Applied Non-parametric Methods MAS-123-6B

# Load necessary package

library(pacman)  
p\_load(rmarkdown,officer,knitr,kableExtra,latex2exp,xtable,flextable,tidyverse, rstatix, ggsignif)

#### Load necessary package

# 1

### (a) What is meant by nonparametric statistics?

**Nonparametric statistics refers to a set of statistical methods that do not make assumptions about the specific form of the underlying probability distribution from which the data are drawn. Unlike parametric methods, which require the assumption of normality or other distributional properties, nonparametric methods focus on the ranks, medians, or other characteristics of the data. These methods are particularly useful when data do not meet the stringent assumptions required by parametric tests, such as normal distribution, homogeneity of variance, or linearity.**

### (b) When should nonparametric statistics be used?

Nonparametric statistics should be used under the following conditions:

1. **Violation of Assumptions for Parametric Tests**: When the data do not meet the assumptions of parametric tests, such as normality or homogeneity of variance.
2. **Ordinal or Nominal Data**: For data measured on ordinal (ranked) or nominal (categorical) scales, where the intervals between values are not meaningful or cannot be assumed to be equal.
3. **Small Sample Sizes**: In cases where the sample size is small, and parametric methods may not be reliable due to insufficient data to validate the assumptions of normality or other conditions.
4. **Presence of Outliers or Nonlinear Relationships**: Nonparametric methods are more robust against outliers, skewed distributions, or nonlinear relationships, which may affect the validity of parametric methods.
5. **Unknown or Unclear Distribution**: When there is uncertainty or a lack of knowledge about the distribution of the data, nonparametric methods provide a distribution-free alternative.

### (c) Why does the term "distribution-free" describe nonparametric procedures?

The term "distribution-free" describes nonparametric procedures because these methods do not depend on specific probability distributions (such as normal, binomial, etc.). In contrast to parametric procedures that assume data follow a particular distribution (e.g., normal distribution), nonparametric procedures are applicable to a wide range of data types and distributions. They operate independently of the assumption about the underlying distribution, making them flexible and widely applicable to many different data scenarios.

(d) How to measurement and measurement scales in nonparametric statistics?

**Measurement and measurement scales are fundamental to the use of nonparametric statistics:**

1. **Nominal Scale: Nonparametric methods like chi-square tests can analyze categorical data classified into groups or categories.**
2. **Ordinal Scale: Nonparametric methods such as the Mann-Whitney U test or Spearman’s rank correlation are suitable for data ranked in order without assuming equal intervals.**
3. **Interval and Ratio Scales: While these scales are often analyzed using parametric methods, nonparametric alternatives (e.g., Kruskal-Wallis test) are useful when the data violate parametric assumptions.**

(e) What are the advantages and disadvantages of nonparametric statistics?

**Advantages:**

1. **Fewer Assumptions: Nonparametric methods do not require data to follow a specific distribution (such as normality) or have equal variances. This makes them suitable for data that do not meet the strict assumptions of parametric tests.**
2. **Versatility: Nonparametric tests can be applied to data from various measurement scales, including nominal, ordinal, interval, and ratio scales. This flexibility allows them to handle a wide range of data types.**
3. **Robustness: Nonparametric methods are less sensitive to outliers, skewed data, or violations of assumptions like linearity. They can still produce reliable results even when the data contain such issues.**
4. **Usable with Small Samples: Nonparametric tests are often more reliable for small sample sizes where parametric assumptions (like normality) may not be met or may be difficult to verify.**

**Disadvantages:**

1. **Lower Statistical Power: Nonparametric tests often have lower statistical power compared to parametric tests, meaning they may have a higher chance of failing to detect a true effect (Type II error) when one exists. This is especially true in cases where the data actually meet the assumptions for parametric testing.**
2. **Loss of Information: Nonparametric methods often involve converting data into ranks or medians, which means that some information about the magnitude of differences between values is lost. For example, in comparing two distributions, nonparametric methods do not consider the exact values, only their relative ordering.**
3. **Interpretation: Results from nonparametric methods can be more challenging to interpret compared to parametric tests. For example, the meaning of the test statistic might not be as straightforward as the t-statistic in a t-test, making interpretation more complex.**
4. **Limited Methods for Complex Analyses: Although nonparametric statistics are versatile, there are fewer nonparametric alternatives for certain complex analyses, such as multiple regression or ANOVA with many groups. The lack of these methods can be limiting when analyzing complex relationships between variables.**

# 2

You are doing research for an article on the waterfalls on our planet. You want to make a statement about the heights of waterfalls on three continents. Three random samples of waterfall heights (in feet) are shown.

# Create dataset  
waterfall\_data <- tibble(  
 Continent = rep(c("North America", "Africa", "Asia"), each = 6),  
 Height = c(600, 1200, 182, 620, 1170, 442,   
 406, 508, 630, 726, 480, 2014,  
 330, 830, 614, 1100, 885, 330)  
)  
kable(waterfall\_data, caption = "Sorted Waterfall Heights")

Sorted Waterfall Heights

| Continent | Height |
| --- | --- |
| North America | 600 |
| North America | 1200 |
| North America | 182 |
| North America | 620 |
| North America | 1170 |
| North America | 442 |
| Africa | 406 |
| Africa | 508 |
| Africa | 630 |
| Africa | 726 |
| Africa | 480 |
| Africa | 2014 |
| Asia | 330 |
| Asia | 830 |
| Asia | 614 |
| Asia | 1100 |
| Asia | 885 |
| Asia | 330 |

#### (a) What questions are you trying to answer?

Is there a significant difference in the heights of waterfalls among the three countries being compared?

#### b) What nonparametric test would you use to find the answer?

The most appropriate nonparametric test for this scenario is the Kruskal-Wallis H-test. This test is suitable because it is used to compare the medians of three or more independent groups when the data does not meet the assumptions required for parametric tests (such as normality or equal variances). Since we are comparing the heights of waterfalls across three countries, the Kruskal-Wallis H-test is the ideal choice to determine if there are statistically significant differences in the medians of the groups.Since the data may not be normally distributed, we use the Kruskal-Wallis test, a nonparametric alternative to ANOVA.

#### (c) What are the hypotheses?

* **Null Hypothesis (H0):**: The median waterfall heights are the same across all three continents.
* **Alternative Hypothesis (H1):**: At least one continent has a different median waterfall height.

## Rank and Tie Case

You can also embed plots, for example:

ranked\_data <- waterfall\_data |>   
 mutate(Rank = rank(Height),   
 Tie\_Case = duplicated(Height) | duplicated(Height, fromLast = TRUE)) |>   
 arrange(Rank)  
  
# Display ranked data  
kable(ranked\_data, caption = "Sorted Waterfall Heights, Ranks, and Tie Cases")

Sorted Waterfall Heights, Ranks, and Tie Cases

|  |  |  |  |
| --- | --- | --- | --- |
| Continent | Height | Rank | Tie\_Case |
| North America | 182 | 1.0 | FALSE |
| Asia | 330 | 2.5 | TRUE |
| Asia | 330 | 2.5 | TRUE |
| Africa | 406 | 4.0 | FALSE |
| North America | 442 | 5.0 | FALSE |
| Africa | 480 | 6.0 | FALSE |
| Africa | 508 | 7.0 | FALSE |
| North America | 600 | 8.0 | FALSE |
| Asia | 614 | 9.0 | FALSE |
| North America | 620 | 10.0 | FALSE |
| Africa | 630 | 11.0 | FALSE |
| Africa | 726 | 12.0 | FALSE |
| Asia | 830 | 13.0 | FALSE |
| Asia | 885 | 14.0 | FALSE |
| Asia | 1100 | 15.0 | FALSE |
| North America | 1170 | 16.0 | FALSE |
| North America | 1200 | 17.0 | FALSE |
| Africa | 2014 | 18.0 | FALSE |

####(d) Select a significance level and run the test. What is the H value? Compute H Statistic Manually

# Compute total rank per continent  
rank\_sums <- ranked\_data %>%   
 group\_by(Continent) %>%   
 summarise(Rank\_Sum = sum(Rank), n = n())  
kable(rank\_sums, caption = "Rank Sums Calculate H statistic ")

Rank Sums Calculate H statistic

|  |  |  |
| --- | --- | --- |
| Continent | Rank\_Sum | n |
| Africa | 58 | 6 |
| Asia | 56 | 6 |
| North America | 57 | 6 |

Compute H statistic

Where: - is the total number of observations. - is the rank sum for each group. - is the number of observations in each group.

To Compute the Test Statistic doing Ties Correction (C\_H)

The ties correction is applied to adjust for tied ranks. The formula for the ties correction is:

Where: - is the number of tied ranks in each group of ties. - is the total number of observations.

N <- nrow(ranked\_data)  
H <- (12 / (N \* (N + 1))) \* sum((rank\_sums$Rj^2) / rank\_sums$nj) - 3 \* (N + 1)  
  
# Compute tie correction factor  
tie\_counts <- ranked\_data |> count(Rank) |> filter(n > 1) |> pull(n)  
C <- 1 - sum(tie\_counts^3 - tie\_counts) / (N^3 - N)  
  
# Adjust H for ties  
H\_corrected <- H / C  
  
# Display results  
kable(tibble(H\_original = H, Tie\_Correction = C, H\_corrected = H\_corrected),   
 caption = "Kruskal-Wallis H Statistic with Tie Correction")

Kruskal-Wallis H Statistic with Tie Correction

|  |  |  |
| --- | --- | --- |
|  | H-statistic manunal |  |
|  | 0.9942 |  |

Kruskal-Wallis Test Computation

# Perform Kruskal-Wallis test  
kruskal\_result <- kruskal\_test(waterfall\_data, Height ~ Continent)  
p\_value <- kruskal\_result$p # p-value  
  
# Display result  
kable(kruskal\_result, caption = "Kruskal-Wallis Test Results")

Kruskal-Wallis Test Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| .y. | n | statistic | df | p | method |
| Height | 18 | 0.011708 | 2 | 0.9942 | Kruskal-Wallis |

####(e) What is your conclusion?

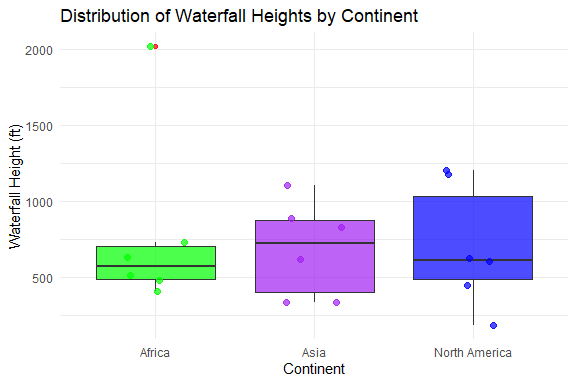
p\_value <- kruskal\_result$p  
  
# Check significance  
if (p\_value < 0.05) {  
 cat("The result is significant (p =", round(p\_value, 3), "). There is evidence that at least one continent has a different median waterfall height.\n")  
} else {  
 cat("The result is not significant (p =", round(p\_value, 3), "). We do not have enough evidence to conclude that waterfall heights differ by continent.\n")  
}

## The result is not significant (p = 0.9942 ). We do not have enough evidence to conclude that waterfall heights differ by continent.

**Decision and conclusion**: Since p-value ( 0.9942 ) > α ( 0.05 ), H0 will not be rejected. Since H0 was not rejected, there is significant evidence, with significant level ( 0.05 ), to conclude that the median of all heights of waterfalls in three countries are equal

Visualization: Boxplot of Waterfall Heights by Continent

ggplot(waterfall\_data, aes(x = Continent, y = Height, fill = Continent)) +  
 geom\_boxplot(outlier.color = "red", alpha = 0.7) +  
 geom\_jitter(aes(color = Continent), width = 0.2, size = 2, alpha = 0.7) +  
 labs(title = "Distribution of Waterfall Heights by Continent",  
 x = "Continent",  
 y = "Waterfall Height (ft)") +  
 theme\_minimal() +  
 scale\_fill\_manual(values = c("North America" = "blue", "Africa" = "green", "Asia" = "purple")) +  
 scale\_color\_manual(values = c("North America" = "blue", "Africa" = "green", "Asia" = "purple")) +  
 theme(legend.position = "none")



# 3(a)

As a biologist, you wish to see if there is a relationship between the heights of tall trees and their diameters. You find the following data for the diameter (in inches) of the tree at 4.5 feet from the ground and the corresponding heights (in feet). #### What question are you trying to answer? (i) Research Question Is there a relationship between the diameter of a tree and its height? #### What type of nonparametric analysis could be used to answer the question? (ii) Nonparametric Test Selection To analyze the relationship between the two continuous variables (tree diameter and tree height), the Spearman’s rank correlation test is the most appropriate nonparametric method. This test is used to assess the strength and direction of the association between two variables when the data does not meet the assumptions of parametric tests (such as normality or linearity). Spearman’s rank correlation is particularly useful for ordinal data or when the relationship between variables is monotonic but not necessarily linear. #### Is there a relationship between the two at the 0.05 level of significance? (iii) - **Null Hypothesis (H0):**: There is no monotonic relationship between tree diameter and height. (𝜌=0

* **Alternative Hypothesis (H1):**: There is a significant monotonic relationship between tree diameter and height. (𝜌≠0)

# Create dataset  
tree\_data <- tibble(  
 Diameter = c(1024, 950, 451, 505, 761, 644, 707, 586, 442, 546),  
 Height = c(261, 321, 219, 281, 159, 83, 191, 141, 232, 108)  
)  
kable(tree\_data, caption = "Sorted Tree Data")

Sorted Tree Data

|  |  |
| --- | --- |
| Diameter | Height |
| 1024 | 261 |
| 950 | 321 |
| 451 | 219 |
| 505 | 281 |
| 761 | 159 |
| 644 | 83 |
| 707 | 191 |
| 586 | 141 |
| 442 | 232 |
| 546 | 108 |

Rank transformation and check for ties

ranked\_tree\_data <- tree\_data |>   
 mutate(Rank\_Diameter = rank(Diameter),   
 Rank\_Height = rank(Height),   
 Tie\_Case\_Diameter = duplicated(Diameter) | duplicated(Diameter, fromLast = TRUE),  
 Tie\_Case\_Height = duplicated(Height) | duplicated(Height, fromLast = TRUE)) |>   
 arrange(Rank\_Diameter)  
# Display ranked data  
kable(ranked\_tree\_data, caption = "Sorted Tree Data with Ranks and Tie Cases")

Sorted Tree Data with Ranks and Tie Cases

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Diameter | Height | Rank\_Diameter | Rank\_Height | Tie\_Case\_Diameter | Tie\_Case\_Height |
| 442 | 232 | 1 | 7 | FALSE | FALSE |
| 451 | 219 | 2 | 6 | FALSE | FALSE |
| 505 | 281 | 3 | 9 | FALSE | FALSE |
| 546 | 108 | 4 | 2 | FALSE | FALSE |
| 586 | 141 | 5 | 3 | FALSE | FALSE |
| 644 | 83 | 6 | 1 | FALSE | FALSE |
| 707 | 191 | 7 | 5 | FALSE | FALSE |
| 761 | 159 | 8 | 4 | FALSE | FALSE |
| 950 | 321 | 9 | 10 | FALSE | FALSE |
| 1024 | 261 | 10 | 8 | FALSE | FALSE |

Manunal p value calculation

P= 0.76

Spearman Rank Correlation Computation

# Compute Spearman correlation  
spearman\_result <- cor\_test(tree\_data, Diameter, Height, method = "spearman")  
  
# Display results  
kable(spearman\_result, caption = "Spearman Rank Correlation Results")

Spearman Rank Correlation Results

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| var1 | var2 | cor | statistic | p | method |
| Diameter | Height | 0.11515 | 146 | 0.7588 | Spearman |

Conclusion

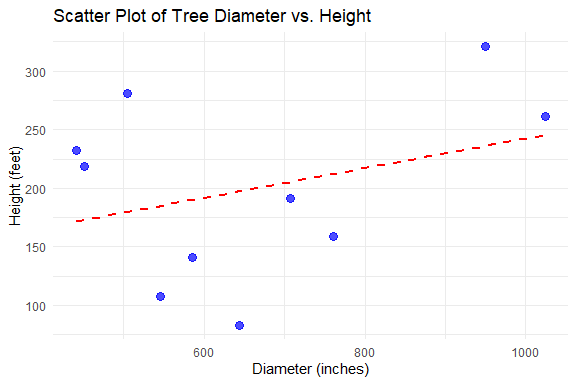
if (spearman\_result$p < 0.05) {  
 cat("The result is significant (p < 0.05). There is evidence of a monotonic relationship between tree diameter and height.\n")  
} else {  
 cat("The result is not significant (p >= 0.05). We do not have enough evidence to conclude a monotonic relationship.\n")  
}

## The result is not significant (p >= 0.05). We do not have enough evidence to conclude a monotonic relationship.

**Decision and conclusion**: Since p-value ( 0.7588 ) > α ( 0.05 ), H0 will not be rejected. Since H0 was not rejected, there is no significant evidence, with significant level ( 0.05 ), to conclude that there is a relationship between the height of the tree and the diameter of the tree

Visualization: Scatter Plot with Rank Correlation Trendline

ggplot(tree\_data, aes(x = Diameter, y = Height)) +  
 geom\_point(color = "blue", size = 3, alpha = 0.7) +  
 geom\_smooth(method = "lm", se = FALSE, color = "red", linetype = "dashed") +  
 labs(title = "Scatter Plot of Tree Diameter vs. Height",  
 x = "Diameter (inches)",  
 y = "Height (feet)") +  
 theme\_minimal()



# 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). Is there a statistical relationship between handedness and eye hand coordination?

# Create dataset  
data <- tibble(  
 Subject = 1:10,  
 Handedness = c(1, 1, 1, 1, 1, 0, 0, 0, 0, 0), # Right-handed = 1, Left-handed = 0  
 Coordination\_Score = c(11, 1, 0, 2, 0, 11, 11, 5, 8, 4) # Higher = better coordination  
)  
  
# Display data  
kable(data, caption = "Dataset: Handedness and Eye-Hand Coordination Scores")

Dataset: Handedness and Eye-Hand Coordination Scores

|  |  |  |
| --- | --- | --- |
| Subject | Handedness | Coordination\_Score |
| 1 | 1 | 11 |
| 2 | 1 | 1 |
| 3 | 1 | 0 |
| 4 | 1 | 2 |
| 5 | 1 | 0 |
| 6 | 0 | 11 |
| 7 | 0 | 11 |
| 8 | 0 | 5 |
| 9 | 0 | 8 |
| 10 | 0 | 4 |

1. Research Question Is there a significant relationship between handedness (X: Right-handed = 1, Left-handed = 0) and eye-hand coordination scores (Y: Higher score = better coordination)?
2. Nonparametric Test Selection Since X is binary (categorical) and Y is continuous, the Point-Biserial Correlation is the appropriate statistical test.
3. Hypotheses Null Hypothesis (H₀): There is no correlation between handedness and eye-hand coordination. (𝑟𝑝𝑏=0) Alternative Hypothesis (H₁): There is a significant correlation between handedness and eye-hand coordination. (𝑟𝑝𝑏≠0)
4. Compute Point-Biserial Correlation

# Compute Point-Biserial Correlation  
pb\_correlation <- cor\_test(data, Handedness, Coordination\_Score, method = "pearson")  
  
# Display results  
kable(pb\_correlation, caption = "Point-Biserial Correlation Results")

Point-Biserial Correlation Results

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| var1 | var2 | cor | statistic | p | conf.low | conf.high | method |
| Handedness | Coordination\_Score | -0.57 | -1.964185 | 0.0851 | -0.88293 | 0.0924229 | Pearson |

conculation

if (pb\_correlation$p < 0.05) {  
 cat("The result is significant (p < 0.05). There is evidence of a relationship between handedness and eye-hand coordination.\n")  
} else {  
 cat("The result is not significant (p >= 0.05). We do not have enough evidence to conclude a relationship between handedness and eye-hand coordination.\n")  
}

## The result is not significant (p >= 0.05). We do not have enough evidence to conclude a relationship between handedness and eye-hand coordination.

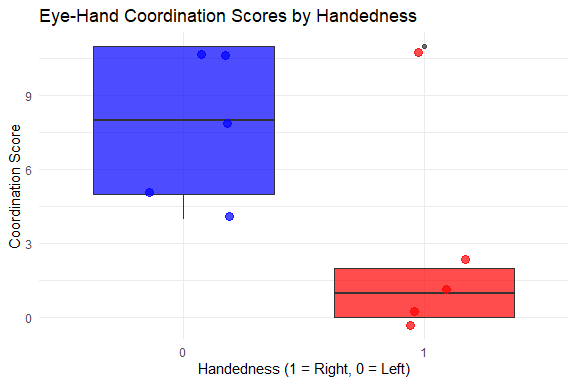
**Determine the level of significance**: α = 0.05

**Decision rule**:

If p-value < α, reject H0. Otherwise accept H0. Decision and conclusion: Since p-value ( 0.08511 ) > α ( 0.056 ), H0 will not be rejected. Since H0 was not rejected, there is no significant evidence, with significant level ( 0.05 ), to support the claim that there is a relationship between handedness and eye-hand coordination performance

Visualization: Boxplot of Coordination Scores by Handedness

ggplot(data, aes(x = as.factor(Handedness), y = Coordination\_Score, fill = as.factor(Handedness))) +  
 geom\_boxplot(alpha = 0.7) +  
 geom\_jitter(aes(color = as.factor(Handedness)), width = 0.2, size = 3, alpha = 0.7) +  
 labs(title = "Eye-Hand Coordination Scores by Handedness",  
 x = "Handedness (1 = Right, 0 = Left)",  
 y = "Coordination Score") +  
 scale\_fill\_manual(values = c("0" = "blue", "1" = "red")) +  
 scale\_color\_manual(values = c("0" = "blue", "1" = "red")) +  
 theme\_minimal() +  
 theme(legend.position = "none")



# 4(a)

The gender of 20 consecutive patients who register at the emergency room of a local hospital is recorded below (where: M = Male; F = Female). F F F M M M F F M M F M F M F M M M F F Do the data suggest that the gender distribution of entering patients is random?

# Create dataset  
data <- tibble(  
 Patient = 1:20,  
 Gender = c("F", "F", "F", "M", "M", "M", "F", "F", "M", "M",  
 "F", "M", "F", "M", "F", "M", "M", "M", "F", "F")  
)  
  
# Display data  
kable(data, caption = "Recorded Gender Data for ER Patients")

Recorded Gender Data for ER Patients

|  |  |
| --- | --- |
| Patient | Gender |
| 1 | F |
| 2 | F |
| 3 | F |
| 4 | M |
| 5 | M |
| 6 | M |
| 7 | F |
| 8 | F |
| 9 | M |
| 10 | M |
| 11 | F |
| 12 | M |
| 13 | F |
| 14 | M |
| 15 | F |
| 16 | M |
| 17 | M |
| 18 | M |
| 19 | F |
| 20 | F |

1. Research Question Is the gender sequence of emergency room patients random, or does it show a pattern?
2. Statistical Test Selection The Runs Test for randomness is used to check whether the sequence of genders is random.
3. **Hypotheses Null Hypothesis (H₀):** The sequence of gender arrivals is random. **Alternative Hypothesis (H₁):** The sequence of gender arrivals is not random.
4. Compute Runs Test

# Convert Gender to Binary (F = 0, M = 1)  
data <- data %>% mutate(Gender\_Binary = ifelse(Gender == "M", 1, 0))  
kable(data, caption = "Gender\_Binary")

Gender\_Binary

|  |  |  |
| --- | --- | --- |
| Patient | Gender | Gender\_Binary |
| 1 | F | 0 |
| 2 | F | 0 |
| 3 | F | 0 |
| 4 | M | 1 |
| 5 | M | 1 |
| 6 | M | 1 |
| 7 | F | 0 |
| 8 | F | 0 |
| 9 | M | 1 |
| 10 | M | 1 |
| 11 | F | 0 |
| 12 | M | 1 |
| 13 | F | 0 |
| 14 | M | 1 |
| 15 | F | 0 |
| 16 | M | 1 |
| 17 | M | 1 |
| 18 | M | 1 |
| 19 | F | 0 |
| 20 | F | 0 |

Runs test function Compute the Expected Number of Runs and Standard Deviation

The expected number of runs () and the standard deviation () are given by:

where: - = number of **M** (Male occurrences) - = number of **F** (Female occurrences) -

# Runs test function  
runs\_test <- function(sequence) {  
 n1 <- sum(sequence == 1) # Count of M  
 n2 <- sum(sequence == 0) # Count of F  
 N <- n1 + n2 # Total count  
   
 # Count the number of runs  
 runs <- sum(diff(sequence) != 0) + 1  
   
 # Mean and standard deviation of runs  
 mean\_runs <- (2 \* n1 \* n2) / N + 1  
 sd\_runs <- sqrt((2 \* n1 \* n2 \* (2 \* n1 \* n2 - N)) / (N^2 \* (N - 1)))  
   
 # Compute Z-score  
 z\_score <- (runs - mean\_runs) / sd\_runs  
 p\_value <- 2 \* pnorm(-abs(z\_score)) # Two-tailed test  
   
 return(tibble(Runs = runs, Expected\_Runs = mean\_runs, Z\_Score = z\_score, P\_Value = p\_value))  
}  
  
# Perform the Runs Test  
runs\_test\_result <- runs\_test(data$Gender\_Binary)  
  
# Display results  
kable(runs\_test\_result, caption = "Runs Test Results")

Runs Test Results

|  |  |  |  |
| --- | --- | --- | --- |
| Runs | Expected\_Runs | Z\_Score | P\_Value |
| 11 | 11 | 0 | 1 |

conclution

if (runs\_test\_result$P\_Value < 0.05) {  
 cat("The result is significant (p < 0.05). The gender sequence does not appear to be random.\n")  
} else {  
 cat("The result is not significant (p >= 0.05). We do not have enough evidence to conclude that the gender sequence is non-random.\n")  
}

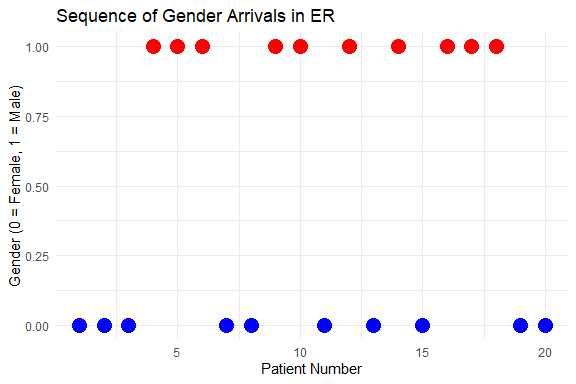
## The result is not significant (p >= 0.05). We do not have enough evidence to conclude that the gender sequence is non-random.

**p-value**: 1

**Determine the level of significance**: α = 0.05 Decision rule: If p-value < α, reject H0. Otherwise accept H0. Decision and conclusion: Since p-value ( 1 ) > α ( 0.05 ), H0 will not be rejected. Since H0 was not rejected, there is significant evidence, with significant level ( 0.05 ), to conclude that the gender distribution of entering patients is random

Visualization: Gender Sequence Plot

ggplot(data, aes(x = Patient, y = Gender\_Binary, color = Gender)) +  
 geom\_point(size = 5) +  
 scale\_color\_manual(values = c("F" = "blue", "M" = "red")) +  
 labs(title = "Sequence of Gender Arrivals in ER",  
 x = "Patient Number",  
 y = "Gender (0 = Female, 1 = Male)") +  
 theme\_minimal() +  
 theme(legend.position = "none")



# 4(b)

A sample of consecutively produced bolts is selected from an assembly line and measured. The following are the deviations in thousandths of an inch of the lengths of the bolts from 3.000 inches:

bolt\_deviations <- c(4, 2, -2, 3, 8, 4, 3, 3, 1, 4, 1,   
 1, 5, 3, -2, 6, 1, 3, -11, -10, 12, 5,   
 3, 7, 8, -9, 3, 3, -2, 10, -1, 4, -5,   
 6, -3, 1, 5, 3, 5, 3, -1, -5, 3, -7,   
 -4, 4, -2, -1, -2, -1, 10, -5, -5, 5, -2,   
 1, -7, 4, 4, -5)  
kable(bolt\_deviations, caption = "Bolt Length Deviation Data")

Bolt Length Deviation Data

|  |
| --- |
| x |
| 4 |
| 2 |
| -2 |
| 3 |
| 8 |
| 4 |
| 3 |
| 3 |
| 1 |
| 4 |
| 1 |
| 1 |
| 5 |
| 3 |
| -2 |
| 6 |
| 1 |
| 3 |
| -11 |
| -10 |
| 12 |
| 5 |
| 3 |
| 7 |
| 8 |
| -9 |
| 3 |
| 3 |
| -2 |
| 10 |
| -1 |
| 4 |
| -5 |
| 6 |
| -3 |
| 1 |
| 5 |
| 3 |
| 5 |
| 3 |
| -1 |
| -5 |
| 3 |
| -7 |
| -4 |
| 4 |
| -2 |
| -1 |
| -2 |
| -1 |
| 10 |
| -5 |
| -5 |
| 5 |
| -2 |
| 1 |
| -7 |
| 4 |
| 4 |
| -5 |

data <- tibble(  
 Measurement = bolt\_deviations,  
 Category = ifelse(bolt\_deviations >= 0, 1, 0) # 1 = Above or equal to 3.000, 0 = Below  
)  
  
# Display first few rows  
kable(head(data, 10), caption = "Bolt Length Deviation Data")

Bolt Length Deviation Data

|  |  |
| --- | --- |
| Measurement | Category |
| 4 | 1 |
| 2 | 1 |
| -2 | 0 |
| 3 | 1 |
| 8 | 1 |
| 4 | 1 |
| 3 | 1 |
| 3 | 1 |
| 1 | 1 |
| 4 | 1 |

Hypothesis:

H0: The sequence of deviation is random

H1: The sequence of deviation is not random

Compute the Expected Number of Runs and Standard Deviation

The expected number of runs () and the standard deviation () are given by:

where: - = number of **M** (Male occurrences) - = number of **F** (Female occurrences) -

runs\_test <- function(sequence) {  
 n1 <- sum(sequence == 1) # Count of Above (1)  
 n2 <- sum(sequence == 0) # Count of Below (0)  
 N <- n1 + n2 # Total count  
   
 # Count the number of runs  
 runs <- sum(diff(sequence) != 0) + 1  
   
 # Mean and standard deviation of runs  
 mean\_runs <- (2 \* n1 \* n2) / N + 1  
 sd\_runs <- sqrt((2 \* n1 \* n2 \* (2 \* n1 \* n2 - N)) / (N^2 \* (N - 1)))  
   
 # Compute Z-score  
 z\_score <- (runs - mean\_runs) / sd\_runs  
 p\_value <- 2 \* pnorm(-abs(z\_score)) # Two-tailed test  
   
 return(tibble(Runs = runs, Expected\_Runs = mean\_runs, Z\_Score = z\_score, P\_Value = p\_value))  
}

# Perform the Runs Test  
runs\_test\_result <- runs\_test(data$Category)  
  
# Display results  
kable(runs\_test\_result, caption = "Runs Test Results")

Runs Test Results

|  |  |  |  |
| --- | --- | --- | --- |
| Runs | Expected\_Runs | Z\_Score | P\_Value |
| 30 | 28.86667 | 0.3181517 | 0.7503699 |

Conclusion

if (runs\_test\_result$P\_Value < 0.05) {  
 cat("The result is significant (p < 0.05). The sequence of bolt deviations does not appear to be random.\n")  
} else {  
 cat("The result is not significant (p >= 0.05). We do not have enough evidence to conclude that the sequence is non-random.\n")  
}

## The result is not significant (p >= 0.05). We do not have enough evidence to conclude that the sequence is non-random.Determine the level of significance: α = 0.05

**Decision rule**: If p-value < α, reject H0. Otherwise accept H0. Decision and conclusion: Since p-value ( 0.7503699) > α ( 0.05 ), H0 will not be rejected. Since H0 was not rejected, there is significant evidence, with significant level ( 0.05 ), to said that the sequence of the deviation is random

# 5(a)

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 eyehand coordination. One hundred randomly selected females and 100 randomly selected males are 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). The following summarizes the results of the study. Do the data indicate that there are gender differences in performance?

#### Hypotheses:

* **Null Hypothesis (H0):** There is no association between gender and scoring above/below the median.
* **Alternative Hypothesis (H1):** There is an association between gender and scoring above/below the median.

data\_eye\_hand <- tibble(  
 Gender = rep(c("Male", "Female"), each = 2),  
 Score = rep(c("Above Median", "Below Median"), 2),  
 Count = c(30, 70, 60, 40)  
)  
  
kable(data\_eye\_hand)

|  |  |  |
| --- | --- | --- |
| Gender | Score | Count |
| Male | Above Median | 30 |
| Male | Below Median | 70 |
| Female | Above Median | 60 |
| Female | Below Median | 40 |

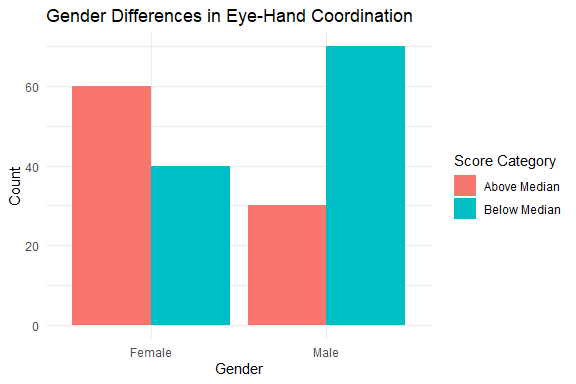
### Perform Chi-Square Test

table\_eye\_hand <- matrix(c(30, 70, 60, 40), nrow = 2, byrow = TRUE)  
chisq\_test\_result <- chisq.test(table\_eye\_hand)  
chisq\_test\_result

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

### Visualization

ggplot(data\_eye\_hand, aes(x = Gender, y = Count, fill = Score)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 theme\_minimal() +  
 labs(title = "Gender Differences in Eye-Hand Coordination",  
 x = "Gender",  
 y = "Count",  
 fill = "Score Category")



# 5(b)

A researcher conducts a study in order to evaluate the effect of noise on altruistic behavior. Each of the 12 subjects who participate in the experiment is randomly assigned to one of two experimental conditions. Subjects in both conditions are given a one-hour test which is ostensibly a measure of intelligence. During the test the six subjects in Group 1 are exposed to continual loud noise, which they are told is due to a malfunctioning generator. The six subjects in Group 2 are not exposed to any noise during the test. Upon completion of this stage of the experiment, each subject on leaving the room is confronted by a middle-aged man whose arm is in a sling. The man asks the subject if she would be willing to help him carry a heavy package to his car. In actuality, the man requesting help is an experimental confederate (i.e., working for the experimenter). The number of subjects in each group who help the man is recorded. One of the six subjects who were exposed to noise elects to help the man, while five of the six subjects who were not exposed to noise elect to help the man. Do the data indicate that altruistic behavior is influenced by noise?

## Fisher’s Exact Test for Noise Effect on Altruistic Behavior

### Hypotheses:

* **Null Hypothesis (H0):** There is no association between noise exposure and altruistic behavior.
* **Alternative Hypothesis (H1):** There is an association between noise exposure and altruistic behavior.

data\_altruism <- matrix(c(1, 5, 5, 1), nrow = 2, byrow = TRUE,  
 dimnames = list(Noise = c("Exposed", "Not Exposed"),  
 Help = c("Yes", "No")))  
data\_altruism

## Help  
## Noise Yes No  
## Exposed 1 5  
## Not Exposed 5 1

fisher\_test\_result <- fisher.test(data\_altruism)  
fisher\_test\_result

##   
## Fisher's Exact Test for Count Data  
##   
## data: data\_altruism  
## p-value = 0.04004  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.0000000 0.8940503  
## sample estimates:  
## odds ratio   
## 0.06022953

**Determine the level of significance**: α = 0.05 **Decision rule**: If p-value < α, reject H0. Otherwise accept H0. Decision and conclusion: Since p-value ( 0.04 ) < α ( 0.05 ), H0 will be rejected. Since H0 was rejected, there is significant evidence, with significant level ( 0.05 ), to support the claim that there is association between altruistic behavior and noise exposure

### Visualization

data\_altruism\_long <- as\_tibble(as.table(data\_altruism))  
colnames(data\_altruism\_long) <- c("Noise", "Help", "Count")  
  
ggplot(data\_altruism\_long, aes(x = Noise, y = Count, fill = Help)) +  
 geom\_bar(stat = "identity", position = "dodge") +  
 theme\_minimal() +  
 labs(title = "Effect of Noise on Altruistic Behavior",  
 x = "Noise Exposure",  
 y = "Count",  
 fill = "Help Response")

