Individual Assignment II

Group(1) Kyaw Min Khaing

### Load necessary packages

library(pacman)  
p\_load(tidyverse,rstatix,randtests,ggpubr,knitr)

#### No-1(a) A physician who specializes in genetic diseases develops a theory which predicts that two-thirds of the people who develop a disease called cyclomeiosis will be males. She randomly selects 300 people who are afflicted with cyclomeiosis and observes that 140 of them are females. Is the physician’s theory supported? Above the median (Males30 Females60) Below the median (Males70 Females60)

## Hypothesis Statements

* **Null Hypothesis (H0):** The proportion of males afflicted is **2/3**.
* **Alternative Hypothesis (HA):** The proportion of males afflicted is **not 2/3**. The Chi-Square test statistic () is calculated using the formula:

where:  
- = Observed frequency  
- = Expected frequency

# Given observed frequencies  
observed <- c(Male = 160, Female = 140)  
  
# Expected proportions  
expected\_prop <- c(Male = 2/3, Female = 1/3)  
  
# Sample size  
N <- sum(observed)  
  
# Expected frequencies  
expected <- expected\_prop \* N  
  
# Combine into a tibble  
data <- tibble(  
 Category = names(observed),  
 Observed = observed,  
 Expected = expected,  
 Chi\_Square\_Contribution = (Observed - Expected)^2 / Expected  
)  
kable(data, caption = "chi\_square\_contribution")

chi\_square\_contribution

|  |  |  |  |
| --- | --- | --- | --- |
| Category | Observed | Expected | Chi\_Square\_Contribution |
| Male | 160 | 200 | 8 |
| Female | 140 | 100 | 16 |

Performing the Chi-Square Test

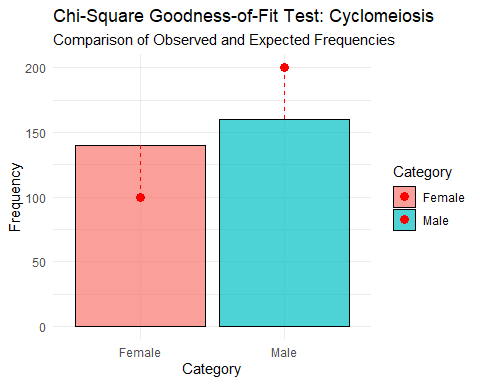
chi\_sq\_stat <- sum(data$Chi\_Square\_Contribution)  
  
# Degrees of freedom (df = categories - 1)  
df <- length(observed) - 1  
  
# Critical value at alpha = 0.05  
critical\_value <- qchisq(0.05, df, lower.tail = FALSE)  
  
# Decision Rule  
decision <- ifelse(chi\_sq\_stat >= critical\_value, "Reject H0", "Fail to Reject H0")  
  
# Display results  
tibble(Chi\_Square\_Statistic = chi\_sq\_stat, Critical\_Value = critical\_value, Decision = decision)

## # A tibble: 1 × 3  
## Chi\_Square\_Statistic Critical\_Value Decision   
## <dbl> <dbl> <chr>   
## 1 24 3.84 Reject H0

Conclusion: there is no supportive predictions of the physician’s theory, stating that the proportion of males among those afflicted is not 2/3 .

#### Visualize observed vs. expected counts

library(ggplot2)  
  
ggplot(data, aes(x = Category, y = Observed, fill = Category)) +  
 geom\_bar(stat = "identity", color = "black", alpha = 0.7) +  
 geom\_point(aes(y = Expected), color = "red", size = 3) +  
 geom\_segment(aes(xend = Category, y = Expected, yend = Observed), color = "red", linetype = "dashed") +  
 labs(  
 title = "Chi-Square Goodness-of-Fit Test: Cyclomeiosis",  
 subtitle = "Comparison of Observed and Expected Frequencies",  
 y = "Frequency",  
 x = "Category"  
 ) +  
 theme\_minimal()



#### Interpretation

Based on the computed Chi-Square statistic and critical value, we conclude: - If the test statistic is **greater than or equal to** the critical value, we **reject H0**. - Otherwise, we **fail to reject H0**.

This result indicates whether the physician’s theory is supported or not.

## No-1(b) A study is conducted to determine whether five-year old females are more likely than five-year old males to score above the population median on a standardized test of eye-hand coordination. One hundred randomly selected females and 100 randomly selected males are © 2000 by Chapman & Hall/CRC administered the test of eye-hand coordination, and categorized with respect to whether they score above or below the overall population median (i.e., the 50th percentile for both males and females). Table 16.11 summarizes the results of the study. Do the data indicate that there are gender differences in performance? Above the median (Males30 Females60) Below the median (Males70 Females60)

# Create a dataframe for the study  
study\_data <- data.frame(  
 Gender = c("Males", "Females"),  
 Above\_Median = c(30, 60),  
 Below\_Median = c(70, 40)  
)  
# Print the data  
kable(study\_data, caption = "study\_data")

study\_data

|  |  |  |
| --- | --- | --- |
| Gender | Above\_Median | Below\_Median |
| Males | 30 | 70 |
| Females | 60 | 40 |

We use a Chi-Square test for independence to determine if there is a significant association between gender and test performance.

# Create a contingency table  
contingency\_table <- as.table(rbind(  
 c(30, 70), # Males  
 c(60, 40) # Females  
))  
rownames(contingency\_table) <- c("Males", "Females")  
colnames(contingency\_table) <- c("Above Median", "Below Median")  
# Perform the Chi-Square test  
chisq\_test <- chisq.test(contingency\_table)  
# Display the results  
chisq\_test

##   
## Pearson's Chi-squared test with Yates' continuity correction  
##   
## data: contingency\_table  
## X-squared = 16.99, df = 1, p-value = 3.758e-05

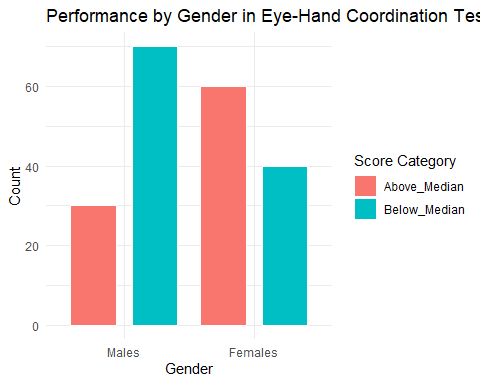
### Report significance

alpha <- 0.05 # Significance level  
if (chisq\_test$p.value < alpha) {  
 cat("The result is significant at α =", alpha,   
 "\n(p-value =", round(chisq\_test$p.value, 4), ").",  
 "\nThere is evidence of a relationship between gender and performance.\n")  
} else {  
 cat("The result is not significant at α =", alpha,   
 "\n(p-value =", round(chisq\_test$p.value, 4), ").",  
 "\nThere is no evidence of a relationship between gender and performance.\n")  
}

## The result is significant at α = 0.05   
## (p-value = 0 ).   
## There is evidence of a relationship between gender and performance.

### Visualization

# Transform the data for plotting  
plot\_data <- study\_data %>%  
 pivot\_longer(cols = c(Above\_Median, Below\_Median),  
 names\_to = "Score\_Category",  
 values\_to = "Count") %>%  
 mutate(Score\_Category = factor(Score\_Category, levels = c("Above\_Median", "Below\_Median")))  
  
# Plot  
gender\_plot <- ggbarplot(  
 plot\_data, x = "Gender", y = "Count", fill = "Score\_Category",  
 color = "white", position = position\_dodge(),  
 xlab = "Gender", ylab = "Count",  
 legend.title = "Score Category",  
 title = "Performance by Gender in Eye-Hand Coordination Test"  
) +  
 theme\_minimal()  
gender\_plot



## No-2(a) Doctor Radical, a math instructor at Logarithm University, has three classes in advanced calculus. There are five students in each class. The instructor uses a programmed textbook in Class 1, a conventional textbook in Class 2, and his own printed notes in Class 3. At the end of the semester, in order to determine if the type of instruction employed influences student performance, Dr. Radical has another math instructor, Dr. Root, rank the 15 students in the three classes with respect to math ability. The rankings of the students in the three classes follow: Class 1: 9.5, 14.5, 12.5, 14.5, 12.5; Class 2: 6, 9.5, 3, 9.5, 3; and Class 3: 1, 9.5, 6, 3, 6 (assume the lower the rank, the better the student). At α = 0.05, is there a difference three classes in advanced calculus.?

### To input Dataset

The rankings of the students in the three classes are as follows: Class 1 (Programmed Textbook): 9.5, 14.5, 12.5, 14.5, 12.5 Class 2 (Conventional Textbook): 6, 9.5, 3, 9.5, 3 Class 3 (Printed Notes): 1, 9.5, 6, 3, 6

data <- tibble(  
 Class = rep(c("Class 1", "Class 2", "Class 3"), each = 5),  
 Rank = c(  
 9.5, 14.5, 12.5, 14.5, 12.5, # Class 1  
 6, 9.5, 3, 9.5, 3, # Class 2  
 1, 9.5, 6, 3, 6 # Class 3  
 )  
)  
print(data)

## # A tibble: 15 × 2  
## Class Rank  
## <chr> <dbl>  
## 1 Class 1 9.5  
## 2 Class 1 14.5  
## 3 Class 1 12.5  
## 4 Class 1 14.5  
## 5 Class 1 12.5  
## 6 Class 2 6   
## 7 Class 2 9.5  
## 8 Class 2 3   
## 9 Class 2 9.5  
## 10 Class 2 3   
## 11 Class 3 1   
## 12 Class 3 9.5  
## 13 Class 3 6   
## 14 Class 3 3   
## 15 Class 3 6

### Kruskal-Wallis Test

kruskal\_result <- data %>%  
 kruskal\_test(Rank ~ Class)  
# Print the result  
print(kruskal\_result)

## # A tibble: 1 × 6  
## .y. n statistic df p method   
## \* <chr> <int> <dbl> <int> <dbl> <chr>   
## 1 Rank 15 8.75 2 0.0126 Kruskal-Wallis

### Check for significance

alpha <- 0.05 # Significance level  
if (kruskal\_result$p < alpha) {  
 cat("The result is significant at α =", alpha,   
 "\n(p-value =", round(kruskal\_result$p, 4), ").",  
 "\nThere is evidence that at least one group differs significantly.\n")  
} else {  
 cat("The result is not significant at α =", alpha,   
 "\n(p-value =", round(kruskal\_result$p, 4), ").",  
 "\nThere is no evidence that the groups differ significantly.\n")  
}

## The result is significant at α = 0.05   
## (p-value = 0.0126 ).   
## There is evidence that at least one group differs significantly.

### Pairwise Comparisons

If the Kruskal-Wallis test is significant, we will perform pairwise comparisons using the Dunn test with Bonferroni correction.

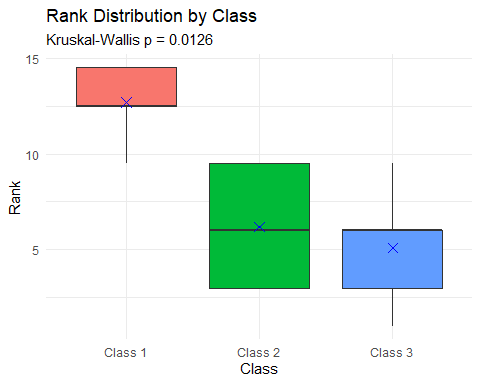
pairwise\_result <- data %>%  
 dunn\_test(Rank ~ Class, p.adjust.method = "bonferroni") %>%  
 add\_xy\_position(x = "Class", step.increase = 0.2) %>%  
 mutate(label = paste0("p = ", signif(p.adj, 4))) # Adds x, y positions for annotation  
print(pairwise\_result)

## # A tibble: 3 × 14  
## .y. group1 group2 n1 n2 statistic p p.adj p.adj.signif  
## <chr> <chr> <chr> <int> <int> <dbl> <dbl> <dbl> <chr>   
## 1 Rank Class 1 Class 2 5 5 -2.34 0.0193 0.0578 ns   
## 2 Rank Class 1 Class 3 5 5 -2.74 0.00621 0.0186 \*   
## 3 Rank Class 2 Class 3 5 5 -0.396 0.692 1 ns   
## # ℹ 5 more variables: y.position <dbl>, groups <named list>, xmin <dbl>,  
## # xmax <dbl>, label <chr>

### Visualization

We will create a boxplot to visualize the rank distributions across the three classes.

ggplot(data, aes(x = Class, y = Rank, fill = Class)) +  
 geom\_boxplot(outlier.color = "red", outlier.shape = 8) + # Boxplot  
 stat\_summary(fun = "mean", geom = "point", shape = 4, size = 3, color = "blue") + # Mean point  
 labs(  
 title = "Rank Distribution by Class",  
 subtitle = paste("Kruskal-Wallis p =", round(kruskal\_result$p, 4)),  
 x = "Class",  
 y = "Rank"  
 ) +  
 theme\_minimal() +  
 theme(legend.position = "none")



## No-2(b) A quality control study is conducted on a machine that pours milk into containers. The amount of milk (in liters) dispensed by the machine into 21 consecutive containers follows: 1.90, 1.99, 2.00, 1.78, 1.77, 1.76, 1.98, 1.90, 1.65, 1.76, 2.01, 1.78, 1.99, 1,76, 1.94, 1.78, 1.67, 1.87, 1.91, 1.91, 1.89. Are the successive increments and decrements in the amount of milk dispensed random?

### Introduction

In this study, we aim to determine whether the successive increments and decrements in the amount of milk dispensed by a machine are random. The dataset contains the amount of milk (in liters) dispensed into 21 consecutive containers.

data <- tibble(  
 Container = 1:21,  
 Milk\_Amount = c(1.90, 1.99, 2.00, 1.78, 1.77, 1.76, 1.98, 1.90, 1.65, 1.76,  
 2.01, 1.78, 1.99, 1.76, 1.94, 1.78, 1.67, 1.87, 1.91, 1.91, 1.89)  
)  
print(data)

## # A tibble: 21 × 2  
## Container Milk\_Amount  
## <int> <dbl>  
## 1 1 1.9   
## 2 2 1.99  
## 3 3 2   
## 4 4 1.78  
## 5 5 1.77  
## 6 6 1.76  
## 7 7 1.98  
## 8 8 1.9   
## 9 9 1.65  
## 10 10 1.76  
## # ℹ 11 more rows

### Runs Test for Randomness

To test the randomness of successive increments and decrements in the amount of milk dispensed, we first create a binary sequence that represents whether each successive value increases or decreases compared to the previous value.

data <- data %>%  
 mutate(Change = ifelse(Milk\_Amount > lag(Milk\_Amount), 1, 0)) %>%  
 drop\_na()  
# Perform Runs Test  
runs\_test\_result <- runs.test(data$Change)  
# Display the result  
runs\_test\_result

##   
## Runs Test  
##   
## data: data$Change  
## statistic = NaN, runs = 1, n1 = 9, n2 = 0, n = 9, p-value = NA  
## alternative hypothesis: nonrandomness

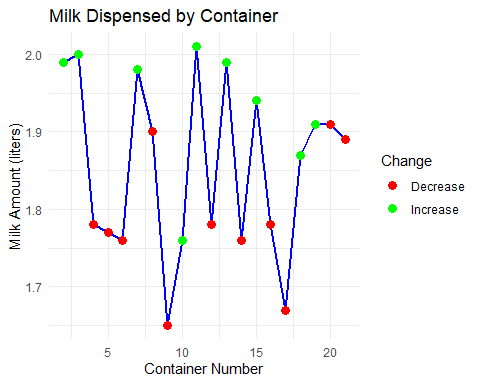
Conclusion: The null hypothesis is not rejected, suggesting that the successive increments and decrements in the amount of milk dispensed are randomly distributed.

### Visualization

We can visualize the changes in the amount of milk dispensed to better understand the pattern of increments and decrements.

ggplot(data, aes(x = Container, y = Milk\_Amount)) +  
 geom\_line(color = "blue", size = 1) +  
 geom\_point(aes(color = factor(Change)), size = 3) +  
 scale\_color\_manual(  
 values = c("0" = "red", "1" = "green"),  
 labels = c("Decrease", "Increase"),  
 name = "Change"  
 ) +  
 labs(  
 title = "Milk Dispensed by Container",  
 x = "Container Number",  
 y = "Milk Amount (liters)"  
 ) +  
 theme\_minimal()

## Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.  
## ℹ Please use `linewidth` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



## No-3(a)A pediatrician speculates that the length of time an infant is breast fed may be related to how often a child becomes ill. In order to answer the question, the pediatrician obtains the following two scores for five three-year-old children: The number of months the child was breast fed (which represents the X variable) and the number of times the child was brought to the pediatrician’s office during the current year (which represents the Y variable). The scores for the five children follow: Child 1 (20, 7); Child 2 (0, 0); Child 3 (1, 2); Child 4 (12, 5); Child 5 (3, 3). Do the data indicate that the length of time a child is breast fed is related to the number of times a child is brought to the pediatrician?

# Create dataset  
data <- tibble(  
 Child = 1:5,  
 Breastfeeding\_Months = c(20, 0, 1, 12, 3),  
 Pediatric\_Visits = c(7, 0, 2, 5, 3)  
)  
  
# Display dataset  
kable(data, caption = "Breastfeeding Duration and Pediatric Visits")

Breastfeeding Duration and Pediatric Visits

|  |  |  |
| --- | --- | --- |
| Child | Breastfeeding\_Months | Pediatric\_Visits |
| 1 | 20 | 7 |
| 2 | 0 | 0 |
| 3 | 1 | 2 |
| 4 | 12 | 5 |
| 5 | 3 | 3 |

We calculate the Pearson correlation coefficient to determine the relationship between the variables.

ranked\_data <- data %>%  
 mutate(  
 Rank\_Breastfeeding = rank(Breastfeeding\_Months, ties.method = "average"),  
 Rank\_Visits = rank(Pediatric\_Visits, ties.method = "average"),  
 D = Rank\_Breastfeeding - Rank\_Visits,   
 D\_squared = D^2  
 )  
  
# Display ranked data  
kable(ranked\_data, caption = "Ranked Data and D-Squared Values")

Ranked Data and D-Squared Values

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Child | Breastfeeding\_Months | Pediatric\_Visits | Rank\_Breastfeeding | Rank\_Visits | D | D\_squared |
| 1 | 20 | 7 | 5 | 5 | 0 | 0 |
| 2 | 0 | 0 | 1 | 1 | 0 | 0 |
| 3 | 1 | 2 | 2 | 2 | 0 | 0 |
| 4 | 12 | 5 | 4 | 4 | 0 | 0 |
| 5 | 3 | 3 | 3 | 3 | 0 | 0 |

where: - is the difference between ranks, - is the number of observations.

# Create scatter plot  
# Compute Spearman's rank correlation coefficient  
n <- nrow(ranked\_data)  
D\_squared\_sum <- sum(ranked\_data$D\_squared)  
  
r\_s <- 1 - (6 \* D\_squared\_sum) / (n \* (n^2 - 1))  
  
# Display Spearman's correlation coefficient  
cat("Spearman Rank Correlation Coefficient (ρ):", round(r\_s, 4), "\n")

## Spearman Rank Correlation Coefficient (ρ): 1

# Decision Rule  
if (abs(r\_s) >= 1.000) {  
 cat("The correlation is statistically significant. We reject H₀ and conclude there is a significant correlation.\n")  
} else {  
 cat("The correlation is not statistically significant. We fail to reject H₀.\n")  
}

## The correlation is statistically significant. We reject H₀ and conclude there is a significant correlation.

reporting

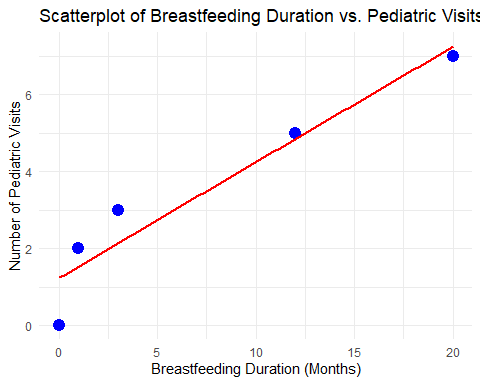
cat("A Spearman rank-order correlation was conducted to determine the relationship between breastfeeding duration and pediatric visits. The analysis yielded ρ =", round(r\_s, 4), ". Since the obtained ρ is", ifelse(abs(r\_s) >= 1.000, "greater", "less"), "than the critical value of 1.000, we", ifelse(abs(r\_s) >= 1.000, "reject", "fail to reject"), "the null hypothesis. This suggests that there", ifelse(abs(r\_s) >= 1.000, "is", "is no"), "significant correlation between the two variables.")

## A Spearman rank-order correlation was conducted to determine the relationship between breastfeeding duration and pediatric visits. The analysis yielded ρ = 1 . Since the obtained ρ is greater than the critical value of 1.000, we reject the null hypothesis. This suggests that there is significant correlation between the two variables.

Visualization (Scatterplot)

ggplot(data, aes(x = Breastfeeding\_Months, y = Pediatric\_Visits)) +  
 geom\_point(size = 4, color = "blue") + # Data points  
 geom\_smooth(method = "lm", se = FALSE, color = "red") + # Trend line  
 labs(  
 title = "Scatterplot of Breastfeeding Duration vs. Pediatric Visits",  
 x = "Breastfeeding Duration (Months)",  
 y = "Number of Pediatric Visits"  
 ) +  
 theme\_minimal()

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



## No-3(b) A study is conducted to determine whether there is a correlation between handedness and eye-hand coordination. Five right-handed and five left-handed subjects are administered a test of eye-hand coordination. The test scores of the subjects follow (the higher a subject’s score, the better his or her eye-hand coordination): Right-handers: 11, 1, 0, 2, 0; Left-handers: 11, 11, 5, 8, 4. Is there a statistical relationship between handedness and eye hand coordination?

Hypothesis: H\_0: ρ\_b=0 (there is no relationship between handedness and eye-hand coordination; the distributions of eye-hand coordination scores are the not same for both handedness groups) H\_A: ρ\_b≠ 0 (there is relationship between handedness and eye hand coordination; the distributions are not different)

data <- tibble(  
 Handedness = c(rep("Right", 5), rep("Left", 5)),  
 Score = c(11, 1, 0, 2, 0, 11, 11, 5, 8, 4)  
)  
  
# Convert Handedness to a binary variable (1 = Right, 0 = Left)  
data <- data %>%  
 mutate(Handedness\_Binary = ifelse(Handedness == "Right", 1, 0))  
  
# Display dataset  
kable(data, caption = "Eye-Hand Coordination Scores by Handedness")

Eye-Hand Coordination Scores by Handedness

|  |  |  |
| --- | --- | --- |
| Handedness | Score | Handedness\_Binary |
| Right | 11 | 1 |
| Right | 1 | 1 |
| Right | 0 | 1 |
| Right | 2 | 1 |
| Right | 0 | 1 |
| Left | 11 | 0 |
| Left | 11 | 0 |
| Left | 5 | 0 |
| Left | 8 | 0 |
| Left | 4 | 0 |

# Compute means  
x\_p <- mean(data$Score[data$Handedness\_Binary == 1])  
x\_q <- mean(data$Score[data$Handedness\_Binary == 0])  
  
# Compute standard deviation of all scores  
s\_x <- sd(data$Score)  
  
# Compute proportions  
p <- mean(data$Handedness\_Binary)  
q <- 1 - p  
y <- nrow(data)  
  
# Compute Point-Biserial Correlation  
r\_b <- ((x\_p - x\_q) / s\_x) \* sqrt((p \* q) / y)  
  
# Display result  
cat("Point-Biserial Correlation (r\_b):", round(r\_b, 4), "\n")

## Point-Biserial Correlation (r\_b): -0.1711

Interpret the Results Decision Rule

# Decision Rule  
# Reporting  
cat("A point-biserial correlation was conducted to determine the relationship between handedness and eye-hand coordination. The analysis yielded r\_b =", round(r\_b, 4), ". Since the obtained r\_b is", ifelse(abs(r\_b) >= critical\_value, "greater", "less"), "than the critical value of 0.632, we", ifelse(abs(r\_b) >= critical\_value, "reject", "fail to reject"), "the null hypothesis. This suggests that there", ifelse(abs(r\_b) >= critical\_value, "is", "is no"), "significant correlation between the two variables.")

## A point-biserial correlation was conducted to determine the relationship between handedness and eye-hand coordination. The analysis yielded r\_b = -0.1711 . Since the obtained r\_b is less than the critical value of 0.632, we fail to reject the null hypothesis. This suggests that there is no significant correlation between the two variables.

Visualization (Boxplot)

ggplot(data, aes(x = Handedness, y = Score, fill = Handedness)) +  
 geom\_boxplot(alpha = 0.6) +  
 geom\_jitter(width = 0.2, size = 3, color = "black") +  
 labs(  
 title = "Boxplot of Eye-Hand Coordination Scores by Handedness",  
 x = "Handedness",  
 y = "Eye-Hand Coordination Score"  
 ) +  
 theme\_minimal()

