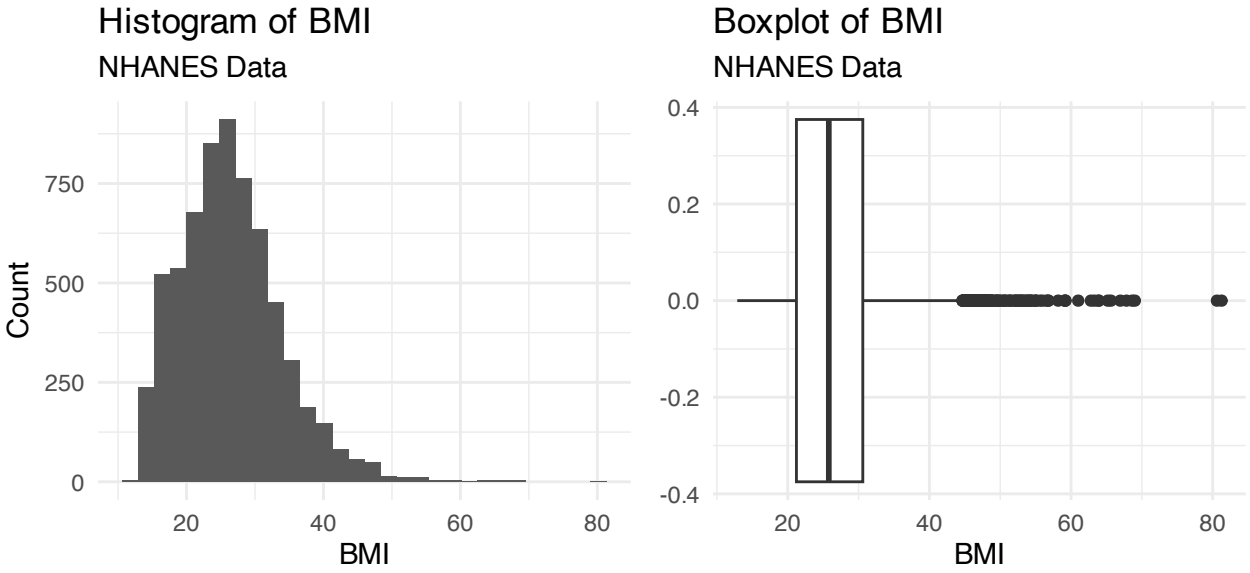
STA 5363, Homework 3

Carson Slater *Baylor University*

Suppose we are interested in studying central tendency of the distribution of BMI in your own NHANES data.

> An aside: It was discovered that some individuals in the NHANES dataset (identified by `ID`) had multiple, conflicting records. A decision was made to resolve this by retaining only the single record for each person that contained the fewest missing (NA) values, a strategy that prioritizes the most complete observation for each participant. Additionally, participants without BMI data were not considered in this assignment.

1. Produce histogram and boxplot to explore BMI and report your finding.



These graphics show the distribution of BMI values from NHANES data using a histogram and boxplot. The histogram reveals a right-skewed distribution with most values concentrated between BMI 20-35 and a long tail extending to higher values. The boxplot confirms this pattern, showing a median around BMI 27-28 with numerous outliers at higher BMI values, indicating that while most individuals fall in normal to overweight categories, there's a substantial portion with obesity.

2. Consider the population mean μ as a measure of central tendency.

(a) Find an estimate for μ. Call this estimate $\hat{\mu}$.

From the data we have that $\bar{x} = \hat{\mu}_{\text{BMI}} \approx 26.488$.

```r
(mu <- mean(nhanes$BMI, na.rm = TRUE) |> round(3))
```

```
## [1] 26.488
```

(b) Provide an estimates of standard error of $\hat{\mu}$.

We have that $\hat{\sigma}_{\bar{x}} = \sqrt{\hat{\sigma}^2/n}$, so then we compute that using R to be approximately 0.094.

```r
(se_mu <- (sd(nhanes$BMI, na.rm = TRUE)/sqrt(sum(!is.na(nhanes$BMI))))
 |> round(3))
```

```
## [1] 0.094
```

(c) Estimate the standard error of $\hat{\mu}$ using the bootstrap and compare it to the estimate obtained from (b).

Method	Standard_Error
Analytical (part b)	0.094
Bootstrap (part c)	0.092
Difference	0.002

(d) Construct a 95% confidence interval for μ using the estimated standard error obtained from (b).

For a 95% confidence interval with large sample size, we use the normal approximation:

$$\text{CI}_{95\%} = \hat{\mu} \pm z_{\alpha/2} \cdot \text{SE}(\hat{\mu})$$

where $z_{\alpha/2} = z_{0.025} = 1.96$ for a 95% confidence level. Therefore, the 95% confidence interval for μ is

$$[26.304,\ 26.672],$$

using the estimated standard error for $\hat{\mu}$.

(e) Construct an approximate 95% confidence interval for μ using the estimated standard error obtained from (c). You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2 \cdot \textbf{SE}(\hat{\mu}), \hat{\mu} + 2 \cdot \textbf{SE}(\hat{\mu})]$.

Using the bootstrap standard error for $\hat{\mu}$ in (c), the 95% confidence interval for μ is

$$[26.304,\ 26.672].$$

Although they are the same as the normal theory approximation, we reiterate that we used the formula given in the question to compute this interval.

3. Consider the population median μ_{med} as a measure of central tendency.

(a) Find an estimate for μ_{med}. Call this estimate $\hat{\mu}_{\text{med}}$.

```
(mu_med <- median(nhanes$BMI, na.rm = TRUE))
```

```
## [1] 25.8
```

We estimate μ_{med} using the sample median, which is 25.8.

(b) Estimate the standard error of $\hat{\mu}_{\text{med}}$ using the bootstrap.

```
## [1] 0.109
```

The standard error of $\hat{\mu}_{\text{med}}$ was found to be 0.109.

(c) Construct an approximate 95% confidence interval for μ_{med} using the estimated standard error obtained from (b).

Using the bootstrap standard error for $\hat{\mu}_{\text{med}}$ in (b), the 95% confidence interval for μ_{med} is

$$[25.582,\ 26.018]$$

4. Consider the population mode μ_{mod} as a measure of central tendency.

(a) Find an estimate for μ_{mod}. Call this estimate $\hat{\mu}_{\text{mod}}$.

```
## [1] 23.2
```

We estimate μ_{mod} to be 23.2.

(b) Estimate the standard error of $\hat{\mu}_{\text{mod}}$ using the bootstrap.

Because it is possible to estimate more than one mode using the estimator we have employed, which is the most frequently occurring value, we elect to, upon this instance, randomly sample from those modes and pick that as the estimator.

```
## [1] 7.814
```

The standard error of $\hat{\mu}_{\text{med}}$ was found to be 7.814.

(c) Construct an approximate 95% confidence interval for μ_{mod} using the estimated standard error obtained from (b).

Using the bootstrap standard error for $\hat{\mu}_{\text{mod}}$ in (b), the 95% confidence interval for μ_{mod} is

$$[7.252,\ 39.148]$$

5. Summarize your findings. Do the parameters provide similar information about the tendency? Do the estimates provide similar uncertainty? Which parameter and estimate are the best in explaining central tendency of the distribution of BMI?

The mean BMI was estimated to be $\hat{\mu} = 26.488$, which reflects the arithmetic average but is sensitive to extreme values. The median BMI was found to be $\hat{\mu}_{\text{med}} = 25.8$, offering a robust measure of central tendency that is less affected by skewness and outliers. The mode, estimated as $\hat{\mu}_{\text{mod}} = 23.2$, represents the most frequently occurring value in the data, though its usefulness is limited in continuous distributions where repeated values are rare.

In terms of uncertainty, the standard error for the mean was estimated analytically and via bootstrap as approximately 0.094 and 0.092, respectively, resulting in a narrow 95% confidence interval of $[26.304,\ 26.672]$. The bootstrap standard error for the median was slightly larger at 0.109, yielding a 95% confidence interval of $[25.582,\ 26.018]$. However, the mode exhibited extreme variability, with a bootstrap standard error of 7.814 and a resulting 95% confidence interval of $[7.252,\ 39.148]$, indicating substantial uncertainty and instability in its estimation.

Given the right-skewed nature of the BMI distribution, the median provides the most accurate representation of the typical individual in the dataset. Nevertheless, from an inferential perspective, the mean offers the most precise and reliable estimate due to its smaller standard error and narrower confidence interval. In contrast, the mode is unsuitable for describing the central tendency in this context due to its high variability and wide confidence bounds. Therefore, while the median best explains the central tendency conceptually, the mean is the best estimated statistically. The mode should be avoided in this setting.

Code Appendix

```
knitr::opts_chunk$set(dev = "cairo_pdf",
                      fig.width = 6.75,
                      fig.height = 3.25,
                      fig.align = 'center',
                      echo = FALSE,
                      message = FALSE,
                      warning = FALSE,
                      error = FALSE,
                      cache = TRUE)
library("tidymodels")
library("patchwork")
library("glue")
library("scales")
library("extrafont")
library("tinytex")
library("patchwork")
library("gridExtra")
library("tidyr")
library("latex2exp")
library("GGally")
library("leaps")
library("broom")
library("pls")
library("knitr")
library("rprojroot")
theme_set(theme_minimal(base_family = "Roboto Condensed"))

conflicted::conflicts_prefer(
  readr::col_factor(),
  purrr::discard(),
  rstan::extract(),
  dplyr::lag(),
  rstan::traceplot(),
  viridis::viridis_pal(),
  readr::parse_date
)


# Problem 1 ---------------------------------------------------------
library("NHANES")
data("NHANES")

# Found multiple people with duplicate rows and conflicting information.
# Selected row for each person with least amount of NA values
```

```r
nhanes <- NHANES |>
  distinct() |>
  group_by(ID) |>
  group_split() |>
  map_dfr(~ .x[which.min(rowSums(is.na(.x))), ])

nhanes_hist <- nhanes |>
  ggplot(aes(BMI)) +
  geom_histogram() +
  labs(
    x = "BMI",
    y = "Count",
    title = "Histogram of BMI",
    subtitle = "NHANES Data"
  ) +
  theme_minimal()

nhanes_boxplot <- nhanes |>
  ggplot(aes(x = BMI)) +
  geom_boxplot() +
  labs(
    # x = "BMI",
    title = "Boxplot of BMI",
    subtitle = "NHANES Data"
  ) +
  theme_minimal()


nhanes_hist + nhanes_boxplot


# Problem 2a ----------------------------------------------------------
(mu <- mean(nhanes$BMI, na.rm = TRUE) |> round(3))



# Problem 2b ----------------------------------------------------------
(se_mu <- (sd(nhanes$BMI, na.rm = TRUE)/sqrt(sum(!is.na(nhanes$BMI))))
 |> round(3))


# Problem 2c ----------------------------------------------------------
library("boot")
set.seed(613)  # For reproducibility

# Define statistic function for boot()
mean_stat <- function(data, indices) {
```

```r
  return(mean(data[indices]))
}

# Perform bootstrap with boot()
bootstrap_results <- boot(data = nhanes$BMI[!is.na(nhanes$BMI)],
                          statistic = mean_stat,
                          R = 1000)
# Extract bootstrap standard error
se_mu_bootstrap <- sd(bootstrap_results$t) |> round(3)

# Compare results
comparison <- data.frame(
  Method = c("Analytical (part b)", "Bootstrap (part c)", "Difference"),
  Standard_Error = c(se_mu, se_mu_bootstrap, abs(se_mu - se_mu_bootstrap) |> r
)
kable(comparison)



# Problem 2d ----------------------------------------------------------
# For large samples, use normal approximation with z-score
alpha <- 0.05
z_critical <- qnorm(1 - alpha/2)  # 1.96 for 95% CI

# Calculate confidence interval bounds
ci_lower <- mu - z_critical * se_mu
ci_upper <- mu + z_critical * se_mu

# CI
ci_95 <- c(lower = ci_lower, upper = ci_upper) |> round(3)



# Problem 2e ----------------------------------------------------------
# Calculate confidence interval bounds
ci_lower_boot <- mu - 2 * se_mu_bootstrap
ci_upper_boot <- mu + 2 * se_mu_bootstrap

# CI
ci_95_boot <- c(lower = ci_lower_boot, upper = ci_upper_boot) |>
  round(3)


# Problem 3a ----------------------------------------------------------
(mu_med <- median(nhanes$BMI, na.rm = TRUE))
```

```r
# Problem 3b ----------------------------------------------------------
median_stat <- function(data, indices) {
   return(median(data[indices]))
}

# Perform bootstrap with boot()
bootstrap_results_med <- boot(data = nhanes$BMI[!is.na(nhanes$BMI)],
                          statistic = median_stat,
                          R = 1000)
# Extract bootstrap standard error
(se_med_bootstrap <- sd(bootstrap_results_med$t) |> round(3))


# Problem 3c ----------------------------------------------------------
ci_lower_boot <- mu_med - 2 * se_med_bootstrap
ci_upper_boot <- mu_med + 2 * se_med_bootstrap

# CI
ci_95_boot <- c(lower = ci_lower_boot, upper = ci_upper_boot) |>
  round(3)



# Problem 4a ----------------------------------------------------------
# Mode function
mode <- function(data, na.rm = TRUE) {
  if (na.rm) data <- data[!is.na(data)]

  freq_tab <- table(data)

  max_freq <- max(freq_tab)

  # Find all values with maximum frequency
  modes <- names(freq_tab)[freq_tab == max_freq]

  modes |> as.numeric()
}
(mu_mod <- mode(nhanes$BMI))



# Problem 4b ----------------------------------------------------------
mode_stat <- function(data, indices) {
   mode(data[indices]) |> sample(size = 1)
}
# Perform bootstrap with boot()
bootstrap_results_mode <- boot(data = nhanes$BMI[!is.na(nhanes$BMI)],
```

```
                          statistic = mode_stat,
                          R = 1000)
# Extract bootstrap standard error
(se_mode_bootstrap <- sd(bootstrap_results_mode$t) |> round(3))

# Problem 4c ----------------------------------------------------------
23.2 - 2*7.974

23.2 + 2*7.974
```