Applied Linear Modeling

September 3, 2019

To-do

- With your table:
  - *pick a to-do number*
  - *compare your exercise results with the others at the table*
  - *agree on a set of results and work together to write them on the board*
  - *when finished, put all your names on the to-do number and drop in Done jar*
- Create a week-2 folder on your laptop and put the following files in it (from GitHub):
  - *week-2-workshop.Rmd*
  - *nhanes_2011_2012_ch3.csv*
- Install the following R packages:
  - *tableone*

All the things for today

- Discussion of exercises
- Workshop
  - *Bivariate review flow chart activity*
  - *Data management: selecting cases and variables, adding labels*
  - *Descriptive statistics review & conducting in R*
  - *Null hypothesis significance testing (NHST)*
  - *Chi-squared review & R code*
  - *One-sample t-test review & R code*
  - *One-way ANOVA review & R code*

Import and review the data for today

- Note that the file extension should be **.csv** for this code to work

```
# import the data using read.csv
nhanes.2012 <- read.csv(file = "nhanes_2011_2012_ch3.csv")

# examine the data set
summary(object = nhanes.2012)
```

```
##       SEQN            cycle         SDDSRVYR    RIDSTATR
##  Min.   :62161   2011-2012:9364   Min.   :7   Min.   :1.000
##  1st Qu.:64599                    1st Qu.:7   1st Qu.:2.000
##  Median :67025                    Median :7   Median :2.000
##  Mean   :67029                    Mean   :7   Mean   :1.956
##  3rd Qu.:69457                    3rd Qu.:7   3rd Qu.:2.000
##  Max.   :71916                    Max.   :7   Max.   :2.000
##
##     RIAGENDR        RIDAGEYR        RIDAGEMN        RIDRETH1
##  Min.   :1.000   Min.   : 1.00   Min.   :12.00   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:11.00   1st Qu.:15.00   1st Qu.:3.000
##  Median :2.000   Median :28.00   Median :18.00   Median :3.000
##  Mean   :1.502   Mean   :32.72   Mean   :17.94   Mean   :3.242
##  3rd Qu.:2.000   3rd Qu.:53.00   3rd Qu.:21.00   3rd Qu.:4.000
##  Max.   :2.000   Max.   :80.00   Max.   :24.00   Max.   :5.000
##                                  NA's   :9130
##     RIDRETH3        RIDEXMON        RIDEXAGY        RIDEXAGM
##  Min.   :1.000   Min.   :1.000   Min.   : 2.000   Min.   : 12.0
##  1st Qu.:3.000   1st Qu.:1.000   1st Qu.: 5.000   1st Qu.: 57.0
##  Median :3.000   Median :2.000   Median : 9.000   Median :109.0
##  Mean   :3.454   Mean   :1.516   Mean   : 9.641   Mean   :114.5
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:14.000   3rd Qu.:167.0
##  Max.   :7.000   Max.   :2.000   Max.   :20.000   Max.   :239.0
##                  NA's   :408     NA's   :5946     NA's   :5737
##     DMQMILIZ        DMQADFC         DMDBORN4        DMDCITZN
##  Min.   :1.000   Min.   :1.000   Min.   : 1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 1.00   1st Qu.:1.000
##  Median :2.000   Median :1.000   Median : 1.00   Median :1.000
##  Mean   :1.908   Mean   :1.501   Mean   : 1.27   Mean   :1.128
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.: 1.00   3rd Qu.:1.000
##  Max.   :2.000   Max.   :9.000   Max.   :99.00   Max.   :7.000
##  NA's   :3357    NA's   :8813                    NA's   :5
##     DMDYRSUS        DMDEDUC3        DMDEDUC2        DMDMARTL
```

```
##  Min.   : 1.000   Min.   : 0.000   Min.   :1.000   Min.   : 1.000
##  1st Qu.: 3.000   1st Qu.: 2.000   1st Qu.:3.000   1st Qu.: 1.000
##  Median : 5.000   Median : 5.000   Median :4.000   Median : 2.000
##  Mean   : 7.437   Mean   : 6.038   Mean   :3.467   Mean   : 2.749
##  3rd Qu.: 6.000   3rd Qu.: 9.000   3rd Qu.:5.000   3rd Qu.: 5.000
##  Max.   :99.000   Max.   :66.000   Max.   :9.000   Max.   :99.000
##  NA's   :7292     NA's   :6765     NA's   :3804    NA's   :3804
##     RIDEXPRG        SIALANG         SIAPROXY        SIAINTRP
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000
##  Median :2.000   Median :1.000   Median :2.000   Median :2.000
##  Mean   :2.023   Mean   :1.124   Mean   :1.654   Mean   :1.965
##  3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :3.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##  NA's   :8156                    NA's   :4
##     FIALANG         FIAPROXY        FIAINTRP        MIALANG
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :1.000   Median :2.000   Median :2.000   Median :1.000
##  Mean   :1.079   Mean   :1.998   Mean   :1.969   Mean   :1.053
##  3rd Qu.:1.000   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :2.000   Max.   :2.000   Max.   :2.000   Max.   :2.000
##  NA's   :99      NA's   :99      NA's   :99      NA's   :2651
##     MIAPROXY        MIAINTRP        AIALANGA        WTINT2YR
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :  3601
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.: 11762
##  Median :2.000   Median :2.000   Median :1.000   Median : 18490
##  Mean   :1.994   Mean   :1.969   Mean   :1.114   Mean   : 32348
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.000   3rd Qu.: 35510
##  Max.   :2.000   Max.   :2.000   Max.   :3.000   Max.   :220233
##  NA's   :2651    NA's   :2651    NA's   :3610
##     WTMEC2YR         SDMVPSU         SDMVSTRA         INDHHIN2
##  Min.   :     0   Min.   :1.000   Min.   : 90.00   Min.   : 1.00
##  1st Qu.: 11582   1st Qu.:1.000   1st Qu.: 92.00   1st Qu.: 5.00
##  Median : 18606   Median :2.000   Median : 96.00   Median : 7.00
##  Mean   : 32348   Mean   :1.644   Mean   : 95.88   Mean   :11.53
##  3rd Qu.: 36132   3rd Qu.:2.000   3rd Qu.: 99.00   3rd Qu.:14.00
##  Max.   :222580   Max.   :3.000   Max.   :103.00   Max.   :99.00
##                                                    NA's   :78
##     INDFMIN2        INDFMPIR        DMDHHSIZ        DMDFMSIZ
##  Min.   : 1.0   Min.   :0.000   Min.   :1.00   Min.   :1.000
##  1st Qu.: 4.0   1st Qu.:0.870   1st Qu.:2.00   1st Qu.:2.000
##  Median : 7.0   Median :1.640   Median :4.00   Median :4.000
##  Mean   :11.1   Mean   :2.218   Mean   :3.73   Mean   :3.556
##  3rd Qu.:14.0   3rd Qu.:3.615   3rd Qu.:5.00   3rd Qu.:5.000
##  Max.   :99.0   Max.   :5.000   Max.   :7.00   Max.   :7.000
##  NA's   :49     NA's   :805
##     DMDHHSZA        DMDHHSZB        DMDHHSZE        DMDHRGND
##  Min.   :0.0000   Min.   :0.000   Min.   :0.0000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:0.0000   1st Qu.:1.000
##  Median :0.0000   Median :1.000   Median :0.0000   Median :1.000
##  Mean   :0.4852   Mean   :0.946   Mean   :0.4076   Mean   :1.492
##  3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:2.000
```

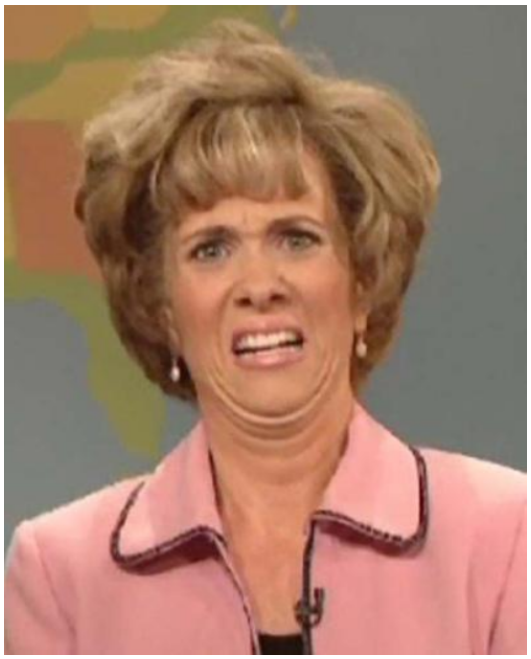


```
##  Max.   :3.0000   Max.   :4.000   Max.   :3.0000   Max.   :2.000
##
##     DMDHRAGE        DMDHRBR4        DMDHREDU        DMDHRMAR
##  Min.   :18.00   Min.   : 1.000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:34.00   1st Qu.: 1.000   1st Qu.:2.000   1st Qu.: 1.000
##  Median :43.00   Median : 1.000   Median :4.000   Median : 1.000
##  Mean   :45.85   Mean   : 1.435   Mean   :3.434   Mean   : 3.166
##  3rd Qu.:57.00   3rd Qu.: 2.000   3rd Qu.:4.000   3rd Qu.: 5.000
##  Max.   :80.00   Max.   :99.000   Max.   :9.000   Max.   :99.000
##                  NA's   :361      NA's   :358     NA's   :128
##     DMDHSEDU        AUQ054          AUQ060          AUQ070
##  Min.   :1.000   Min.   : 1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:3.000   1st Qu.: 1.000   1st Qu.:1.000   1st Qu.:1.000
##  Median :4.000   Median : 2.000   Median :1.000   Median :1.000
##  Mean   :3.594   Mean   : 1.929   Mean   :1.345   Mean   :1.301
##  3rd Qu.:5.000   3rd Qu.: 2.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :99.000   Max.   :9.000   Max.   :9.000
##  NA's   :4716    NA's   :1        NA's   :6459    NA's   :8587
##     AUQ080          AUQ090          AUQ100          AUQ110
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:4.000   1st Qu.:4.000
##  Median :1.000   Median :1.000   Median :5.000   Median :5.000
##  Mean   :1.286   Mean   :1.463   Mean   :4.145   Mean   :4.585
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:5.000
##  Max.   :9.000   Max.   :2.000   Max.   :9.000   Max.   :9.000
##  NA's   :9151    NA's   :9310    NA's   :4689    NA's   :4689
##     AUQ136          AUQ138          AUQ144          AUQ146
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.:3.000   1st Qu.:2.000
##  Median :2.000   Median :2.000   Median :4.000   Median :2.000
##  Mean   :2.019   Mean   :2.001   Mean   :3.838   Mean   :1.987
##  3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:5.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :2.000
##  NA's   :3805    NA's   :3805    NA's   :4689    NA's   :4689
##     AUD148          AUQ152          AUQ154          AUQ191
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:2.000
##  Median :1.000   Median :3.000   Median :2.000   Median :2.000
##  Mean   :1.032   Mean   :3.148   Mean   :1.985   Mean   :1.857
##  3rd Qu.:1.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :2.000   Max.   :5.000   Max.   :2.000   Max.   :9.000
##  NA's   :9301    NA's   :9303    NA's   :4689    NA's   :4689
##     AUQ250          AUQ255          AUQ260          AUQ270
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:2.000   1st Qu.:1.000
##  Median :3.000   Median :3.000   Median :2.000   Median :2.000
##  Mean   :3.219   Mean   :3.016   Mean   :1.883   Mean   :1.617
##  3rd Qu.:5.000   3rd Qu.:5.000   3rd Qu.:2.000   3rd Qu.:2.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##  NA's   :8680    NA's   :8680    NA's   :8680    NA's   :8680
##     AUQ280          AUQ300          AUQ310          AUQ320
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000
```



```
##  Median :2.000   Median :2.000   Median :2.000   Median :3.000
##  Mean   :2.155   Mean   :1.656   Mean   :2.113   Mean   :3.058
##  3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:5.000
##  Max.   :5.000   Max.   :7.000   Max.   :9.000   Max.   :9.000
##  NA's   :8680    NA's   :4689    NA's   :7751    NA's   :7751
##     AUQ330          AUQ340          AUQ350          AUQ360
##  Min.   :1.000   Min.   : 1.00   Min.   :1.000   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.: 3.00   1st Qu.:1.000   1st Qu.: 3.00
##  Median :2.000   Median : 5.00   Median :1.000   Median : 5.00
##  Mean   :1.687   Mean   : 4.77   Mean   :1.366   Mean   : 4.65
##  3rd Qu.:2.000   3rd Qu.: 7.00   3rd Qu.:2.000   3rd Qu.: 7.00
##  Max.   :3.000   Max.   :99.00   Max.   :9.000   Max.   :99.00
##  NA's   :4689    NA's   :7828    NA's   :7828    NA's   :8383
##     AUQ370          AUQ380        file_name       begin_year
##  Min.   :1.000   Min.   : 1.000   AUQ_G:9364   Min.   :2011
##  1st Qu.:2.000   1st Qu.: 5.000                1st Qu.:2011
##  Median :2.000   Median : 5.000                Median :2011
##  Mean   :1.883   Mean   : 4.682                Mean   :2011
##  3rd Qu.:2.000   3rd Qu.: 6.000                3rd Qu.:2011
##  Max.   :2.000   Max.   :77.000                Max.   :2011
##  NA's   :4689    NA's   :4689
##     end_year
##  Min.   :2012
##  1st Qu.:2012
##  Median :2012
##  Mean   :2012
##  3rd Qu.:2012
##  Max.   :2012
##
```


That's a lot of variables and observations



Research questions we can answer with this data

- Are age and sex associated with gun use?
- Are age and sex associated with frequency of gun use among those who have used a gun?
- Is the mean age of a gun user the same as the mean age of all participants?

Data management topic: selecting cases and variables

- Firearm use:
  - *AUQ300: Every used firearms for any reason?*
  - *AUQ310: How many total rounds ever fired?*
- Demographics:
  - *RIAGENDR: Gender of the participant.*
  - *RIDAGEYR: Age in years at screening.*

Let's select the four variables listed and all the cases where AUQ300 (gun use) was answered either Yes or No.

Codebook for AUQ300 and AUQ310

AUQ300 - Ever used firearms for any reason?

Variable Name: AUQ300

SAS Label: Ever used firearms for any reason?

English Text: This next question is about {your/SP's} use of firearms that {you/he/she} may have used for target shooting, hunting, for {your/his/her} job or in military service. {Have you/Has SP} ever used firearms for any reason?

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Yes	1613	1613	
2	No	3061	4674	AUQ330
7	Refused	1	4675	AUQ330
9	Don't know	0	4675	AUQ330
.	Missing	4689	9364	

AUQ310 - How many total rounds ever fired?

Variable Name: AUQ310

SAS Label: How many total rounds ever fired?

English Text: How many total rounds {have you/has SP} ever fired?

English Instructions: READ CATEGORIES IF NECESSARY INTERVIEWER: ONE ROUND EQUALS ONE SHOT. INCLUDE TARGET SHOOTING, HUNTING, YOUR JOB AND MILITARY SERVICE.

Target: Both males and females 20 YEARS - 69 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	1 to less than 100 rounds	701	701	
2	100 to less than 1000 rounds	423	1124	
3	1000 to less than 10,000 rounds	291	1415	
4	10,000 to less than 50,000 rounds	106	1521	
5	50,000 rounds or more	66	1587	
7	Refused	0	1587	
9	Don't know	26	1613	
.	Missing	7751	9364	

Data management topic: selecting cases and variables

```
# open the tidyverse for data management
library(package = "tidyverse")

# make a smaller data frame with four variables
nhanes.2012.cleaned <- nhanes.2012 %>%
    select(AUQ300, AUQ310, RIAGENDR, RIDAGEYR)

# check the new data frame
summary(object = nhanes.2012.cleaned)
```

```
##      AUQ300          AUQ310          RIAGENDR        RIDAGEYR
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:11.00
##  Median :2.000   Median :2.000   Median :2.000   Median :28.00
##  Mean   :1.656   Mean   :2.113   Mean   :1.502   Mean   :32.72
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:53.00
##  Max.   :7.000   Max.   :9.000   Max.   :2.000   Max.   :80.00
##  NA's   :4689    NA's   :7751
```

Data management topic: selecting cases and variables

```
# add code to keep values of AUQ300 that are Yes and No
nhanes.2012.cleaned <- nhanes.2012 %>%
    select(AUQ300, AUQ310, RIAGENDR, RIDAGEYR) %>%
    filter(AUQ300 <= 2)

# check the new data frame
summary(object = nhanes.2012.cleaned)
```

```
##      AUQ300          AUQ310         RIAGENDR        RIDAGEYR
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :20.00
##  1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:31.00
##  Median :2.000   Median :2.000   Median :2.000   Median :43.00
##  Mean   :1.655   Mean   :2.113   Mean   :1.506   Mean   :43.74
##  3rd Qu.:2.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:56.00
##  Max.   :2.000   Max.   :9.000   Max.   :2.000   Max.   :69.00
##                  NA's   :3061
```


Codebook for RIAGENDR

RIAGENDR - Gender

Variable Name: RIAGENDR

SAS Label: Gender

English Text: Gender of the participant.

Target: Both males and females 0 YEARS - 150 YEARS

Code or Value	Value Description	Count	Cumulative	Skip to Item
1	Male	4856	4856	
2	Female	4900	9756	
.	Missing	0	9756	

Data cleaning: Fix data types and add labels

```
# add to the smaller data frame with four variables
# to keep values of AUQ300 that are Yes and No
nhanes.2012.cleaned <- nhanes.2012 %>%
    select(AUQ300, AUQ310, RIAGENDR, RIDAGEYR) %>%
    filter(AUQ300 <= 2) %>%
    mutate(AUQ300 = recode_factor(.x = AUQ300,
                                  `1` = 'Yes',
                                  `2` = 'No')) %>%
    mutate(AUQ310 = recode_factor(.x = AUQ310,
                                  `1` = "1 to less than 100",
                                  `2` = "100 to less than 1000",
                                  `3` = "1000 to less than 10k",
                                  `4` = "10k to less than 50k",
                                  `5` = "50k or more",
                                  `7` = "Don't know",
                                  `9` = "Refused")) %>%
    mutate(RIAGENDR = recode_factor(.x = RIAGENDR,
                                    `1` = 'Male',
                                    `2` = 'Female')) %>%
    rename(gun.use = AUQ300) %>%
    rename(rounds.fired = AUQ310) %>%
    rename(sex = RIAGENDR) %>%
    rename(age = RIDAGEYR)


# check the recoding
summary(object = nhanes.2012.cleaned)
```


Data cleaning: Fix data types and add labels

```
##  gun.use                  rounds.fired      sex           age
##  Yes:1613   1 to less than 100   : 701   Male  :2311   Min.   :20.00
##  No :3061   100 to less than 1000: 423   Female:2363   1st Qu.:31.00
##             1000 to less than 10k: 291                 Median :43.00
##             10k to less than 50k : 106                 Mean   :43.74
##             50k or more          :  66                 3rd Qu.:56.00
##             NA's                 :3087                 Max.   :69.00
```

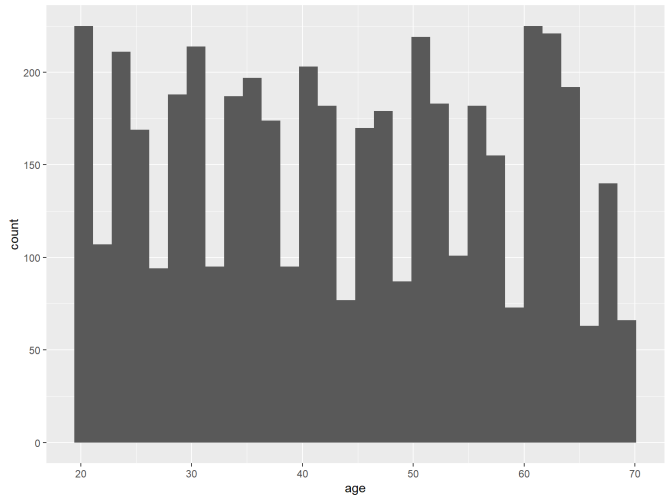
Descriptive statistics review & conducting in R

- Continuous variables
  - *Mean and standard deviation*
  - *Median and interquartile range (IQR)*
- Categorical variables
  - *Frequencies and percentages*

Choosing mean or median

- Means are useful when a variable is normally distributed (or close to normal)
- Medians are useful when a variable is not normally distributed

```
# examine age distribution
nhanes.2012.cleaned %>%
    ggplot(aes(x = age)) +
    geom_histogram()
```



Find the mean, median, std dev, IQR in tidyverse

- The median and IQR would be more appopriate given the distribution
- The `summarize()` function can compute descriptive statistics for continuous variables

```
# descriptive statistics for continuous variables
nhanes.2012.cleaned %>%
    drop_na(age) %>%
    summarize(mean.age = mean(x = age),
              sd.age = sd(x = age),
              med.age = median(x = age),
              iqr.age = IQR(x = age))
```

```
##   mean.age   sd.age med.age iqr.age
## 1 43.74369 14.39535      43      25
```


Interpreting the results

Participants in the 2012 NHANES survey who responded Yes or No to the gun use question had a median age of 43 years old (IQR = 25).

Find frequencies and percentages in base R

- Base R functions can be used to get simple frequencies and percents

```
# descriptive statistics for categorical variables
table(nhanes.2012.cleaned$gun.use)
```

```
##
##  Yes   No
## 1613 3061
```

```
prop.table(x = table(nhanes.2012.cleaned$gun.use))
```

```
##
##       Yes        No
## 0.3451006 0.6548994
```


Find frequencies and percentages in tidyverse

- To stay consistent with tidyverse, use `group_by()`

```
# descriptive stats for categorical in tidyverse
nhanes.2012.cleaned %>%
  group_by(gun.use) %>%
  summarize(freq.gun.use = n()) %>%
  mutate(perc.gun.use = 100*(freq.gun.use / sum(freq.gun.use)))
```

```
## # A tibble: 2 x 3
##   gun.use freq.gun.use perc.gun.use
##   <fct>          <int>        <dbl>
## 1 Yes             1613         34.5
## 2 No              3061         65.5
```


Interpreting the frequencies and percentages

Fewer participants in the NHANES 2012 survey had ever used a gun (n = 1613; 34.5%) compared to not ever used a gun (n = 3061; 65.5%).

Using tableone to get all the descriptive stats

```
# open tableone package
library(package = "tableone")
```

```
# create and print table
gun.use.table <- CreateTableOne(data = nhanes.2012.cleaned)
print(x = gun.use.table)
```

```
##
##                          Overall
##   n                       4674
##   gun.use = No (%)        3061 (65.5)
##   rounds.fired (%)
##      1 to less than 100    701 (44.2)
##      100 to less than 1000 423 (26.7)
##      1000 to less than 10k 291 (18.3)
##      10k to less than 50k  106 ( 6.7)
##      50k or more            66 ( 4.2)
##   sex = Female (%)        2363 (50.6)
##   age (mean (SD))        43.74 (14.40)
```

Update the table with median instead of mean

```
# create table
gun.use.table <- CreateTableOne(data = nhanes.2012.cleaned)

# show all levels for categorical
# specify non-normal age variable
print(x = gun.use.table,
      showAllLevels = TRUE,
      nonnormal = 'age')
```

```
##
##                      level                Overall
##   n                                        4674
##   gun.use (%)        Yes                   1613 (34.5)
##                      No                    3061 (65.5)
##   rounds.fired (%)   1 to less than 100     701 (44.2)
##                      100 to less than 1000  423 (26.7)
##                      1000 to less than 10k  291 (18.3)
##                      10k to less than 50k   106 ( 6.7)
##                      50k or more             66 ( 4.2)
##   sex (%)            Male                  2311 (49.4)
##                      Female                2363 (50.6)
##   age (median [IQR])                       43.00 [31.00, 56.00]
```


Null Hypothesis Significance Testing

- Step 1: Write the null and alternate hypothesis
- Step 2: Calculate the test statistic
- Step 3: Calculate the probability that your test statistic is at least as big as it is if there were no relationship (i.e., if the null hypothesis is true)
- Step 4: If the probability that the null is true is very small (usually less than 5%) reject the null hypothesis
- Step 5: If the probability that the null is true is not small (usually 5% or greater) retain the null hypothesis

Step 1: Chi-squared null and alternate hypotheses

- Chi-squared is usually used to examine associations between *two categorical variables*

Research question: Is sex associated with gun use? Is there a difference between males and females in having used a gun?

- H0: There is no association between sex and gun use
- HA: There is an association between sex and gun use

Step 2: Chi-squared test statistic

```
# compute test statistic with CrossTable
chisq.test(x = nhanes.2012.cleaned$sex,
           y = nhanes.2012.cleaned$gun.use)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  nhanes.2012.cleaned$sex and nhanes.2012.cleaned$gun.use
## X-squared = 532.48, df = 1, p-value < 2.2e-16
```

Step 3: Calculate the probability that your test statistic is at least this big

- The probability that your chi-squared statistic is 532.48 or bigger under the null hypothesis is shown in the output as p < 2.2e-16
- This is scientific notation for .00000000000000022
- So, there is a tiny probability that you'd have a chi-squared of 532.48 or bigger *if there were no association between sex and gun use*

Step 4 & 5: Reject or retain the null hypothesis

The null hypothesis that there is no association between sex and gun use is rejected (chi-squared = 532.48; p < .001). There is a statistically significant association between sex and gun use.

Ok, but what is the relationship?

- Examine percentages or make a graph to demonstrate

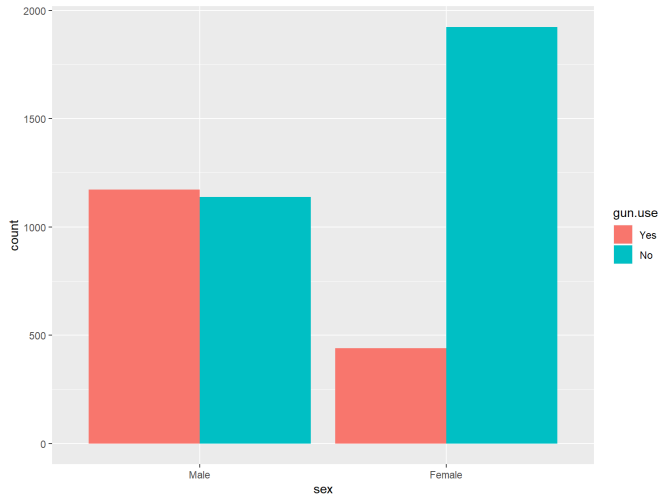
```
# gun use by sex percentages
prop.table(x = table(nhanes.2012.cleaned$sex,
                     nhanes.2012.cleaned$gun.use),
           margin = 1)
```

```
##
##                Yes        No
##   Male   0.5075725 0.4924275
##   Female 0.1862040 0.8137960
```

Among NHANES 2012 participants, a higher percentage of males (50.8%) compared to females (18.6%) had ever used a gun.

Ok, but what is the relationship? (graph)

```
# gun use by sex
nhanes.2012.cleaned %>%
    ggplot(aes(x = sex, fill = gun.use)) +
    geom_bar(position = "dodge")
```



Chi-squared assumptions

- Both variables are categorical (either nominal or ordinal)
  - *If not met, use another statistical test*
- A minimum of 5 observations in at least 80% of groups
  - *If not met, use Fisher's Exact Test (in R: `fisher.test()`)*
- Independent observations
  - *If not met, McNemar's test if the variables are binary and the observations are paired*
  - *If not met, Cochran's Q-test if one variable is binary and the other has 3+ categories and the observations are paired*
  - *If not met, clean up the data and use descriptives if this results from sloppy data collection (e.g., mistakenly surveyed the same person twice)*

One-sample t-test review

- One-sample t-tests are used to compare a mean from a sample to a hypothesized or population mean

Research question: Is the mean age of gun users the same as the mean age of everyone (m_{age} = 43.7)?

Step 1: One-sample t-test null and alternate hypotheses

H0: There is no difference between the mean age of gun users and the mean age of 43.7 years old in the full sample (i.e., gun users have a mean age of 43.7 years old)

HA: There is a difference between the mean age of gun users and the mean age of 43.7 years old in the full sample (i.e., gun users *do not* have a mean age of 43.7 years old)

Step 2: Calculate the test statistic

```
# select the gun users from the clean data
nhanes.2012.gun.users <- nhanes.2012.cleaned %>%
    filter(gun.use == "Yes")

# conduct the t-test
t.test(x = nhanes.2012.gun.users$age, mu = 43.7)
```

```
##
##  One Sample t-test
##
## data:  nhanes.2012.gun.users$age
## t = 0.26907, df = 1612, p-value = 0.7879
## alternative hypothesis: true mean is not equal to 43.7
## 95 percent confidence interval:
##  43.09988 44.49094
## sample estimates:
## mean of x
##  43.79541
```

Step 3: Calculate the probability that your test statistic is at least this big

- The probability that your t-statistic is .27 under the null hypothesis is shown in the output as p = .79
- So, there is a large probability that you'd have a t-statistic this big or bigger *if the mean age was no different from 43.7* (i.e., the null hypothesis was true)

Step 4 & 5: Reject or retain the null hypothesis

- The null hypothesis is retained. The mean age of gun users was 43.8 years old. The mean age of gun users was *not* statistically significantly different from the hypothesized mean age of 43.7 years old (t = .27; p = .79).

Assumptions of one-sample t-test

- The continuous variable is normally distributed
  - *If failed, use the **sign test** to compare medians instead*
- Independent observations
  - *If failed, use descriptive statistics instead*

One-way ANOVA review

- One-way ANOVA is used to compare means across more than two groups

Research question: Is mean age different by gun usage?

Step 1: ANOVA null and alternate hypotheses

H0: Mean age does not differ by gun usage

HA: Mean age differs by gun usage

Step 2: Calculate the test statistic

```
# conduct the ANOVA
oneway.test(formula = age ~ rounds.fired,
            data = nhanes.2012.gun.users,
            var.equal = TRUE)
```

```
##
##  One-way analysis of means
##
## data:  age and rounds.fired
## F = 2.5864, num df = 4, denom df = 1582, p-value = 0.03539
```


Step 3: Calculate the probability that your test statistic is at least this big

- The probability that your F-statistic is 2.59 under the null hypothesis is shown in the output as p = .035
- So, there is a small probability that you'd have a t-statistic this big or bigger *if mean age was no different from 43.7* (i.e., the null hypothesis was true)

Step 4 & 5: Reject or retain the null hypothesis

- The null hypothesis is rejected. The mean age of gun users differs by the number of rounds fired (F = 2.59; p = .035). ## Add some context
- Use mutate to get group means

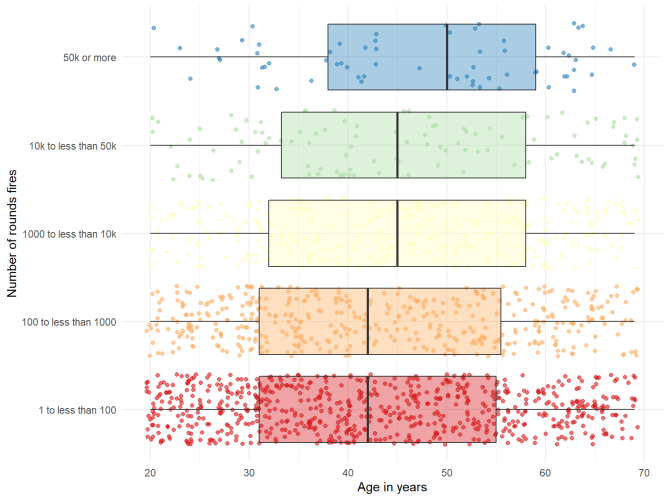
```
# means by group
nhanes.2012.gun.users %>%
    drop_na(age) %>%
    drop_na(rounds.fired) %>%
    group_by(rounds.fired) %>%
    summarize(mean.age = mean(age),
              sd.age = sd(age))
```

```
## # A tibble: 5 x 3
##   rounds.fired          mean.age sd.age
##   <fct>                    <dbl>  <dbl>
## 1 1 to less than 100        42.9   14.1
## 2 100 to less than 1000     43.2   14.2
## 3 1000 to less than 10k     44.9   14.2
## 4 10k to less than 50k      45.2   14.6
## 5 50k or more               47.2   13.2
```

Add some context

- Use ggplot to examine boxplots

```
# make a fancy graph
nhanes.2012.gun.users %>%
  drop_na(age) %>%
  drop_na(rounds.fired) %>%
  ggplot(aes(y = age, x = rounds.fired)) +
  geom_jitter(aes(color = rounds.fired), alpha = .6) +
  geom_boxplot(aes(fill = rounds.fired), alpha = .4) +
  scale_fill_brewer(palette = "Spectral", guide = FALSE) +
  scale_color_brewer(palette = "Spectral", guide = FALSE) +
  theme_minimal() +
  coord_flip() +
  labs(x = "Number of rounds fires", y = "Age in years")
```



ANOVA assumptions

- Continuous variable and three or more independent groups
- Independent observations
- Data are normally distributed by group
- Variances are equal by group (homogeneity of variances)

The End