R Markdown Yahriel Salinas-Reyes HW 2

Yahriel Salinas-Reyes

2024-09-25

# Question 1

## Q.1.1 State the Ho, Ha Hypothesis

Null hypothesis (Ho): The dietary supplement has no effect on blood pressure. That is, the mean blood pressure after taking the supplement is the same as before (mu\_before=mu\_after)

Alternative hypothesis (Ha): The dietary supplement reduces blood pressure. That is, the mean blood pressure after taking the supplement is less than before. (mu\_before > mu\_after)

## Q.1.2.

# Blood pressure before taking the supplement  
before <- c(130, 125, 140, 135, 138, 132, 137, 142, 129, 136, 133, 139, 140, 150)  
# Blood pressure after taking the supplement for 8 weeks  
after <- c(122, 118, 133, 127, 129, 124, 130, 135, 121, 128, 126, 132, 120, 123)  
  
# Run a paired t-test  
t\_test\_result <- t.test(after, before, paired = TRUE, alternative = "less")

## Q.1.3

# View the results  
print(t\_test\_result)

##   
## Paired t-test  
##   
## data: after and before  
## t = -6.1753, df = 13, p-value = 1.674e-05  
## alternative hypothesis: true mean difference is less than 0  
## 95 percent confidence interval:  
## -Inf -7.030364  
## sample estimates:  
## mean difference   
## -9.857143

Based on the Output, we get the following results

t-statistic = -6.1753 Degrees of freedom (df) = 13 p-value = .00001674 95% CI for the difference in means: (-INF, -7.030364)

## Q.1.4

# If the Ho is true, the probability of having a statistic as extreme as -6.1753 or more extreme, is .001674%.

# Since the value of Ho is outside of the 95% interval # we reject the null hypothesis.

# We conclude that there is a significant reduction in blood pressure after taking the dietary supplement.

# Question 2

## Q.2.1

* Null Hypothesis (Ho): There is no significant difference in productivity before and after the training program (mu\_before = mu\_after).
* Alternative Hypothesis (Ha): There is a significant difference in productivity before and after the training program (mu\_before =/ mu\_after)

## Q.2.2

Since we have two sets of related measurements that are paired (before and after), we can use a paired t-test to compare the means. The asumptions for a paired t-test include: - The difference between the pairs of data (before-after) should be approximately normally distributed. We can check this by doing a normality test (Shapiro-Wilk test) and then performing a paired t-test.

# Data  
before <- c(20, 22, 19, 5, 23, 21, 25, 22, 24, 10)  
after <- c(15, 26, 21, 10, 27, 23, 60, 26, 27, 50)  
  
# Differences  
diff <- before - after  
  
# Check normality of the differences (Shapiro-Wilk test)  
shapiro.test(diff)

##   
## Shapiro-Wilk normality test  
##   
## data: diff  
## W = 0.68162, p-value = 0.0005278

# Perform a paired t-test  
t.test(before, after, paired = TRUE)

##   
## Paired t-test  
##   
## data: before and after  
## t = -1.9665, df = 9, p-value = 0.08079  
## alternative hypothesis: true mean difference is not equal to 0  
## 95 percent confidence interval:  
## -20.213227 1.413227  
## sample estimates:  
## mean difference   
## -9.4

## Q.2.3

The P-Value = 0.08079

## Q.2.4

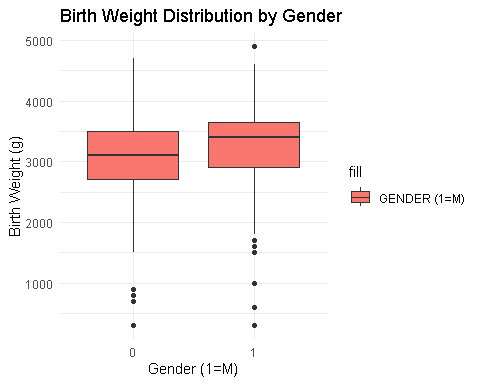
Based on the p-value 0.08079, explain the result using the following language # If the Ho is true, the probability of having a difference of this magnitude # or more extreme, is 8.079% # Since the value of p value 0.08079 is greater than .05 # we fail to reject the null hypothesis. # We conclude that there is not a significant difference in productivity before and after the training program.

# Question 3

library(ggplot2)  
library(dplyr)  
library(readxl)  
Births <- read\_excel("Births.xlsx")  
#View(Births)  
  
  
  
# Convert GENDER to a factor  
Births$`GENDER (1=M)` <- as.factor(Births$`GENDER (1=M)`)  
levels(Births$`GENDER (1=M)`) <- c("0", "1")  
  
# Check the structure of the data  
str(Births)

## tibble [400 × 8] (S3: tbl\_df/tbl/data.frame)  
## $ FACILITY : chr [1:400] "Albany Medical Center Hospital" "Albany Medical Center Hospital" "Albany Medical Center Hospital" "Albany Medical Center Hospital" ...  
## $ INSURANCE : chr [1:400] "Insurance Company" "Blue Cross" "Blue Cross" "Insurance Company" ...  
## $ GENDER (1=M) : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 1 2 1 2 ...  
## $ LENGTH OF STAY: num [1:400] 2 2 36 5 2 4 3 2 2 2 ...  
## $ ADMITTED : chr [1:400] "FRI" "FRI" "WED" "MON" ...  
## $ DISCHARGED : chr [1:400] "SUN" "SUN" "THU" "SAT" ...  
## $ BIRTH WEIGHT : num [1:400] 3500 3900 800 2800 3700 2400 4200 4000 3100 3400 ...  
## $ TOTAL CHARGES : num [1:400] 13986 3632 359091 8536 3632 ...

# Visualize the birth weight distribution  
ggplot(Births, aes(x = `GENDER (1=M)`, y = `BIRTH WEIGHT`, fill = 'GENDER (1=M)')) +  
 geom\_boxplot() +  
 labs(title = "Birth Weight Distribution by Gender",  
 x = "Gender (1=M)",  
 y = "Birth Weight (g)") +  
 theme\_minimal()



# Perform normality test for each group (Shapiro-Wilk Test)  
shapiro.test(Births$`BIRTH WEIGHT`[Births$`GENDER (1=M)` == "0"])

##   
## Shapiro-Wilk normality test  
##   
## data: Births$`BIRTH WEIGHT`[Births$`GENDER (1=M)` == "0"]  
## W = 0.93861, p-value = 1.291e-07

shapiro.test(Births$`BIRTH WEIGHT`[Births$`GENDER (1=M)` == "1"])

##   
## Shapiro-Wilk normality test  
##   
## data: Births$`BIRTH WEIGHT`[Births$`GENDER (1=M)` == "1"]  
## W = 0.91619, p-value = 4.358e-09

# Since the data may not be normally distributed, use the Mann-Whitney U test  
wilcox.test(`BIRTH WEIGHT` ~ `GENDER (1=M)`, data = Births, alternative = "two.sided")

##   
## Wilcoxon rank sum test with continuity correction  
##   
## data: BIRTH WEIGHT by GENDER (1=M)  
## W = 15416, p-value = 7.465e-05  
## alternative hypothesis: true location shift is not equal to 0

# Display summary statistics  
Births %>%  
 group\_by(`GENDER (1=M)`) %>%  
 summarise(median\_bw = median(`BIRTH WEIGHT`),  
 mean\_bw = mean(`BIRTH WEIGHT`),  
 sd\_bw = sd(`BIRTH WEIGHT`))

## # A tibble: 2 × 4  
## `GENDER (1=M)` median\_bw mean\_bw sd\_bw  
## <fct> <dbl> <dbl> <dbl>  
## 1 0 3100 3037. 706.  
## 2 1 3400 3273. 660.

To test the claim that girls and boys have the same median birth weight using a 0.05 significance level, we can conduct a Mann-Whitney U test (A.K.A. Wilcoxon rank-sum test) since it’s a non-parametric test suitable when testing the difference in medians between two independent groups. This test is used instead of a t-test if the assumption of normality is violated.

## Hypothesis

* Null Hypothesis (H0): The median birth weights of boys and girls are the same.
* Alternative Hypothesis (Ha): The median birth weights of boys and girls are different.

## Assumption Checks

* Independence: The birth weight of one child does not influence another child’s birth weight.
* Non-normality: Birth weight data may not be normally distributed. This can be checked visually (using histograms or boxplots) or by a normality test (e.g., Shapiro-Wilk test). If the data is not normal, the Mann-Whitney U test will be appropriate.

## R Output Interpretations

* Boxplot: Visualizes if any obvious difference in birth weights between boys and girls.
* Shapiro-Wilk Test: Checks if birth weights for both genders are normally distributed. If p < 0.05, the data is not normally distributed. FOr both groups, the p-value is less than alpha and therefore not normally distributed.
* Mann-Whitney U Test: This test outputs a p-value. If p < 0.05, you reject the null hypothesis.

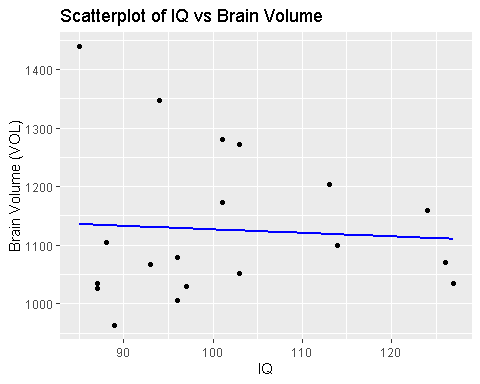
## Conclusion

If p < 0.05: There is sufficient evidence to reject the null hypothesis, meaning there is a significant difference in the median birth weight between boys and girls. Since the p-value is for both male and female is less than 0.05, we reject the null hypothesis and therefore they are not normally distributed and a significant difference in the median birth weights.

# Question 4

library(readxl)  
IQ\_Brain\_Size <- read\_excel("IQ\_Brain\_Size.xlsx")  
data <- IQ\_Brain\_Size   
  
library(ggplot2)  
  
# Assumption check: Scatter plot  
ggplot(data, aes(x = IQ, y = VOL)) +  
 geom\_point() +  
 geom\_smooth(method = "lm", se = FALSE, color = "blue") +  
 labs(title = "Scatterplot of IQ vs Brain Volume", x = "IQ", y = "Brain Volume (VOL)")

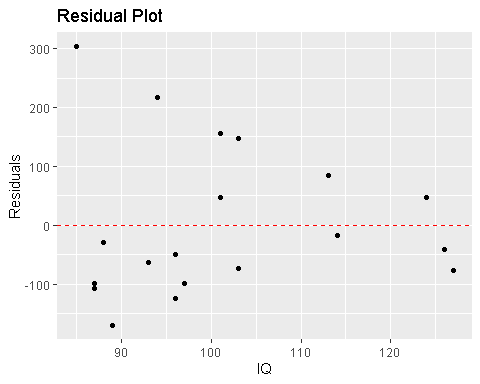
## `geom\_smooth()` using formula = 'y ~ x'



# Pearson correlation test  
cor\_test <- cor.test(data$IQ, data$VOL)  
  
# Output correlation coefficient and p-value  
cor\_test

##   
## Pearson's product-moment correlation  
##   
## data: data$IQ and data$VOL  
## t = -0.26949, df = 18, p-value = 0.7906  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## -0.4921080 0.3900712  
## sample estimates:  
## cor   
## -0.06339202

# Plot residuals for assumption check  
model <- lm(VOL ~ IQ, data = data)  
ggplot(data, aes(x = IQ, y = model$residuals)) +  
 geom\_point() +  
 geom\_hline(yintercept = 0, color = "red", linetype = "dashed") +  
 labs(title = "Residual Plot", x = "IQ", y = "Residuals")



# Summary of the linear regression model (optional)  
summary(model)

##   
## Call:  
## lm(formula = VOL ~ IQ, data = data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -170.14 -82.11 -45.45 57.19 303.46   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1186.5013 226.5056 5.238 5.56e-05 \*\*\*  
## IQ -0.5995 2.2246 -0.269 0.791   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 128.1 on 18 degrees of freedom  
## Multiple R-squared: 0.004019, Adjusted R-squared: -0.05131   
## F-statistic: 0.07263 on 1 and 18 DF, p-value: 0.7906

## Hypothesis:

* Null Hypothesis (Ho): There is no linear correlation between IQ and Brain Volume (VOL), (correlation coeff.=0).
* Alternative Hypothesis (Ha): There is a linear correlation between IQ and Brain Volume (VOL), (correlation coeff. =/ 0).

## Assumption Checks:

* Both IQ and VOL should be continuous variables.
* Data should follow a bivariate normal distribution.
* Check linearity through scatter plots.

## Interpretation of Output

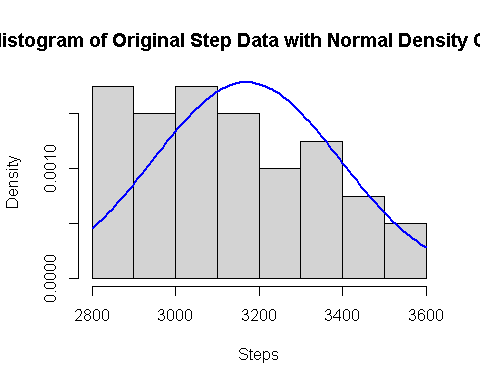
* Correlation Coefficient (r): Since the Correlation Coeff. = -0.06339202 , we can see that there is a negative correlation between the two groups.
* p-value: Since the p-value = 0.7906, we fail to reject Null and conclude that there is not a significant linear correlation between IQ and Brain Volume.

## Conclusions

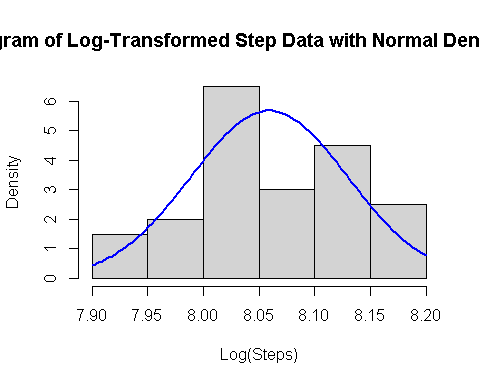
* Since the p-value is greater than 0.05, we fail to reject Null and conclude that there is not enough evidence to support the claim of a linear correlation.

# Question 5

# Original data  
set.seed(123)  
steps <- c(3200, 2900, 3100, 3000, 2800, 3500, 3100, 3000, 3300, 3400,  
 2800, 2900, 3200, 3000, 3100, 3300, 3400, 3600, 3500, 3400,  
 3000, 2900, 3100, 3200, 3100, 3000, 2800, 2900, 3200, 3100,  
 3000, 3200, 3300, 3400, 3500, 3600, 3400, 3300, 3200, 3100)  
  
## Q.5.1  
# Plotting histogram of original data  
hist(steps, prob = TRUE, main = "Histogram of Original Step Data with Normal Density Curve", xlab = "Steps")  
curve(dnorm(x, mean = mean(steps), sd = sd(steps)), add = TRUE, col = "blue", lwd = 2)



## Q.5.2  
# Log transformation  
log\_steps <- log(steps)  
  
# Plotting histogram of log-transformed data  
hist(log\_steps, prob = TRUE, main = "Histogram of Log-Transformed Step Data with Normal Density Curve", xlab = "Log(Steps)")  
curve(dnorm(x, mean = mean(log\_steps), sd = sd(log\_steps)), add = TRUE, col = "blue", lwd = 2)



## Q.5.3  
# Shapiro-Wilk test for original data  
shapiro\_original <- shapiro.test(steps)  
  
# Shapiro-Wilk test for log-transformed data  
shapiro\_log <- shapiro.test(log\_steps)  
  
# Display results  
shapiro\_original

##   
## Shapiro-Wilk normality test  
##   
## data: steps  
## W = 0.95932, p-value = 0.1588

shapiro\_log

##   
## Shapiro-Wilk normality test  
##   
## data: log\_steps  
## W = 0.96127, p-value = 0.1851

## Q.5.4

Interpretation and Conclusion: - Visual Inspection (Histogram): The histogram of the original data likely shows a skewed distribution, while the log-transformed data may appear closer to normal (depending on how the data looks). - Normality Test (Shapiro-Wilk): The p-value from the Shapiro-Wilk test helps determine if the data follows a normal distribution. A p-value greater than 0.05 would indicates the data is normally distributed. Comparing the p-values of the original and log-transformed data shows that normality improved.

# Question 6

## Q.6.1 Hypothesis

* Null Hypothesis (Ho): The mean arsenic levels in rice from Arkansas, California, and Texas are the same.
* Alternative Hypothesis (Ha): At least one of the states has a different mean arsenic level in rice.

## Q.6.2 Assumptions for One-way ANOVA

* Normality: The arsenic levels within each state should be normally distributed. We can use Q-Q plots or the Shapiro-Wilk test to check this.
* Homogeneity of variances: The variances of arsenic levels across the groups should be equal. We can check this using Levene’s test.
* Independent observations: The samples from each state should be independent of each other.

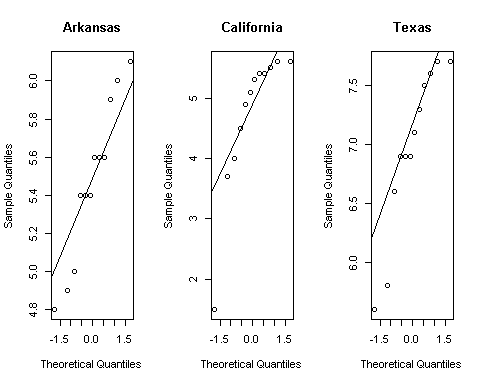
library(readxl)  
arsenic\_data <- read\_excel("Arsenic.xlsx")  
  
# Convert STATE to a factor  
arsenic\_data$STATE <- as.factor(arsenic\_data$STATE)  
  
# Check the structure of the data  
str(arsenic\_data)

## tibble [36 × 2] (S3: tbl\_df/tbl/data.frame)  
## $ STATE : Factor w/ 3 levels "Arkansas","California",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ ARSENIC: num [1:36] 4.8 4.9 5 5.4 5.4 5.4 5.6 5.6 5.6 5.9 ...

# Q.6.2. Assumption Checks  
# Check for homogeneity of variances using Levene's Test  
leveneTest(ARSENIC ~ STATE, data = arsenic\_data)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 1.305 0.2848  
## 33

# Check for normality using QQ plots and Shapiro-Wilk test for each group  
par(mfrow = c(1, 3)) # Set up a 1x3 plotting area  
with(arsenic\_data, {  
 qqnorm(ARSENIC[STATE == "Arkansas"], main = "Arkansas")  
 qqline(ARSENIC[STATE == "Arkansas"])  
 qqnorm(ARSENIC[STATE == "California"], main = "California")  
 qqline(ARSENIC[STATE == "California"])  
 qqnorm(ARSENIC[STATE == "Texas"], main = "Texas")  
 qqline(ARSENIC[STATE == "Texas"])  
})



shapiro.test(arsenic\_data$ARSENIC[arsenic\_data$STATE == "Arkansas"])

##   
## Shapiro-Wilk normality test  
##   
## data: arsenic\_data$ARSENIC[arsenic\_data$STATE == "Arkansas"]  
## W = 0.94086, p-value = 0.5093

shapiro.test(arsenic\_data$ARSENIC[arsenic\_data$STATE == "California"])

##   
## Shapiro-Wilk normality test  
##   
## data: arsenic\_data$ARSENIC[arsenic\_data$STATE == "California"]  
## W = 0.75277, p-value = 0.002857

shapiro.test(arsenic\_data$ARSENIC[arsenic\_data$STATE == "Texas"])

##   
## Shapiro-Wilk normality test  
##   
## data: arsenic\_data$ARSENIC[arsenic\_data$STATE == "Texas"]  
## W = 0.88129, p-value = 0.09105

# Q.6.3. Perform the one-way ANOVA  
anova\_model <- aov(ARSENIC ~ STATE, data = arsenic\_data)  
  
  
# Get the summary of the ANOVA (provides MS Between, MS Within, and F value)  
summary(anova\_model)

## Df Sum Sq Mean Sq F value Pr(>F)   
## STATE 2 31.65 15.83 22.95 5.68e-07 \*\*\*  
## Residuals 33 22.76 0.69   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

cat("Q.6.3. Mean Squared Between (MS Between) = 15.83")

## Q.6.3. Mean Squared Between (MS Between) = 15.83

cat("Q.6.4. Mean Squared Between (Ms Within) = 0.69")

## Q.6.4. Mean Squared Between (Ms Within) = 0.69

cat("Q.6.5. F-Value = 22.95")

## Q.6.5. F-Value = 22.95

# Q.6.6. Interpretation of the p-value  
p\_value <- .000000568  
  
# Q.6.7. Conclusion based on the p-value  
cat("Reject Ho: The mean arsenic levels differ between states since the p-value < 0.005\n")

## Reject Ho: The mean arsenic levels differ between states since the p-value < 0.005

# Q.6.8. Post-hoc test (Tukey's HSD)  
tukey\_test <- TukeyHSD(anova\_model)  
print(tukey\_test)

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = ARSENIC ~ STATE, data = arsenic\_data)  
##   
## $STATE  
## diff lwr upr p adj  
## California-Arkansas -0.7666667 -1.5985757 0.06524234 0.0757234  
## Texas-Arkansas 1.4916667 0.6597577 2.32357567 0.0003075  
## Texas-California 2.2583333 1.4264243 3.09024234 0.0000004

## Conclusions

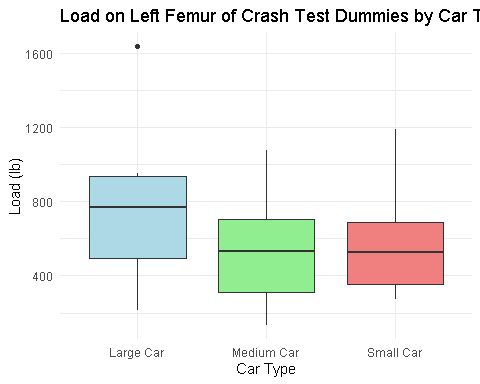
Based on the Post-hoc test, we test the claim that these states have the same amount of arsenic. Based on the difference, we can conclude that they do not… Therefore we reject the Null. diff California-Arkansas -0.7666667 Texas-Arkansas 1.4916667  
Texas-California 2.2583333

# Question 7

## Q.7.1

* Null Hypothesis (Ho): The median femur loads are the same across the different car categories (small car, medium car, large car).
* Alternatice Hypothesis (Ha): The median femur loads differ across at least one car category.

# Load necessary libraries  
library(ggplot2)  
  
# Define the data  
smallcar <- c(548,782,1188,707,324,320,634,501,274,437)  
mcar <- c(194,280,1076,411,617,133,719,656,874,445)  
largecar <- c(215,937,953,1636,937,472,882,562,656,433)  
  
# Combine the data into a single dataframe  
car\_data <- data.frame(  
 load = c(smallcar, mcar, largecar),  
 car\_type = factor(rep(c("Small Car", "Medium Car", "Large Car"), each = 10))  
)  
  
## Q.7.2 Descriptive statistics plot: Boxplot  
ggplot(car\_data, aes(x = car\_type, y = load)) +  
 geom\_boxplot(fill = c("lightblue", "lightgreen", "lightcoral")) +  
 theme\_minimal() +  
 labs(title = "Load on Left Femur of Crash Test Dummies by Car Type",   
 x = "Car Type",   
 y = "Load (lb)")



## Q.7.3 Perform Kruskal-Wallis Test  
kruskal\_test <- kruskal.test(load ~ car\_type, data = car\_data)  
print(kruskal\_test)

##   
## Kruskal-Wallis rank sum test  
##   
## data: load by car\_type  
## Kruskal-Wallis chi-squared = 2.3514, df = 2, p-value = 0.3086

## Q.7.4 Interpretation of results  
if(kruskal\_test$p.value < 0.05) {  
 cat("Reject the null hypothesis: There is a significant difference in medians between car types.\n")  
} else {  
 cat("Fail to reject the null hypothesis: There is no significant difference in medians between car types.\n")  
}

## Fail to reject the null hypothesis: There is no significant difference in medians between car types.

## Q.7.5 Discuss safety  
if(median(largecar) < median(smallcar)) {  
 cat("Based on the medians, larger cars appear to exert less load on the femur, suggesting they might be safer.")  
} else {  
 cat("There is no evidence to suggest larger cars are safer based on this data.")  
}

## There is no evidence to suggest larger cars are safer based on this data.

# Question 8

# Question 9