STAT 351 HW 6

Christian Panici

4/3/2020

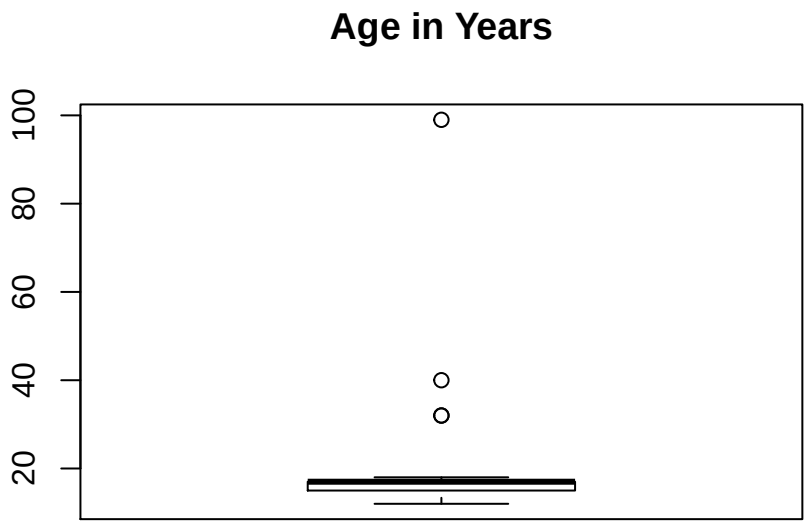
1. (a)

This dataset contains a variety of information about 9th-12th graders in Illinois. One characteristic I found interesting had to do with the age distribution of the students. While the majority of students fell within the same span of ages (which makes sense given their grades), there were a few outliers, including two 32-year olds and a 40-year old. There was also a 99-year old, but this seems more attributable to an error in the survey than a near-centenarian sitting in algebra class. This person also recorded 'Boat' as their mode of transportation, suggesting that they may not have taken the survey all that seriously.

Looking at means of transportation, there were a couple others that reported taking a boat to school, which I struggle to believe as a lifelong Illinois resident. Based on some of these results, we should be careful when working with this data, as it may not all be accurate.

Beyond these two I highlighted, there are a variety of other tidbits collected such as their favorite subjects, seasons, and activities as well as their feelings on issues like climate change. Certainly, there could be some interesting analysis done here given that the data is properly inspected for legitimacy.

```
boxplot(ilschool[6], main = "Age in Years")
```



```
summary(ilschool[12])
```

```
##                   Travel_to_School
##                                  : 21
##  Bicycle                         :  5
##  Boat                            :  6
##  Bus                             : 69
##  Car                             :346
##  Rail (Train/Tram/Subway):  7
##  Walk                            : 46
```

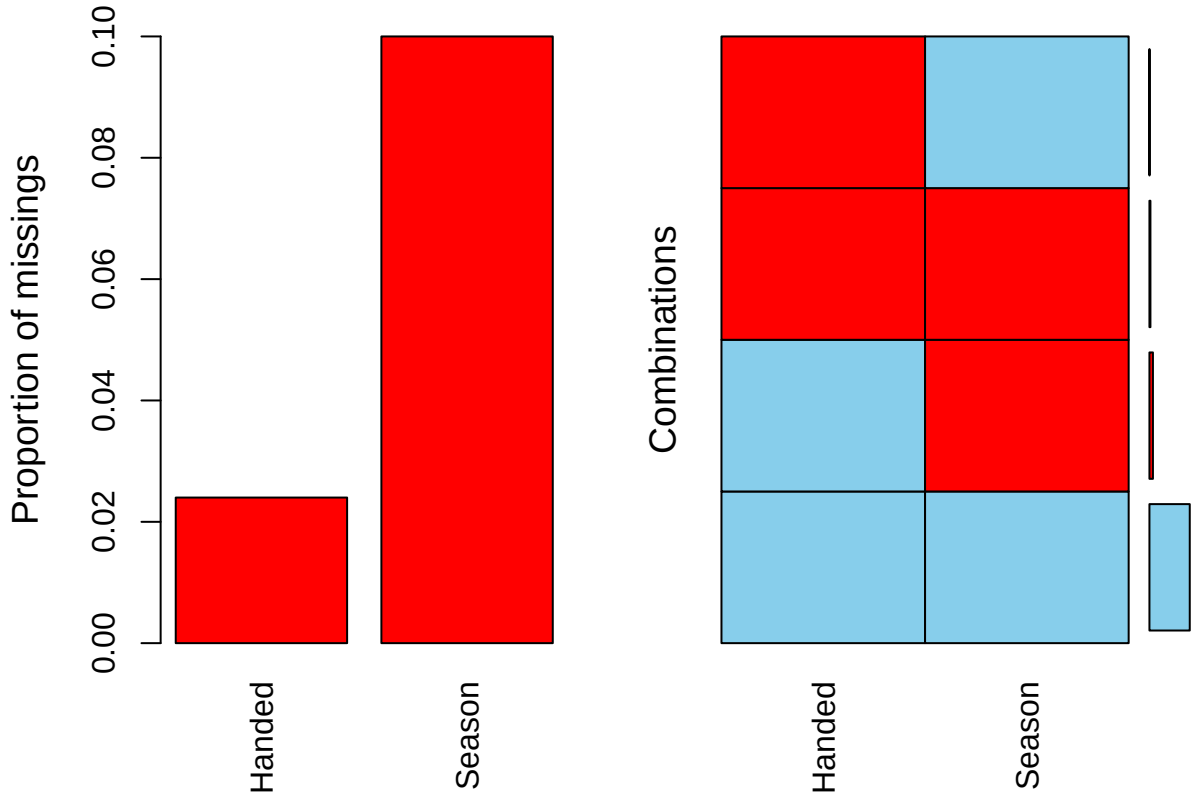
(b) H0: There is no significant association between handedness and favorite season.

HA: There is a significant association between handedness and favorite season.

It seems okay to remove rows with missing observations since we're still left with about 90% of the data left after removal. The missingness seems MCAR as well, unrelated to any particular variable.

We are performing a test for association with a contingency table here, but some of the cell frequencies are small, so we should use the chi-squared permutation test to calculate an exact distribution rather than use an approximation.

```
ilschool[ilschool == ""] = NA
hand_season <- data.frame(ilschool$Handed, ilschool$Favorite_Season)
colnames(hand_season) = c("Handed", "Season")
aggr(hand_season)
```

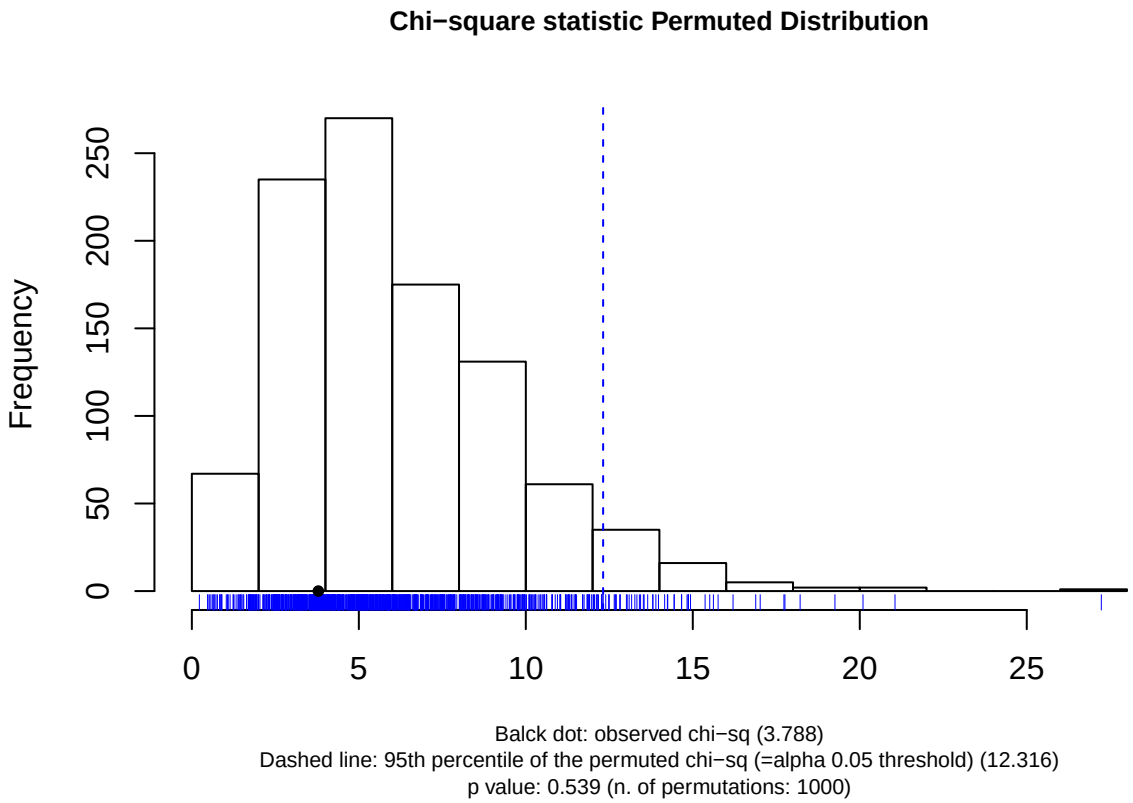


```r
missing_removed = hand_season[complete.cases(hand_season),]

tab = table(missing_removed$Handed, missing_removed$Season)
# Still getting extra rows for some reason
tab = tab[-1,-1]
tab
```

```
##
##                Fall Spring Summer Winter
##   Ambidextrous    7      2      4      3
##   Left-Handed    14      8     17      5
##   Right-Handed  157     47    146     39
```

```r
chiperm(tab,B = 1000)
```




```
## NULL
```

At alpha = .05, we fail to reject the null. There is not sufficient evidence to suggest a significant association between handedness and favorite season.

(c) I cleaned the data manually in Excel since there didn't seem to really be a standard pattern to the way that things were inputted incorrectly. Along with height and armspan, I included age and footspan since they might be helpful when imputing.

With CART imputation:

```
tempdata = mice(height_arm,maxit=50,meth='cart',seed=500)
```

```
fit = with(tempdata, lm(Height ~ Armspan))
pool1 = summary(pool(fit))
# Intercept Estimate
pool1$estimate[1]
```

```
## [1] 48.51369
```

```
# Slope Estimate
pool1$estimate[2]
```

```
## [1] 0.7159695
```

```
# Intercept Std Err
pool1$std.error[1]
```

```
## [1] 7.21623
```

```
#Slope Std Err
pool1$std.error[2]
```

```
## [1] 0.04276844
```

With Random Forest imputation:

```
tempdata2 = mice(height_arm,maxit=50,meth='rf',seed=500)
```

```
fit2 = with(tempdata2, lm(Height ~ Armspan))
pool2 = summary(pool(fit2))
# Intercept Estimate
pool2$estimate[1]
```

```
## [1] 47.43427
```

```
# Slope Estimate
pool2$estimate[2]
```

```
## [1] 0.7222957
```

```
# Intercept Std Err
pool2$std.error[1]
```

```
## [1] 6.77752
```

```r
#Slope Std Err
pool2$std.error[2]
```

```
## [1] 0.04040624
```