Intro to Clinical Research

Madelyn Houser

2/18/2021

Table of Contents

# This will document the version of R and the operating system that I am using for this analysis.  
sessionInfo()

## R version 4.0.2 (2020-06-22)  
## Platform: x86\_64-w64-mingw32/x64 (64-bit)  
## Running under: Windows 10 x64 (build 19041)  
##   
## Matrix products: default  
##   
## locale:  
## [1] LC\_COLLATE=English\_United States.1252   
## [2] LC\_CTYPE=English\_United States.1252   
## [3] LC\_MONETARY=English\_United States.1252  
## [4] LC\_NUMERIC=C   
## [5] LC\_TIME=English\_United States.1252   
##   
## attached base packages:  
## [1] stats graphics grDevices utils datasets methods base   
##   
## loaded via a namespace (and not attached):  
## [1] compiler\_4.0.2 magrittr\_1.5 tools\_4.0.2 htmltools\_0.5.0  
## [5] yaml\_2.2.1 stringi\_1.5.3 rmarkdown\_2.4 knitr\_1.30   
## [9] stringr\_1.4.0 xfun\_0.18 digest\_0.6.25 rlang\_0.4.8   
## [13] evaluate\_0.14

# This will load the packages needed for this analysis and document the version of each that I'm using  
  
# If you have not previously installed any of these packages, install them using the function  
## install.packages("PACKAGENAME")  
  
# Because a line of code in this Markdown produces HTML content and we want to be able to knit to Word still,  
## make sure the following packages are installed on your machine. They do not have to be loaded each time.  
# install.packages("webshot")  
# webshot::install\_phantomjs()  
  
library("Hmisc")

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

## Loading required package: ggplot2

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:base':  
##   
## format.pval, units

packageVersion("Hmisc")

## [1] '4.4.1'

library("tidyverse")

## -- Attaching packages ------------------------------------------------------------------------------------ tidyverse 1.3.0 --

## v tibble 3.0.3 v dplyr 1.0.2  
## v tidyr 1.1.2 v stringr 1.4.0  
## v readr 1.4.0 v forcats 0.5.0  
## v purrr 0.3.4

## -- Conflicts --------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()  
## x dplyr::src() masks Hmisc::src()  
## x dplyr::summarize() masks Hmisc::summarize()

packageVersion("tidyverse")

## [1] '1.3.0'

library("desctable")

##   
## Attaching package: 'desctable'

## The following objects are masked from 'package:stats':  
##   
## chisq.test, fisher.test, IQR

packageVersion("desctable")

## [1] '0.1.7'

library("ggpubr")  
packageVersion("ggpubr")

## [1] '0.4.0'

# Import and merge data

# Import the data sets of your choice into R from files saved locally on your computer  
# Replace the path here with the path to your local file  
demo <- sasxport.get("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/NHANES/DEMO\_J.XPT")

## Processing SAS dataset DEMO\_J ..

cbc <- sasxport.get("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/NHANES/CBC\_J.XPT")

## Processing SAS dataset CBC\_J ..

hepc <- sasxport.get("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/NHANES/HEPC\_J.XPT")

## Processing SAS dataset HEPC\_J ..

# Alternatively, you can download the files from NHANES directly into R  
# Replace the URLs with the URLs of the data sets of your choice  
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/DEMO\_J.XPT", tf <- tempfile(), mode="wb")  
demo2 <- foreign::read.xport(tf)  
  
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/CBC\_J.XPT", tf <- tempfile(), mode="wb")  
cbc2 <- foreign::read.xport(tf)  
  
download.file("https://wwwn.cdc.gov/Nchs/Nhanes/2017-2018/HEPC\_J.XPT", tf <- tempfile(), mode="wb")  
hepc2 <- foreign::read.xport(tf)

# You can remove variables that you don't want to keep in your environment  
rm(tf, cbc2, demo2, hepc2)

# Select only the variables of interest from the data sets  
demo <- demo %>% select(seqn, riagendr, ridageyr, ridagemn)

# Merge into one data set, keeping only rows with matching identifiers in demographic data set  
df <- merge(x=demo, y=cbc, by="seqn", all.x=TRUE)  
df <- merge(x=df, y=hepc, by="seqn", all.x=TRUE)

# Explore and Summarize data

# Report class of the data set  
class(df)

## [1] "data.frame"

# This will provide some basic information on each variable in the data set  
describe(df)

## df   
##   
## 28 Variables 9254 Observations  
## --------------------------------------------------------------------------------  
## seqn : Respondent sequence number   
## n missing distinct Info Mean Gmd .05 .10   
## 9254 0 9254 1 98330 3085 94166 94628   
## .25 .50 .75 .90 .95   
## 96016 98330 100643 102031 102493   
##   
## lowest : 93703 93704 93705 93706 93707, highest: 102952 102953 102954 102955 102956  
## --------------------------------------------------------------------------------  
## riagendr : Gender   
## n missing distinct Info Mean Gmd   
## 9254 0 2 0.75 1.508 0.4999   
##   
## Value 1 2  
## Frequency 4557 4697  
## Proportion 0.492 0.508  
## --------------------------------------------------------------------------------  
## ridageyr : Age in years at screening   
## n missing distinct Info Mean Gmd .05 .10   
## 9254 0 81 1 34.33 29.18 1 3   
## .25 .50 .75 .90 .95   
## 11 31 58 71 79   
##   
## lowest : 0 1 2 3 4, highest: 76 77 78 79 80  
## --------------------------------------------------------------------------------  
## ridagemn : Age in months at screening - 0 to 24 mos   
## n missing distinct Info Mean Gmd .05 .10   
## 597 8657 25 0.998 10.44 8.157 0 1   
## .25 .50 .75 .90 .95   
## 4 10 17 21 22   
##   
## lowest : 0 1 2 3 4, highest: 20 21 22 23 24  
## --------------------------------------------------------------------------------  
## lbxwbcsi : White blood cell count (1000 cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 149 1 7.383 2.519 4.3 4.8   
## .25 .50 .75 .90 .95   
## 5.8 7.0 8.5 10.1 11.3   
##   
## lowest : 1.9 2.3 2.4 2.5 2.6, highest: 22.8 38.1 41.9 74.2 400.0  
##   
## Value 0 5 10 15 20 25 40 75 400  
## Frequency 5 4330 3022 162 4 1 2 1 1  
## Proportion 0.001 0.575 0.401 0.022 0.001 0.000 0.000 0.000 0.000  
##   
## For the frequency table, variable is rounded to the nearest 5  
## --------------------------------------------------------------------------------  
## lbxlypct : Lymphocyte percent (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 600 1 34.11 11.81 18.61 21.60   
## .25 .50 .75 .90 .95   
## 26.70 33.10 40.20 48.10 53.70   
##   
## lowest : 4.4 5.0 5.2 5.6 6.5, highest: 76.3 84.3 88.8 89.5 89.7  
## --------------------------------------------------------------------------------  
## lbxmopct : Monocyte percent (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 167 1 8.2 2.332 5.3 5.8   
## .25 .50 .75 .90 .95   
## 6.7 7.9 9.3 10.9 12.0   
##   
## lowest : 0.7 1.1 1.2 1.9 2.0, highest: 21.6 22.5 24.1 24.4 57.2  
## --------------------------------------------------------------------------------  
## lbxnepct : Segmented neutrophils percent (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 633 1 54.09 12.65 33.50 38.90   
## .25 .50 .75 .90 .95   
## 47.40 55.10 61.90 67.50 71.09   
##   
## lowest : 8.4 8.8 10.3 10.6 11.3, highest: 87.7 88.6 89.2 91.9 92.8  
## --------------------------------------------------------------------------------  
## lbxeopct : Eosinophils percent (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 170 1 2.975 2.24 0.7 1.0   
## .25 .50 .75 .90 .95   
## 1.5 2.4 3.7 5.6 7.3   
##   
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 21.2 22.0 22.1 27.4 29.1  
## --------------------------------------------------------------------------------  
## lbxbapct : Basophils percent (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 34 0.989 0.755 0.3438 0.3 0.4   
## .25 .50 .75 .90 .95   
## 0.5 0.7 0.9 1.1 1.3   
##   
## lowest : 0.1 0.2 0.3 0.4 0.5, highest: 3.1 3.2 3.3 3.4 4.8  
## --------------------------------------------------------------------------------  
## lbdlymno : Lymphocyte number (1000 cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 86 0.998 2.502 1.134 1.2 1.4   
## .25 .50 .75 .90 .95   
## 1.8 2.3 2.9 3.6 4.2   
##   
## lowest : 0.4 0.5 0.6 0.7 0.8, highest: 10.6 32.1 37.5 65.9 358.8  
##   
## Value 0 1 2 3 4 5 6 7 8 9 10  
## Frequency 6 768 3995 1855 637 148 76 21 8 1 3  
## Proportion 0.001 0.102 0.531 0.247 0.085 0.020 0.010 0.003 0.001 0.000 0.000  
##   
## Value 11 32 38 66 359  
## Frequency 1 1 1 1 1  
## Proportion 0.000 0.000 0.000 0.000 0.000  
##   
## For the frequency table, variable is rounded to the nearest 1  
## --------------------------------------------------------------------------------  
## lbdmono : Monocyte number (1000 cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 24 0.969 0.5878 0.2142 0.3 0.4   
## .25 .50 .75 .90 .95   
## 0.5 0.6 0.7 0.8 1.0   
##   
## lowest : 0.1 0.2 0.3 0.4 0.5, highest: 2.0 2.6 2.8 2.9 6.7  
## --------------------------------------------------------------------------------  
## lbdneno : Segmented neutrophils num (1000 cell/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 121 1 4.035 1.836 1.8 2.1   
## .25 .50 .75 .90 .95   
## 2.8 3.8 4.9 6.2 7.1   
##   
## lowest : 0.4 0.5 0.7 0.8 0.9, highest: 14.1 14.7 15.3 15.7 35.2  
## --------------------------------------------------------------------------------  
## lbdeono : Eosinophils number (1000 cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 23 0.918 0.2155 0.1692 0.0 0.1   
## .25 .50 .75 .90 .95   
## 0.1 0.2 0.3 0.4 0.5   
##   
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 1.8 1.9 2.0 2.6 3.2  
## --------------------------------------------------------------------------------  
## lbdbano : Basophils number (1000 cells/uL)   
## n missing distinct Info Mean Gmd   
## 7523 1731 6 0.755 0.05033 0.05172   
##   
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 0.1 0.2 0.3 0.4 0.5  
##   
## Value 0.0 0.1 0.2 0.3 0.4 0.5  
## Frequency 3802 3664 53 1 2 1  
## Proportion 0.505 0.487 0.007 0.000 0.000 0.000  
## --------------------------------------------------------------------------------  
## lbxrbcsi : Red blood cell count (million cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 327 1 4.734 0.5317 4.00 4.17   
## .25 .50 .75 .90 .95   
## 4.42 4.71 5.03 5.32 5.54   
##   
## lowest : 2.32 2.53 2.79 2.87 2.90, highest: 6.78 6.80 6.99 7.04 7.84  
## --------------------------------------------------------------------------------  
## lbxhgb : Hemoglobin (g/dL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 118 1 13.73 1.681 11.5 12.0   
## .25 .50 .75 .90 .95   
## 12.8 13.7 14.7 15.7 16.2   
##   
## lowest : 6.4 6.7 6.9 7.0 7.2, highest: 18.6 18.7 19.1 19.2 19.9  
## --------------------------------------------------------------------------------  
## lbxhct : Hematocrit (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 277 1 40.85 4.666 34.6 35.9   
## .25 .50 .75 .90 .95   
## 38.0 40.7 43.6 46.3 47.8   
##   
## lowest : 23.0 23.1 23.3 23.4 24.2, highest: 55.4 55.6 56.4 57.8 58.8  
## --------------------------------------------------------------------------------  
## lbxmcvsi : Mean cell volume (fL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 434 1 86.56 7.179 75.4 78.6   
## .25 .50 .75 .90 .95   
## 82.9 87.1 90.9 94.2 96.3   
##   
## lowest : 35.4 50.8 51.6 52.2 53.6, highest: 110.5 111.2 111.8 113.1 114.6  
## --------------------------------------------------------------------------------  
## lbxmchsi : Mean cell hemoglobin (pg)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 193 1 29.07 2.727 24.64 26.10   
## .25 .50 .75 .90 .95   
## 27.78 29.30 30.70 31.90 32.60   
##   
## lowest : 12.4 13.2 14.3 15.2 16.2, highest: 37.7 38.1 38.2 38.4 39.2  
## --------------------------------------------------------------------------------  
## lbxmc : Mean Cell Hgb Conc. (g/dL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 82 0.999 33.57 1.019 32.0 32.4   
## .25 .50 .75 .90 .95   
## 33.0 33.6 34.2 34.7 35.0   
##   
## lowest : 25.2 27.7 27.9 28.1 28.6, highest: 36.5 36.6 36.8 37.2 38.3  
## --------------------------------------------------------------------------------  
## lbxrdw : Red cell distribution width (%)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 110 0.999 13.78 1.232 12.4 12.6   
## .25 .50 .75 .90 .95   
## 13.0 13.5 14.2 15.2 16.1   
##   
## lowest : 11.3 11.4 11.6 11.7 11.8, highest: 24.0 25.5 25.8 26.0 29.2  
## --------------------------------------------------------------------------------  
## lbxpltsi : Platelet count (1000 cells/uL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 424 1 259.3 77.19 160 180   
## .25 .50 .75 .90 .95   
## 211 251 300 349 384   
##   
## lowest : 8 54 57 61 63, highest: 648 662 696 697 818  
## --------------------------------------------------------------------------------  
## lbxmpsi : Mean platelet volume (fL)   
## n missing distinct Info Mean Gmd .05 .10   
## 7528 1726 66 0.999 8.094 1.019 6.7 7.0   
## .25 .50 .75 .90 .95   
## 7.5 8.0 8.7 9.3 9.7   
##   
## lowest : 5.4 5.5 5.7 5.8 5.9, highest: 11.6 11.8 11.9 12.3 13.0  
## --------------------------------------------------------------------------------  
## lbxnrbc : Nucleated red blood cells   
## n missing distinct Info Mean Gmd .05 .10   
## 7523 1731 12 0.793 0.0851 0.07441 0.0 0.0   
## .25 .50 .75 .90 .95   
## 0.0 0.1 0.1 0.2 0.2   
##   
## lowest : 0.0 0.1 0.2 0.3 0.4, highest: 0.7 0.8 1.1 1.5 2.0  
##   
## Value 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 1.1 1.5  
## Frequency 2404 4192 731 136 20 21 8 3 5 1 1  
## Proportion 0.320 0.557 0.097 0.018 0.003 0.003 0.001 0.000 0.001 0.000 0.000  
##   
## Value 2.0  
## Frequency 1  
## Proportion 0.000  
## --------------------------------------------------------------------------------  
## lbxhcr : Hepatitis C RNA   
## n missing distinct Info Mean Gmd   
## 6690 2564 3 0.065 2.97 0.05869   
##   
## Value 1 2 3  
## Frequency 51 98 6541  
## Proportion 0.008 0.015 0.978  
## --------------------------------------------------------------------------------  
## lbdhci : Hepatitis C Antibody (confirmed)   
## n missing distinct Info Mean Gmd   
## 6674 2580 4 0.059 2.988 0.05402   
##   
## Value 1 2 3 4  
## Frequency 49 33 6541 51  
## Proportion 0.007 0.005 0.980 0.008  
## --------------------------------------------------------------------------------  
## lbxhcg : Hepatitis C Genotype   
## n missing distinct Info Mean Gmd   
## 49 9205 8 0.861 2.612 2.221   
##   
## lowest : 1 2 3 4 5, highest: 4 5 6 8 9  
##   
## Value 1 2 3 4 5 6 8 9  
## Frequency 25 7 1 6 6 1 1 2  
## Proportion 0.510 0.143 0.020 0.122 0.122 0.020 0.020 0.041  
## --------------------------------------------------------------------------------

# Data cleanup and wrangling

Let’s check the format of our variables and label the levels of categorical variables and make sure they’re classified as factors.

# Check class of variable  
class(df$riagendr)

## [1] "labelled" "integer"

# Make variable a factor and define levels and labels of levels  
df$riagendr <- factor(df$riagendr, levels=c("1", "2"), labels=c("Male", "Female"))  
# Change the name of the variable  
df <- rename(df, "Sex" = "riagendr")  
# Check the class of the variable again to be sure that the changes were made correctly  
class(df$Sex)

## [1] "factor"

# Check class and change label of another variable  
class(df$ridageyr)

## [1] "labelled" "integer"

label(df$ridageyr) <- "Age (years)"  
df <- rename(df, "Age" = "ridageyr")  
  
# Spot check class of lab test variables  
class(df$lbxwbcsi)

## [1] "labelled" "numeric"

class(df$lbxlypct)

## [1] "labelled" "numeric"

class(df$lbxnrbc)

## [1] "labelled" "numeric"

# Fix formatting of Hep C RNA variable  
class(df$lbxhcr)

## [1] "labelled" "integer"

df$lbxhcr <- factor(df$lbxhcr, levels=(c("1", "2", "3")), labels=c("Positive", "Negative", "Negative Ab Screening"))  
df <- rename(df, "HepC\_RNA" = "lbxhcr")  
class(df$HepC\_RNA)

## [1] "factor"

# Check factor levels of the variable  
levels(df$HepC\_RNA)

## [1] "Positive" "Negative" "Negative Ab Screening"

# Rename some other variables  
df <- rename(df, "RBCs" = "lbxrbcsi")  
df <- rename(df, "Leukocytes" = "lbxwbcsi")  
df <- rename(df, "Lymphocytes" = "lbxlypct")

Let’s filter the data so that we’re only analyzing data for adults

dfad <- df %>% filter(Age >= 18)  
  
summary(dfad$Age)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 18.00 33.00 51.00 49.89 65.00 80.00

# Visualize data for variables of interest

## Table - Categorical Variables by Sex

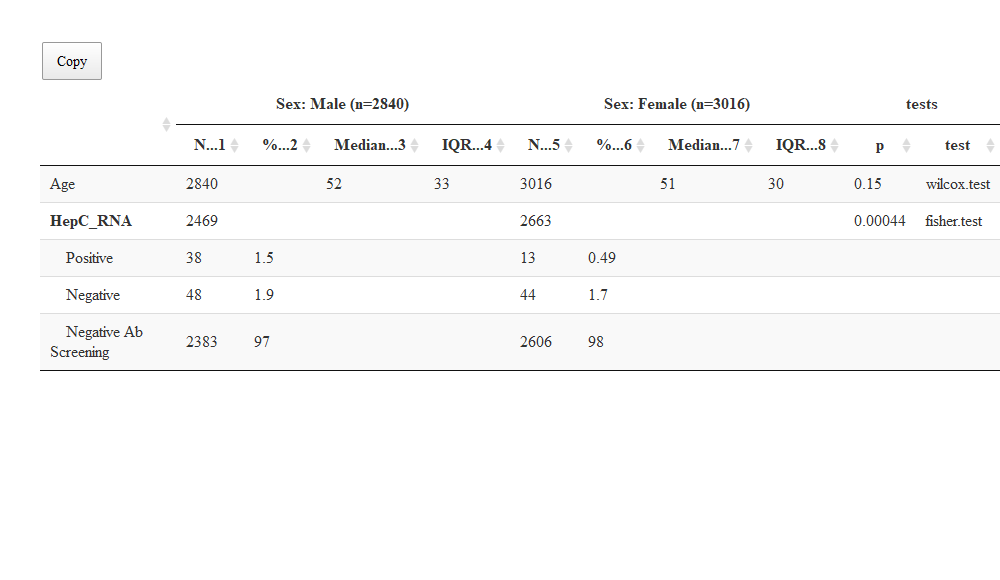
# Create a table summarizing some variables of interest by sex  
Tab1 <- dfad %>% group\_by(Sex) %>% select(Age, HepC\_RNA) %>% desctable() %>% datatable()

## Adding missing grouping variables: `Sex`

## New names:  
## \* N -> N...1  
## \* `%` -> `%...2`  
## \* Median -> Median...3  
## \* IQR -> IQR...4  
## \* N -> N...5  
## \* ...

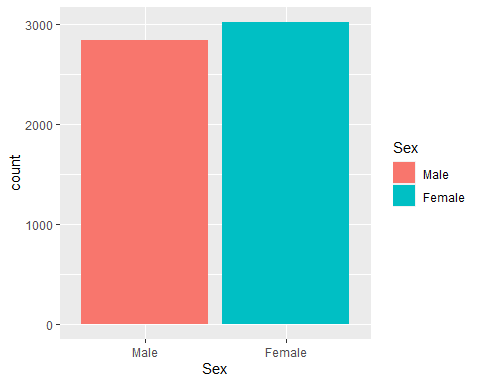
## New names:  
## \* N -> N...2  
## \* `%` -> `%...3`  
## \* Median -> Median...4  
## \* IQR -> IQR...5  
## \* N -> N...6  
## \* ...

Tab1

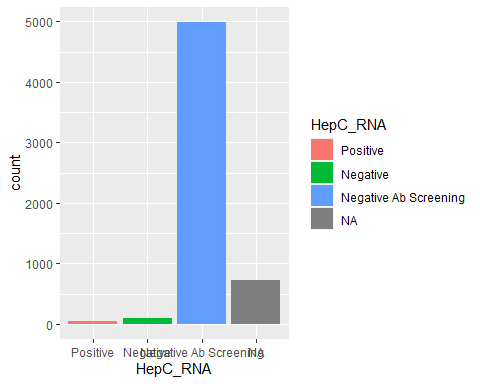


## Variable graphs

# Plot counts of each level of a categorical variable  
ggplot(data = dfad) +   
 geom\_bar(mapping = (aes(x = Sex, fill=Sex)))

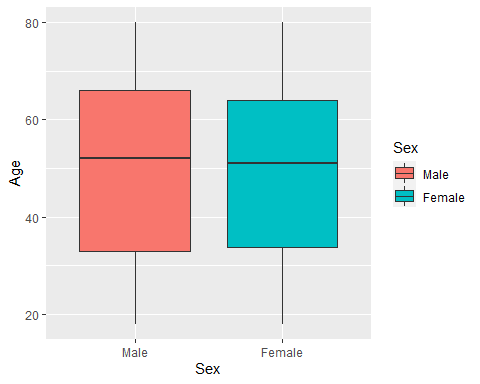


ggplot(data = dfad) +   
 geom\_bar(mapping = (aes(x = HepC\_RNA, fill=HepC\_RNA)))



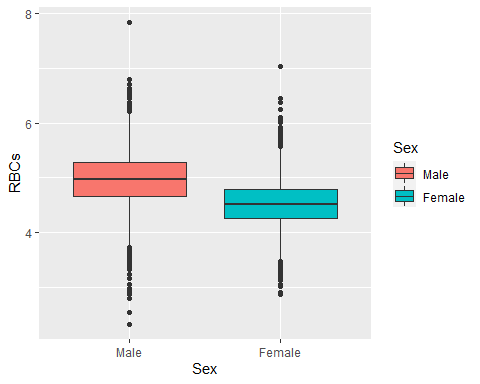
# Plot Age split by Sex  
ggplot(data = dfad) +  
 geom\_boxplot(mapping=aes(x=Sex, y=Age, fill=Sex))

## Don't know how to automatically pick scale for object of type labelled/integer. Defaulting to continuous.



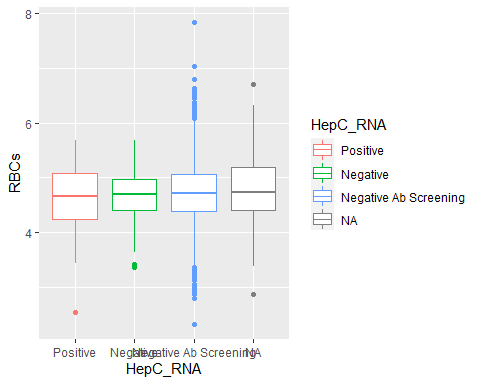
# Create a boxplot of RBCs data split by Sex  
ggplot(data=dfad, mapping=aes(x=Sex, y=RBCs, fill=Sex)) +  
 geom\_boxplot()

## Warning: Removed 585 rows containing non-finite values (stat\_boxplot).



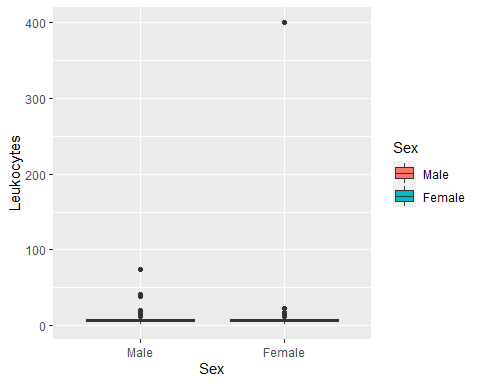
# Create a boxplot of RBCs data split by HepC\_RNA  
ggplot(data=dfad, mapping=aes(x=HepC\_RNA, y=RBCs, color=HepC\_RNA)) +  
 geom\_boxplot()

## Warning: Removed 585 rows containing non-finite values (stat\_boxplot).



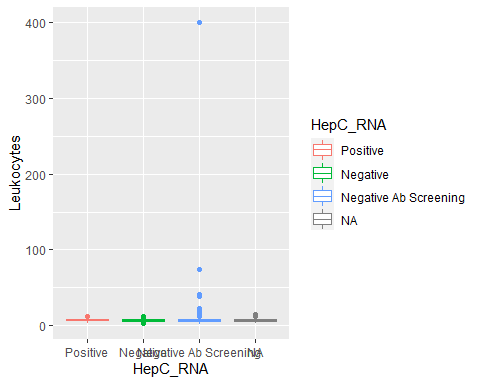
# Repeat with other continuous variables  
ggplot(data=dfad, mapping=aes(x=Sex, y=Leukocytes, fill=Sex)) +  
 geom\_boxplot()

## Warning: Removed 585 rows containing non-finite values (stat\_boxplot).



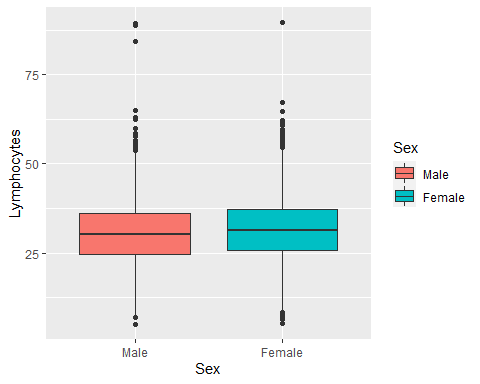
ggplot(data=dfad, mapping=aes(x=HepC\_RNA, y=Leukocytes, color=HepC\_RNA)) +  
 geom\_boxplot()

## Warning: Removed 585 rows containing non-finite values (stat\_boxplot).



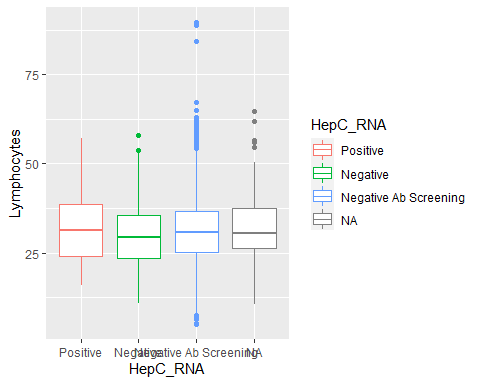
ggplot(data=dfad, mapping=aes(x=Sex, y=Lymphocytes, fill=Sex)) +  
 geom\_boxplot()

## Warning: Removed 589 rows containing non-finite values (stat\_boxplot).

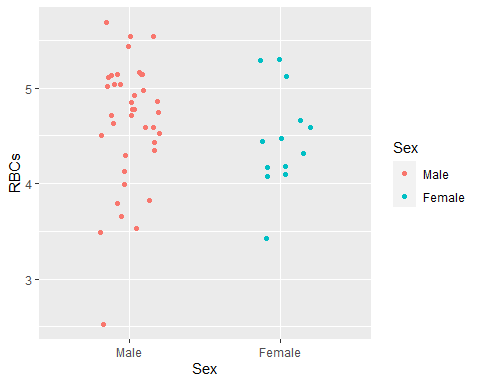


ggplot(data=dfad, mapping=aes(x=HepC\_RNA, y=Lymphocytes, color=HepC\_RNA)) +  
 geom\_boxplot()

## Warning: Removed 589 rows containing non-finite values (stat\_boxplot).



# Let's make sure we have enough data to work with within our HepC positive group if we split it by sex  
  
# Create a data set with only subjects positive for HepC RNA  
dfadpos <- dfad %>% filter(HepC\_RNA == "Positive")  
  
ggplot(data=dfadpos, mapping=aes(x=Sex, y=RBCs, color=Sex)) +  
 geom\_jitter(width=.2)



#### We observe that there is an extremely high value in the Leukocytes variable data.

## Check distribution of continuous variables

### Density and Q-Q plots

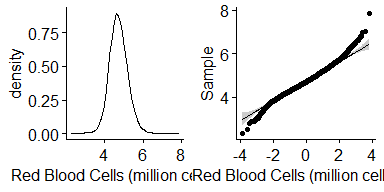
Let’s plot the distribution of some continuous variables.

# Create an object consisting of a plot of the distribution of the RBCs variable  
denplot <- ggdensity(df$RBCs, xlab = "Red Blood Cells (million cells/uL)")  
# Create an object consisting of a Q-Q plot of the RBCs variable  
qqplot <- ggqqplot(df$RBCs, xlab = "Red Blood Cells (million cells/uL)")  
# Show the plots side by side  
ggarrange(denplot, qqplot, ncol=2, nrow=1)

## Warning: Removed 1726 rows containing non-finite values (stat\_density).

## Warning: Removed 1726 rows containing non-finite values (stat\_qq).

## Warning: Removed 1726 rows containing non-finite values (stat\_qq\_line).  
  
## Warning: Removed 1726 rows containing non-finite values (stat\_qq\_line).

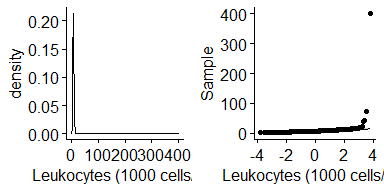


# Repeat for the next variable of interest  
denplot <- ggdensity(df$Leukocytes, xlab = "Leukocytes (1000 cells/uL)")  
qqplot <- ggqqplot(df$Leukocytes, xlab = "Leukocytes (1000 cells/uL)")  
  
ggarrange(denplot, qqplot, ncol=2, nrow=1)

## Warning: Removed 1726 rows containing non-finite values (stat\_density).

## Warning: Removed 1726 rows containing non-finite values (stat\_qq).

## Warning: Removed 1726 rows containing non-finite values (stat\_qq\_line).  
  
## Warning: Removed 1726 rows containing non-finite values (stat\_qq\_line).

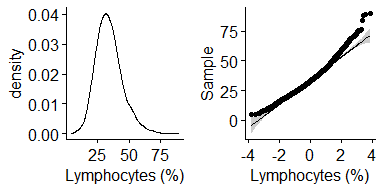


# Repeat for the next variable of interest  
denplot <- ggdensity(df$Lymphocytes, xlab = "Lymphocytes (%)")  
qqplot <- ggqqplot(df$Lymphocytes, xlab = "Lymphocytes (%)")  
  
ggarrange(denplot, qqplot, ncol=2, nrow=1)

## Warning: Removed 1731 rows containing non-finite values (stat\_density).

## Warning: Removed 1731 rows containing non-finite values (stat\_qq).

## Warning: Removed 1731 rows containing non-finite values (stat\_qq\_line).  
  
## Warning: Removed 1731 rows containing non-finite values (stat\_qq\_line).



#### Results: The distributions of the RBCs, Leukocytes, and Lymphocytes variables appear roughly normal with a few unusually high values.

# Remove outlier

There is one extremely high value in the Leukocytes variable data that is not at all comparable to the other values. Let’s exclude that subject from our analysis.

# Return the value in column 1 that has the maximum value in the Leukocytes column in the dfad data set  
# This relies on indexing, which means referring to a specific position in the dataset  
# In R, indexing takes the format: data[row#, column#]  
print(dfad[which.max(dfad$Leukocytes),1])

## Respondent sequence number   
## [1] 102389

# Remove the row with the outlier (keep all data that do not match the selected ID number in the seqn column)  
dfad <- dfad %>% filter(seqn != "102389")

# Is there a sex difference in Hepatitis C RNA positivity?

Let’s run a chi-square test to address this question.

# Chi-square test  
stats::chisq.test(x=dfad$Sex, y=dfad$HepC\_RNA)

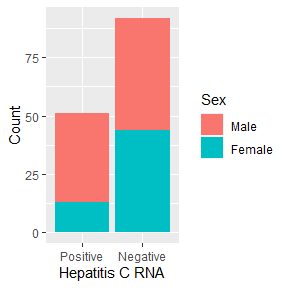
##   
## Pearson's Chi-squared test  
##   
## data: dfad$Sex and dfad$HepC\_RNA  
## X-squared = 15.071, df = 2, p-value = 0.0005338

# Let's focus on just the positives and negatives in people who got an RNA test  
# Let's also just drop the NAs in the HepC\_RNA column to simplify graphing  
dfadHepC <- dfad %>% drop\_na(HepC\_RNA) %>% filter(HepC\_RNA != "Negative Ab Screening")  
  
# Run the Chi-square test on just the RNA test subjects  
stats::chisq.test(x=dfadHepC$Sex, y=dfadHepC$HepC\_RNA)

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: dfadHepC$Sex and dfadHepC$HepC\_RNA  
## X-squared = 5.9286, df = 1, p-value = 0.0149

Let’s make a graph to visualize this comparison.

ggplot(data=dfadHepC, mapping=aes(x=HepC\_RNA, fill=Sex)) +  
 geom\_bar() +  
 xlab("Hepatitis C RNA") +  
 ylab("Count")



# Save this plot  
ggsave("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/HepCRNA\_by\_Sex.pdf")

## Saving 3 x 3 in image

### Results: HepC RNA positivity by sex

#### Positivity in an RNA test for hepatitis C does differ significantly by sex. Males are more likely than females to test positive for hepatitis C RNA.

# Do leukocyte counts differ in subjects who test positive for hepatitis C from those that test negative?

Let’s run a one-way ANOVA to address this question.

# One-way ANOVA   
res <- aov(Leukocytes ~ HepC\_RNA, dfad)  
summary(res)

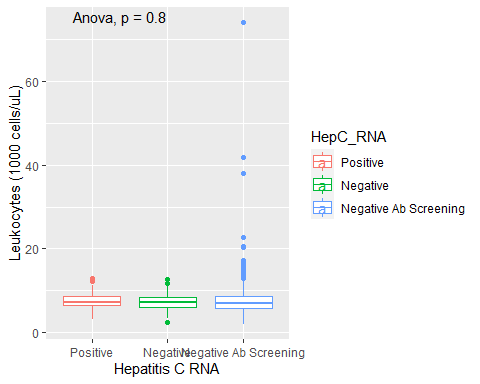
## Df Sum Sq Mean Sq F value Pr(>F)  
## HepC\_RNA 2 3 1.287 0.218 0.804  
## Residuals 5125 30246 5.902   
## 727 observations deleted due to missingness

Let’s make a graph to visualize this comparison.

# Remove NAs in HepC\_RNA variable for graphing purposes  
dfadHepC2 <- dfad %>% drop\_na(HepC\_RNA)  
  
ggplot(data=dfadHepC2, mapping=aes(x=HepC\_RNA, y=Leukocytes, color=HepC\_RNA)) +  
 geom\_boxplot() +  
 xlab("Hepatitis C RNA") +  
 ylab("Leukocytes (1000 cells/uL)") +  
 stat\_compare\_means(method="aov")

## Warning: Removed 3 rows containing non-finite values (stat\_boxplot).

## Warning: Removed 3 rows containing non-finite values (stat\_compare\_means).



ggsave("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/Leukocytes\_by\_HepCRNA.pdf")

## Saving 5 x 4 in image

## Warning: Removed 3 rows containing non-finite values (stat\_boxplot).  
  
## Warning: Removed 3 rows containing non-finite values (stat\_compare\_means).

### Results: Leukocyte concentrations by HepC RNA result

#### Leukocyte concentrations do not differ significantly based on hepatitis C RNA positivity status.

# Do lymphocyte and RBC levels correlate with one another?

Let’s do a Pearson correlation to address this question.

# Pearson correlation   
cor.test(x=dfad$Lymphocytes, y=dfad$RBCs, method="pearson")

##   
## Pearson's product-moment correlation  
##   
## data: dfad$Lymphocytes and dfad$RBCs  
## t = 0.4836, df = 5264, p-value = 0.6287  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.02034841 0.03366933  
## sample estimates:  
## cor   
## 0.006665322

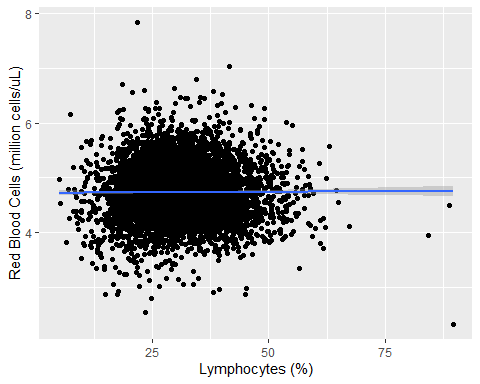
Let’s make a graph to visualize this comparison.

ggplot(data=dfad, mapping=aes(x=Lymphocytes, y=RBCs)) +  
 geom\_point() +  
 xlab("Lymphocytes (%)") +  
 ylab("Red Blood Cells (million cells/uL)") +  
 geom\_smooth(method="lm")

## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 589 rows containing non-finite values (stat\_smooth).

## Warning: Removed 589 rows containing missing values (geom\_point).



stat\_cor(method="pearson")

## geom\_text: parse = TRUE, na.rm = FALSE  
## stat\_cor: label.x.npc = left, label.y.npc = top, label.x = NULL, label.y = NULL, label.sep = , , method = pearson, alternative = two.sided, output.type = expression, digits = 2, r.digits = 2, p.digits = 2, r.accuracy = NULL, p.accuracy = NULL, cor.coef.name = R, na.rm = FALSE  
## position\_identity

ggsave("C:/Users/mecho/Documents/Emory/Post-doc/Presentations/IntroToClinicalResearch\_GuestLecture/Lymphocytes\_by\_RBCs.pdf")

## Saving 5 x 4 in image  
## `geom\_smooth()` using formula 'y ~ x'

## Warning: Removed 589 rows containing non-finite values (stat\_smooth).  
  
## Warning: Removed 589 rows containing missing values (geom\_point).

### Results: Lymphocyte and RBC concentrations

#### Lymphocyte and RBC concentrations are not correlated with one another.