Problem\_Set\_3

Amy Richardson

3/16/2021

#Initiate packages

library(ggplot2)  
library(tidyr)  
library (dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(granova)

## Loading required package: car

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':  
##   
## recode

library(car)  
library(Rcmdr)

## Loading required package: splines

## Loading required package: RcmdrMisc

## Loading required package: sandwich

## Loading required package: effects

## Registered S3 methods overwritten by 'lme4':  
## method from  
## cooks.distance.influence.merMod car   
## influence.merMod car   
## dfbeta.influence.merMod car   
## dfbetas.influence.merMod car

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

## The Commander GUI is launched only in interactive sessions

##   
## Attaching package: 'Rcmdr'

## The following object is masked from 'package:base':  
##   
## errorCondition

library(pastecs)

##   
## Attaching package: 'pastecs'

## The following objects are masked from 'package:dplyr':  
##   
## first, last

## The following object is masked from 'package:tidyr':  
##   
## extract

library(multcomp)

## Loading required package: mvtnorm

## Loading required package: survival

## Loading required package: TH.data

## Loading required package: MASS

##   
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':  
##   
## select

##   
## Attaching package: 'TH.data'

## The following object is masked from 'package:MASS':  
##   
## geyser

library(compute.es)  
library(WRS2)  
library(gmodels)

Read in Data

sat\_df <- read.csv("SAT.csv", header = TRUE)

Change Character columns (Race.Ethnicity and Sex) to factors.

sat\_df <- sat\_df %>% mutate(Sex\_factor = as.factor(Sex))  
sat\_df <- sat\_df %>% mutate(Race\_factor = as.factor(Race.Ethnicity))  
sat\_df <- sat\_df %>% mutate(MATH = as.numeric(MATH))

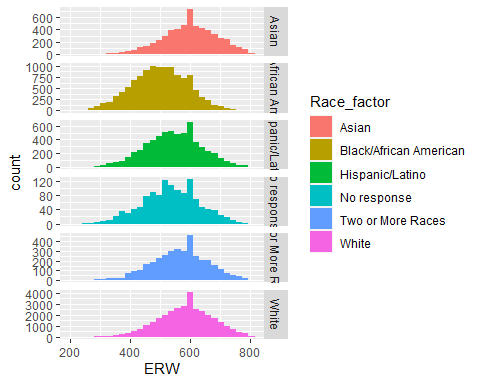
## Part 1

### There is no difference between scores on the English section of the SAT among racial/ethnic groups.

Plot histograms of English Scores among racial/ethnic groups

sat\_df %>%  
 ggplot(aes(x = ERW, fill = Race\_factor)) +  
 geom\_histogram () +  
 facet\_grid(Race\_factor ~., scales = "free")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



Run descriptive statistics for each group

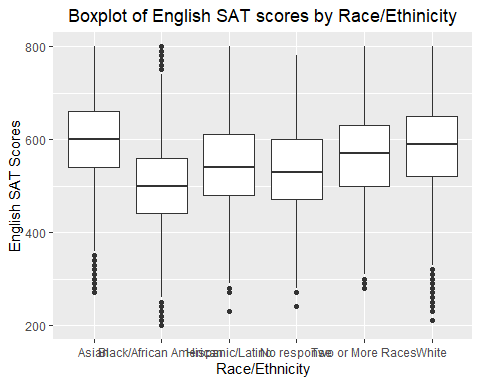
by(sat\_df$ERW, sat\_df$Race\_factor, stat.desc)

## sat\_df$Race\_factor: Asian  
## nbr.val nbr.null nbr.na min max range   
## 5.541000e+03 0.000000e+00 0.000000e+00 2.700000e+02 8.000000e+02 5.300000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 3.318810e+06 6.000000e+02 5.989551e+02 1.205387e+00 2.363031e+00 8.050839e+03   
## std.dev coef.var   
## 8.972647e+01 1.498050e-01   
## ------------------------------------------------------------   
## sat\_df$Race\_factor: Black/African American  
## nbr.val nbr.null nbr.na min max range   
## 1.169900e+04 0.000000e+00 0.000000e+00 2.000000e+02 8.000000e+02 6.000000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 5.848080e+06 5.000000e+02 4.998786e+02 8.602215e-01 1.686178e+00 8.657038e+03   
## std.dev coef.var   
## 9.304320e+01 1.861316e-01   
## ------------------------------------------------------------   
## sat\_df$Race\_factor: Hispanic/Latino  
## nbr.val nbr.null nbr.na min max range   
## 6.157000e+03 0.000000e+00 0.000000e+00 2.300000e+02 8.000000e+02 5.700000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 3.356030e+06 5.400000e+02 5.450755e+02 1.185232e+00 2.323469e+00 8.649202e+03   
## std.dev coef.var   
## 9.300109e+01 1.706206e-01   
## ------------------------------------------------------------   
## sat\_df$Race\_factor: No response  
## nbr.val nbr.null nbr.na min max range   
## 1.231000e+03 0.000000e+00 0.000000e+00 2.400000e+02 7.800000e+02 5.400000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 6.592800e+05 5.300000e+02 5.355646e+02 2.705593e+00 5.308087e+00 9.011205e+03   
## std.dev coef.var   
## 9.492737e+01 1.772473e-01   
## ------------------------------------------------------------   
## sat\_df$Race\_factor: Two or More Races  
## nbr.val nbr.null nbr.na min max range   
## 3.694000e+03 0.000000e+00 0.000000e+00 2.800000e+02 8.000000e+02 5.200000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 2.093500e+06 5.700000e+02 5.667298e+02 1.520664e+00 2.981424e+00 8.542079e+03   
## std.dev coef.var   
## 9.242337e+01 1.630819e-01   
## ------------------------------------------------------------   
## sat\_df$Race\_factor: White  
## nbr.val nbr.null nbr.na min max range   
## 3.263500e+04 0.000000e+00 0.000000e+00 2.100000e+02 8.000000e+02 5.900000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 1.910021e+07 5.900000e+02 5.852677e+02 5.062344e-01 9.922380e-01 8.363478e+03   
## std.dev coef.var   
## 9.145205e+01 1.562568e-01

### The data appears to be normally distributed. The makes sense as the data set is very large and standardized tests such as SAT are designed by normal distribution.

Plot boxplots of English Scores among Racial/ethnic groups

sat\_df %>%   
 ggplot(aes(x = Race\_factor, y = ERW)) +  
 geom\_boxplot() +  
 labs(title = "Boxplot of English SAT scores by Race/Ethinicity",  
 x ="Race/Ethnicity",  
 y = "English SAT Scores") +  
 theme(plot.title = element\_text(hjust = 0.5))



### There appears to be little variation between groups as the boxes and error bars overlap.

Levene’s Test - Assumption of Homogeneity of variance.

leveneTest(sat\_df$ERW, sat\_df$Race\_factor, center = median)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 5 2.2886 0.04329 \*  
## 60951   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

### This output indicates that Levene’s test is significant, the variances are different accross groups. However, the sample size is quite large and in large samples even small differences in variances might be deemed significant. I will contintue with the one-way independent ANOVA test.

Run ANOVA

anova\_model <- aov(ERW ~ Race\_factor, data = sat\_df)  
summary(anova\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## Race\_factor 5 72860267 14572053 1726 <2e-16 \*\*\*  
## Residuals 60951 514679583 8444   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Post-Hoc tests Will use Bonferrioni (samples sizes are not equal, going to control for Type I error)

pairwise.t.test(sat\_df$ERW, sat\_df$Race\_factor, p.adjust.method = "bonferroni")

##   
## Pairwise comparisons using t tests with pooled SD   
##   
## data: sat\_df$ERW and sat\_df$Race\_factor   
##   
## Asian Black/African American Hispanic/Latino  
## Black/African American <2e-16 - -   
## Hispanic/Latino <2e-16 <2e-16 -   
## No response <2e-16 <2e-16 0.014   
## Two or More Races <2e-16 <2e-16 <2e-16   
## White <2e-16 <2e-16 <2e-16   
## No response Two or More Races  
## Black/African American - -   
## Hispanic/Latino - -   
## No response - -   
## Two or More Races <2e-16 -   
## White <2e-16 <2e-16   
##   
## P value adjustment method: bonferroni

Effect Size using values from ANOVA results omega^2 = (SSM - df \* MSR) / (SST + MSR) omega^2 = (72860267- 5\*8444) / (72860267 + 514679583 + 8444) omega = 0.35

## Part 2

The mean of the scores in the sample for the English section on the SAT is 550. (meaning not significantly different from 550)

Run one sample t-test

num\_3\_ttest <- t.test(sat\_df$ERW, mu = 550)  
num\_3\_ttest

##   
## One Sample t-test  
##   
## data: sat\_df$ERW  
## t = 35.049, df = 60956, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 550  
## 95 percent confidence interval:  
## 563.1576 564.7164  
## sample estimates:  
## mean of x   
## 563.937

effect size

t<-num\_3\_ttest$statistic[[1]]  
df<-num\_3\_ttest$parameter[[1]]  
r <- sqrt(t^2/(t^2+df))  
round(r, 3)

## [1] 0.141

statistically the mean does not vary significantly from 550 however practically there is a small effect size (between 0.1 < r < 0.3)

Assumptions for t-test + Normality - descriptions and plots + Independent - assuming that no one took the test twice + Continuous - the English scores are continuous

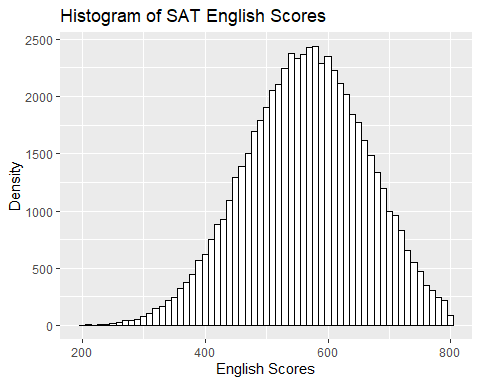
Descriptive statistics

round(stat.desc(sat\_df$ERW, basic = TRUE, norm = FALSE), digits = 3)

## nbr.val nbr.null nbr.na min max range   
## 60957.000 0.000 0.000 200.000 800.000 600.000   
## sum median mean SE.mean CI.mean.0.95 var   
## 34375910.000 570.000 563.937 0.398 0.779 9638.753   
## std.dev coef.var   
## 98.177 0.174

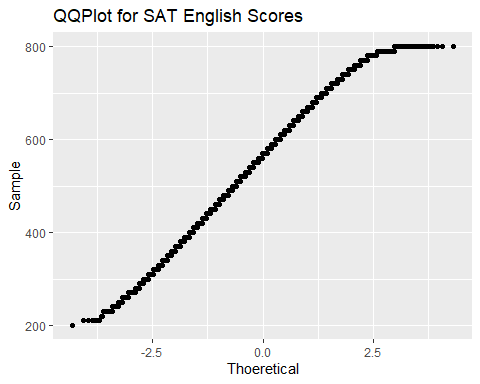
Histogram of English scores

hist\_ENG\_all <- ggplot(sat\_df, aes (ERW))+ theme(legend.position = "none") +  
 geom\_histogram(color = "black", fill = "white", binwidth = 10) +  
 labs(title = "Histogram of SAT English Scores", x = "English Scores", y = "Density")  
  
hist\_ENG\_all



Q-Q Plot of Math scores

qqplot.satENG <- qplot(sample = sat\_df$ERW, stat="qq")  
qqplot.satENG + labs(title = "QQPlot for SAT English Scores", x = "Thoeretical", y = "Sample")



## Part 3

The observed distribution of values of the mathematics section is approximately normal.

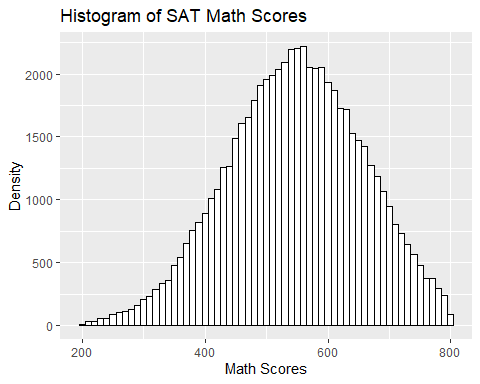
Descriptive statistics

round(stat.desc(sat\_df$MATH, basic = TRUE, norm = FALSE), digits = 3)

## nbr.val nbr.null nbr.na min max range   
## 60957.000 0.000 0.000 200.000 800.000 600.000   
## sum median mean SE.mean CI.mean.0.95 var   
## 33328170.000 550.000 546.749 0.443 0.869 11972.713   
## std.dev coef.var   
## 109.420 0.200

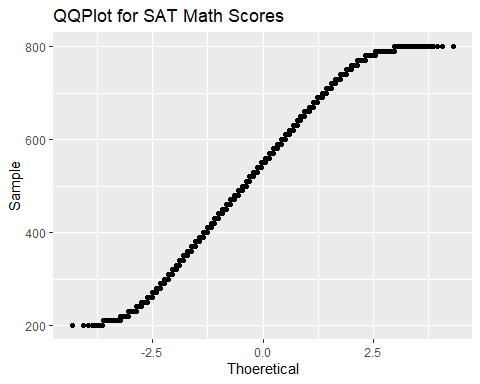
Histogram of Math scores

hist\_satMath <- ggplot(sat\_df, aes (MATH))+ theme(legend.position = "none") +  
 geom\_histogram(color = "black", fill = "white", binwidth = 10) +  
 labs(title = "Histogram of SAT Math Scores", x = "Math Scores", y = "Density")  
  
hist\_satMath



Q-Q Plot of Math scores

qqplot.satMATH <- qplot(sample = sat\_df$MATH, stat="qq")  
qqplot.satMATH+ labs(title = "QQPlot for SAT Math Scores", x = "Thoeretical", y = "Sample")



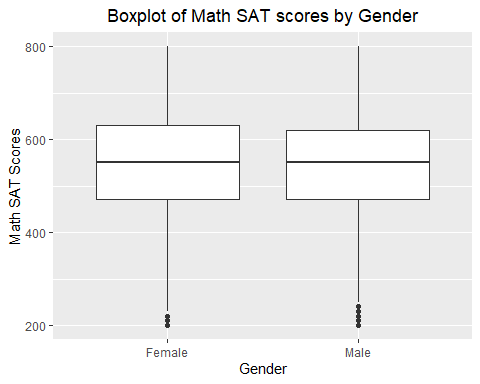
Cannot run Shapiro Wilk’s test since sample size is so big. Assume data is normally distributed.

## Part 4

There is no difference between male and female scores on the mathematics section of the SAT.

Create an errorbar plot

sat\_df %>%   
 ggplot(aes(x = Sex\_factor, y = MATH)) +   
 geom\_boxplot() +  
 labs(title = "Boxplot of Math SAT scores by Gender",  
 x ="Gender",  
 y = "Math SAT Scores") +  
 theme(plot.title = element\_text(hjust = 0.5))

 ### There appears to be little variation between groups as the boxes and error bars overlap.

General Statistics

by(sat\_df$MATH, sat\_df$Sex\_factor, stat.desc, )

## sat\_df$Sex\_factor: Female  
## nbr.val nbr.null nbr.na min max range   
## 3.230700e+04 0.000000e+00 0.000000e+00 2.000000e+02 8.000000e+02 6.000000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 1.766572e+07 5.500000e+02 5.468078e+02 6.101938e-01 1.196003e+00 1.202907e+04   
## std.dev coef.var   
## 1.096771e+02 2.005771e-01   
## ------------------------------------------------------------   
## sat\_df$Sex\_factor: Male  
## nbr.val nbr.null nbr.na min max range   
## 2.865000e+04 0.000000e+00 0.000000e+00 2.000000e+02 8.000000e+02 6.000000e+02   
## sum median mean SE.mean CI.mean.0.95 var   
## 1.566245e+07 5.500000e+02 5.466824e+02 6.447415e-01 1.263724e+00 1.190957e+04   
## std.dev coef.var   
## 1.091310e+02 1.996241e-01

Independent t-test for Math Scores comparing Genders

ind.t.test <- t.test(MATH ~ Sex\_factor, data = sat\_df)  
ind.t.test

##   
## Welch Two Sample t-test  
##   
## data: MATH by Sex\_factor  
## t = 0.14131, df = 60156, p-value = 0.8876  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## -1.614475 1.865353  
## sample estimates:  
## mean in group Female mean in group Male   
## 546.8078 546.6824

Effect sizes

t<-ind.t.test$statistic[[1]]  
df<-ind.t.test$parameter[[1]]  
r <- sqrt(t^2/(t^2+df))  
round(r, 3)

## [1] 0.001

## Part 5

There is no relationship between the sex of the test-taker and whether the test-taker completed the essay component.

Method: Pearson’s Chi-Squared Test

Run Pearson’s Chi-squared Test

CrossTable(sat\_df$Sex\_factor, sat\_df$Took.Essay., fisher = TRUE, chisq = TRUE, expected = TRUE, prop.c = FALSE, prop.t = FALSE, prop.chisq = FALSE, sresid = TRUE, format = "SPSS")

##   
## Cell Contents  
## |-------------------------|  
## | Count |  
## | Expected Values |  
## | Row Percent |  
## | Std Residual |  
## |-------------------------|  
##   
## Total Observations in Table: 60957   
##   
## | sat\_df$Took.Essay.   
## sat\_df$Sex\_factor | No | Yes | Row Total |   
## ------------------|-----------|-----------|-----------|  
## Female | 17125 | 15182 | 32307 |   
## | 17122.599 | 15184.401 | |   
## | 53.007% | 46.993% | 53.000% |   
## | 0.018 | -0.019 | |   
## ------------------|-----------|-----------|-----------|  
## Male | 15182 | 13468 | 28650 |   
## | 15184.401 | 13465.599 | |   
## | 52.991% | 47.009% | 47.000% |   
## | -0.019 | 0.021 | |   
## ------------------|-----------|-----------|-----------|  
## Column Total | 32307 | 28650 | 60957 |   
## ------------------|-----------|-----------|-----------|  
##   
##   
## Statistics for All Table Factors  
##   
##   
## Pearson's Chi-squared test   
## ------------------------------------------------------------  
## Chi^2 = 0.001524477 d.f. = 1 p = 0.9688549   
##   
## Pearson's Chi-squared test with Yates' continuity correction   
## ------------------------------------------------------------  
## Chi^2 = 0.0009557168 d.f. = 1 p = 0.9753376   
##   
##   
## Fisher's Exact Test for Count Data  
## ------------------------------------------------------------  
## Sample estimate odds ratio: 1.000635   
##   
## Alternative hypothesis: true odds ratio is not equal to 1  
## p = 0.9740578   
## 95% confidence interval: 0.9690991 1.033171   
##   
## Alternative hypothesis: true odds ratio is less than 1  
## p = 0.5188167   
## 95% confidence interval: 0 1.027891   
##   
## Alternative hypothesis: true odds ratio is greater than 1  
## p = 0.4876649   
## 95% confidence interval: 0.9741291 Inf   
##   
##   
##   
## Minimum expected frequency: 13465.6