SAI Median & Mean Differences

Median Diff. Group

2025-04-24

Introduction

This report focuses on data found within the SAI data set- median and mean differences in anxiety scores between one group consuming a drug (caffeine) when compared to a placebo group (i.e., Drug vs. Placebo). We utilized bootstrapping and data visualization techniques to assess how our statistial results can be both accurate and intentionally misleading. Additionally, to prepare our data we had to utilize the mutate function in order to examine the relationship between the Drug and Placebo variables.

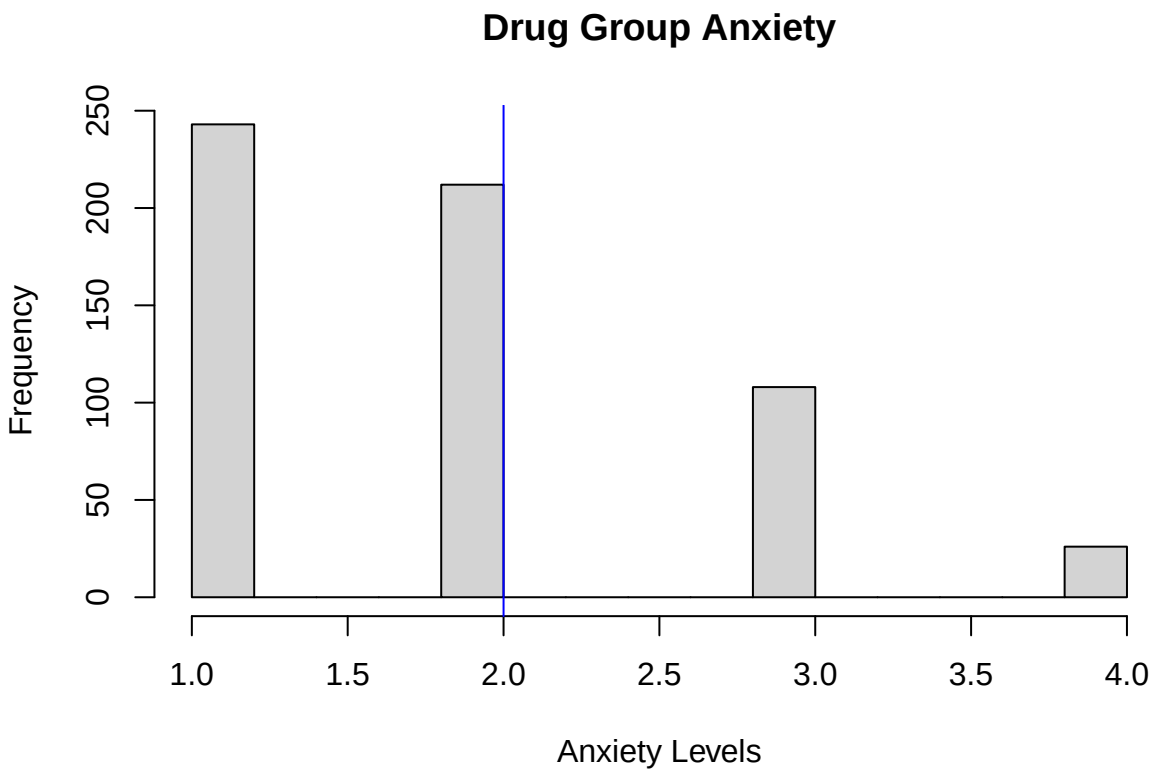
Data Preparation

```
data("sai")
data(sai)
sai <- sai %>%
  mutate(grouped_study = recode(study,
    "AGES" = "DRUG", "CITY" = "DRUG", "EMIT" = "DRUG",
    "SALT" = "DRUG", "XRAY" = "DRUG",
    "Cart" = "PLACEBO", "Fast" = "PLACEBO",
    "Shed" = "PLACEBO", "Raft" = "PLACEBO",
    "Shop" = "PLACEBO"
  )) %>%
  filter(grouped_study %in% c("DRUG", "PLACEBO"), time == "1")

sai_drug <- filter(sai, grouped_study == "DRUG")
sai_placebo <- filter(sai, grouped_study == "PLACEBO")
```

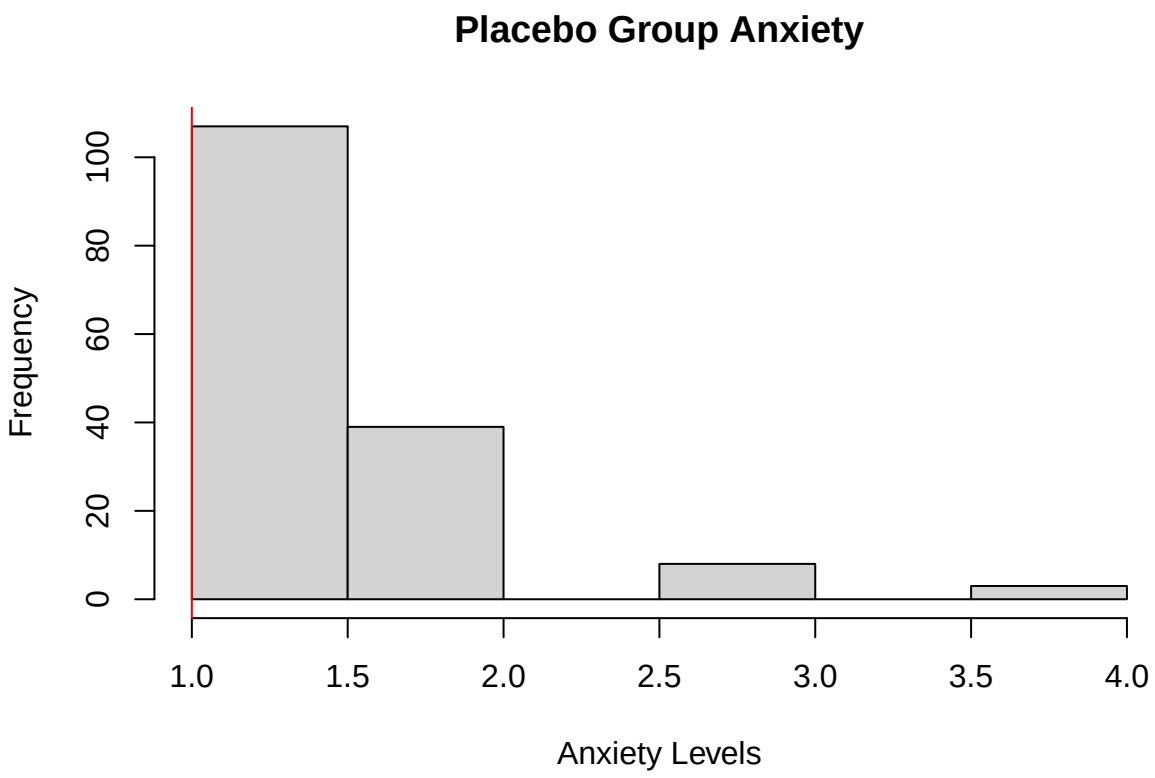
Anxiety Median Analysis

Our first analysis assesses the difference in anxiety medians between the drug and placebo groups by creating histograms to visualize our dataset.

```
med_drug <- median(sai_drug$anxious, na.rm = TRUE)
med_placebo <- median(sai_placebo$anxious, na.rm = TRUE)

med_diff <- med_drug - med_placebo
hist(sai_drug$anxious, main = "Drug Group Anxiety", xlab = "Anxiety Levels")
abline(v = med_drug, col="blue")
```




```
hist(sai_placebo$anxious, main = "Placebo Group Anxiety", xlab = "Anxiety Levels")
abline(v = med_placebo, col="red")
```




```r
anxious_diff <- sai_drug$anxious - sai_placebo$anxious
```

```
## Warning in sai_drug$anxious - sai_placebo$anxious: longer object length is not
## a multiple of shorter object length
```

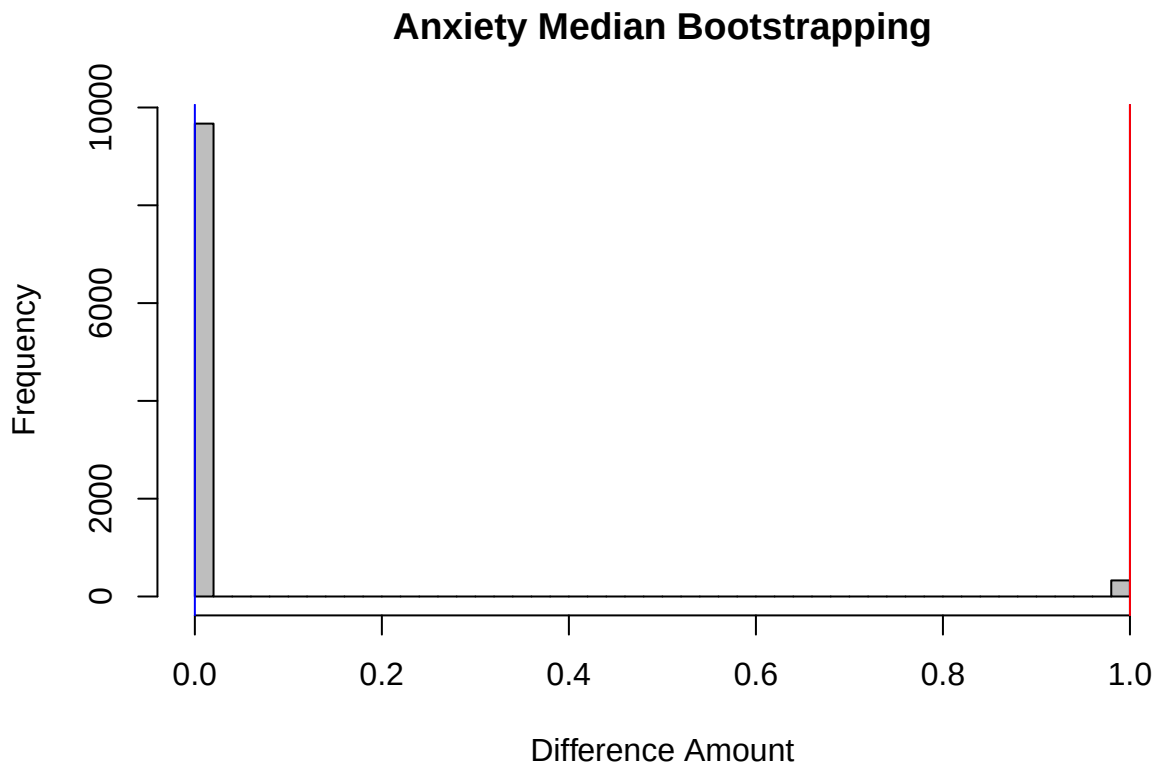
```r
med_diff <- median(anxious_diff, na.rm = TRUE)
```

The histograms show a right-skewed distribution for the drug group with higher variability and a median around 2.0. However, our placebo group shows a much tighter clustering of anxiety scores around 1.0. This suggests that the placebo group experienced consistently lower anxiety. In contrast, the drug group had more varied responses which might explain the prevalence of higher anxiety levels.

Bootstrapping Median of Observed Anxiety

```r
# Remove NAs
drug_anxious <- na.omit(sai_drug$anxious)
placebo_anxious <- na.omit(sai_placebo$anxious)

# Calculate observed median difference
med_diff <- median(drug_anxious) - median(placebo_anxious)

# Bootstrap median differences
```

```r
boot_med_diff <- numeric(10000)
for (reps in 1:10000) {
  # Resample with replacement from each group
  drug_sample <- sample(drug_anxious, replace = TRUE)
  placebo_sample <- sample(placebo_anxious, replace = TRUE)

  # Creation of possible differences from sampled values
  pairwise_diff <- outer(drug_sample, placebo_sample, `-`)

  # Median of all pairwise differences
  boot_med_diff[reps] <- median(pairwise_diff)
}

# Plot histogram
hist(boot_med_diff,
     breaks = 50,
     main = "Anxiety Median Bootstrapping",
     xlab = "Difference Amount",
     ylab = "Frequency",
     col = "gray")

# Add 95% confidence interval and observed difference line
abline(v = quantile(boot_med_diff, c(0.025, 0.975)), col = "blue")
abline(v = med_diff, col = "red")
```

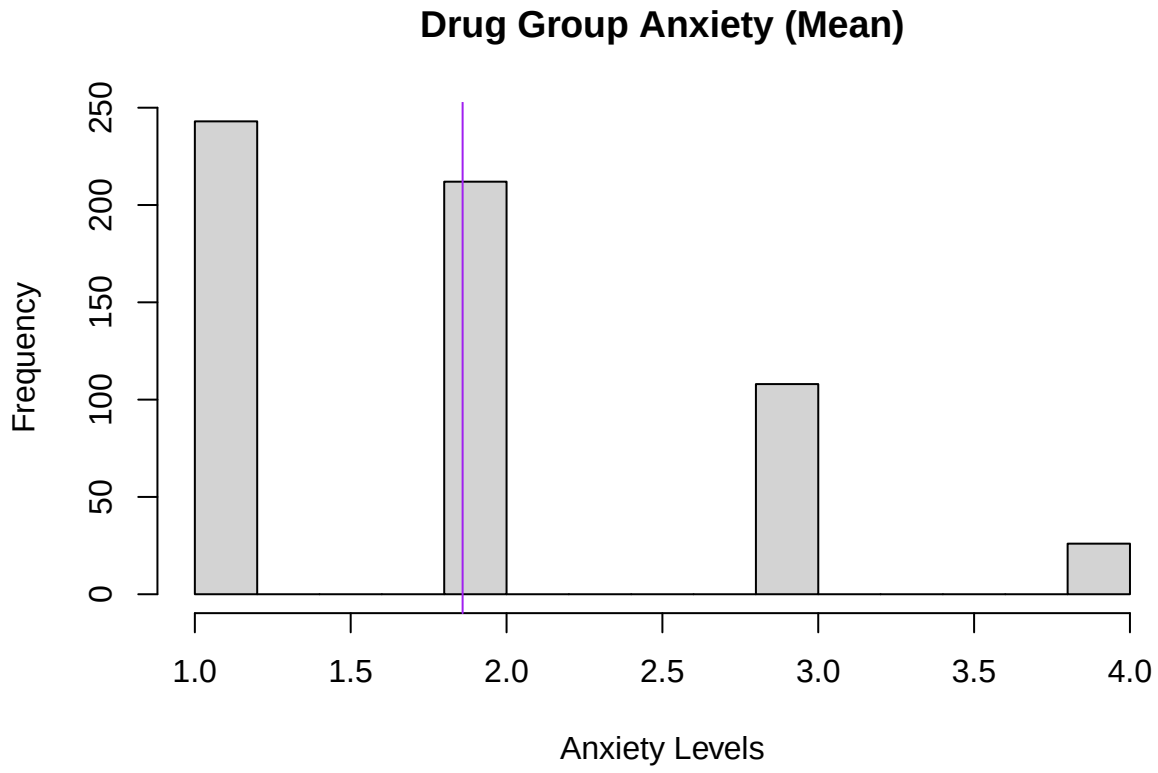


When bootstrapping the median of observed anxiety between the drug and placebo groups, the observed

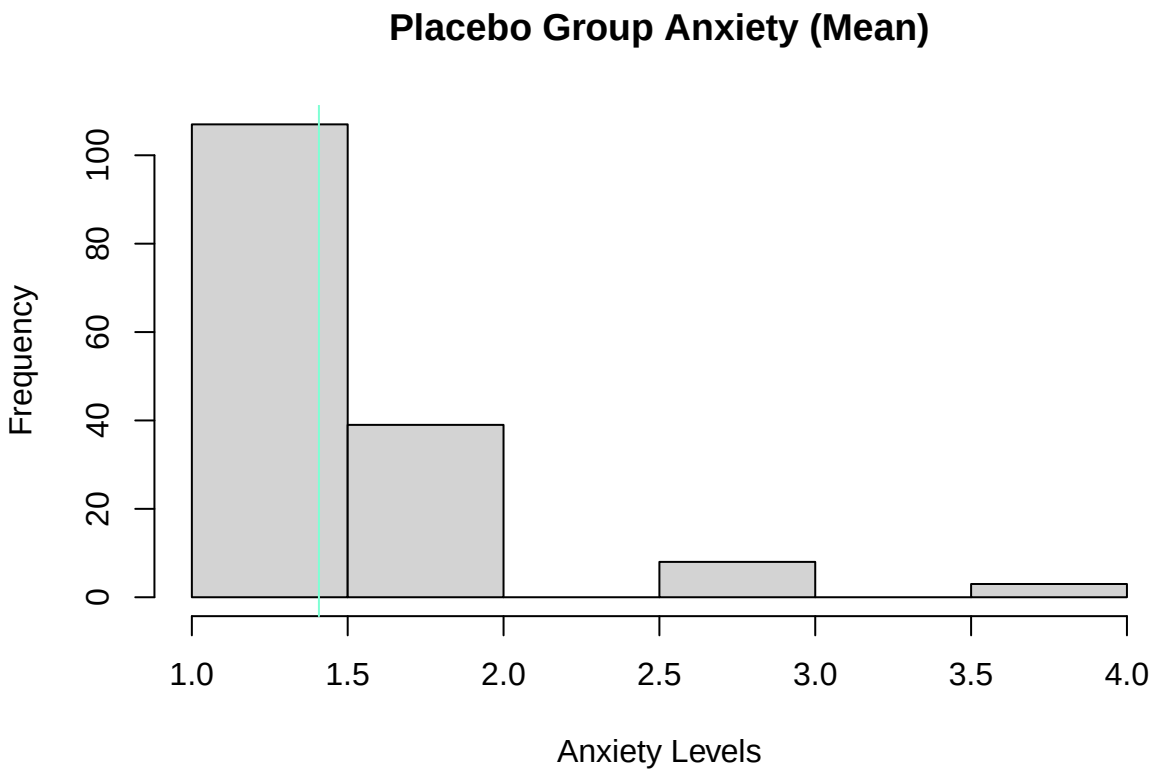
right-handed skew is not normally distributed. Bootstrapping reveals that a significant number of our values have minute differences. Additionally, our observed median difference between drug and placebo groups sits approximately at 1, and our CI lower bound (blue line) is close to zero while our CI upper bound (red line) sits near the obtained value. This distribution might reveal that our obtained median anxiety differences are potentially unreliable due to confounds such as discrete underlying data.

Anxiety or Anxiousness Mean Analysis

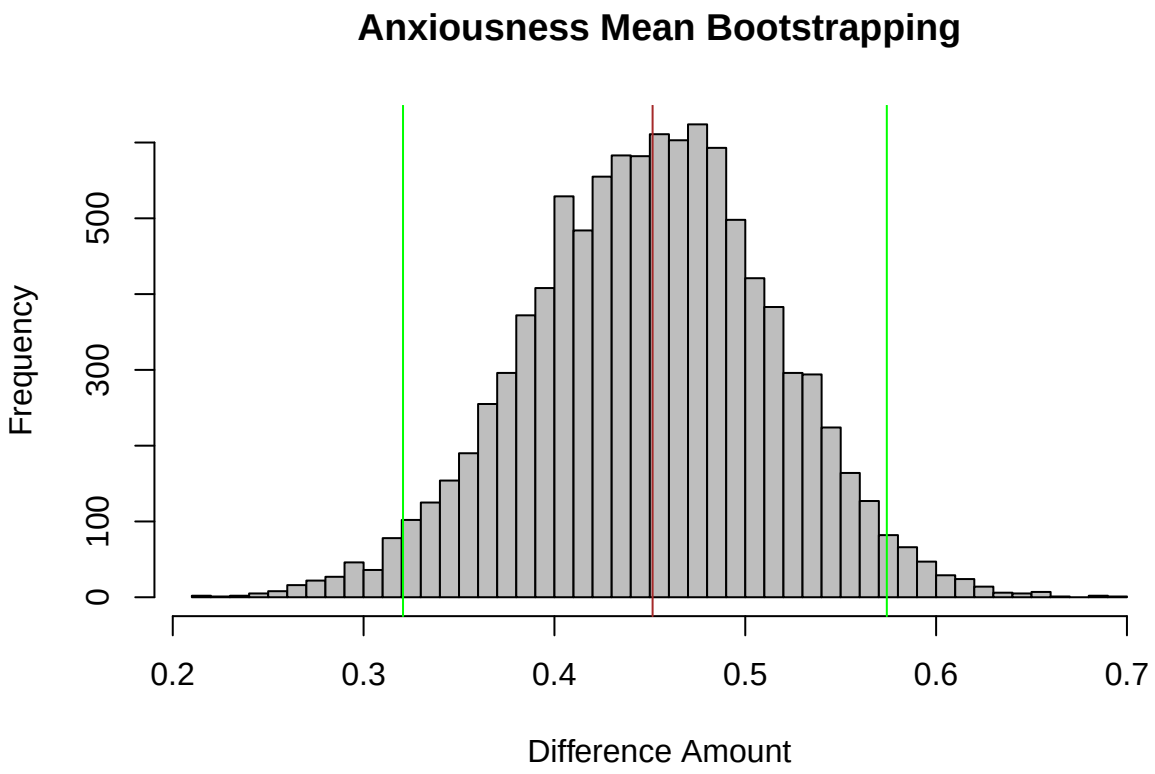
```
mean_drug <- mean(sai_drug$anxious, na.rm = TRUE)
mean_placebo <- mean(sai_placebo$anxious, na.rm = TRUE)

hist(sai_drug$anxious, main = "Drug Group Anxiety (Mean)", xlab = "Anxiety Levels")
abline(v = mean_drug, col = "purple")
```




```
hist(sai_placebo$anxious, main = "Placebo Group Anxiety (Mean)", xlab = "Anxiety Levels")
abline(v = mean_placebo, col = "aquamarine")
```




```r
# Observed mean difference
mean_diff <- mean(drug_anxious) - mean(placebo_anxious)

# Bootstrap mean differences
boot_mean_diff <- numeric(10000)
for (reps in 1:10000) {
  drug_sample <- sample(drug_anxious, replace = TRUE)
  placebo_sample <- sample(placebo_anxious, replace = TRUE)
  boot_mean_diff[reps] <- mean(drug_sample) - mean(placebo_sample)
}

hist(boot_mean_diff, breaks = 50,
     main = "Anxiousness Mean Bootstrapping",
     xlab = "Difference Amount", ylab = "Frequency", col = "gray")
abline(v = quantile(boot_mean_diff, c(0.025, 0.975)), col = "green")
abline(v = mean_diff, col = "brown")
```



While examining the means values within the drug and placebo groups, we observed that our drug group possesses both higher variability and higher average anxiety scores. In contrast, our placebo groups cluster around low anxiety scores. Our bootstrapped mean difference distribution also appears to support this finding and shows these observations are likely not due to chance.

Thought Experiment

Let's say an energy drink manufacturer is planning to promote their product as "incapable of causing anxiousness associated with caffeine over-consumption." To support this claim, they also run the same experiment with some 'tweaks'.

```r
# Use only median differences
mislead_med_diff <- median(drug_anxious) - median(placebo_anxious)

# Bootstrap again for medians
mislead_boot <- numeric(10000)
for (i in 1:10000) {
  drug_sample <- sample(drug_anxious, replace = TRUE)
  placebo_sample <- sample(placebo_anxious, replace = TRUE)
  pairwise_diff <- outer(drug_sample, placebo_sample, `-`)
  mislead_boot[i] <- median(pairwise_diff)
}

# 80 % CI Instead of 95 %
ci_bounds <- quantile(mislead_boot, c(0.10, 0.90))
```

```r
# Plot
hist(mislead_boot,
     breaks = 50,
     col = "lightgreen",
     border = "gray30",
     xlim = c(-0.05, 0.2),  # centered to make lines visible
     main = "Our Beverage Has No Effect on Anxiety!",
     xlab = "Difference in Anxiety (Drug - Placebo)",
     ylab = "Frequency")

# Add clearer CI lines and observed difference
abline(v = ci_bounds, col = "blue", lty = 2, lwd = 2)
abline(v = mislead_med_diff, col = "red", lwd = 2)

# Add matching legend
legend("topright",
       legend = c("80% CI", "Observed Median Diff"),
       col = c("blue", "red"),
       lty = c(2, 1),
       lwd = 2,
       bty = "n")
```