Linear Methods for Data Science Notes

John Michael Epperson

Contents

About

These are my notes for the Fall 2024 session of STAT 6021: Linear Methods for Data Science at the School of Data Science at the University of Virginia.

The course is taught by Jeffrey Woo.

Syllabus Week

0.1 Agenda

- Welcome
- Live Session
- Some logistical tips
- Meet group members
- Q&A on Protocol and Policies
- Q&A on Modules A & B (Review Material)

Chapter 1

Data Wrangling with R

Cheat Sheet

- pipes "%>%" are interpreted as 'and then' in code
  - can be typed or accessed by Ctrl+Alt+M

1.1 Data Wrangling Using Base R Functions

```
Data<-read.csv("datasets/ClassDataPrevious.csv",header = TRUE)
dim(Data)
```

```
## [1] 298   8
```

```
colnames(Data)
```

```
## [1] "Year"     "Sleep"    "Sport"    "Courses"  "Major"    "Age"      "Computer"
## [8] "Lunch"
```

```
Data[1,2]
```

```
## [1] 8
```

```
Data[c(1,3,4),c(1,5,8)]
```

```
##     Year                           Major Lunch
## 1 Second                        Commerce    11
## 3 Second Cognitive science and psychology    10
## 4  First                        Pre-Comm     4
```

To view a column

```
Data$Year
Data[,1]
Data[,-c(2:8)]
```

```
which(Data$Sport=="Soccer")
```

```
##  [1]   3  20  25  26  31  32  33  38  44  46  48  50  51  64  67  71  87  92  98
## [20]  99 118 122 124 126 128 133 136 137 143 146 153 159 165 174 197 198 207 211
## [39] 214 226 234 241 255 259 260 266 274 278 281 283 294 295
```

```
SoccerPeeps<-Data[which(Data$Sport=="Soccer"),]
dim(SoccerPeeps)
```

```
## [1] 52  8
```

```
SoccerPeeps_2nd<-Data[which(Data$Sport=="Soccer" & Data$Year=="Second"),]
dim(SoccerPeeps_2nd)
```

```
## [1] 25  8
```

```
Sleepy<-Data[which(Data$Sleep>8),]
```

1.1.1 Changing Column Names

```
names(Data)[c(1,7)]<-c("Yr","Computer")
```

Find and remove missing data

```
is.na(Data)
```

```
Data[!complete.cases(Data),]
```

```
##         Yr Sleep      Sport Courses                                      Major
## 103 Second    NA Basketball       7 psychology and youth and social innovation
## 206 Second     8       None       4                           Cognitive Science
##     Age Computer Lunch
## 103  19      Mac    10
## 206  19      Mac    NA
```

1.1.2 Summarizing Variables

```
apply(Data[,c(2,4,6,8)],2,mean)
```

```
##     Sleep   Courses       Age     Lunch
##        NA  5.016779 19.573826        NA
```

To not include missing values, use arg na.rm

```
apply(Data[,c(2,4,6,8)],2,mean,na.rm=T)
```

```
##      Sleep    Courses        Age      Lunch
## 155.559259   5.016779  19.573826 156.594175
```

In apply(), the second argument specifies whether to summarize row (put 1) or column (put 2) values Since some of the means are very high, we can use the median instead to be a little more informative.

```
apply(Data[,c(2,4,6,8)],2,median,na.rm=T)
```

```
##   Sleep Courses     Age   Lunch
##     7.5     5.0    19.0     9.0
```

1.1.3 Summarizing variable by groups

use tapply()

```
tapply(Data$Sleep,Data$Yr,median,na.rm=T)
```

```
##  First Fourth Second  Third
##    8.0    7.0    7.5    7.0
```

```
Data$Yr<-factor(Data$Yr,levels=c("First","Second","Third","Fourth"))
levels(Data$Yr)
```

```
## [1] "First"  "Second" "Third"  "Fourth"
```

```
tapply(Data$Sleep,Data$Yr,median,na.rm=T)
```

```
##  First Second  Third Fourth
##    8.0    7.5    7.0    7.0
```

```
tapply(Data$Sleep,list(Data$Yr,Data$Computer),median,na.rm=T)
```

```
##           Mac   PC
## First  NA 8.0 7.50
## Second  7 7.5 7.50
## Third  NA 7.5 7.00
## Fourth NA 7.0 7.25
```

1.1.4 Create a new variable based on existing variable

```
sleep_mins<-Data$Sleep*60
deprived<-ifelse(Data$Sleep<7,"yes","no")
```

Create catagorical variable based on numerical value

```
CourseLoad<-cut(Data$Courses,breaks=c(-Inf,3,5,Inf),labels=c("light","regular","heavy"))
```

Collapse levels into upperclassmen and lowerclassmen

```
levels(Data$Yr)
```

```
## [1] "First"  "Second" "Third"  "Fourth"
```

```
new.levels<-c("und","und","up","up")
Year2<-factor(new.levels[Data$Yr])
levels(Year2)
```

```
## [1] "und" "up"
```

1.1.5 Combine data frames

```
Data<-data.frame(Data,sleep_mins,deprived,CourseLoad,Year2)
head(Data)
```

```
##       Yr Sleep      Sport Courses                           Major Age Computer
## 1 Second     8 Basketball       6                       Commerce  19      Mac
## 2 Second     7     Tennis       5                      Psychology  19      Mac
## 3 Second     8     Soccer       5 Cognitive science and psychology  21      Mac
## 4  First     9 Basketball       5                        Pre-Comm  19      Mac
## 5 Second     4 Basketball       6                      Statistics  19       PC
## 6  Third     7       None       4                      Psychology  20       PC
##   Lunch sleep_mins deprived CourseLoad Year2
## 1    11        480       no      heavy   und
## 2    10        420       no    regular   und
## 3    10        480       no    regular   und
## 4     4        540       no    regular   und
## 5     0        240      yes      heavy   und
## 6    11        420       no    regular    up
```

Can use cbind() alternatively for same result

```
Data2<-cbind(Data,sleep_mins,deprived,CourseLoad,Year2)
```

When combining data frames which have different observations but the same columns, we can merge them using rbind()

```
dat1<-Data[1:3,1:3]
dat3<-Data[6:8,1:3]
res.dat2<-rbind(dat1,dat3)
head(res.dat2)
```

```
##       Yr Sleep      Sport
## 1 Second     8 Basketball
## 2 Second     7     Tennis
## 3 Second     8     Soccer
## 6  Third     7       None
## 7 Second     7 Basketball
## 8  First     7 Basketball
```

**Export data frame to csv

```
write.csv(Data,file="exports/newdata.csv",row.names=FALSE)
```

1.1.6 Sort data frame by column values

to sort in ascending order by age, then descending

```
Data_by_age<-Data[order(Data$Age),]
Data_by_age_des<-Data[order(-Data$Age),]
```

To sort ascending by age then by sleep:

```
Data_by_age_sleep<-Data[order(Data$Age, Data$Sleep),]
```

1.2 Data Wrangling Using dplyr Functions

First we'll clear our environment using rm(list=ls()), then load tidyverse, which contains the dplyr functions:

```
rm(list=ls())
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ----------------- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ---------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become err
```

```
Data<-read.csv("datasets/ClassDataPrevious.csv",header=TRUE)
```

1.2.1 Select Specific Columns of a Data Fram

use select() function, two ways to do it

Or use **Pipes**

Pipes can be typed using either %>% or Ctrl+Shift+M on keyboard. To thing of the operations above, we can read the code as:

1. Take the data frame called Data
2. and then select the column named Year

In this way, we can interpret a pipe as "and then". Commands after pipe should be placed on a new line. Pipes are especially useful for lots of sequential commands.

1.2.2 Select observations by conditions

The **filter()** function allows us to subset our data based on some conditions, for example, to select students whose favorite sport is soccer:

or use a pipe to store results in a new variable "SoccerPeeps"

```
SoccerPeeps<-Data %>%
  filter(Sport=="Soccer")

SoccerPeeps_2nd<-Data %>%
  filter(Sport=="Soccer" & Year=="Second")

Sleepy<-Data %>%
  filter(Sleep>8)

Sleepy_or_Soccer<-Data %>%
  filter(Sport=="Soccer"|Sleep>8)
```

1.2.3 Change Column Name

Changing the names of columns is easy with dplyr, use rename() function

```
Data<-Data %>%
  rename(Yr=Year,Comp=Computer)
```

1.2.4 Summarizing Variables

The summarize() function allows us to summarize a column. Suppose we want to find the mean value of the numeric columns: Sleep, Courses, Age, Lunch:

```
Data %>%
  summarize(mean(Sleep,na.rm=T),mean(Courses),mean(Age),mean(Lunch,na.rm=T))
```

```
##   mean(Sleep, na.rm = T) mean(Courses) mean(Age) mean(Lunch, na.rm = T)
## 1               155.5593      5.016779  19.57383                156.5942
```

This output is cumbersome, but we can give names to each summary:

```
Data %>%
  summarize(avgSleep=mean(Sleep,na.rm = T),avgCourse=mean(Courses,na.rm = T),avgAge=mean(Age,na.rm=T)
```

```
##   avgSleep avgCourse   avgAge   avgLun
## 1 155.5593  5.016779 19.57383 156.5942
```

As previously seen, some of these variables are suspiciously high, we can use the median instead of mean to get more informative results:

```
Data %>%
  summarize(avgSleep=median(Sleep,na.rm = T),avgCourse=median(Courses,na.rm = T),avgAge=median(Age,na
```

```
##   avgSleep avgCourse avgAge avgLun
## 1      7.5         5     19      9
```

1.2.5 Summarizing Variable by Groups

If we want to find the median amount of sleep for 1st, 2nd, 3rd, and 4th years, we can use the 'group_by()' function.

```
Data %>%
  group_by(Yr) %>%
  summarize(medSleep=median(Sleep,na.rm=T))
```

```
## # A tibble: 4 x 2
##   Yr     medSleep
##   <chr>     <dbl>
## 1 First       8
## 2 Fourth      7
## 3 Second      7.5
## 4 Third       7
```

The way we can read the code is: 1. Get the data frame called Data, 2. and then group the observations by Yr, 3. and the find the median amount of sleep by each Yr and store the median in a vector called medSleep

The order of the factor levels is in alphabetical, which isn't very useful. We can use the 'mutate()' function whenever we want to transform or create a new variable. In this case, we are transforming the variable 'Yr' by reordering the factor levels with the 'fct_relevel()' function:

```
Data<-Data %>%
  mutate(Yr<-Yr %>%
           fct_relevel(c("First","Second","Third","Fourth")))
```

which reads: 1. Get data frame called 'Data', 2. and then transform the variable called 'Yr', 3. and then reorder the factor levels

then we use pipes, the 'group_by()', and 'summarize()' functions like before:

```
Data %>%
  group_by(Yr) %>%
  summarize(medSleep=median(Sleep,na.rm=T))
```

```
## # A tibble: 4 x 2
##   Yr     medSleep
##   <chr>     <dbl>
## 1 First       8
## 2 Fourth      7
## 3 Second      7.5
## 4 Third       7
```

This output makes a lot more sense for this context.

To summarize a variable on groups formed by more than one variable, we just add the other variables in the 'group_by()' function:

```
Data %>%
  group_by(Yr,Comp) %>%
  summarize(medSleep=median(Sleep,na.rm=T))
```

```
## `summarise()` has grouped output by 'Yr'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 9 x 3
## # Groups:   Yr [4]
##   Yr     Comp  medSleep
##   <chr>  <chr>    <dbl>
## 1 First  "Mac"     8
## 2 First  "PC"      7.5
## 3 Fourth "Mac"     7
## 4 Fourth "PC"      7.25
## 5 Second ""        7
## 6 Second "Mac"     7.5
## 7 Second "PC"      7.5
## 8 Third  "Mac"     7.5
## 9 Third  "PC"      7
```

1.2.6 Create a New Variable Based on Existing Variable(s)

As mentioned previously, the 'mutate()' function is used to transform a variable or to create a new variable. There are a few variations on this task, based on the type of variable you want to create, and the type of variable it is based on.

Create a numeric variable based on another numerica variable

The variable 'Sleep' is in number of hours. Suppose we need to convert the values of 'Sleep' to number of minutes, we can simply perform the following mathematical operation:

```
Data<-Data %>%
  mutate(Sleep_mins=Sleep*60)
```

and store the transformed variable called 'Sleep_mins' and add 'Sleep_mins' to the data frame called 'Data'.

Create a binary variable based on a numeric variable

Create binary variable called 'deprived'. An observation will obtain a value of 'yes' if they sleep less than 7 hours a night, and 'no' otherwise. We will then add this variable to the data frame called 'Data':

```
Data<-Data %>%
  mutate(deprived=ifelse(Sleep<7,"yes","no"))
```

Create a categorical variable based on a numeric variable

Suppose we want to create a categorical variable based on the number of courses a student takes. We will call this new variable 'CourseLoad', which takes on the following variables

- 'light' if 3 courses or less
- 'regular' if 4 or 5 courses
- 'heavy' if more than 5 courses

and then add 'CourseLoad to the data frame 'Data'. We can use the 'case_when()' function from the 'dplyr' package, instead of the 'cut()' function:

```
Data<-Data %>%
  mutate(CourseLoad=case_when(Courses<=3 ~ "light",
                              Courses>3 & Courses <=5 ~"regular",
                              Courses>5 ~ "heavy"))
```

Note how the category names are suppied after a specific condition is specified.

Collapsing Levels

Sometimes a categorical variable has more levels than needed, so we may want to collapse the levels, as such in this case: collapsing 'Year' into upperclassmen and lowerclassmen.

```
Data<-Data %>%
  mutate(UpUnder=fct_collapse(Yr,under=c("First","Second"),up=c("Third","Fourth")))
```

here we've created a new variable called 'UpUnder', which is done by collapsing 'First' and 'Second' into a new factor called 'under', and collapsing 'Third' and 'Fourth' into a new factor called 'up'. 'UpUnder' is also added to the dataframe 'Data'.

1.2.7 Combine Data Frames

To combine data frames which have different observations but the same columns, we can combine them using 'bind_rows()':

```
dat1<-Data[1:3,1:3]
dat3<-Data[6:8,1:3]
res.dat2<-bind_rows(dat1,dat3)
head(res.dat2)
```

```
##       Yr Sleep      Sport
## 1 Second     8 Basketball
## 2 Second     7     Tennis
## 3 Second     8     Soccer
## 4  Third     7       None
## 5 Second     7 Basketball
## 6  First     7 Basketball
```

'bind_rows()' works the same way as 'rbind()'. Likewise, we can use 'bind_cols()' instead of 'cbind()'.

1.2.8 Sort data frame by column values

To sort your data frame in ascending order by 'Age':

```
Data_by_age<-Data %>%
  arrange(Age)
```

To sort in descending order by 'Age':

```
Data_by_age_des<-Data %>%
  arrange(desc(Age))
```

To sort in ascending order by 'Age' first, then 'Sleep':

```
Data_by_age_sleep<-Data %>%
  arrange(Age,Sleep)
```

1.2.9 More About Combining Datasets

```
rm(list = ls())
#################
##load packages##
#################

library(nycflights13)
library(tidyverse)
```

```
##see dataframes from packages
##View(flights)

##check documentation
##?flights

##View(airlines)
##?airlines

##merge data frames that share one column with the same name
flight_airlines<-flights%>%
  inner_join(airlines,by="carrier")

##merge data frames with multiple shared common columns
##View(weather)
flights_weather<-flights%>%
  inner_join(weather, by=c("year","month","day","hour","origin"))

##merge data frames when columns have different names but same content
##View(airports)
flights_airports<-flights%>%
  inner_join(airports,by=c("dest"="faa"))

##similar function in base R, merge()

flight_airlines2<-merge(flights,airlines, by="carrier")
##View(flight_airlines2)

flights_weather2<-merge(flights,weather, by=c("year","month","day","hour","origin"))
##View(flights_weather2)

##not sure if you notice, the merge() function takes longer to run.
##use proc.time() to time how long your code takes to run

begin<-proc.time()
flight_airlines2<-merge(flights,airlines, by="carrier")
proc.time()-begin
```

```
##    user  system elapsed
##    2.06    0.02    2.10
```

```
begin<-proc.time()
flight_airlines<-flights%>%
  inner_join(airlines,by="carrier")
proc.time()-begin
```

```
##    user  system elapsed
##    0.00    0.05    0.05
```

Basically, use innerjoin rather than merge because it's an order of magnitude faster

1.2.10 A Note About Missing Values

Let's go over some standard missing values that R recognizes and how to handle nonstandard ones that R will not recognize

```
rm(list = ls())
library(tidyverse)
Data<-read.csv("datasets/missing.csv",header=TRUE)
Data
```

```
##   Height Weight
## 1     62    135
## 2     66    190
## 3     70    230
## 4     65    130
## 5     NA    260
## 6    NaN    250
## 7     70
## 8     72     na
## 9     63    N/A
```

Some of these observations are missing values

We can apply is.na() to dataframe to see which entries are viewed as missing:

```
is.na(Data)
```

```
##       Height Weight
##  [1,]  FALSE  FALSE
##  [2,]  FALSE  FALSE
##  [3,]  FALSE  FALSE
##  [4,]  FALSE  FALSE
##  [5,]   TRUE  FALSE
##  [6,]   TRUE  FALSE
##  [7,]  FALSE  FALSE
##  [8,]  FALSE  FALSE
##  [9,]  FALSE  FALSE
```

As we can see, R only recognized entries with NA and NaN as missing. These are the standard values for missing entries. Any other way is not recognized.

Note:

- NaN represents undefined number
- NA represents missing value

We can convert the non standard missing values to 'standard missing values' using the 'replace()' function within the 'mutate()' function:

```
Data<-Data %>%
  mutate(Weight = replace(Weight, Weight == "na", NA))%>%
  mutate(Weight = replace(Weight, Weight == "N/A", NA))%>%
  mutate(Weight = replace(Weight, Weight == "", NA))

is.na(Data)
```

```
##       Height Weight
##  [1,]  FALSE  FALSE
##  [2,]  FALSE  FALSE
##  [3,]  FALSE  FALSE
##  [4,]  FALSE  FALSE
##  [5,]   TRUE  FALSE
##  [6,]   TRUE  FALSE
##  [7,]  FALSE   TRUE
##  [8,]  FALSE   TRUE
##  [9,]  FALSE   TRUE
```

And just like that, the rest of the missing values are recognized as missing as they should be.

Module 1 Guided Question Set

```
students<-read.table("datasets/students.txt",header=TRUE)
```

1. Student ID will likely not tell us anything of interest. I will now remove it.

```
students<-students[-1]
```

2. Students in dataset:

```
nrow(students)
```

```
## [1] 249
```

3. Students with missing entry in at least one column?

```
nrow(students[!complete.cases(students),])
```

```
## [1] 12
```

4. Median values of numeric variables

```
students %>%
  summarize(medGPA=median(GPA,na.rm=T),medParty=median(PartyNum,na.rm=T),medBeer=median(DaysBeer,na.r
```

```
##   medGPA medParty medBeer medStudy
## 1    3.2        8       8       14
```

5. Mean and stddev of Study hours for female and male students.

```
tapply(students$StudyHrs,students$Gender,mean,na.rm=T)
```

```
##   female     male
## 15.40690 14.70192
```

```
tapply(students$StudyHrs,students$Gender,sd,na.rm=T)
```

```
##    female      male
##  8.972564 10.198877
```

From this we can see the mean Study hours for females is slightly higher than for males in this sample. However, the standard deviation shows that this is likely not statistically significant.

6. Time to party...

```
students<-students %>%
  mutate(PartyAnimal=ifelse(PartyNum>8,"yes","no"))
```

7. New catagorical variable GPA.cat where

- 'low' if GPA < 3.0
- 'moderate' if $3.0 <$ GPA < 3.5
- 'high' if GPA > 3.5

```
students<-students %>%
  mutate(GPA.cat=case_when(GPA<3.0~"low",
                           GPA>=3.0 & GPA<3.5~"moderate",
                           GPA>=3.5~"high"))

table(students$GPA.cat)
```

```
##
##     high      low moderate
##       70       87       85
```

8. Create data frame containing students with low GPA (<3.0), party more than 8 days a month, and study little (<15 hours a week). How many students fit these criteria?

```
slackers<-students %>%
  filter(GPA<3.0&PartyNum>8&StudyHrs<15)
nrow(slackers)
```

```
## [1] 29
```

9. Add 'PartyAnimal' and 'GPA.cat' to 'students' and export as .csv file.

These variables were added in the mutate function in their respective questions.

```
write.csv(students,"datasets/new_students.csv")
```

Module 1 Homework

Name: John Michael Epperson
Course: STAT 6021 | Fall 2024
Professor: Jeffrey Woo, PhD
Date: 05-Sep-2024

1. County level data

a. create dataset 'latest' that has:

- only rows pertaining to data from June 3 2021
- remove rows pertaining to "Unknown" counties
- remove columns 'date' and 'fips'
- order by county then state alphabetically

```
latest<-data %>%
  filter(date=='2021-06-03' & county!="Unknown") %>%
  mutate(date=NULL,fips=NULL) %>%
  arrange(county,state)


head(latest)
```

```
##      county          state cases deaths
## 1 Abbeville South Carolina  2599     41
## 2    Acadia      Louisiana  6703    195
## 3  Accomack       Virginia  2862     43
## 4       Ada          Idaho 52964    475
## 5     Adair           Iowa   873     32
## 6     Adair       Kentucky  1944     54
```

b. Calculate case fatality rate and store as 'death.rate'

```
latest<-latest %>%
  mutate(death.rate=round((deaths/cases)*100,2))

head(latest)
```

```
##      county          state cases deaths death.rate
```

```
## 1 Abbeville South Carolina  2599     41       1.58
## 2    Acadia      Louisiana  6703    195       2.91
## 3  Accomack       Virginia  2862     43       1.50
## 4       Ada          Idaho 52964    475       0.90
## 5     Adair           Iowa   873     32       3.67
## 6     Adair       Kentucky  1944     54       2.78
```

c. Top 10 Counties for Number of Cases

```
latest %>%
  arrange(desc(cases)) %>%
  head(10)
```

```
##            county      state   cases deaths death.rate
## 1     Los Angeles California 1245127  24375       1.96
## 2   New York City   New York  949986  33257       3.50
## 3            Cook   Illinois  554390  10893       1.96
## 4        Maricopa    Arizona  551509  10084       1.83
## 5      Miami-Dade    Florida  501925   6472       1.29
## 6          Harris      Texas  401345   6462       1.61
## 7          Dallas      Texas  303533   4082       1.34
## 8       Riverside California  300879   4614       1.53
## 9  San Bernardino California  298599   4760       1.59
## 10      San Diego California  280410   3760       1.34
```

d. Top 10 Counties for Number of Deaths

```
latest %>%
  arrange(desc(deaths)) %>%
  head(10)
```

```
##            county      state   cases deaths death.rate
## 1   New York City   New York  949986  33257       3.50
## 2     Los Angeles California 1245127  24375       1.96
## 3            Cook   Illinois  554390  10893       1.96
## 4        Maricopa    Arizona  551509  10084       1.83
## 5      Miami-Dade    Florida  501925   6472       1.29
## 6          Harris      Texas  401345   6462       1.61
## 7          Orange California  272242   5070       1.86
## 8           Wayne   Michigan  164612   5048       3.07
## 9  San Bernardino California  298599   4760       1.59
## 10      Riverside California  300879   4614       1.53
```

e. Top 10 Counties for Case Fatality Rate

```
latest %>%
  arrange(desc(death.rate)) %>%
  head(10)
```

```
##          county      state cases deaths death.rate
## 1         Grant   Nebraska    41      4       9.76
## 2        Sabine      Texas   524     45       8.59
## 3       Harding New Mexico    12      1       8.33
## 4     Petroleum    Montana    12      1       8.33
## 5         Foard      Texas   124     10       8.06
## 6       Hancock    Georgia   928     68       7.33
## 7      Glascock    Georgia   269     19       7.06
## 8        Motley      Texas   116      8       6.90
## 9       Candler    Georgia   978     67       6.85
## 10 Throckmorton      Texas    73      5       6.85
```

- I notice that the counties with the highest case fatality rates all had relatively few cases (<1000).

f. Top 10 Counties for Case Fatality Rate (>100,000 Cases)

```
latest %>%
  filter(cases>=100000) %>%
  arrange(desc(death.rate)) %>%
  head(10)
```

```
##           county         state  cases deaths death.rate
## 1  New York City      New York 949986  33257       3.50
## 2          Wayne      Michigan 164612   5048       3.07
## 3      Middlesex Massachusetts 134980   3761       2.79
## 4         Bergen    New Jersey 104301   2868       2.75
## 5         Macomb      Michigan 100190   2441       2.44
## 6   Philadelphia  Pennsylvania 153521   3692       2.40
## 7      St. Louis      Missouri 100195   2249       2.24
## 8      Fairfield   Connecticut 100093   2198       2.20
## 9           Pima       Arizona 116997   2406       2.06
## 10       Oakland      Michigan 118035   2368       2.01
```

g. Cases for Albemarle & Charlottesville City

```
latest %>%
  filter(state=="Virginia" & (county=="Albemarle" | county=="Charlottesville city"))
```

```
##                 county    state cases deaths death.rate
## 1            Albemarle Virginia  5801     83       1.43
## 2 Charlottesville city Virginia  4014     57       1.42
```

2. State level data

a. create data frame 'state.level' that:

- has 55 rows, 1 for each state, DC, and territory
- 3 columns: name of state, case count, death count
- ordered alphabetically by state
- only has data from June 3 2021

```
state.level<-data %>%
  filter(date=="2021-06-03") %>%
  mutate(fips=NULL,date=NULL,county=NULL) %>%
  group_by(state) %>%
  summarise(cases=sum(cases),deaths=sum(deaths)) %>%
  arrange(state)

head(state.level)
```

```
## # A tibble: 6 x 3
##   state        cases deaths
##   <chr>        <int>  <int>
## 1 Alabama     545028  11188
## 2 Alaska       69826    352
## 3 Arizona     882691  17653
## 4 Arkansas    341889   5842
## 5 California 3793055  63345
## 6 Colorado    547961   6746
```

b. Calculate case fatality rate 'state.rate' for each state.

```
state.level<-state.level %>%
  mutate(state.rate=round((deaths/cases)*100,2))

head(state.level)
```

```
## # A tibble: 6 x 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Alabama     545028  11188       2.05
## 2 Alaska       69826    352       0.5
## 3 Arizona     882691  17653       2
## 4 Arkansas    341889   5842       1.71
## 5 California 3793055  63345       1.67
## 6 Colorado    547961   6746       1.23
```

c. Case fatality rate in Virginia?

```
state.level %>%
  filter(state=="Virginia")
```

```
## # A tibble: 1 x 4
##   state     cases deaths state.rate
##   <chr>     <int>  <int>      <dbl>
## 1 Virginia 676041  11216       1.66
```

- The case fatality rate for Virginia on June 3 2021 was 1.66%.

d. Case fatality rate in Puerto Rico?

```
state.level %>%
  filter(state=="Puerto Rico")
```

```
## # A tibble: 1 x 4
##   state        cases deaths state.rate
##   <chr>        <int>  <int>      <dbl>
## 1 Puerto Rico 172414     NA         NA
```

- We don't have any information on the case fatality rate on June 3 2021 in Puerto Rico.

e. Which states have the 10 highest case fatality rates?

```
state.level %>%
  arrange(desc(state.rate)) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    state                 cases deaths state.rate
##    <chr>                 <int>  <int>      <dbl>
##  1 New Jersey           1017044  26253       2.58
##  2 Massachusetts         707523  17893       2.53
##  3 New York             2102003  52811       2.51
##  4 Connecticut           347748   8245       2.37
##  5 District of Columbia   49041   1136       2.32
##  6 Mississippi           318048   7324       2.3
##  7 Pennsylvania         1208879  27349       2.26
##  8 Louisiana             472617  10605       2.24
##  9 New Mexico            203330   4275       2.1
## 10 Maryland              460406   9626       2.09
```

- New Jersey, Massachusetts, New York, Connecticut, District of Columbia, Mississipi, Pennsylvania, Louisiana, New Mexico, and Maryland - in descending order - had the highest case fatality rates on June 3 2021.

f. Which states had the 10 lowest case fatality rates?

```
state.level %>%
  arrange(state.rate) %>%
  head(10)
```

```
## # A tibble: 10 x 4
##    state                     cases deaths state.rate
##    <chr>                     <int>  <int>      <dbl>
##  1 Alaska                    69826    352       0.5
##  2 Utah                     406895   2308       0.57
##  3 Virgin Islands             3512     28       0.8
##  4 Vermont                   24240    255       1.05
##  5 Nebraska                 223517   2385       1.07
##  6 Idaho                    192704   2103       1.09
##  7 Northern Mariana Islands    183      2       1.09
```

```
##  8 Wisconsin                675152   7923       1.17
##  9 Wyoming                   60543    720       1.19
## 10 Colorado                 547961   6746       1.23
```

- In order from lowest case fatality rate to highest: Alaska, Utah, Virgin Islands, Vermont, Nebraska, Idaho, Northern Mariana Islands, Wisconsin, Wyoming, and Colorado had the lowest case fatality rates on June 3 2021.

g. Export dataset to file called *stateCovid.csv*

```
write.csv(state.level,"datasets/stateCovid.csv")
```

Chapter 2

Data Visualization with R

This is where Module 2 notes will go!

Appendix A

Review of Statistical Inference

Cheat Sheet

Central Limit Theorem: Tells us that with a large enough sample size, we can use the normal distribution to find probabilities associated with sample means

Confidence Intervals

- Given by $\bar{x} \pm t_{1-\alpha,k} \frac{s}{\sqrt{n}}$
- $t_{1-\alpha,k}$ is given by the R function qt(percentile,df)

A.1 Introduction to the Lesson

In many statistical studies or experiments, we want to get answers to questions regarding a population of interest. For example, what is the average annual income of American adults? In this example, the population of interest is American adults. Ideally, we would like to obtain the data from every single American adult. However, due to constraints such as time and money, we are unable to obtain the data from every person who makes up the population. We then typically collect data from a random sample of American adults. A sample is ideally a subset and is representative of the population. We then collect data from the sample, and then use the characteristics of the sample, called statistics, to make an inference about the characteristics of the population, called parameters.

Consider the sample mean annual income among 500 American adults is \$52,000. Does this mean the population mean annual income among all American adults is \$52,000? Probably not. The sample mean, even if it comes from a representative and large sample, is probably close to the population mean, but unlikely to be exactly equal to the population mean. This uncertainty is simply due to the variability associated with the sample mean. Another random sample of 500 American adults may result in a sample mean with a different value, for example, \$51,000. This is where statistical inference comes in. Statistical inference allows us to quantify the variability associated with statistics and allows us to make inferences about the parameter. The main inferential methods we will use are confidence intervals and hypothesis tests.

A.2 Sampling Distributions

A probability density function (pdf) is a mathematical representation of the distribution of data and must

- be non-negative, and
- integrate to 1

Common Probability Density Functions

- Normal Distribution
  - The one we'll mostly look at
- t distribution
- χ^2 distribution
- F distribution

A.2.1 Normal Distribution

A normal distribution is a symmetric, bell shaped-distribution. A normal distribution with a mean μ and standard deviation σ is denoted by $N(\mu, \sigma)$. Its pdf is

$$f(x) = \frac{1}{(\sigma\sqrt{2\pi}e^{(\frac{1}{2}(\frac{x-\mu}{\sigma})^2)}} \tag{A.1}$$

If (A.1) is a good approximation for the distribution of data, we can estimate probabilities by integrating (A.1) over the relevant range(s).

- A normal distribution with mean 0 and standard deviation 1 is called a **standard normal distribution**.
- It turns out that any normal distribution X with mean and standard deviation can be standardized by:

$$Z = \frac{X - \mu}{\sigma} \tag{A.2}$$

- Then Z follows a standard normal distribution.
  - This (A.2) is also called the *Z-score*.

A.2.2 Population & Samples

A.2.2.1 Motivation

In many studies, we want to get answers to questions regarding a population of interest. For example, what is the average income of American adults? - Ideally, we would like to obtain the data from every single American adult - however, due to constraints (e.g. time and money), we are unable to obtain the data from every single American adult - We then typically collect data from a random sample of American adults - We then use characteristics of the sample to estimate the characteristics of the population

The above is the basic way we conduct statistical analysis.

- Population: The group of all items in our study.
- Sample: The items from which we actually collect data.

A.2.2.2 Example

A manufacturing company produces 5 million parts. To estimate the proportion of parts that are defective, 300 parts are randomly selected and carefully inspected for defects. What is the: - population of interest? - All 5 million parts - sample? - the 300 randomly selected parts

A.2.2.3 Parameter vs. Statistic

- A **parameter** is a number describing a characteristic of the population. Parameters are fixed values, but in practice we do not know their numerical values
- A **statistic** is a number describing a characteristic of a sample. Statistics vary from sample to sample.

We often use a statistic to estimate an unknown parameter.

One could take many sample groups together to estimate the population characteristics.

Each time we take a random sample from a population, we are likely to get a different set of individuals and calculate a different statistic. There is **variability** in the statistics.

Question: Can we quantify this variability without having to obtain many different random samples?

- Yes, we can take lots of random samples of the same size from a given population, the distribution of the sample statistics, **the sampling distribution**, will follow a predictable shape.
- Under some circumstances, the sampling distribution can be well-approximated by a specific distribution and its pdf.
- The variance of the statistics generally decreases as the sample size increases.
  - $variance = \sigma^2$

A.2.3 Sampling Distribution of Sample Means

When a continuous variable, X, in a population follows a $N(\ ,\ )$ distribution, the sampling distribution of the sample mean, $\bar{x}$, for all possible samples of size n is $N(\ ,\frac{\sigma}{\sqrt{n}})$.

- 1st random sample of size 50, $\bar{x} = 63.7$
- 2nd is $\bar{x} = 64.3$
- 3rd is $\bar{x} = 65.8$

Mean is the same, but standard deviation is reduced by $\sqrt{(n)}$

A.2.3.1 Central Limit Theorem

Consider a quantitative variable, X, in a population that has mean μ and standard deviation σ, and is not necessarily normally distributed. If n is **large enough**, the sampling distribution of the sample mean, $\bar{x}$, for all possible samples of size n is approximately $N(\mu, \frac{\sigma}{\sqrt{n}})$.

This is known as the **Central Limit Theorem**.

Implication: With a large enough sample size, we can use the normal distribution to find probabilities associated with sample means.

- **Large enough** is relative, many say 25 or 30, but 24 is still better than 23, ya dig?

A.2.4 Worked Example: Textbook Spending

Question: Based on data from Spring 2017 semester, the mean amount spent on textbooks for the semester is \$405.17 with standard deviation \$210.59. The histogram for the variable amount spent on textbooks that semester is displayed below. How would you describe the shape of this histogram?

The distribution is right skewed, can't use a normal distribution to describe this.

Question: Suppose we take repeated samples of size 25. What do we expect the sampling distribution for the sample mean to be? How about if we take repeated samples of size 50?

$n = 25, \bar{x}_n = 25\ N(\$405.17, \frac{210.59}{\sqrt{25}}) = \$42.118$

$n = 50, \bar{x}_n = 25\ N(\$405.17, \frac{210.59}{\sqrt{50}}) = \$29.782$

Question: Suppose I have a random sample of 25 students. What is the probability that the sample mean is less than \$415?

```
n<-25
sigma<-210.59
mu<-405.17
guess<-415

sample_sig<-sigma/sqrt(n)
zscore<-(guess-mu)/sample_sig
pnorm(zscore)
```

```
## [1] 0.5922714
```

What if I have a random sample of 50 students instead?

```
n<-50

sample_sig<-sigma/sqrt(n)
zscore<-(guess-mu)/sample_sig
pnorm(zscore)
```

```
## [1] 0.6293249
```

So, the probability that the sample mean is less than $415 increases with more samples.

A.2.4.1 Practice Exercise

Question: Suppose I have a random sample of 50 students. What is the probability that the sample mean is more than $400?

```
n<-50
guess<-400

sample_sig<-sigma/sqrt(n)
zscore<-(guess-mu)/sample_sig
1-pnorm(zscore)
```

```
## [1] 0.5689082
```

A.2.4.2 Where do we go from here?

- We know that the sample mean, $\bar{x}$, describes our particular sample. However, if we select another random sample, the sample mean will probably be different.
- We do know that with a large enough sample size, the distribution of the sample means can be approximated by a normal distribution.
- We also know that with larger sample size, the sample means will be closer to the population mean, on average.

Reality: we will not know the value of the population mean, . So how do we use the sample mean, $\bar{x}$, to estimate ?

This brings us to Confidence Intervals…

A.3 Confidence Intervals

A.3.1 Intro to Confidence Intervals

Goals of Confidence Intervals - Provide and estimate for the unknown parameter of interest - Provide a **range of plausible values** for the unknown paramter of interest - Provide a measure of **uncertainty**

A.3.1.1 General Form of Confidence Intervals

Confidence intervals generally take the following form:

$$\text{Estimate} \pm \text{margin of error.} \tag{A.3}$$

The **margin of error** reflects how precise we believe our estimate is, and is calculated using the confidence level $C = 1 - \alpha$.

$C = 0.95$ is considered the standard.

A.3.1.2 Confidence Levels and Margin of Error

- **Confidence Level**: If we obtain many random samples of the same sample size n, and construct a confidence interval with $C\%$ confidence level based on each sample, $C\%$ of samples will have a confidence interval that contains the population mean .
- **Margin of Error**: Suppose we obtain many random samples of the same size n, and construct a confidence interval with $C\%$ confidence level based on each sample. The difference between the sample mean and population mean in $C\%$ of samples will be no greater than the value of the margin of error.

To illustrate these concepts, consider samples of number of hours of sleep for college students, with margin of error = 0.2:

- 1^{st} sample: $\bar{x_1} = 5.3$, CI = (5.1,5.5)
- 2^{nd} sample: $\bar{x_1} = 5.5$, CI = (5.3,5.7)
- 3^{rd} sample: $\bar{x_1} = 5.2$, CI = (5.0,5.4)
- and many more…

Out of all these CIs, 95% of these samples contain the true population mean.

Difference between actual mean, say 5.47, and the sample mean will be no greater than margin of error in 95% of samples.

A.3.1.3 Confidence Interval for Population Mean

The confidence interval for population mean is given by:

$$\bar{x} \pm z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \tag{A.4}$$

- $z_{1-\alpha/2}$ denotes the value of the standard normal distribution that corresponds to the $(1 - \frac{\alpha}{2})^{\text{th}}$ percentile. In a confidence interval, this is also called a **multiplier**.

- Generally speaking, the margin of error can be viewed as multiplier ×standard deviation of estimate.

A.3.2 Finding Multipliers

A.3.2.1 Finding Multiplier in CI

Recall from (A.4), $z_{1-\alpha/2}$ denotes the value of the standard normal distribution that corresponds to the $(1-\frac{\alpha}{2})^{\text{th}}$ percentile.

We want a CI at 1- confidence. is typically 0.05.

Since we know that $\bar{x} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$, we can rewrite the equation for *Z-score*, (A.2), as:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \tag{A.5}$$

Now we can apply these facts to find out why equation is the way it is.

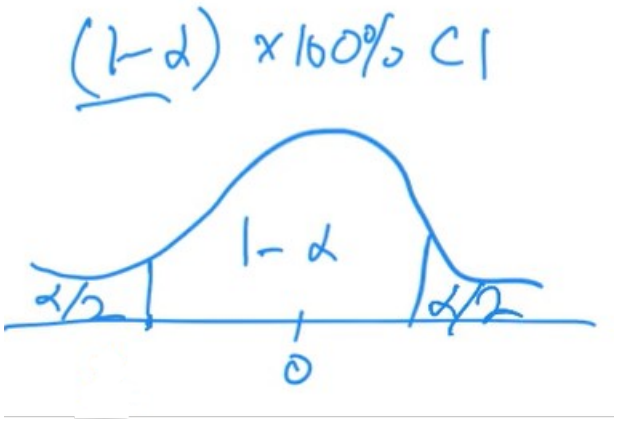

Figure A.1: Standard Normal PDF

Looking at Figure A.1, we can see that the section in the middle corresponds to the $1-\alpha$ percentile, that's our 95% CI.

The sections on the left and right must be equal to $\alpha/2$ since for a standard normal pdf, the whole thing must add up to exactly 1.

These two $\alpha/2$ sections on the left and right of Figure A.1 refer to the $Z_{\alpha/2}$ and $Z_{1-\alpha/2}$ percentiles, respectively.

Note: Due to symmetry, the two /2 z-scores will be the same magnitude, i.e. $-Z_{\alpha/2} = Z_{1-\alpha/2}$. We can use this fact to our advantage.

One could say that the probability we are in the middle area of Figure A.1 is:

$$P(Z_{\alpha/2} \leq Z \leq Z_{1-\alpha/2}) = 1-\alpha \tag{A.6}$$

Equation (A.5) implies that:

$$P(Z_{\alpha/2} \leq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \leq Z_{1-\alpha/2}) = 1 - \alpha \tag{A.7}$$

To isolate μ in the above equation, we can substitute $Z_{\alpha/2}$ with $-Z_{1-\alpha/2}$ from earlier, then multiply by $-\sigma/\sqrt{n}$, and finally add $\bar{x}$ to get:

$$P(\bar{x} + Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha \tag{A.8}$$

So this means that the population mean must be within the interval given by $\bar{x} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$, which is the definition of our confidence interval.

So how do we get the numerical value of the Z-score?

Type qnorm(percentile) in R to find z_percentile.

So, if $1 - \alpha = 0.95$, then $1 - \alpha/2 = 0.975$, then to R we go…

```
qnorm(0.975)
```

```
## [1] 1.959964
```

≈ 2

Voila!

pnorm is the inverse of qnorm:

pnorm(z-score) = percentile qnorm(percentile) = z-score

A.3.2.2 Exercise

1. Find the z multiplier at 90% confidence

for 90% confidence, $1 - \alpha = 0.90$, so $1 - \alpha/2 = 0.95$, so plugging into R

```
qnorm(0.95)
```

```
## [1] 1.644854
```

2. Find the z multiplier at 98% confidence

for 98% confidence $1 - \alpha = 0.98$, so $1 - \alpha/2 = 0.99$

```
qnorm(0.99)
```

```
## [1] 2.326348
```

3. Find the z multiplier at 99% confidence

for 99% confidence $1 - \alpha = 0.99$, so $1 - \alpha/2 = 0.995$

```
qnorm(0.995)
```

```
## [1] 2.575829
```

Question: Do you notice a trend in the z multiplier as confidence level increases? Does this make sense?

- The multiplier increases as the confidence level increases. This makes sense because as the confidence level increases, the range that $1-\alpha$ encompasses must expand, meaning the z multiplier must also increase in kind. In other words, we're more confident our sample mean contains the population mean because we increased our margin of error.

Looking back at Equation (A.8), is anything strange?

$\bar{x} \pm Z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$

σ represents population variance, which is rarely known! So how can we apply this formula?

Using sample variance is the right direction.

A.3.3 t distributions

Recall that the population variance is

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} \tag{A.9}$$

and the sample variance is

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} \tag{A.10}$$

When σ is **unknown**, we use the sample standard deviation, s, to estimate σ.

- Previously we computed the standard deviation of sample mean, $sd(\bar{x})$, as $\frac{\sigma}{\sqrt{n}}$.
- When σ is unknown, we compute the **standard error** of the sample mean: $se(\bar{x}) = \frac{s}{\sqrt{n}}$.

When the standard deviation of a statistic is estimated from the data, the result is the **standard error of the statistic**.

In reality, we'll calculate the standard error much more frequently.

Scenario: A random sample of size n is drawn from $N(\mu, \sigma)$.

- When σ is known, $\bar{x} \approx N(\mu, \frac{\sigma}{\sqrt{n}})$, and so $Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \approx N(0,1)$
- When σ is known and estimated using s, the sampling distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ is approximated by a t **distribution with degrees of freedom** n-1.
- If we do not have a normal proportion, the approximation to the t distribution works well if we have a large enough sample size.

A.3.3.1 Degrees of Freedom

- t distributions are specified by their **degrees of freedom**.
- We specify t distributions using t_k, where k is the degrees of freedom.

So CI for μ, df: $k = n - 1$ since we lose one degree of freedom because the equality $\frac{\sum x_i}{n} = \bar{x}$ must hold.

A.3.3.2 t Distribution vs Standard Normal

Both distributions are centered at 0, symmetric, and bell-shaped. Their differences are: - t_k has associated degrees of freedom. - t_k has slightly **larger spread**. As the sample size (degrees of freedom) increases, t_k approaches the standard normal.

t distribution has more area at extremes or tails of pdf.

A.3.3.3 Confidence Interval for Population Mean

We use s to estimate σ when it is unknown. THe level C CI for a population mean becomes:

$$\bar{x} \pm t_{1-\alpha/2,k} \frac{s}{\sqrt{n}} \tag{A.11}$$

Where $t_{1-\alpha/2,k}$ is the value from the t_k curve with area C between $t_{\alpha/2,k}$ and $t_{1-\alpha/2,k}$. The degrees of freedom is $k = n - 1$.

A.3.3.4 Finding Multiplier

In R, type qt(percentile, df) to find *t_{percentile,df}*. 1. Find the t multiplier at 90% confidence with 10df

```
qt(.95,10)
```

```
## [1] 1.812461
```

2. Find the t multiplier at 92% confidence with 35df

```
qt(.96,35)
```

```
## [1] 1.803024
```

3. Find the t multiplier at 98% confidence with 50df

```
qt(.99,50)
```

```
## [1] 2.403272
```

A.3.3.5 Worked Example: Banks' Loan-to-Deposit Ratio (LTDR)

Question: The sample mean LTDR for 110 randomly selected American banks is 76.7 and the sample standard deviation is 12.3. Compute a 95% CI for the population mean LTDR. Based on this CI, is it reasonable to say that the average LTDR is less than 80 for the population?

So, stating our variables:

```
n<-110
xbar<-76.7
s<-12.3
```

Using Equation (A.11):

$$\bar{x} \pm t_{1-\alpha/2,k} \frac{s}{\sqrt{n}} \tag{A.12}$$

and plugging in our variables we get:

$$76.7 \pm t_{1-\alpha/2,k} \frac{12.3}{\sqrt{110}} \tag{A.13}$$

and $t_{1-\alpha/2,k}$ is found using qt(percentile,df), where percentile is $0.95 + \frac{1-0.95}{2} = 0.975$ and $df = n-1 = 110-1 = 109$:

```
qt(0.975,109)
```

```
## [1] 1.981967
```

So plugging this back into Equation (A.11) we get margin of error equal to:

```
qt(0.975,n-1)*s/sqrt(n)
```

```
## [1] 2.32437
```

So our 95% CI is given by:

76.7 ± 2.3 or $(74.4, 79.0)$

So, it is reasonable to say that the average LTDR is less than 80 for the population. In other words we're 95% sure that the average LTDR for the population is less than 80.

Or say:

Yes, since the entire CI is less than 80.

A.4 Hypothesis Testing

A.4.1 Hypotheses

A.4.1.1 Motivation

The general approach to hypothesis testing is the following: we perform probability calculations to distinguish patterns seen in data between those that are due to **chance** and those that **reflect a real feature** of the phenomenon under study.

A.4.1.2 Example

You are in charge of quality control in your food company. You randomly sample 40 apcks of cherry tomatoes, each labeled 1/2 lb. (227 g), and find their average weight is 226.5 g. Obviously, we cannot expect boxes filled with whole tomatoes to all weight exactly half a pound. Thus, - is the weight in our sample due to chance? - is the weight in our sample evidence the machine that sorts the tomatoes needs revision?

A.4.1.3 Stating Hypotheses

Hypothesis testing uses sample data to decide on the validity of a hypothesis. A **hypothesis** is an assumption or a theory about the characteristics of one or more variables in one or more populations. - What you want to know: does the calibrating machine that sorts cherry tomatoes into packs need revision? - The same question reframed statistically: Is the population mean μ for the distribution of weights of cherry tomato packages different from 227 g (i.e., half a pound)?

The statement being tested in a test of significance is called the **null hypothesis**, H_0. The test of significance is designed to assess the strength of the evidence against the null hypothesis. The null hypothesis is usually a statement of "no effect" or "no difference."

The **alternative hypothesis**, H_a is the statement we suspect is true instead of the null hypothesis.

- $H_0 : \mu = 227g$
- $H_a : \mu \neq 227g$

A.4.1.4 One-sided and Two-sided Tests

- A two-sided test of the population mean has the following hypotheses
  - $H_0 : \mu =$ specific number (μ_0)
  - $H_a : \mu \neq$ specific number (μ_0)
- A one-sided test of the population mean has the following hypotheses
  - $H_0 : \mu =$ specific number (μ_0)
  - $H_a : \mu <$ specific number (μ_0)

  OR
  - $H_0 : \mu =$ specific number (μ_0)
  - $H_a : \mu >$ specific number (μ_0)

What determines the choice of a one-sided versus a two-sided test is what we know about the problem **before** we perform a test of statistical significance.

Question: You are in charge of quality control in your food company. You randomly sample 40 apcks of cherry tomatoes, each labeled 1/2 lb. (227 g), and find their average weight is 226.5 g. A consumer advocacy group is trying to claim that consumers are being cheated by the food company. What should the null and alternative hypotheses be in this scenario?

- $H_0 : \mu = 227g$
- $H_a : \mu < 227g$

Consumers would only be cheated if the weight of tomatoes they got was **less than** the amount on the package.

A.4.2 Evaluating Evidence: p-values

Next, we evaluate the evidence our data provides **against** the null hypothesis. This evaluation is done by - assuming the null hypothesis is true - computing a test statistic to measure how dissimilar our sample is with the null hypothesis - comparing our test statistic with a benchmark to decide if we have enough evidence against the null hypothesis We then end by making a relevant conclusion

A.4.2.1 Test Statistics

- The test statistic measures how dissimilar our sampled data is with the null hypothesis.
- In a hypothesis test for a mean the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \tag{A.14}$$

  where μ_0 represents the value in the null hypothesis H_0
- The larger (in magnitude) the test statistic, the more evidence we have against the null hypothesis.

A.4.2.2 Statistic Level

Before deciding if our test statistic provides enough evidence against the null hypothesis, we first decide on an appropriate **significance level**, α. THe scientific standard is 0.05, although this value should change based on the context of your problem

- The significance level, α, is the probability of wrongly rejecting the null hypothesis (when the null hypothesis is true, a false positive).
- The benchmark with which we decide if we have enough evidence against the null hypothesis is based on α. There are actually two, equivalent, approaches.

A.4.2.3 The p-value Approach

p-value: The probability of obtaining your particular random sample result (or more extreme) if the null hypothesis, H_0, were true.

- A high p-value implies that a random sample result is consistent with H_0.
- A small p-value implies that a random variation alone is unlikely to account for the difference between H_0 and the observation from our random sample. Our sample is inconsistent with H_0.
- The smaller the p-value, the stronger the evidence against H_0.
- With a small p-value we reject H_0, and say that our data support H_a. We reject H_0 when the p-value is **less than** the significance level, α.

A.4.2.4 Distribution of a Statistic

Since the test statistic for testing a mean is Equation (A.14),

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{(refeq:hyptstmean revisited)}$$

our test statistic is compared with a t_k distribution.

A.4.2.5 Finding the p-value

The p-value is represented by the area under the sampling distribution for values at least as extreme, in the direction of H_a, as that of our random sample.

- In this case, the *sampling distribution* is a t-distribution.

Assuming the H_0 is true, $t \approx t_k$ where $k = n - 1$.

- $H_a : \mu \neq \mu_0$ (two sided test)

To find area under distribution for all points greater than the magnitude of t, or $|t|$ in R, use *pt$(-|t|, df)$ then multiply by two for both tails of the distribution.

- One Sided test
  - $H_a : \mu < \mu_0$, use pt(t, df), since you just want all area to the left of t.
  - You are looking for the probability that the population mean is less than the sample mean.
  - $H_a : \mu > \mu_0$, use 1−pt(t, df), since you want all area to the right of t.
  - You are looking for the probability that the population mean is greater than the sample mean.

A.4.2.6 Decision

We compare the p-value with the **significance level**, α.

- If the p-value is less than or equal to α, we reject H_0. Our data support H_a.
- If the p-value is greater than α, we fail to reject H_0. Our data do not support H_a.

– Does not mean we support the null hypothesis, just weren't able to reject it.

Rejecting H_0 is said to be a "statistically significant result". Failing to reject H_0 is said to be a "statistically insignificant result".

A.4.3 Evaluating Evidence: Critical Value Approach

Critical value: the value of the test statistic that results in a p-value equal to the significance level

- The larger the test statistic, the more evidence we have against H_0.
- If the magnitude of the test statistic is larger than the critical value, we reject H_0.

A.4.3.1 Finding the Critical Value

Critical value, t^*: the value of t-statistic whose p-value is equal to significance level, α.

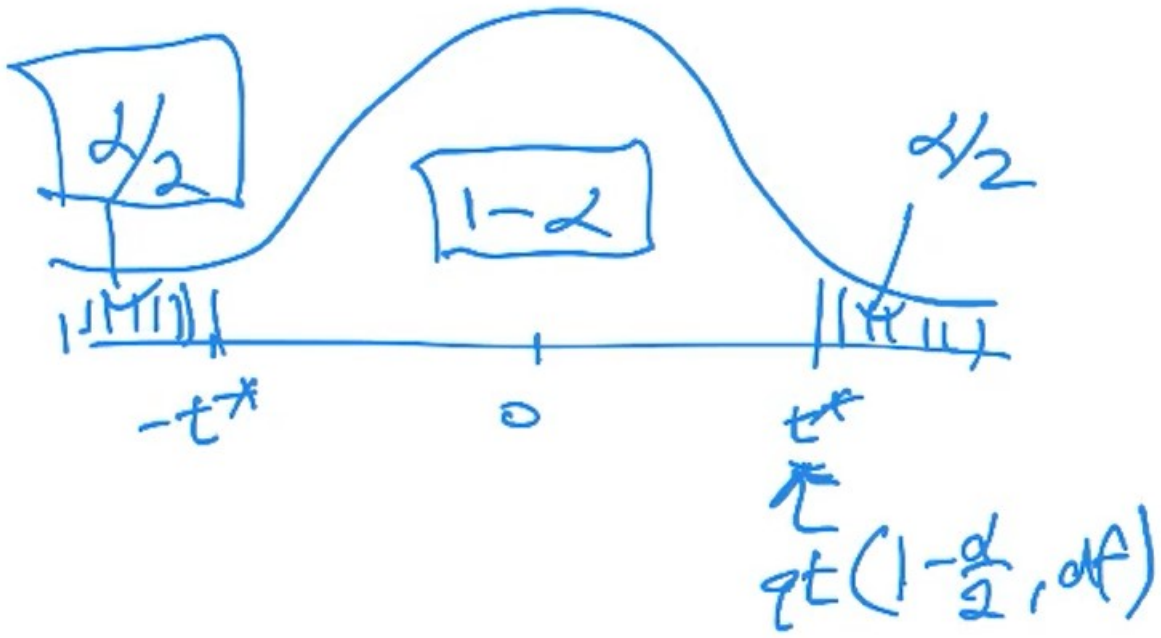

Figure A.2: t* for 2-Sided Critical Value

As seen in Figure A.2, the value of t^* for a two-sided test will be given by typing into R the following:

qt(1-$\frac{\alpha}{2}$,df)

This is because t^* will be the area in the region $1 - \alpha + \alpha/2 = 1 - \alpha/2$

Nota bene: Remember that qt and pt are the approximations of qnorm and pnorm, respectively, which take into account the degrees of freedom.

For a one sided test, we have two possible scenarios

- $H_a : \mu > \mu_0$ in this case, the area to the right of t^* will be represented by α, so we will type qt(1-α) to find t^*.
- $H_a : \mu < \mu_0$ in this case, the area to the left of - t^* will be represented by α, so we will type qt(α) to find - t^*, or -qt(α) to find t^*

Note: By symmetry notice that -qt(α) = qt(1-α,df), thus typically for a one-sided test we will typically use qt(1-α,df).

In review:

- Two-sided test
  - qt(1-$\frac{\alpha}{2}$,df)
- One-sided test
  - qt(1-α,df)

A.4.3.2 Exercises

1. Find critical value of two-sided t-test at $\alpha = 0.1$ with 10 df.

```
qt(1-0.1/2,10)
```

```
## [1] 1.812461
```

2. Find critical value of one-sided t-test at $\alpha = 0.02$ with 55 df.

```
qt(1-0.02,55)
```

```
## [1] 2.103607
```

3. Find critical value of one-sided t-test at $\alpha = 0.01$ with 70 df.

```
qt(1-0.01,70)
```

```
## [1] 2.380807
```

A.4.4 Summary

- Based on your question of interest, write H_0 and H_a.
- Evaluate how dissimilar your sample is from H_0 by calculating the test statistic.
- Compare you sample with a benchmark. Two equivalent approaches:
  1. Compare p-value with significance level α.
  2. Compare your test statistic with the critical value.
- Write relevant conclusion for your analysis

A.4.5 Worked Examples

Question 1: You are in charge of quality control in your food company. You randomly sample 40 packs of cherry tomatoes, each labeled 1/2 lb. (227 g), and find their average weight is 226.5 g. Is the weight in our sample evidence that the machine that sorts the tomatoes needs revision? Suppose the sample standard deviation is 1.5 g.

I think I'd want a two-sided test, so first we find the test statistic using Equation (A.14),

$t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$,

plug that into pt(t,df), then compare to significance value of 0.05.

So,

$t = \frac{226.5-227}{1.5/\sqrt{40}} = -2.108$

```
pt((226.5-227)/(1.5/sqrt(40)),39)
```

```
## [1] 0.020746
```

Multiplying this by 2 (to account for both sides of the distribution) to get our p-value gives us p-value $\approx$ 0.04.

Since this is less than the significance level of 0.05, we reject the null hypothesis. This is enough evidence that the machine needs revision. Stated statistically, the probability, or p-value, that we would get a sample mean of 226.5 g or less if the true population mean, or μ, was actually equal to 227 g is so low ($\approx 4\%$), that we cannot say this is a chance happening and something is amiss with the sorting machine.

Using the Critical Value Method

- This time we find t^* using the significance level $\alpha = 0.05$ and compare to the absolute value of our *t-stat* which we previously found to be -2.108.
- use qt(1-alpha/2,df)

```
qt(1-0.05/2,39)
```

```
## [1] 2.022691
```

Since $t^* < t$, we reject the null hypothesis. Our data support the claim that the population average weight of cherry tomato packs is different from 227 g.

Question 2: A consumer advocacy group is trying to claim that consumers are being cheated by the food company. Carry out an appropriate hypothesis test.

- $H_a : \mu > \mu_0$

Only difference here from Question 1 is that now $t = 0.95$ since for a one-sided test $t = 1 - \alpha$. So,

```
qt(0.95,39)
```

```
## [1] 1.684875
```

A.5 Practice Questions

A.5.1 Sampling Distributions

1. Statistical theory tells us the distribution of the sample means with a fixed sample size, under certain circumstances. The sampling distribution is an approximation of the density histogram of the sample means. WE know the sample means vary from sample to sample. The sampling distribution tells us the expected value (mean) of the distribution, and the standard deviation of the sample means.

a. Suppose the variable X follows a normal distribution with mean μ and standard deviation σ. Consider taking random samples, each with size n, repeatedly. What is the sampling distribution of the mean?

The sampling distribution is $N(\mu, \frac{\sigma}{\sqrt{n}})$ since $X \approx N(\mu, \sigma)$

b. Suppose the variable X has an unknown distribution but known mean μ and known standard deviation σ. What is the name of the statistical theory that informs us that the sampling distribution of the sample mean, $\bar{x}$, can be well-approximated by a normal distribution?

The Central Limit Theorem

2. An automatic machine in a manufacturing process produces sub-components. The lengths of the sub-components follow a normal distribution of the sample mean of 116 cm and a standard deviation of 4.8 cm.

a. Find the probability that one selected sub-component is longer than 118 cm.

Z-score here is $\frac{118-116}{4.8}$

```
1-pnorm(2/4.8)
```

```
## [1] 0.3384611
```

b. Find the probability that if 3 sub-components are randomly selected, their mean length exceeds 118 cm.

z-score here is $\frac{118-116}{4.8/\sqrt{3}}$

```
zscore<-(118-116)/(4.8/sqrt(3))
1-pnorm(zscore)
```

```
## [1] 0.2352432
```

A.5.2 Confidence Intervals

3. What are the goals of constructing a confidence interval?

- A confidence interval shows in what range we expect to find a sample mean (range of plausible values)
- provide estimate for unknown parameter of interest
- provide measure of uncertainty

4. How does increasing the confidence level affect the margin of error and the width of the confidence interval? Hint: sketching the standard normal distribution will be helpful

- When the confidence level increases, the margin of error and the width of the confidence interval increase as well. This is because we can be more certain the unknown parameter is contained within the confidence interval when it is wider, but this also increases the range within which it can be found, leading to a larger margin of error.

5. How does increasing the sample size affect the margin of error and width of the confidence interval? Briefly explain.

- Increasing the sample size decreases the margin of error and width of confidence interval. This is because the margin of error and width of confidence interval are proportional to $\frac{1}{\sqrt{n}}$ where n is the sample size. So increasing n decreases the MOE and CI width.

6. Use R to find the value of the t-multiplier when constructing a confidence interval for the mean in the following situations:

(a) 94% CI with $n = 49$.

```
qt(.97,48)
```

```
## [1] 1.926298
```

(b) 86% CI with $n = 82$.

```
qt(.93,81)
```

```
## [1] 1.490412
```

(a) 74% CI with $n = 150$.

```
qt(.87,149)
```

```
## [1] 1.130695
```

7. A random sample of 100 students had a mean grade point average (GPA) of 3.2 with a standard deviation of 0.2.

a. Calculate a 97% CI for the mean GPA for all students

- Confidence interval is given by $\bar{x} \pm t_{mult} \times \frac{s}{\sqrt{n}}$
- Where the right half is the MOE. So,

```
qt(0.985,99)*0.2/sqrt(100)
```

```
## [1] 0.04403637
```

- So confidence interval is 3.2 ± 0.044 or (3.156,3.244)

b. What is the margin of error for the confidence interval found in the previous part? what is the margin of error telling us?

- The margin of error is about 0.044 grade points. This tells us the range within which we expect the true mean GPA to be; it also gives us a measure of uncertainty of the mean GPA.

c. Based on this confidence interval, is it reasonable to say that the mean GPA of all students is 3.25 or greater?

- Because 3.25 falls outside of our confidence interval, it is unlikely that that the true mean GPA of all students is 3.25 or greater.

A.5.3 Hypothesis Testing

8. What is the goal of conducting a hypothesis test?

- The goal of hypothesis testing is to determine the likelihood of obtaining a particular sample mean
- distinguish if sample statistic is due to random chance or reflects a real feature of phenomenon under study

9. Hypothesis statements are always about the **population / sample statistic** (choose one) of interest.

- sample statistic

10. For each of the situations, state the appropriate null and alternative hypotheses, in symbols and in words. Sketch how you would find the p-value based on the calculated test-statistic.

a. David's car averages 29 miles per gallon on the highway. He just switched to a new motor oil that is advertised as increasing gas mileage. He wants to investigate if the advertisement is accurate.

- $H_0 : \mu = 29$ mpg
- $H_a : \mu < 29$ mpg

b. The diameter of a spindle in a small motor is supposed to be 4 millimeters. If the spindle is too small or too large, the motor will not function properly. The manufacturer wants to investigate further whether the mean diameter is moved away from the target.

- $H_0 : \mu = 4$ mm
- $H_a : \mu \neq 4$ mm

c. The average time in traffic between 2 points of a congested highway used to be 2 hours. The government invested money to improve travel times by building extra lanes and overpasses. Citizens want to access if travel times have improved, on average.

- $H_0 : \mu = 2$ hr
- $H_a : \mu < 2$ hr

11. To have more evidence against the null hypothesis, our test statistic should be **larger / smaller** (choose one) in magnitude. Briefly explain.

- Larger, since the test statistic is a measure of how dissimilar the sampled data is from the null hypothesis, a larger test statistic is more evidence against the null hypothesis.

12. How does increasing the difference between the sample mean and the population mean under the null hypothesis affect the test statistic and the evidence against the null hypothesis?

- Increasing the difference between the sample mean and population mean will increase the test statistic and provide more evidence against the null hypothesis.

13. How does increasing the sample size affect the test statistic and the evidence against the null hypothesis?

- Test statistic is given by $t = \frac{\bar{x}-\mu}{s/\sqrt{n}} = \frac{(\bar{x}-\mu)\sqrt{n}}{s}$, so an increase in sample size will also increase the test-statistic and provide more evidence against the null hypothesis.

14. Use R to obtain the critical values of the following hypothesis tests:

-find t-stat whose p-value is equal to the significance level α.

a. $H_0 : \mu = 3.5, H_a : \mu \neq 3.5,$ with $\alpha = 0.8$ and $n = 96$

```
qt(0.96,95)
```

```
## [1] 1.769615
```

b. $H_0 : \mu = 75, H_a : \mu < 75,$ with $\alpha = 0.12$ and $n = 43$

```
qt(0.88,42)
```

```
## [1] 1.191879
```

c. $H_0 : \mu = 10, H_a : \mu > 10,$ with $\alpha = 0.045$ and $n = 132$

```
qt(1-0.045,131)
```

```
## [1] 1.708027
```

15. Use R to obtain the p-values of the following hypothesis tests:

a. $H_0 : \mu = 48, H_a : \mu \neq 48,$ with $t - stat = 2.14$ and $n = 50$.

- for a two sided test, multiply p-value by 2

```
(1-pt(2.14,49))*2
```

```
## [1] 0.03735955
```

b. $H_0 : \mu = 3, H_a : \mu > 3,$ with $t - stat = 0.78$ and $n = 316$.

```
1-pt(0.78,315)
```

```
## [1] 0.2179883
```

c. $H_0 : \mu = 12, H_a : \mu < 12,$ with $t - stat = 1.57$ and $n = 34$.

```
pt(1.57,33)
```

```
## [1] 0.9370224
```

16. The 10-year historical average yield of corn in the United States is 160 bushels per acre. A survey of 50 farmers this year gives a sample mean yield of 158.4 bushels per acre, with a standard deviation of 5 bushels per acre. Does this sample provide evidence that the yield of corn has decreased from the 10-year historical average? Conduct an appropriate hypothesis test.

a. State the null and alternative hypotheses.

- $H_0 : \mu = 160$ bushels per acre
- $H_a : \mu < 160$ bushels per acre

b. Calculate the test-statistic.

- Given by $t^* = \frac{\bar{x}-\mu}{s/\sqrt{n}}$ so,

```
tstar16<-(158.4-160)/((5/sqrt(50)))
tstar16
```

```
## [1] -2.262742
```

c. Find the p-value and the critical value.

- Usign significance level, α, of 0.05:
- **p-value approach**

```
pt(tstar16,49)
```

```
## [1] 0.01405941
```

- **critical value approach**

```
qt(0.95,49)
```

```
## [1] 1.676551
```

d. State a conclusion in context.

- Because the p-value is less than the significance level, and because the critical value is less than the magnitude of the t-statistic, we must reject the null hypothesis. Our data support the claim that the yield of corn this year has decreased from the 10-year historical average.

e. How would you interpret the calculated p-value?

- There is a probability of 0.014 that we will obtain a sample mean of 158.4 bushels per acre if the average yield this year is truly 160 bushels per acre.

A.5.4 General Questions

17. Obtain the critical value of a hypothesis test where $H_0 : \mu = 145, H_a : \mu \neq 145,$ with significance level $\alpha = 0.02$. Suppose the sample size is 50.

- This is a **two-sided** test, so I'll use qt($1 - \alpha/2$,df)

```
qt(0.99,49)
```

```
## [1] 2.404892
```

18. Obtain the t-multiplier for a 98% confidence interval. Suppose the sample size is 50.

- To find t-multiplier, I'll use qt($1 - \alpha/2$,df)

```
qt(0.99,49)
```

```
## [1] 2.404892
```

19. Compare the critical value and the t-multiplier found in the previous two parts. What is the implication based on this comparison?

- The two values are equal. The implication is that conclusions from a two-sided hypothesis test conducted at significance level α will be consistent with conclusions from a $(1 - \alpha)\times$ 100% confidence interval.

20. Suppose the hypothesis test in question 17 is carried out and the p-value is 0.043. Which of the following confidence intervals is/are possible?

- (143.2,144.5)
- (151.3,154.6)
- (144.5,163.5)
- Since the p-value is higher than the significance level, $\alpha = 0.02$, we fail to reject the null hypothesis. Thus 145 must fall within the confidence interval, so only the CI (144.5,163.5) is possible.

21. A random sample of 85 banded archerfish were collected, and their lengths were measured and recorded. Their average length was 20cm with a standard deviation of 3cm.

a. Construct a 95% confidence interval for the population mean length of banded archerfish.

- So $\alpha = 0.05$ since the CI is 95%. Need to find t-multiplier and variance $\frac{s}{\sqrt{n}}$

```
qt(0.975,84)*3/sqrt(85)
```

```
## [1] 0.647085
```

- CI is 20 ± 0.647 or (19.353,20.647)

b. Based on your confidence interval, is it plausible that the population mean length of banded archerfish is 21cm? Briefly explain.

- It is not plausible that the population mean length is 21cm since 21cm falls outside of the 95% confidence interval.

c. Suppose you conduct the following hypothesis test. $H_0 : \mu = 21, H_a : \mu \neq 21$. Without actually performing any additional calculations, what do you expect the p-value of this hypothesis test will be? Briefly explain.

- greater than 0.05
- less than 0.05
- The p-value will be smaller than 0.05, since this means that if the the probability that a sample mean will be 21.

d. Conduct the hypothesis test to verify your answer to the previous part.

- Find p-value using pt()

```
t21<-(20-21)/(3/sqrt(85))
t21
```

```
## [1] -3.073181
```

```
2*pt(t21,84)
```

```
## [1] 0.002855487
```

- Find critical value

```
qt(0.975,84)
```

```
## [1] 1.98861
```

- Since this is lower than $\alpha = 0.05$ and the magnitude of the critical value ($1.989 < |\text{-}3.073|$) we reject the null hypothesis. Our data supports the claim that the population mean does not equal 21cm.

Appendix B

Basics of R

B.1 Getting Started with R

B.1.1 Question 1

(a)

```
cars.df<-mtcars
```

(b) According to the environment window, there are 32 observations of 11 variables in the dataset *mtcars*.

B.1.2 Question 2

(a)

```
students.df<-read.table("datasets/students.txt",header=TRUE)
```

(b) According to the environment window, there are 249 observations of 9 variables in the dataset *students.txt*.

B.1.3 Question 3

(a) - (h)

The packages *tidyverse, faraway, MASS, leaps, ROCR, nycflights13, gapminder, palmerpenguins* were installed.

B.1.4 Question 4

```
library(faraway)
corn.df<-cornnit
```

B.2 Topic B.3: Data Types & Structures in R

B.2.1 Question 5

(a) 2020_Major Valid

(b) .2020.Age Invalid, number follows .

(c) #Courses.2020 Invalid, # not allowed

(d) _courses_2020 Invalid, cannot start with underscore

(e) Fav_Sport20 Valid

(f) major 2020 Invalid, space not allowed

(g) age(2020) Invalid, parentheses not allowed

(h) FavSport_2020 Valid

B.2.2 Question 6

```
practice<-c(13,91,36,95,9,3,61,20,22,97)
class(practice)
```

```
## [1] "numeric"
```

B.2.3 Question 7

(a) practice[5]==5 False

```
practice[5]==5
```

```
## [1] FALSE
```

(b) practice[10]!=97 False

```
practice[10]!=97
```

```
## [1] FALSE
```

(c) (practice[1]+practice[2])<104 False

```
(practice[1]+practice[2])<104
```

```
## [1] FALSE
```

(d) (practice[1]+practice[2])<=104 True

```
(practice[1]+practice[2])<=104
```

```
## [1] TRUE
```

(e) (practice[2]==91) & (practice[9]==22) True * True=True

```
(practice[2]==91) & (practice[9]==22)
```

```
## [1] TRUE
```

(f) (practice[5]<9) | (practice[6]>=4) False + False = False

```
(practice[5]<9) | (practice[6]>=4)
```

```
## [1] FALSE
```

B.2.4 Question 8

```
Mat.A<-matrix(c(4,6,1,2,3,1),nrow = 2,ncol = 3)

Mat.A
```

```
##      [,1] [,2] [,3]
## [1,]    4    1    3
## [2,]    6    2    1
```

(a)

```
colnames(Mat.A)<-c("Huey","Dewey","Louie")

Mat.A
```

```
##      Huey Dewey Louie
## [1,]    4     1     3
## [2,]    6     2     1
```

(b) Output of Mat.A[2,1] would be 6.

```
Mat.A[2,1]
```

```
## Huey
##    6
```

(c) Output of dim(Mat.A) would be (2,3).

```
dim(Mat.A)
```

```
## [1] 2 3
```

B.2.5 Question 9

```
factor(practice)
```

```
##  [1] 13 91 36 95 9  3  61 20 22 97
## Levels: 3 9 13 20 22 36 61 91 95 97
```

The order of the levels in the factor *practice* are: 3 9 13 20 22 36 61 91 95 97

B.3 R Markdown

B.3.1 Question 10

As evidenced by the above, my answers were typed up using R Markdown, and an HTML file was created.