Statistical inference with the GSS data

Setup

Load packages

```
library(ggplot2)
library(dplyr)
library(statsr)
```

Load data

```
load("gss.Rdata")
```

Part 1: Data

The General Social Survey (GSS) is a nationwide American poll conducted since 1972. It collects data on contemporary adult American society in order to study and explain trends in opinions, attitudes and behaviors, and to compare American society to other societies around the world.

The GSS data were collected randomly via face-to-face interviews, web through Computer-assisted personal interviewing (CAPI), or over the phone. The method used to collect the data was through random selection (simple random sampling) of households from across the United States. Therefore, the result of the study or the scope of inference can be generalized to all adults, aged 18 years or older, living in households in the United States who speak either English or Spanish. Residents of institutions and group quarters are out-of-scope. The sample size is 57061. As all households from across the country had an equal chance of being selected for this survey, bias were not introduced into the data set or at least bias were minimized.

Moreover, causal relationships cannot be established since the participants were not in a randomized experiment. Therefore, the study is an observational study. This allows us to make correlation statements between the variables. In other words, causation cannot be implied in this study because the researchers did not conduct an experiment.

Part 2: Research question

Although having a child is an important decision in life, not everyone has his first child born at the same age. Perhaps people from a specific race become parents at a younger age than people from other race. As United Stated is home for people of diverse races, we want to determine whether there are statistically significant differences in the age people of difference races are when their first child is born.

Part 3: Exploratory data analysis

- SUMMARY STATISTICS:

```
gss %>%
  filter(!is.na(race)) %>%
  filter(!is.na(agekdbrn)) %>%
  group_by (race) %>%
  summarise (n = n(), mean = mean(agekdbrn), sd = sd(agekdbrn), min = min(agekdbrn), max = max(a
         gekdbrn))
```

```
## # A tibble: 3 x 6
##   race      n  mean    sd   min   max
## * <fct> <int> <dbl> <dbl> <int> <int>
## 1 White 14006  24.2  5.35    11    65
## 2 Black  2799  21.6  5.08    12    52
## 3 Other  1314  23.7  5.87     9    56
```

Table 1: Summary statistics of ages people have their first child born, split by people's races. The variability appears to be approximately constant across groups; nearly constant variance across groups is an important assumption that must be satisfied before we consider the ANOVA approach.

- SIDE-BY-SIDE BOX PLOT:

```
ggplot(gss[!is.na(gss$race),], aes(x = race, y = agekdbrn)) + geom_boxplot()
```

```
## Warning: Removed 38942 rows containing non-finite values (stat_boxplot).
```

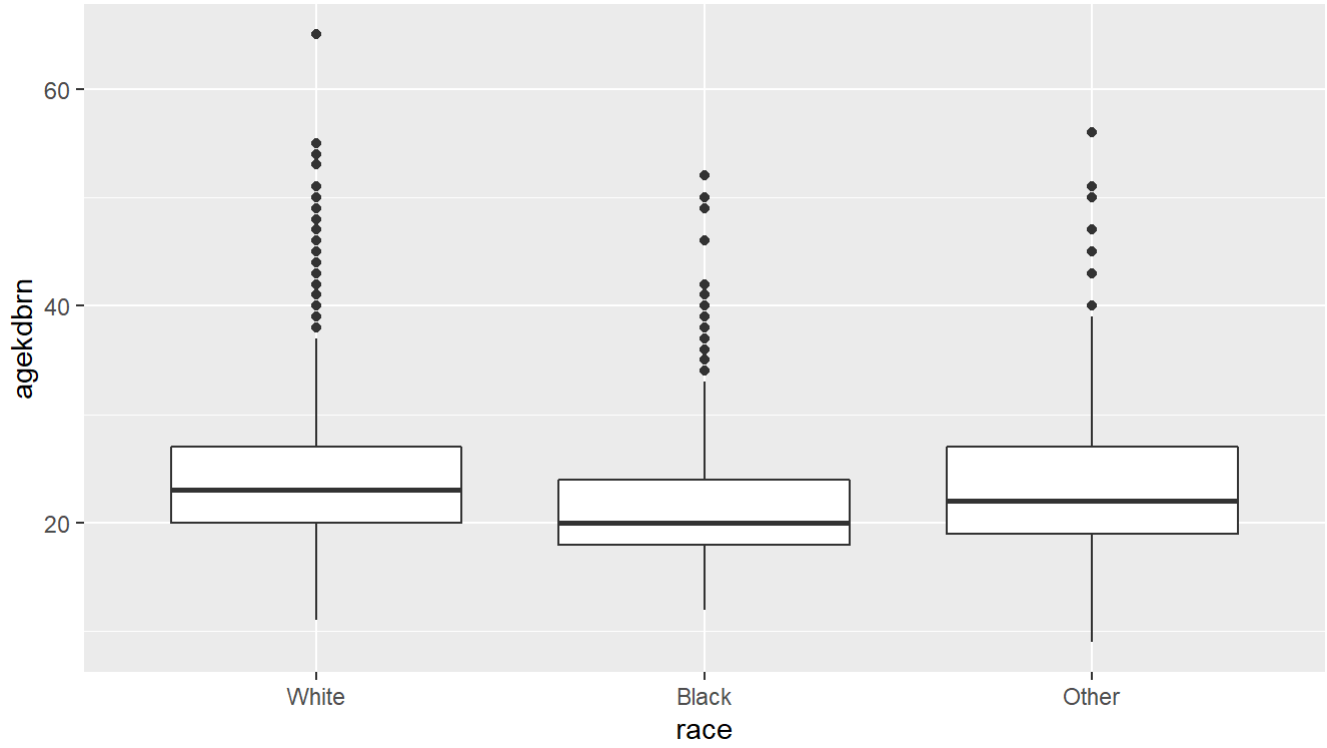

Figure 1: Side-by-side box plot of ages people have their first child born for 18119 people across races. For all races the data is skewed to the right. The box plots also show approximately equal variability, which can be verified in Table 1, supporting the constant variance assumption.

- THE NORMAL PROBABILITY PLOTS FOR EACH GROUPS.

```
par(mfrow = c(1, 3), mar = c(4.1, 4.1, 3.1, 0.1), mgp = c(2, 1, 0))
qqnorm(gss[gss$race == "White" ,"agekdbrn"], main = "White")
qqline(gss[gss$race == "White" ,"agekdbrn"])

qqnorm(gss[gss$race == "Black" ,"agekdbrn"], main = "Black")
qqline(gss[gss$race == "Black" ,"agekdbrn"])

qqnorm(gss[gss$race == "Other" ,"agekdbrn"], main = "Other")
qqline(gss[gss$race == "Other" ,"agekdbrn"])
```

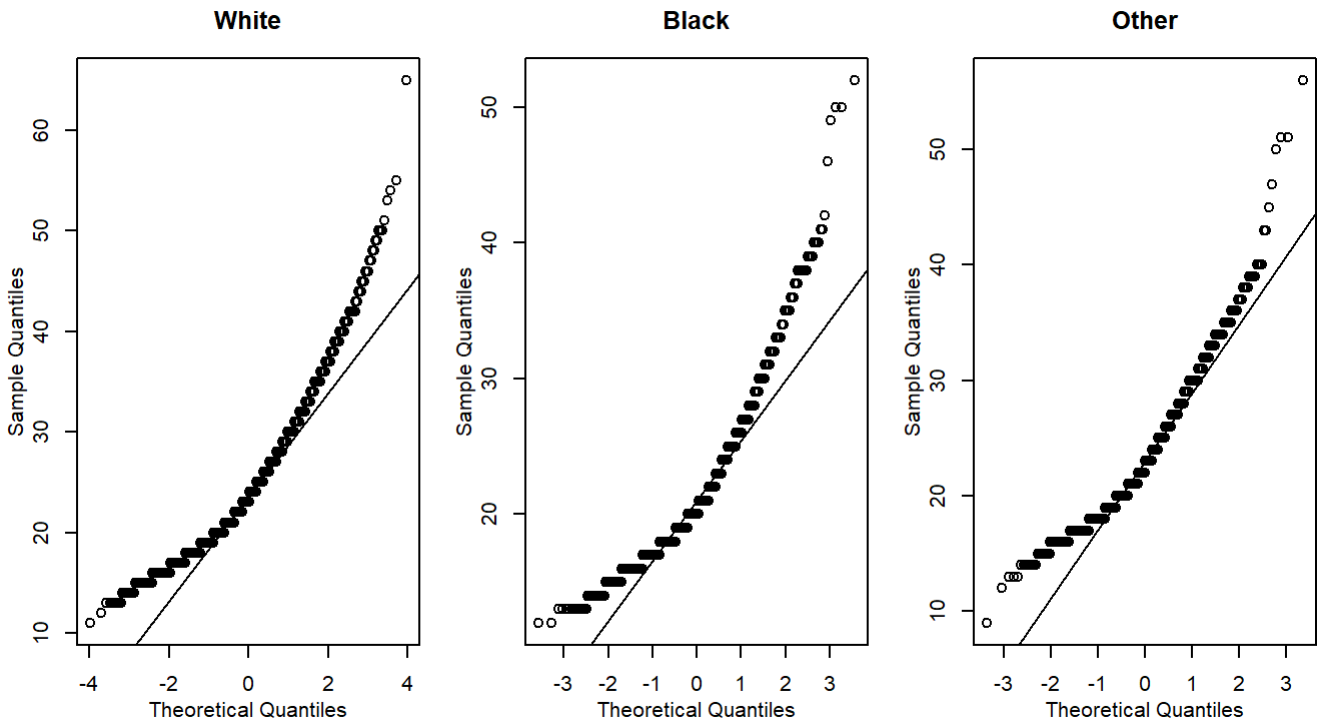

Figure 2: Normal probability plot of ages people have their first child for white, black, and people of other races.

Part 4: Inference

1. STATE HYPOTHESES:

H0:The average age people have their fist child is identical across races. Any observed difference is due to chance. Notationally, we write μw = μb = μo.

HA:The average age people have their fist child varies across some (or all) groups of races.

2. CHECK CONDITIONS: Generally we must check three conditions on the data before performing ANOVA:

- Independence: It seems reasonable to assume that the samples are independent since the data are a simple random sample from less 10% of the population. The observations are independent within and across groups.

- Approximately normal: The normal probability plots for each group of races and ages people become parents for the first time are shown in Figure 2; there is some deviation from normality for white, black and other races, but this is not a substantial concern since there are about 14006, 2799, 1314 observations for each group respectively, and the outliers are not extreme.
- Constant variance: The variability across the groups is about equal. Figure 1 shows a side-by-side box plot of the outcomes across the groups. In this case, the variability is similar in the three groups. Additionally, we see the summary statistics in table 1 that the standard deviation (sd) varies very little from one group to the next supporting the constant variance assumption.

3. STATE THE METHOD(S) TO BE USED AND WHY AND HOW:

In this project, we will use the method called analysis of variance (ANOVA) and the test statistic called F. With ANOVA we will use a single hypothesis test to check whether the means of these groups are equal. Additionally, we are not going to conduct confidence Intervals because this is an analysis of variance that uses a single hypothesis test.

If we reject the hypothesis null (HO), we will conclude that the different means across the three groups is not simply due to chance. Then, we will determine which differences across groups are statistically significant. Therefore, we are going to do pairwise comparisons using the Bonferroni correction and report any significant differences.

4. PERFORM INFERENCE:

```
aov_agekdbrn = aov(agekdbrn ~ race, gss)
summary(aov_agekdbrn)
```

```
##                Df Sum Sq Mean Sq F value Pr(>F)
## race            2  16892    8446   295.5 <2e-16 ***
## Residuals   18116 517734      29
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 38942 observations deleted due to missingness
```

Table 2: ANOVA summary for testing whether the average of ages people became parents for the first time differs across races.

5. INTERPRET RESULTS:

The p-value of the test is <2e-16, less than the default significance level of 0.05. Therefore, we reject the null hypothesis and conclude that people of different races tend to have their first child born at different ages.

6. MULTIPLE COMPARISONS AND THE BONFERRONI CORRECTION FOR THE SIGNIFICANCE LEVEL(α).

There is strong evidence that the different means in each of the three groups of races is not simply due to chance. We might wonder, which of the races are actually different. As we learned in our OpenIntro Statistics book, a two-sample t-test could be used to test for differences in each possible pair. However, when we run so many tests, the Type 1 Error rate increase. This issue is resolved by using a modified significance level.

The Bonferroni correction suggests that a more stringent significance level is more appropriate for these tests:

$\alpha^* = \alpha / k$

α*=0.05/3 = 0.0167

To find which groups are actually different, we will compute a two-sample t-test to test for differences in each possible pair of groups using the Bonferroni correction. First, we are going to analyze the data through histograms. Second, we will analyze conditions for using the t-distribution for a difference in means. Then, we will calculate the two-sided p-value for each possible pairwise using the Bonferroni correction. Finally, we are going to interpret results.

6.1. ANALYSIS OF THE DATA

- HISTOGRAMS

```
hist(gss[gss$race =="White", "agekdbrn"], main = "White", xlab = "age first child born")
```

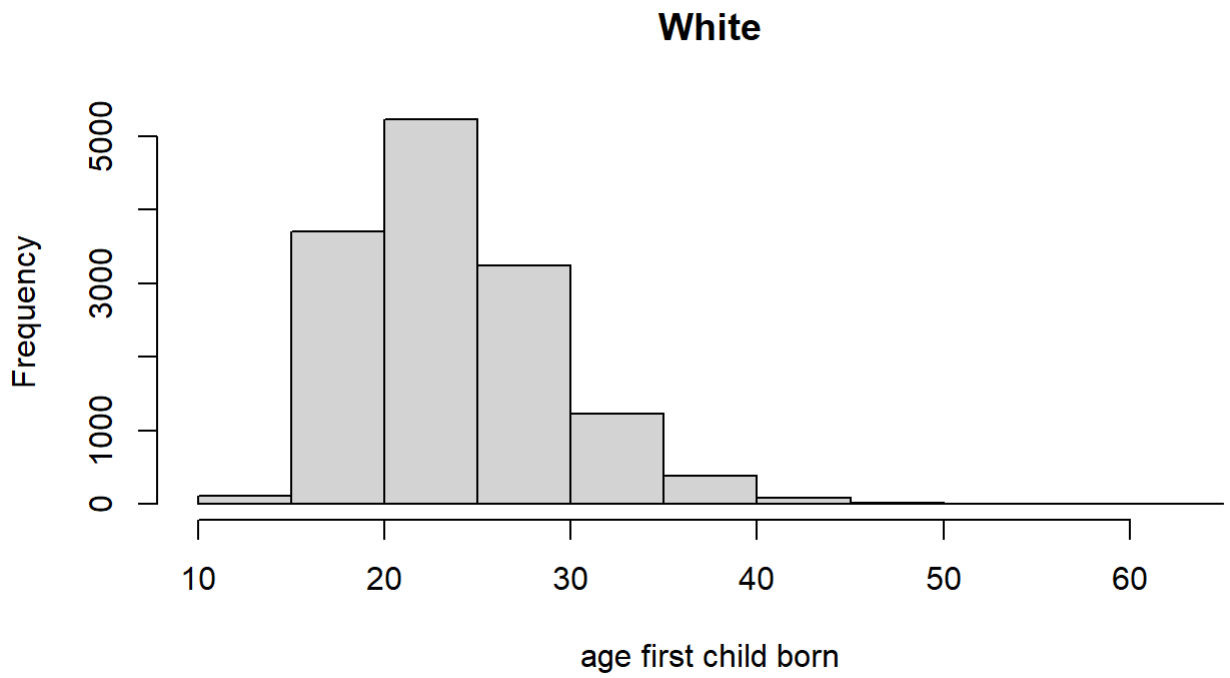

Figure 3: Histogram of white people's age when their first child was born. The data is strongly skewed to the right.

```
hist(gss[gss$race =="Black", "agekdbrn"], main = "Black", xlab = "age first child born")
```

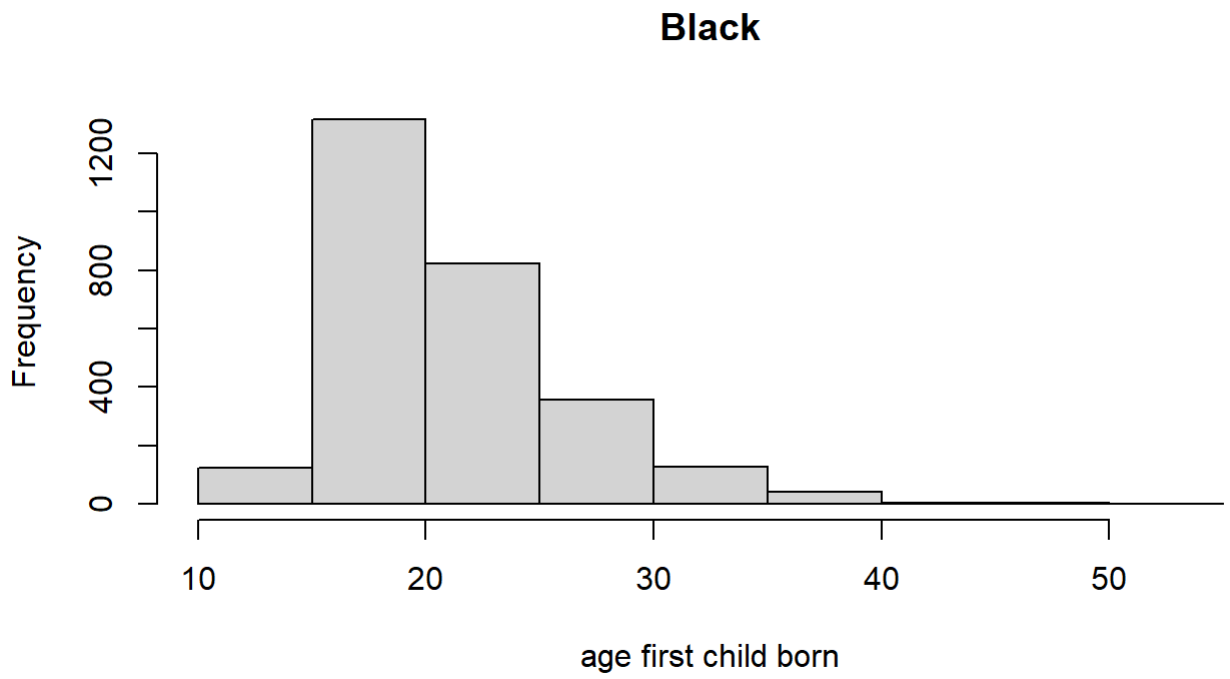

Figure 4: Histogram of black people's age when their first child was born. The data is strongly skewed to the right.

```
hist(gss[gss$race =="Other", "agekdbrn"], main = "Other", xlab = "age first child born")
```

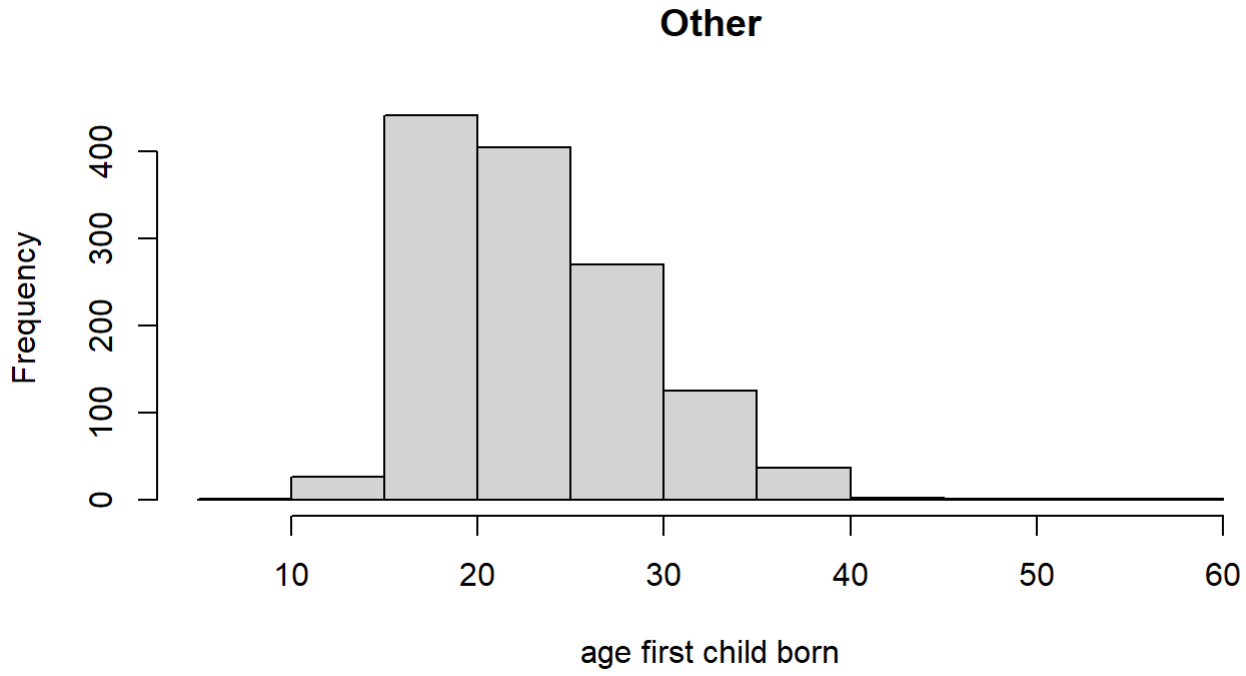

Figure 5: Histogram of other race of people's age when their first child was born. The data is strongly skewed to the right.

6.2. CHECKING CONDITION FOR USING THE T-DISTRIBUTION

- Independence of observations: Because the data come from a simple random sample and consist of less than 10% of all such cases, the observations are independent.
- Observations come from a nearly normal distribution: while each distribution is strongly skewed, the sample size of 14006 whites, 2799 blacks, and 1314 observation for other races would make it reasonable to model each mean separately using a t-distribution. Since both conditions are satisfied, the difference in sample means may be modeled using a t-distribution.

6.3. CALCULATING THE TWO-SIDED P-VALUE FOR EACH POSSIBLE GROUP USING THE BONFERRONI CORRECTION.

We will use the R-package “lsmeans” to compare means across each groups.

```
library(lsmeans)
```

```
## Warning: package 'lsmeans' was built under R version 4.0.5
```

```
## Loading required package: emmeans
```

```
## Warning: package 'emmeans' was built under R version 4.0.5
```

```
## The 'lsmeans' package is now basically a front end for 'emmeans'.
## Users are encouraged to switch the rest of the way.
## See help('transition') for more information, including how to
## convert old 'lsmeans' objects and scripts to work with 'emmeans'.
```

```
lsmeans(aov_agekdbrn, pairwise ~ race, adjust = "bon")
```

```
## $lsmeans
##  race  lsmean      SE    df lower.CL upper.CL
##  White  24.24 0.04517 18116    24.15    24.33
##  Black  21.55 0.10105 18116    21.35    21.75
##  Other  23.71 0.14748 18116    23.42    24.00
##
## Confidence level used: 0.95
##
## $contrasts
##  contrast      estimate    SE    df t.ratio p.value
##  White - Black    2.690 0.111 18116  24.305 <.0001
##  White - Other    0.534 0.154 18116   3.465 0.0016
##  Black - Other   -2.156 0.179 18116 -12.058 <.0001
##
## P value adjustment: bonferroni method for 3 tests
```

Table 3: Multiple comparisons table using the Bonferroni correction for α. The top summary shows confidence interval for each individual group. The bottom summary shows three possible pairwise comparisons using the Bonferroni correction.

INTERPRET RESULTS:

White versus Black: The estimated difference and standard error are 2.69 and 0.111, respectively. This results in a T-score of 24.305 on df=18116 and a two-sided p-value of <.0001. This p-value is less than α=0.05, so there is strong evidence of a difference in the means of white and black people.

White versus Other: The estimated difference and standard error are 0.534 and 0.154, respectively. This results in a T-score of 3.465 on df=18116 and a two-sided p-value of 0.0016. This p-value is less than α=0.05, so there is strong evidence of a difference in the means of white and other races of people.

Black versus Other: The estimated difference and standard error are -2.156 and 0.179, respectively. This results in a T-score of -12.058 on df=18116 and a two-sided p-value of <.0001. This p-value is less than α=0.05, so there is strong evidence of a difference in the means black and other races of people.

Important note: Take into account that we compare across 0.05 because the p-values are already adjusted using the Bonferroni comparison (the p-values were multiplied by the number of comparisons).