# CS130 - Statistical Modeling: Prediction and Causal Inference

Minerva University

CS130 - Assignment 1: R Competency

Prof. Diamond

January 14, 2023

## Table of Contents

# Assignment 1: R Competency

## Executive Summary

Our comprehensive analysis of the strep_tb dataset, incorporating 107 tuberculosis patients, evaluates the efficacy of streptomycin treatment. We applied statistical tests to compare radiologic outcomes between female patients receiving different streptomycin doses. The results show a significant improvement in radiologic scores with a higher dose of streptomycin, supported by a Welch Two Sample t-test (p-value = 0.0149) and a Wilcoxon rank sum test (p-value = 0.008021). A rigorous check for normality affirmed the data's non-normal distribution, underscoring the relevance of non-parametric testing. A visualization of the data revealed a skewed distribution of radiologic scores toward higher efficacy at the 2-gram dose, indicating its potential benefit over the placebo.

## Data Exploration

*Please, check Appendix A.1 for Code.*

### Research Question:

**Objective:** To assess the efficacy of streptomycin vs. placebo in tuberculosis treatment.

**Participants:** 107 young tuberculosis patients.

**Design:** A randomized, placebo-controlled, 2-arm trial.

**Outcomes:** Streptomycin resistance after 6 months, radiological response at 6 months, overall improvement.

### Dataset Structure:

**Dimensions:** 107 rows (patients) and 13 columns (variables).

### Selected Variables:

**Baseline_cavitation:** Indicates presence (1_yes) or absence (0_no) of cavitation.

**Strep_resistance:** Categorizes streptomycin resistance level post-6 months therapy (1_sens_0-8, 2_mod_8-99, 3_resist_100+).

### Causal Inference:

**Modeling Approach:** Utilizing methods like propensity score matching to control for confounders and validate the randomness of treatment assignment.

**Statistical Significance:** Assessment of the treatment effect significance through hypothesis testing.

### Predictive Modeling:

**Techniques:** Logistic regression for binary outcomes, survival analysis for time-to-event data.

**Model Validation:** Employing cross-validation, ROC curves, and AUC metrics to evaluate predictive accuracy.

### Methodological Evolution:

**Historical Context:** Exploration of how clinical trial design and ethical considerations have evolved since 1948.

**Ethical Considerations:** The lack of modern ethical standards in the original study provides a historical perspective on the evolution of clinical research ethics.

### Statistical Optimization:

**Machine Learning Applications:** Using algorithms like Random Forests or SVM to identify patient subgroups most responsive to treatments.

**Decision Analysis:** Cost-effectiveness analysis or decision tree analysis to optimize treatment strategies.

### Additional Considerations:

**Implications for Current Research:** Using this data to inform current tuberculosis treatment research and clinical trial design, highlighting the importance of rigorous randomized control trials and ethical considerations in medical research.

## Data Analysis

*Please, check Appendix A.2 for Code.*

### Results:

1. Number of Male patients: 48
2. Number of Female patients: 59
3. Number of improved patients: 7
4. Percentage of improved patients: 22.58%
5. Median of radnum for females with dose_strep_g = 0 and dose_PAS_g = 0: 3
6. IQR of radnum for females with dose_strep_g = 0 and dose_PAS_g = 0: 4
7. Median of radnum for females with dose_strep_g = 2 and dose_PAS_g = 0: 5

8. IQR of radnum for females with dose_strep_g = 2 and dose_PAS_g = 0: 3.5

**Plausibility:**

Results are cross-verified with the raw data for confirmation.

**Number of Male and Female Patients:** The gender distribution (48 males, 59 females) seems reasonable for a clinical study with 107 participants.

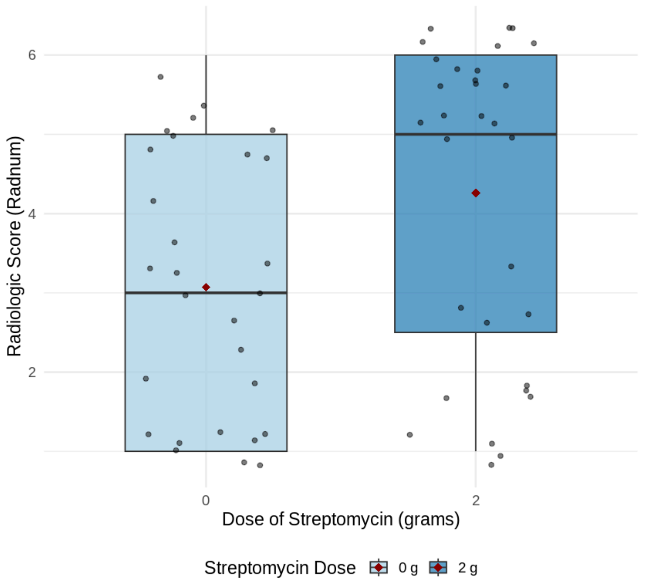
**Number and Percentage of Improved Patients:** The calculated number (7 improved patients) and percentage (22.58%) of improvement among female patients with a "Poor" baseline condition seem consistent with typical clinical study results. The percentage is calculated correctly as $\frac{7 * 100}{total\ \#\ of\ F\ patients\ w/\ Poor\ C}$ .

**Median and IQR of radnum for Different Treatment Groups:** The median and IQR values for 'radnum' in the specified subgroups of females (with different treatment doses) suggest variability in the radiologic response, which is expected in clinical data. The specific values (medians of 3 and 5, IQRs of 4 and 3.5) are consistent with a scenario where the response to the higher dose of Streptomycin (2 grams) is generally better (higher median radnum).

**Data visualization:**

*Please, check Appendix A.3 for Code.*

Radiologic Score Distribution by Streptomycin Dose in Female



Data sourced from the 'medicaldata' package. Points represent individual patients.

This boxplot illustrates the radiologic scores (rad_num) distribution for female patients, categorized by two Streptomycin doses: 0 grams and 2 grams. The interquartile range (IQR) depicted by the boxes captures the central 50% of scores, with the median—marked by a line inside each box—indicating the midpoint. For the 2-gram group, a median near the upper quartile suggests a positively skewed distribution.

Individual patient scores are shown as points; however, outliers are not displayed due to the exclusion setting in the code. The distribution's skewness is inferred from the data point spread around the median.

Red diamonds signify the group means, with the 2-gram group showing a higher mean compared to the 0-gram group, indicating potential greater efficacy at the higher dose.

The plot employs a colorblind-friendly palette and maintains transparency to aid in distinguishing overlapping points. Enhanced readability is achieved through a more extensive base font size and careful legend placement, ensuring data clarity without visual obstruction.

The boxplot is an exploratory tool to assess the effects of different Streptomycin doses on radiologic outcomes in female patients. A comprehensive analysis needs statistical tests to confirm apparent differences observed in the plot.

## Statistical Tests

*Please, check Appendix A.4 for Code.*

The statistical tests conducted on the strep_tb dataset provide information about the difference in radiologic scores (rad_num) between two groups of female patients treated with different doses of Streptomycin (0 grams and 2 grams).

**T-Test:** The Welch Two Sample t-test results in a t-statistic of -2.5109 and a p-value of 0.0149. Since the p-value is less than 0.05, this suggests a statistically significant difference in the means of rad_num between the two Streptomycin doses. The confidence interval for the difference in means does not include 0 (ranging from approximately -2.13 to -0.24), further supporting the evidence that the means differ. The negative sign of the t statistic indicates that the mean rad_num for the 0-gram group is less than the 2-gram group.

**Wilcoxon Test:** The Wilcoxon rank sum test is used here due to the non-normal distribution of rad_num. The test warns about the inability to compute an exact p-value due to ties in the data, which is common in ordinal or discrete data. Despite this, the test provides a p-value of 0.008021, also below the 0.05 threshold, suggesting a significant difference in the median score of rad_num between the two groups. This non-parametric test's results align with the t-test, confirming a difference in distribution between the two doses.

**Normality Check:** The Shapiro-Wilk normality test has a p-value of approximately 5.284e-06, far below 0.05, indicating that the rad_num data does not follow a normal distribution. This lack of normality justifies using the Wilcoxon test as an alternative to the t-test.

The t-test and the Wilcoxon test indicate a statistically significant difference in radiologic outcomes between the two doses of Streptomycin. The higher dose (2 grams) is associated with higher radiologic scores, suggesting it may be more effective than the lower dose. The assumption check for normality confirms that the data is not normally distributed, making the results of the Wilcoxon test particularly relevant.

## Conclusions

The statistical analyses substantiate the effectiveness of streptomycin in treating tuberculosis, with a clear dosage-response relationship evident in female patients. The higher dose (2 grams) consistently showed greater improvement in radiologic scores, reinforcing the drug's role in clinical therapy. These findings are pivotal for future research, suggesting the necessity to optimize streptomycin dosages for varying patient demographics. While the data robustly endorses the use of streptomycin, further studies are recommended to explore long-term impacts and to refine treatment strategies for personalized medicine.

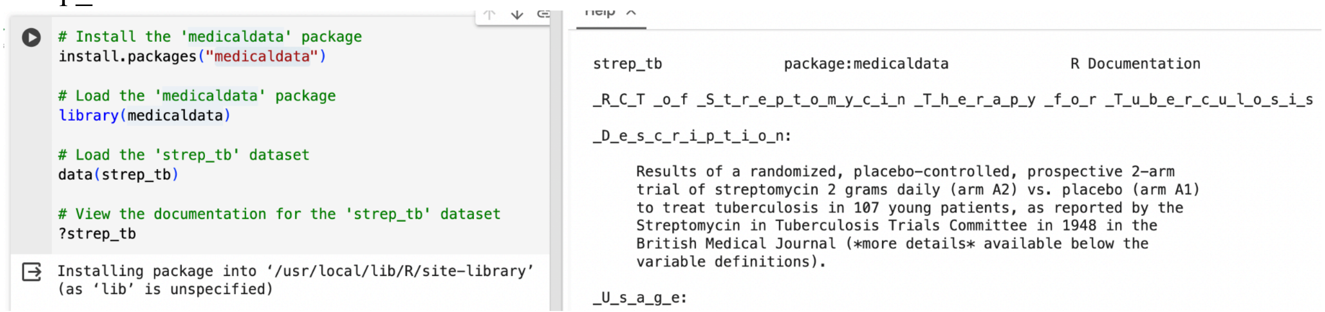
# Appendix

## Appendix A: Code Screenshots & Link

The code can be found through this link: Google Colab.

### Appendix A.1

Code is created to work with the strep_tb dataset and view its documentation by executing ?strep_tb in the R console.



```r
# Install the 'medicaldata' package
install.packages("medicaldata")

# Load the 'medicaldata' package
library(medicaldata)

# Load the 'strep_tb' dataset
data(strep_tb)

# View the documentation for the 'strep_tb' dataset
?strep_tb
```

```
Installing package into ‘/usr/local/lib/R/site-library’
(as ‘lib’ is unspecified)
```

```
strep_tb               package:medicaldata               R Documentation

_R_C_T _o_f _S_t_r_e_p_t_o_m_y_c_i_n _T_h_e_r_a_p_y _f_o_r _T_u_b_e_r_c_u_l_o_s_i_s

_D_e_s_c_r_i_p_t_i_o_n:

     Results of a randomized, placebo-controlled, prospective 2-arm
     trial of streptomycin 2 grams daily (arm A2) vs. placebo (arm A1)
     to treat tuberculosis in 107 young patients, as reported by the
     Streptomycin in Tuberculosis Trials Committee in 1948 in the
     British Medical Journal (*more details* available below the
     variable definitions).

_U_s_a_g_e:
```

### Appendix A.2

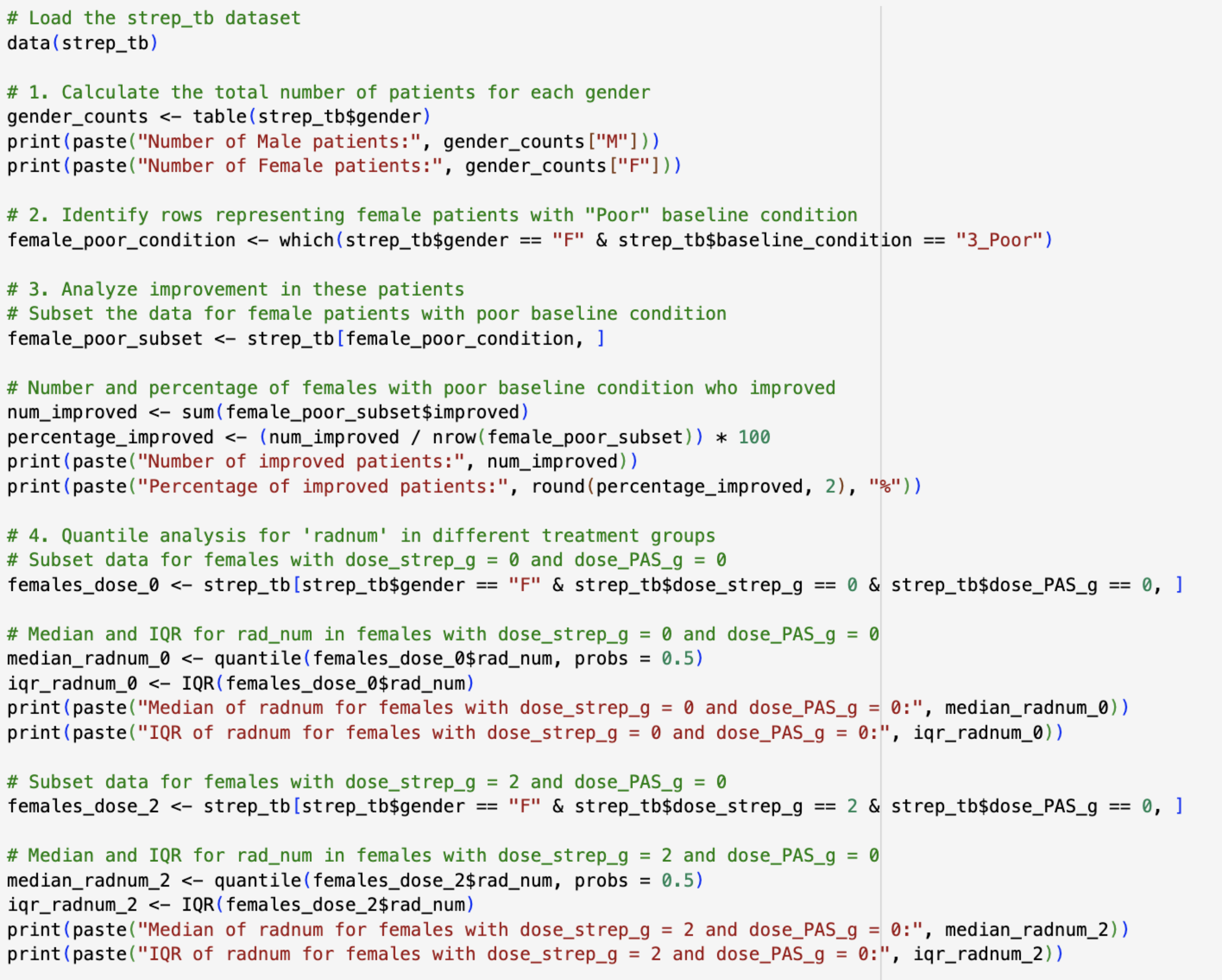
This code performs the following tasks:

1. Calculates the number of patients by gender using the `table` function.

2. Identifies rows where patients are female with a "Poor" baseline condition using the `which` command.

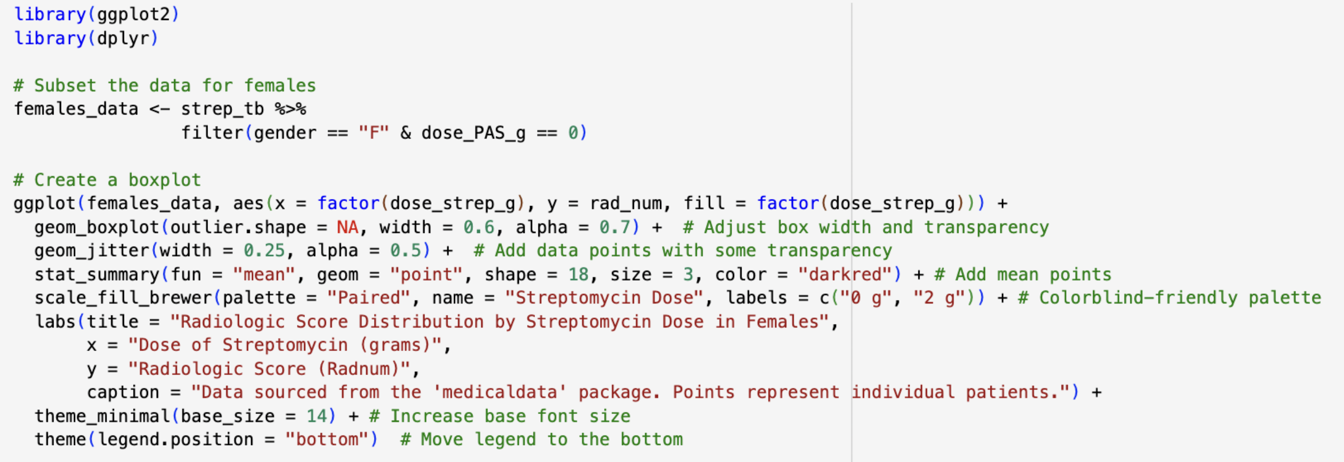
3. Analyzes improvement in these patients by calculating the number and percentage of improved patients.

4. Performs a quantile analysis on the 'radnum' variable for different treatment groups (females receiving different doses of Streptomycin and PAS) to obtain the median and interquartile range (IQR).



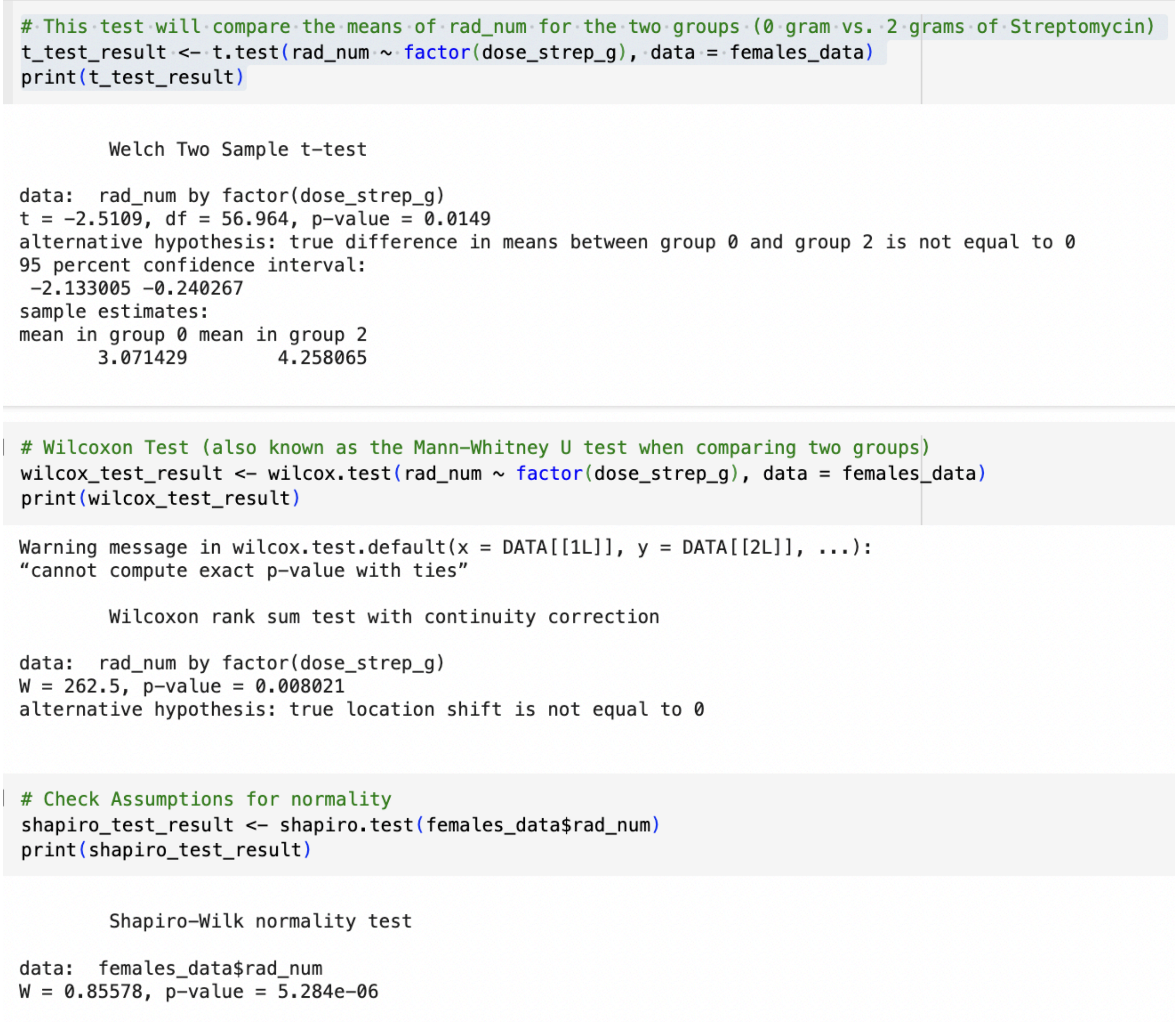
```
# Load the strep_tb dataset
data(strep_tb)

# 1. Calculate the total number of patients for each gender
gender_counts <- table(strep_tb$gender)
print(paste("Number of Male patients:", gender_counts["M"]))
print(paste("Number of Female patients:", gender_counts["F"]))

# 2. Identify rows representing female patients with "Poor" baseline condition
female_poor_condition <- which(strep_tb$gender == "F" & strep_tb$baseline_condition == "3_Poor")

# 3. Analyze improvement in these patients
# Subset the data for female patients with poor baseline condition
female_poor_subset <- strep_tb[female_poor_condition, ]

# Number and percentage of females with poor baseline condition who improved
num_improved <- sum(female_poor_subset$improved)
percentage_improved <- (num_improved / nrow(female_poor_subset)) * 100
print(paste("Number of improved patients:", num_improved))
print(paste("Percentage of improved patients:", round(percentage_improved, 2), "%"))

# 4. Quantile analysis for 'radnum' in different treatment groups
# Subset data for females with dose_strep_g = 0 and dose_PAS_g = 0
females_dose_0 <- strep_tb[strep_tb$gender == "F" & strep_tb$dose_strep_g == 0 & strep_tb$dose_PAS_g == 0, ]

# Median and IQR for rad_num in females with dose_strep_g = 0 and dose_PAS_g = 0
median_radnum_0 <- quantile(females_dose_0$rad_num, probs = 0.5)
iqr_radnum_0 <- IQR(females_dose_0$rad_num)
print(paste("Median of radnum for females with dose_strep_g = 0 and dose_PAS_g = 0:", median_radnum_0))
print(paste("IQR of radnum for females with dose_strep_g = 0 and dose_PAS_g = 0:", iqr_radnum_0))

# Subset data for females with dose_strep_g = 2 and dose_PAS_g = 0
females_dose_2 <- strep_tb[strep_tb$gender == "F" & strep_tb$dose_strep_g == 2 & strep_tb$dose_PAS_g == 0, ]

# Median and IQR for rad_num in females with dose_strep_g = 2 and dose_PAS_g = 0
median_radnum_2 <- quantile(females_dose_2$rad_num, probs = 0.5)
iqr_radnum_2 <- IQR(females_dose_2$rad_num)
print(paste("Median of radnum for females with dose_strep_g = 2 and dose_PAS_g = 0:", median_radnum_2))
print(paste("IQR of radnum for females with dose_strep_g = 2 and dose_PAS_g = 0:", iqr_radnum_2))
```

## Appendix A.3



```
library(ggplot2)
library(dplyr)

# Subset the data for females
females_data <- strep_tb %>%
                filter(gender == "F" & dose_PAS_g == 0)

# Create a boxplot
ggplot(females_data, aes(x = factor(dose_strep_g), y = rad_num, fill = factor(dose_strep_g))) +
  geom_boxplot(outlier.shape = NA, width = 0.6, alpha = 0.7) +  # Adjust box width and transparency
  geom_jitter(width = 0.25, alpha = 0.5) +  # Add data points with some transparency
  stat_summary(fun = "mean", geom = "point", shape = 18, size = 3, color = "darkred") + # Add mean points
  scale_fill_brewer(palette = "Paired", name = "Streptomycin Dose", labels = c("0 g", "2 g")) + # Colorblind-friendly palette
  labs(title = "Radiologic Score Distribution by Streptomycin Dose in Females",
       x = "Dose of Streptomycin (grams)",
       y = "Radiologic Score (Radnum)",
       caption = "Data sourced from the 'medicaldata' package. Points represent individual patients.") +
  theme_minimal(base_size = 14) + # Increase base font size
  theme(legend.position = "bottom")  # Move legend to the bottom
```

## Appendix A.4



```r
# This test will compare the means of rad_num for the two groups (0 gram vs. 2 grams of Streptomycin)
t_test_result <- t.test(rad_num ~ factor(dose_strep_g), data = females_data)
print(t_test_result)
```

```
        Welch Two Sample t-test

data:  rad_num by factor(dose_strep_g)
t = -2.5109, df = 56.964, p-value = 0.0149
alternative hypothesis: true difference in means between group 0 and group 2 is not equal to 0
95 percent confidence interval:
 -2.133005 -0.240267
sample estimates:
mean in group 0 mean in group 2
       3.071429        4.258065
```

```r
# Wilcoxon Test (also known as the Mann-Whitney U test when comparing two groups)
wilcox_test_result <- wilcox.test(rad_num ~ factor(dose_strep_g), data = females_data)
print(wilcox_test_result)
```

```
Warning message in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...):
“cannot compute exact p-value with ties”

        Wilcoxon rank sum test with continuity correction

data:  rad_num by factor(dose_strep_g)
W = 262.5, p-value = 0.008021
alternative hypothesis: true location shift is not equal to 0
```

```r
# Check Assumptions for normality
shapiro_test_result <- shapiro.test(females_data$rad_num)
print(shapiro_test_result)
```

```
        Shapiro-Wilk normality test

data:  females_data$rad_num
W = 0.85578, p-value = 5.284e-06
```