12 11 15 2 2 8 15 10 7 7 12 5 8## [47] 9 6 4 11 9 7 12 6 17 8 13 12 8 9 16 12 9 8 6 10 6 9 10## [70] 9 7 13 10 6 8 10 13 6 4 5 6 6 6 9 9 7 6 7 8 10 9 6## [93] 9 5 8 10 19 16 8 13

If I asked you to tell me something about the number of unique plant species in 20x20m plots in the Evergreen forest, what would you say? Is there anything that stands out to you in this data?

How about if we plot it in a frequency histogram (the higher the bar, the more frequently the count was observed)?

hist(tree.count)



plot of chunk histogram

Let's see the frequencies of the tree counts.

table(tree.count) #computes frequencies

## tree.count## 2 4 5 6 7 8 9 10 11 12 13 15 16 17 18 19 ## 2 4 6 11 10 15 12 10 7 6 6 4 4 1 1 1

The numbers on the top row are the individual observation and the bottom is the number of times it was observed (e.g., a count of 6 trees was observed in 11 of the plots).

If I asked you to describe the data, what would you say. How about if I asked for a single number that described the data? Would you choose 8 because it was the most common, or 9 because it appeared to be near the center of mass of the treatments? How about some description of the spread of the data?

Statisticians have been asking themselves these questions for a long time and they have come up with common metrics to help describe data. The metrics can be categorized as measures of central tendency and measures of spread. If you need a review of those metrics, see below; otherwise feel free to skip to “Examples”:

### Measures of Central Tendency

The central tendency provides information on how the values of the data cluster around some single middle value. There are three measures of central tendency that are used in the analysis of data:

* **Mode**: the most frequently observed value of the data - the number that occurs most often. If there is a tie between two equally frequent numbers, the data set is considered bi-modal and both numbers are reported.
* **Median**: the middle value when the data set is ordered in sequential rank (i.e. highest to lowest, or lowest to highest). If there is an even number of values, take the average of the two middle values.
* **Mean**: average value. The mean is the most commonly used measure of central tendency. It is estimated using the sum of all the individual values (y\_i) divided by the total number of individuals in the sample (n):

Interesting note: You will learn later in the term that we're often trying to minimize the sum of the squared errors (i.e., ). As it turns out that the **mean** is the value of that minimizes the squared errors and the **median** is the value of that minimizes the sum of the absolute errors (i.e,. ).

### Measures of spread or dispersion:

Another set of descriptive statistics describes how spread-out the data are.

* **Range** represents the total spread of the data (i.e., |max - min|).
* **Variance** is calculated as the sum of the differences (or deviations) between each individual value and the mean value (remember, the mean is the value that minimizes the squared deviations). The true population variance (2) is, of course, unobtainable, so we'll have to settle for an estimate, the sample variance ().
* **Standard Deviation (SD)** is defined as the square root of the variance ()
* **Standard Error (SE)** is the standard deviation divided by the square root of the sample size (). It is especially important since it is probably the most commonly used estimate of variance around means for figures in scientific papers in biology. Why? One convenience of the SE is that the 95% CI (confidence interval) often (though sometimes not, depending on sample size) = 1.96 times the standard error (SE).\*
* **Coefficient of variation (CV)**. Measures the variability of values in a sample relative to the magnitude of the sample mean (). The coefficient of variation gives an index of population variability that is comparable across measurements and populations, and is often converted to a percent (%) by multiplying this value by 100. A CV of 10% implies that the standard deviation of measurements in the sample is about +/-10% of the sample mean.

## Examples

Given the following numbers representing a set of tree diameter measurements (3, 3, 4, 5, 6, 6, 6, 6, 7, 8, 10), what is the mode, median, mean, sample variance, standard deviation, and standard error.

*Note, this example is calculated in R. With R and Excel, you need to type in function names while JMP is menu driven.*

Enter the data into R.

tree.data <- c(3, 3, 4, 5, 6, 6, 6, 6, 7, 8, 10)

Determine the **mode**. The mode is the most frequent number so lets see what that is.

table(tree.data) #show frequencies

## tree.data## 3 4 5 6 7 8 10 ## 2 1 1 4 1 1 1

Note: mode() is a R function, but it does not calculate the mode. What does it do?

Six is the most frequent number (it occurs 4 times) in the vector and is therefore the mode.

Let's determine the **median**. This is done by sorting the values and finding the 'middle' one which can be tedious, but in this case there are 11 values and they are already sorted so we know that the median is the 6th value from the left (i.e., 6). Luckily, there are functions in R and Excel for doing this and they are both named *median*.

median(tree.data)

## [1] 6

Let's calculate the **mean**

(3 + 3 + 4 + 5 + 6 + 6 + 6 + 6 + 7 + 8 + 10)/11 #by hand

## [1] 5.818

sum(tree.data)/length(tree.data) #Length gives the count of elements in R; in n Excel, use the 'count function).'

## [1] 5.818

mean(tree.data) #In Excel, you use the 'average' function.

## [1] 5.818

How about the **range**?

abs(max(tree.data) - min(tree.data)) #range

## [1] 7

The **variance** is a bit harder because it depends on the mean. We can calculate the deviations first (), we square those, sum the squares, and then divide by n-1.

sum((tree.data - mean(tree.data))^2)/(length(tree.data) - 1) #More or less by hand

## [1] 4.364

var(tree.data) #In Excel, we'd use the 'var.s' function

## [1] 4.364

Once you have the variance, the **standard deviation** is easy.

sqrt(var(tree.data)) #More or less by hand

## [1] 2.089

sd(tree.data) #In Excel, we'd use the 'stdev.s' function.

## [1] 2.089

There is no function in R or Excel for calculating the **standard error**, but we can quickly create one.

std.error <- function(x) sd(x)/sqrt(length(x))std.error(tree.data)

## [1] 0.6298

If you know the standard deviation and mean, the **coefficient of variation** is easy.

sd(tree.data)/abs(mean(tree.data)) \* 100

## [1] 35.9

### Quantiles

Other common measures of dispersion include quantiles (e.g., 25%, 50%, 75%, etc.) and 95% confidence intervals.

* **Quantiles** generally represent breakpoints in the data (sort of like the median). The 0% quantile is the lowest value and the 100% quantile is the maximum. The 25% quantile is the breakpoint between the lowest 25% of the data and the rest; the 50% quantile is the median, and the 75% quantile is the breakpoint at which 75% of the data is less than this value. If the number of data values is odd, then the median value (or 2nd quartile) is the value found at (#values + 1)/2. Similarly, other quantile boundaries are rounded up when there is an issue with fractions.

quantile(tree.data)

## 0% 25% 50% 75% 100% ## 3.0 4.5 6.0 6.5 10.0

## Exercises

Going back to the 1kcs data, let’s say we want to characterize height and DBH for seven stands (of different ages) by species.

Please do the following:

#### Problem

To look at your data, you might want to sort the data by site (in Excel (use Data > Sort > by site); add a level and also sort by species, add a level and also sort by height. What are the populations of interest in this data set?

Since you are interested in the sites as age classes, you might want to add age as a column to the worksheet (or recode “site” to have age prepended). You can do this in Excel using find/replace, or probably once you get the dataset into R. To find ages of each site, look at the second worksheet in the Excel file.

#### Problem 2

Import the data into R. To do this, you will first need to create a .csv file. Then, you can either use read.csv() (see below for an example) or use the Tools->Import Dataset->from Text File (from the R Studio menu).

tkcs <- read.csv(file="wk6\_stats\_lab\_data.csv",head=TRUE,sep=",")

(note I did not use “1kcs”, since variable names in R cannot begin with a number. If you use Import Dataset from the R studio menu, you will probably want to copy your dataset to some variable name that is easy to type, like tkcs, e.g.,

tkcs = wk9\_stats\_lab\_data\_key

Try summary(tkcs) to see if your data was properly imported into R.

names(tkcs) #tells you what columns are contained in the variable tkcs

To refer to one of the columns in your dataset, you can use, for example, tkcs$Height

For each site (and species) (some equations may require you to type them by hand - you will learn more that way, too!), please calculate the range, mode, median, mean, standard deviation, and standard error for Height or DBH.

You will need to figure out how to do this in R; for help, try <http://stackoverflow.com/questions/5052621/how-do-i-filter-a-data-frame-in-r-by-categorical-variable> It hopefully answers the question: Given a data.frame in R with two columns, one numeric and one categorical, how do I extract a portion of the data.frame for usage?

It’s easy to do this with one categorical variable

Try by(tkcs, tkcs$Species\_ID, summary)

It’s harder to do when you want to get statistics by two categorical variables as we do (Site\_ID and Species\_ID). Worst case, you can always load 8 separate CSV files (one for each site) into R…. and then run, e.g., by(sitePL, sitePL$Species\_ID, summary)

A good prize to anyone who figures this out, and contacts me (so I can tell everyone else)!

Q: Please describe what you did to find out how to solve this problem, what your solution was (if any – even if you used a workaround).

#### Problem 3

Create a nice looking table, with your results (you might want to use Excel to do that, but see what you can do in R), and paste into Word):

#### Problem 4

Since the 50% quantile is the median, calculate the 25% quantiles (aka quartiles) to show the spread around the median for height or DBH for one of the species for all sites; use R.

#### Problem 5

Calculate the CV for each site.

#### Problem 6

Based on your summary statistics of central tendency and spread, which site has the tallest trees?

#### Problem 7

Which site is most variable when it comes to height?

#### Problem 8

Try creating a bar graph of heightfor 2 or 3 of the sites you find most interesting - graph Mean Height ± 1 SE on the y-axis and include the label for site on the x-axis. Please paste your graph below.

#### Problem 9

Just for fun, think about any additional statistics (or diagrams/figures) you could generate using R from this dataset that might be helpful in studying the stands where these data were collected? Write down your top three below (no need to actually calculate or draw unless you really want to, just brainstorm).

1. R Tutorial, Clarkson University Dept of Matematics, Potsdam, NY. Cyclismo.org. [↑](#footnote-ref-1)
2. S is the precursor of R…. S-Plus is the current version of S that is most used today. check it out. Compare the two at <http://staff.washington.edu/tlumley/b514/R-fundamentals.pdf> [↑](#footnote-ref-2)