1(a)

a. # Replace 'path\_to\_file' with the actual file path if the file is not in the working directory

vinegar\_data <- vinegar

# Boxplot to visualize acidity differences between factory locations

boxplot(vinegar\_data$pH ~ vinegar\_data$Site, data = vinegar\_data,

xlab = "Factory Location", ylab = "Acidity",

main = "Acidity Levels across Factory Locations")A chart of a row of boxes

Description automatically generated

Interpretation

1. > median\_pH <- aggregate(pH ~ Site, data = vinegar\_data, FUN = median)

>

> print(median\_pH)

Site pH

1 Birmingham 4.050

2 London 3.470

3 New York 3.285

4 Paris 4.160

5 Sydney 5.560

A screen shot of a computer

Description automatically generated

Birmingham (Median pH = 4.050): Balsamic vinegar produced in Birmingham has a median pH around 4.050. This indicates a moderate acidity level compared to other factories.

London (Median pH = 3.470): The median pH of approximately 3.470 suggests that balsamic vinegar from London tends to be slightly more acidic than the Birmingham factory but less acidic than some other locations.

New York (Median pH = 3.285): Balsamic vinegar produced in New York has a median pH around 3.285, indicating relatively lower acidity compared to other locations in this dataset.

Paris (Median pH = 4.160): The median pH of about 4.160 for balsamic vinegar from Paris suggests a slightly higher acidity level compared to Birmingham and a moderate acidity level overall.

Sydney (Median pH = 5.560): Balsamic vinegar produced in Sydney has the highest median pH among the observed factories (approximately 5.560). This suggests that the balsamic vinegar from Sydney tends to have higher acidity compared to other locations in this dataset.

1. iqr\_values <- tapply(vinegar\_data$pH, vinegar\_data$Site,IQR)

# Calculate IQR for each factory location

iqr\_values <- tapply(vinegar\_data$pH, vinegar\_data$Site,IQR)

View(iqr\_values)

A screenshot of a computer

Description automatically generated

A screenshot of a computer code

Description automatically generated

The IQR values represent the spread or variability of the middle 50% of the pH measurements within each factory location. A higher IQR implies a greater spread of pH values, indicating more variability in acidity measurements.

Interpretation:

Birmingham (IQR = 1.5850): Balsamic vinegar produced in Birmingham has a moderate spread of pH values, suggesting some variability in acidity levels within this factory.

London (IQR = 1.3250): The IQR of approximately 1.3250 for London indicates a slightly lower variability in pH compared to Birmingham but still suggests moderate variability.

New York (IQR = 1.5025): Balsamic vinegar from New York exhibits a moderate spread of pH values, similar to Birmingham and London.

Paris (IQR = 1.0200): The IQR of approximately 1.0200 for Paris suggests a comparatively narrower spread of pH values, indicating a more consistent acidity level within this factory.

Sydney (IQR = 1.6600): Balsamic vinegar from Sydney has the highest IQR among the observed factories, indicating a relatively higher variability in acidity levels compared to other locations in this dataset.

Spread :Interquartile range highest in Sydney ,followed by Birmingham suggesting higher variability in acidity levels

1. Outliers

A screenshot of a computer program

Description automatically generated

Outliers: The identified potential outliers in the pH measurements for balsamic vinegar across different factory locations (Sydney and New York) suggest observations that significantly deviate from the typical range of acidity levels observed within their respective factories. Here's an interpretation:

Sydney Outliers (pH 6.68 and 6.82): These pH values of 6.68 and 6.82, identified as potential outliers in the Sydney factory, represent acidity levels significantly higher than the majority of pH measurements in that factory. Such high pH values may indicate unusual batches or extreme acidity levels compared to the typical range observed for balsamic vinegar produced in Sydney.

New York Outliers (pH 1.89 and 1.86): The pH values of 1.89 and 1.86, identified as potential outliers in the New York factory, are considerably lower than the typical acidity levels observed in that factory. These low pH values might indicate unusual acidity levels or errors in measurement within the New York production site.

1. Overall PH Range :

A screen shot of a computer

Description automatically generated

The overall pH range for the collected balsamic vinegar samples across all factory locations is from 1.86 to 6.82.

This range encapsulates the lowest and highest pH values observed in the dataset, providing a comprehensive understanding of the entire spectrum of acidity levels present in the balsamic vinegar samples from different production locations.

1. Comparative Analysis

Median pH Levels:

Sydney (Median pH = 5.560): Displays the highest median pH among the observed factories, indicating higher acidity levels on average compared to other locations.

Paris (Median pH = 4.160): Shows a moderate median pH, slightly higher than Birmingham and London but lower than Sydney.

Birmingham (Median pH = 4.050): Demonstrates a moderate median pH level, suggesting a balanced acidity level.

London (Median pH = 3.470): Indicates slightly lower acidity levels compared to other locations.

New York (Median pH = 3.285): Exhibits the lowest median pH, indicating comparatively lower acidity levels on average.

Variability (Interquartile Range - IQR):

Sydney (IQR = 1.6600): Shows the highest variability in pH levels among the observed factories, suggesting a wider spread of acidity measurements.

Birmingham (IQR = 1.5850): Displays moderate variability in acidity levels.

New York (IQR = 1.5025): Exhibits a similar moderate variability in pH as Birmingham.

London (IQR = 1.3250): Indicates slightly lower variability compared to Birmingham and New York.

Paris (IQR = 1.0200): Shows the narrowest spread of pH values, suggesting more consistent acidity levels within this factory.

Outliers:

Sydney: Identified two potential outliers with higher pH values (6.68 and 6.82).

New York: Identified two potential outliers with lower pH values (1.89 and 1.86).

Overall pH Range:

The overall pH range across all factories spans from 1.86 to 6.82, indicating the extremes of acidity levels observed within the collected dataset.

Interpretation:

Diversity in Acidity: There's considerable diversity in acidity levels among different factories, with Sydney showing higher median pH and greater variability, while New York demonstrates the lowest median pH and moderate variability.

Consistency: Paris exhibits the narrowest IQR, indicating more consistent acidity levels compared to other locations.

Potential Issues: Outliers identified in Sydney and New York may require further investigation, as they represent extreme pH values deviating from the typical range observed in these factories.

1. Relative pH Levels:

Relative Median pH Levels:

Sydney (Median pH = 5.560): Demonstrates the highest median pH among the observed factories, indicating the highest average acidity level.

Paris (Median pH = 4.160): Shows a moderate median pH, being relatively higher compared to some factories but lower than Sydney.

Birmingham (Median pH = 4.050): Displays a moderate median pH, suggesting a balanced acidity level.

London (Median pH = 3.470): Indicates slightly lower acidity levels compared to other locations.

New York (Median pH = 3.285): Exhibits the lowest median pH, indicating comparatively lower acidity levels on average.

Relative Overall pH Range:

Sydney (1.86 - 6.82): Shows the widest pH range, implying a broad spectrum of acidity levels observed.

Paris (3.55 - 5.61): Displays a narrower pH range compared to Sydney, indicating less variability.

Birmingham (2.08 - 5.25): Shows a moderate pH range, suggesting a moderate variability in acidity.

London (2.11 - 4.26): Indicates a narrower pH range compared to some locations, with less variability.

New York (1.86 - 4.36): Demonstrates a relatively narrower pH range, indicating less variability in acidity compared to Sydney.

Interpretation:

High Acidity Variation: Sydney shows the highest median pH and the widest pH range, indicating greater variability and potentially higher acidity variations in balsamic vinegar produced there.

Moderate to Lower Acidity Levels: Paris, Birmingham, London, and New York exhibit varying levels of acidity, with Paris showing a narrower pH range and moderate median pH, while New York demonstrates comparatively lower acidity levels overall.

1. Potential Observations

1. Acidity Variation: There is significant variability in acidity levels among the factories, with Sydney showing the highest median pH and widest pH range. This indicates substantial variability and potential differences in the acidity of balsamic vinegar produced in different regions.

2. Consistency and Quality Control: Factories like Paris exhibit narrower pH ranges and less variability, suggesting a more consistent acidity level. This consistency might be indicative of stringent quality control measures or standardized production processes.

3. Outliers and Anomalies: Identification of outliers, especially in Sydney and New York, suggests extreme pH values that deviate significantly from the typical acidity range observed in those factories. Further investigation into the reasons behind these outliers is crucial for quality assurance and process optimization.

4. Consumer Preferences: Differences in acidity levels across factories may cater to diverse consumer preferences. Some regions might prefer higher acidity in balsamic vinegar for culinary purposes, while others might prefer milder acidity.

5. Production Process Evaluation: Variations in acidity levels might be due to differences in raw materials, fermentation processes, or aging durations. Evaluating these factors across factories could help optimize production processes for consistency and quality.

6. Potential Quality Assurance Focus: Factories with wider pH ranges and higher variability, such as Sydney, might need focused quality assurance measures to ensure consistency and quality in their products.

7. Market Differentiation: Understanding and leveraging the distinct acidity profiles of balsamic vinegar from different locations can be a marketing strategy, catering to diverse consumer tastes and preferences.

1(b)

1. Hypothesis formulation : We will go with **Alternative Hypothesis (H1):** as There is at least one pair of factory locations with significantly different mean acidity levels.
2. Perform anova test
   1. # Perform one-way ANOVA
   2. anova\_result <- aov(vinegar\_data$pH ~ vinegar\_data$Site, data = vinegar\_data)
   3. # Summary of ANOVA results
   4. summary(anova\_result)

> summary(anova\_result)

A screenshot of a computer

Description automatically generated Df Sum Sq Mean Sq F value Pr(>F)

vinegar\_data$Site 4 24.57 6.143 6.682 0.000534 \*\*\*

Residuals 31 28.50 0.919

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Interpretation:

The one-way ANOVA test results indicate a statistically significant difference in the mean pH levels among the different factory locations for balsamic vinegar. Here's the interpretation of the ANOVA table:

ANOVA Results:

* Between-Group Variation (Factor 'Site'):
  + Degrees of Freedom (Df): 4
  + Sum of Squares (Sum Sq): 24.57
  + Mean Square (Mean Sq): 6.143
  + F-value: 6.682
  + p-value (Pr(>F)): 0.000534 (Significant at the 0.001 level)
* Within-Group Variation (Residuals):
  + Degrees of Freedom (Df): 31
  + Sum of Squares (Sum Sq): 28.50
  + Mean Square (Mean Sq): 0.919

Interpretation:

* The p-value (0.000534) is less than the chosen significance level (e.g., 0.05), indicating strong evidence against the null hypothesis.
* Therefore, we reject the null hypothesis, concluding that there are significant differences in the mean pH levels among at least some factory locations for balsamic vinegar.
* The F-statistic (6.682) further supports this, indicating variability in pH levels among factory locations that cannot be attributed to random chance alone.

Conclusion:

Based on these results, it's evident that there are statistically significant differences in mean pH levels across the different factory locations producing balsamic vinegar. As P value is less than 0.05 , we reject the null hypothesis ,indicating significant differences in mean acidity levels among factory locations.

As there are differences , we go for pairwise tests

> pairwise\_result <- pairwise.t.test(vinegar\_data$pH, vinegar\_data$Site, p.adj = "none")

> pairwise\_result

Pairwise comparisons using t tests with pooled SD

data: vinegar\_data$pH and vinegar\_data$Site

Birmingham London New York Paris

London 0.10843 - - -

New York 0.07349 0.88535 - -

Paris 0.66869 0.04539 0.02842 -

Sydney 0.01358 0.00017 7.6e-05 0.03654

P value adjustment method: none

A screenshot of a computer

Description automatically generated

the value 0.04539 for the comparison between Paris and London.

The p-value (0.04539) is below the conventional significance level of 0.05.

Interpretation: There's evidence to reject the null hypothesis for the pH levels between Paris and London. It suggests a statistically significant difference in mean pH levels between these two locations.

**Interpretation:**

* A lower p-value indicates stronger evidence against the null hypothesis (no difference in means).
* P-values smaller than a significance level (e.g., 0.05) suggest significant differences in mean pH levels between the compared factory locations.

Interpretation (using alpha = 0.05):

* Birmingham vs. London: p = 0.10843 (Not significant at alpha = 0.05)
* Birmingham vs. New York: p = 0.07349 (Not significant at alpha = 0.05)
* Birmingham vs. Paris: p = 0.66869 (Not significant at alpha = 0.05)
* Birmingham vs. Sydney: p = 0.01358 (Significant at alpha = 0.05)

For instance, the comparison between Birmingham and Sydney shows a p-value of 0.01358, indicating a significant difference in mean pH levels between these two factory locations.

It's important to note that no adjustment for multiple comparisons has been applied here (e.g., Bonferroni correction or Tukey's HSD).

A screenshot of a computer screen

Description automatically generated

One way testA white background with black numbers

Description automatically generated

ANOVA Results:

* Between-Group Variation (Factor 'grp'):
  + Degrees of Freedom (Df): 4
  + Sum of Squares (Sum Sq): 24.571
  + Mean Square (Mean Sq): 6.1427
  + F-value: 6.682
  + p-value (Pr(>F)): 0.0005335 (Significant at the 0.001 level)
* Within-Group Variation (Residuals):
  + Degrees of Freedom (Df): 31
  + Sum of Squares (Sum Sq): 28.498
  + Mean Square (Mean Sq): 0.9193

Interpretation:

* The p-value (0.0005335) is less than the chosen significance level (e.g., 0.05), indicating strong evidence against the null hypothesis.
* Therefore, based on this ANOVA test, we reject the null hypothesis, concluding that there are significant differences in the mean pH levels among at least some groups (factory locations) for balsamic vinegar.

This analysis confirms that there are statistically significant differences in mean pH levels across the different factory locations. The F-value of 6.682 also supports the variability observed in pH levels among factory locations,

A larger F-value suggests larger differences between group means relative to the variability within groups.

The p-value obtained (p-value = 0.0005335) represents the probability of observing such an extreme F-statistic if the null hypothesis (no difference in means) were true.

In this case, the p-value is very low (less than the conventional significance level of 0.05), suggesting strong evidence against the null hypothesis.

The obtained p-value of 0.0005335 is less than 0.05 (or any chosen significance level), indicating that there are significant differences in mean acidity levels among the different site locations.

Based on this one-way ANOVA test, there is evidence to suggest that at least one group has a significantly different mean acidity level compared to the others.

1( c )

A screenshot of a computer

Description automatically generated

The Shapiro-Wilk test for normality has been applied to the residuals obtained from the ANOVA model that assessed the pH levels across different factory locations for balsamic vinegar. Here are the results:

Shapiro-Wilk Test Results:

Test Statistic (W): 0.94045

p-value: 0.05246

Interpretation:

The Shapiro-Wilk test assesses the normality assumption of the residuals (deviations from the model's predictions).

The p-value obtained (0.05246) is greater than the commonly chosen significance level of 0.05.

As a result, we do not have sufficient evidence to reject the null hypothesis of normality for the residuals at the 0.05 significance level.

Conclusion:

The Shapiro-Wilk test does not provide strong evidence against the assumption of normality for the residuals at a typical significance level of 0.05. This suggests that the residuals from the ANOVA model reasonably follow a normal distribution. However, it's important to note that while the p-value is slightly above 0.05, it's still relatively close to this threshold, so exercising caution when assuming normality might be advisable, especially with larger datasets where small departures from normality may be detected as statistically significant.

# Checking assumptions - Normality and Homogeneity of Variance

par(mfrow = c(1, 2))

# Residuals vs. Fitted Values (Homogeneity of Variance)

plot(anova\_model, which = 1)

# Normal Q-Q plot (Normality)

qqnorm(residuals)

qqline(residuals)

A comparison of a normal and a normal q-q

Description automatically generated

Levene's Test for Homogeneity of Variances:

Using Levene's Test (leveneTest()):

> levene.test(acidityLevels,site)

Modified robust Brown-Forsythe Levene-type test based on the absolute deviations from the

Median

A screenshot of a computer code

Description automatically generated

the result of a modified robust Brown-Forsythe Levene-type test, which assesses the homogeneity of variances (equality of variances) among different groups or levels of a factor. In this case, the test is conducted on the variable **acidityLevels** across the factor **site** using the median as a measure.

**Levene's Test Results:**

* **Test Statistic**: 0.28038
* **p-value**: 0.8884

**Interpretation:**

* The p-value obtained from the Levene-type test is 0.8884.
* A higher p-value suggests a lack of evidence against the null hypothesis of equal variances among the groups.
* In this case, with a p-value of 0.8884 (which is greater than the common significance level of 0.05), there is insufficient evidence to reject the null hypothesis.

**Conclusion:**

The Levene-type test based on the absolute deviations from the median does not provide significant evidence against the assumption of equal variances among the different groups (sites) regarding the **acidityLevels**. Therefore, based on this test, the assumption of homogeneity of variances across the groups is reasonable, suggesting similar variances in acidity levels among the different factory locations for balsamic vinegar.

Visual Inspection of Variance Homogeneity:

A screenshot of a computer

Description automatically generated

**ANOVA Results:**

* **Between-Group Variation (Site):**
  + Degrees of Freedom (Df): 4
  + Sum of Squares (Sum Sq): 24.57082
  + Mean Square (Mean Sq): 6.143
  + F-value: 6.682
  + p-value (Pr(>F)): 0.000534 (Significant at the 0.001 level)
* **Within-Group Variation (Residuals):**
  + Degrees of Freedom (Df): 31
  + Sum of Squares (Sum Sq): 28.49786
  + Mean Square (Mean Sq): 0.919
  + Residual standard error: 0.9587939

**Interpretation:**

* The p-value obtained from the ANOVA test is 0.000534, which is less than the significance level (e.g., 0.05), indicating strong evidence against the null hypothesis.
* Therefore, based on this ANOVA model, we reject the null hypothesis, concluding that there are significant differences in the mean pH levels among at least some groups (factory locations) for balsamic vinegar.
* The F-value of 6.682 supports the variability observed in pH levels among factory locations that cannot be explained by random chance alone.

This analysis confirms statistically significant differences in mean pH levels across the different factory locations producing balsamic vinegar. The variation observed suggests that the choice of factory location has a significant impact on the acidity levels of the produced balsamic vinegar.

Residuals represent the differences between the observed pH values and the pH values predicted by the ANOVA model. Each value in the list corresponds to a specific observation or sample in your dataset.

**Interpretation of Residuals:**

* Positive Residuals: Indicates observed pH levels higher than predicted by the ANOVA model.
* Negative Residuals: Indicates observed pH levels lower than predicted by the ANOVA model.
* Magnitude of Residuals: Represents the extent of deviation from the predicted pH levels.

For instance:

* Residuals closer to zero indicate observations close to the model's prediction.
* Larger residuals (positive or negative) suggest greater discrepancies between the observed pH values and what the ANOVA model predicted.

The residuals are essential in understanding how well the ANOVA model fits the data. Examining patterns or distribution of residuals can help assess the model's adequacy and identify potential issues like heteroscedasticity or outliers.

Interpretation:

Interpretation of Diagnostic Plots:

1. Residuals vs. Fitted Values Plot (Homogeneity of Variance):
   * the random scattering of points around the horizontal line is zero.
   * No distinct patterns or trends are visible; this confirms for homogeneity of variance.
   * If a cone or fan-like shape is observed, it suggests heteroscedasticity, violating the assumption – no such pattern is observed.
2. Normal Q-Q Plot (Normality):
   * Points fall approximately along the diagonal line.
   * Some of the points points deviate significantly from the line, it indicates departure from normality.
3. Shapiro test for normality test:

The Shapiro-Wilk test for normality of residuals from the ANOVA model yields a p-value of approximately 0.05246. Typically, in statistical practice, the commonly used significance level for hypothesis testing is 0.05.

In this case, the obtained p-value (0.05246) is slightly higher than the typical significance level of 0.05. This suggests that there is no strong evidence to reject the null hypothesis, implying that the residuals may be normally distributed.

Levene's Test Results:

A non-significant result (p > 0.05) from Levene's test suggests that the variances among groups are approximately equal.

Residuals Plot:

The absence of systematic patterns suggests that variances are homogeneous among different groups.

1(d)

Call:

aov(formula = vinegar\_data$pH ~ vinegar\_data$Site, data = vinegar\_data)

Terms:

vinegar\_data$Site Residuals

Sum of Squares 24.57082 28.49786

Deg. of Freedom 4 31

Residual standard error: 0.9587939

Estimated effects may be unbalanced

A screenshot of a computer

Description automatically generated

Interpretation of Anova Results:

* The ANOVA results suggest that there is some variation in pH levels among the different factory locations.
* The significant sum of squares associated with 'vinegar\_data$Site' indicates that the pH levels vary significantly among the factory locations.
* The residual standard error gives an estimate of the variability of the observed pH levels around the mean predicted by the model.

Based on these results, further post-hoc tests (if required) or pairwise comparisons between factory locations can be conducted to determine specific differences in acidity levels among these locations.

A screenshot of a computer

Description automatically generated

Interpretation of Tukey's HSD Results:

* p adj (Adjusted p-value): This value determines the statistical significance of the observed differences after correcting for multiple comparisons. Smaller p-values (< 0.05) indicate significant differences.

Pairwise Comparisons (95% Confidence Level):

1. London vs. Birmingham:
   * Not statistically significant (p = 0.477).
2. New York vs. Birmingham:
   * Not statistically significant (p = 0.363).
3. Paris vs. Birmingham:
   * Not statistically significant (p = 0.992).
4. Sydney vs. Birmingham:
   * Marginally significant (p = 0.092).
5. New York vs. London:
   * Not statistically significant (p = 0.999).
6. Paris vs. London:
   * Not statistically significant (p = 0.252).
7. Sydney vs. London:
   * Statistically significant (p = 0.002).
   * There is a significant difference in acidity levels between Sydney and London.
8. Paris vs. New York:
   * Not statistically significant (p = 0.172).
9. Sydney vs. New York:
   * Statistically significant (p = 0.001).
   * There is a significant difference in acidity levels between Sydney and New York.
10. Sydney vs. Paris:
    * Not statistically significant (p = 0.212).

Key Findings:

1. **Significantly Different Means:**
   * Sydney-Birmingham, Sydney-London, and Sydney-New York comparisons all have significant differences (p < 0.05) based on the adjusted p-values.
   * London-Sydney also displays a notable difference in mean pH levels.
2. **Non-Significant Differences:**
   * Birmingham with other locations (London, New York, Paris) doesn't exhibit significant differences in mean pH levels.
   * Paris shows no significant differences in pH levels concerning Birmingham, London, and New York.
3. **Marginal Differences:**
   * Paris-London and Paris-New York comparisons have p-values slightly above 0.05 (0.25 and 0.17, respectively), indicating marginal differences which are not statistically significant at the conventional 0.05 level but may be worth further investigation.

Conclusion:

Significant differences in pH levels are observed between specific pairs of factory locations, while other comparisons do not demonstrate significant differences.

Consider the adjusted p-values and confidence intervals to assess the significance of the differences after adjusting for multiple comparisons..

Based on the Tukey's HSD test, significant differences in acidity levels were observed between Sydney and both London and New York.

Other comparisons did not reveal statistically significant differences in acidity levels among the factory locations.

These results suggest that the acidity levels of balsamic vinegar between Sydney and both London and New York are significantly different. However, for other pairwise comparisons, no significant differences in acidity were found between the mentioned locations.

1(e)

> p\_values < - posthoc$`vinegar\_data$Site`[, "p adj"]

London-Birmingham New York-Birmingham Paris-Birmingham Sydney-Birmingham

FALSE FALSE FALSE FALSE

New York-London Paris-London Sydney-London Paris-New York

FALSE FALSE FALSE FALSE

Sydney-New York Sydney-Paris

FALSE FALSE

> print(p\_values)

[1] 0.4765222 0.3632112 0.9923747 0.0916315 0.9998934 0.2519132 0.0015038

[8] 0.1723996 0.0006834 0.2118486

> holm\_corrected <- p.adjust(posthoc$`vinegar\_data$Site`[, "p adj"], method = "holm")

> holm\_corrected

London-Birmingham New York-Birmingham Paris-Birmingham Sydney-Birmingham

1.000000000 1.000000000 1.000000000 0.733052161

New York-London Paris-London Sydney-London Paris-New York

1.000000000 1.000000000 0.013533869 1.000000000

Sydney-New York Sydney-Paris

0.006833638 1.000000000

>

> # Apply Bonferroni correction

> bonferroni\_corrected <- p.adjust(posthoc$`vinegar\_data$Site`[, "p adj"], method = "bonferroni")

>

> # Display the corrected p-values

> bonferroni\_corrected

London-Birmingham New York-Birmingham Paris-Birmingham Sydney-Birmingham

1.000000000 1.000000000 1.000000000 0.916315201

New York-London Paris-London Sydney-London Paris-New York

1.000000000 1.000000000 0.015037632 1.000000000

Sydney-New York Sydney-Paris

0.006833638 1.000000000

>

> correction\_results = data.frame(original\_values = p\_values, Holm\_correction = holm\_corrected, bonferroni\_corrected = bonferroni\_corrected)

>

> print(correction\_results)

original\_values Holm\_correction bonferroni\_corrected

London-Birmingham 0.4765222 1.000000000 1.000000000

New York-Birmingham 0.3632112 1.000000000 1.000000000

Paris-Birmingham 0.9923747 1.000000000 1.000000000

Sydney-Birmingham 0.0916315 0.733052161 0.916315201

New York-London 0.9998934 1.000000000 1.000000000

Paris-London 0.2519132 1.000000000 1.000000000

Sydney-London 0.0015038 0.013533869 0.015037632

Paris-New York 0.1723996 1.000000000 1.000000000

Sydney-New York 0.0006834 0.006833638 0.006833638

Sydney-Paris 0.2118486 1.000000000 1.000000000

>

A screenshot of a computer code

Description automatically generated

Observations:

Holm's correction adjusts the p-values to control for multiple comparisons while maintaining a familywise error rate.

The adjusted p-values are generally higher after correction, which may indicate a stricter criterion for identifying significant differences.

Two comparisons (Sydney-London and Sydney-New York) show adjusted p-values below 0.05 after Holm's correction, suggesting significant differences in pH levels between these pairs of locations.

Other comparisons have adjusted p-values greater than 0.05, indicating non-significant differences after Holm's correction.

Conclusion:

Holm's correction has made the criteria for significance more stringent, resulting in fewer significant differences between factory locations in terms of pH levels.

Consider these adjusted p-values to interpret the significance of differences after applying the correction. In this case, Sydney-London and Sydney-New York comparisons remain statistically significant after Holm's correction.

Bonferroni Correction:

Observations:

Bonferroni correction is a conservative method that adjusts the p-values to control for multiple comparisons, maintaining a stricter significance threshold.

After the Bonferroni correction, most adjusted p-values remain above 0.05, suggesting non-significant differences between the majority of the pairs of factory locations regarding pH levels.

Only one comparison (Sydney-London) shows an adjusted p-value below 0.05 after Bonferroni correction, indicating a significant difference in pH levels between these two locations.

Conclusion:

Bonferroni correction tends to be more conservative, requiring a lower threshold for significance compared to Holm's correction.

In this case, after Bonferroni correction, only the Sydney-London comparison remains statistically significant, while other comparisons show non-significant differences in pH levels between factory locations.

Summary:

* In both correction methods, some pairwise comparisons resulted in adjusted p-values lower than 0.05, indicating statistical significance even after multiple comparison adjustments.
* For instance, 'Sydney-London' and 'Sydney-New York' comparisons have adjusted p-values below 0.05, indicating significant differences in acidity levels between these pairs after correction.

1(f)

The experiment displays statistically significant acidity differences between certain factory locations, particularly evident in pairs like Sydney-London (p = 0.0015) and Sydney-New York (p = 0.00068) based on both ANOVA (p < 0.05) and Tukey HSD tests. Enhancements could involve larger sample sizes per location to increase statistical robustness and reliability. Exploring additional factors, such as production methods or environmental variables, may elucidate influential factors impacting acidity variation. Implementing complementary analyses or controlling potential confounders could yield a more comprehensive understanding of acidity disparities among factory locations.

2(a)

# Boxplot of Venom Yield by Body Class and Expression

boxplot(`Yield (mg)` ~ `Body Class` \* `Expression`, data = VenomYield,

xlab = "Body Class and Expression", ylab = "Venom Yield (mg)",

main = "Venom Yield Distribution by Body Class and Expression")

A graph of a graph showing a number of different sizes

Description automatically generated with medium confidence

A graph of body length and body length

Description automatically generated

A diagram of body length

Description automatically generated

A graph showing a number of different sizes of objects

Description automatically generated with medium confidence

Interpretation of the distribution of Venom Yield based on Body Class using the boxplot and violin plot:

**Boxplot Interpretation:**

The boxplot displays the distribution of Venom Yield categorized by Body Class (small and large spiders).

* **Median and Spread:**
  + For small spiders (red boxplot), the median Venom Yield appears to be lower, with the majority of values falling below a certain range.
  + In contrast, for large spiders (blue boxplot), the median Venom Yield is higher, indicating a potentially higher yield among larger spiders.
* **Variability and Outliers:**
  + The range of Venom Yield for large spiders seems wider than for small spiders, suggesting more variability.
  + Outliers, represented by points beyond the whiskers, are observed in both categories, especially in the large spider group, indicating potential extreme values in Venom Yield.

Interpretation of Boxplot of Venom Yield by Body Class and Expression:

1.Venom yield Median values are observed higher for Body Class ‘Large’ and Expression ‘High’.

2.For Body Class ‘Small and Expression ‘High’ , there is a variability of Venom Yield values given by IQR(Inter quartile ranges)

3.For

Body Class ‘Small and Expression ‘Low’ smaller median values for Venom yield are observed as well as min values

4.The outliers which were at 1.5 + IQR are highest for Body Class ‘Large’ and Expression ‘High’.

5.The outliers ,both 1.5+IQR,1.5 -IQR were uniform for Body Class ‘Small and Expression ‘Low’

**Scatter Plot Interpretation:**

1. **Body Length vs. Venom Yield:**
   * **Body Length (x-axis):** The scatter plot displays Body Length on the x-axis, representing the size of individual spiders.
   * **Venom Yield (y-axis):** The y-axis represents the Venom Yield (in mg), showing the amount of venom produced by each spider.
2. **Color-Coding by Body Class:**
   * **Red and Blue Points:** Spiders are color-coded into two categories: red for "small" and blue for "large" based on their Body Class classification.
3. **Observations Based on Scatter Plot:**
   * **Trend:** There seems to be a visible trend indicating a potential positive correlation between Body Length and Venom Yield.
   * **Clustered Distribution:** Spiders of similar Body Class tend to cluster together in certain regions of the plot.
   * **Overlap:** There's some overlap between the two categories, suggesting that Body Length alone may not entirely explain the variation in Venom Yield.
4. **Interpretation:**
   * **Correlation:** The scatter plot implies that larger spiders might exhibit higher Venom Yields compared to smaller ones, as there's a general trend of higher Venom Yields with increasing Body Length.
   * **Variation within Categories:** While the trend is apparent, there's variability within each Body Class. Some smaller spiders produce higher Venom Yields than some larger spiders.

2(b)

A screenshot of a computer

Description automatically generated

Interpretation:

'Body Class' shows a significant effect on 'Yield (mg)' with a very low p-value (p < 0.001). This implies that there is a significant difference in venom yield between the 'Large' and 'Small' body classes of spiders.

The factor "Body Class" shows a significant effect on Venom Yield (mg) with an F-value of 49.160 and a p-value of 3.17e-08 (highly significant).

'Expression' also exhibits a significant effect on 'Yield (mg)' with a very low p-value (p < 0.001). This suggests that differences in gene expression levels ('High' and 'Low') significantly influence venom yield.

**Expression:** The factor "Expression" also exhibits a significant effect on Venom Yield with an F-value of 38.446 and a p-value of 3.75e-07 (highly significant).

However, the interaction between 'Body Class' and 'Expression' does not appear to have a significant effect on 'Yield (mg)' as indicated by a non-significant p-value (p = 0.407). This means that the impact of one factor on venom yield does not depend significantly on the levels of the other factor.

In summary, both 'Body Class' and 'Expression' have a significant individual effect on venom yield, but their interaction does not significantly influence the venom yield of the spiders in this study.

The residual term represents unexplained variability or random error in the model. The residual mean square (1387) indicates the average variability of data points around the fitted values.

These findings indicate that both the physical characteristics (Body Class) and genetic attributes (Expression) independently play essential roles in influencing the Venom Yield of male funnel-web spiders. Further targeted analyses or investigations into each factor's specific impact could provide deeper insights into the mechanisms influencing Venom Yield in these spiders.

2 (c )

A screenshot of a computer code

Description automatically generated

The ANCOVA results examining the effects of Expression (gene Expression levels) and Body Length (as a covariate) on Venom Yield are as follows:

**ANCOVA Summary:**

* **Expression Effect:** The coefficient estimate for the "Expressionlow" variable is -53.046 with a standard error of 23.213 and a t-value of -2.285, resulting in a p-value of 0.0283. This indicates that there is a significant difference in the intercepts between "low" and "high" gene Expression levels concerning Venom Yield.
* **Body Length Effect:** The coefficient estimate for "Body Length (cm)" is 40.678 with a standard error of 4.908 and a t-value of 8.287, resulting in a highly significant p-value of 7.28e-10. This signifies a strong and significant linear relationship between Body Length and Venom Yield.
* **Interaction Effect:** The interaction term ("Expressionlow:Body Length (cm)") has a coefficient estimate of -7.467 with a standard error of 6.990 and a t-value of -1.068, leading to a non-significant p-value of 0.2925. This suggests that the interaction between gene Expression and Body Length does not significantly impact Venom Yield.
* **Model Fit:** The multiple R-squared value is 0.8334, indicating that approximately 83.34% of the variance in Venom Yield is explained by the combined effects of Expression and Body Length.

**Interpretation:**

* **Expression Effect:** There is a statistically significant difference in Venom Yield between spiders with "low" and "high" gene Expression levels. The intercepts of their regression lines differ significantly concerning Venom Yield.
* **Body Length Effect:** Body Length demonstrates a highly significant and positive linear relationship with Venom Yield. For each unit increase in Body Length (cm), Venom Yield increases by approximately 40.678 mg, after accounting for other factors.
* **Interaction Effect:** The interaction between Expression and Body Length does not significantly influence Venom Yield. The relationship between Body Length and Venom Yield remains consistent regardless of gene Expression levels.

**Conclusion:**

The ANCOVA analysis highlights that both gene Expression levels and Body Length significantly affect Venom Yield in male funnel-web spiders. While spiders with different gene Expression levels exhibit varying Venom Yields, Body Length demonstrates a strong and significant linear relationship with Venom Yield. However, there's no significant interaction between gene Expression and Body Length concerning their impact on Venom Yield.

These findings imply that both genetic factors (Expression) and physical characteristics (Body Length) independently contribute to variations in Venom Yield among male funnel-web spiders.

2(d)

Body Length seems to be a suitable covariate for the ANCOVA based on the following reasons:

Relevance to the Outcome Variable (Venom Yield): Body Length often correlates with various biological traits in organisms. In this context, the body size of spiders might affect venom production. A larger body size could accommodate larger venom glands, affecting venom yield.

**Theoretical Grounds:** Body Length should be theoretically justifiable as a covariate affecting the response variable (Venom Yield) and should have a plausible biological or theoretical relationship with the response variable. In this case, there might be biological reasons to assume that larger spiders (with greater Body Length) produce more venom, but this assumption should be supported by prior knowledge or research.

**Correlation:** Body Length should exhibit a linear relationship with the response variable. A strong correlation suggests that changes in Body Length predict changes in Venom Yield. This relationship can be assessed through scatterplots and correlation analysis.

**Homogeneity of Regression Slopes:** The assumption of homogeneity of regression slopes requires that the relationship between the covariate (Body Length) and the response variable (Venom Yield) is consistent across different levels of the categorical variable (Expression in this case). This assumption should be tested, and the absence of a significant interaction term (Expression:Body Length) in the ANCOVA model supports this assumption.

**Residual Analysis:** The ANCOVA assumes that the residuals are normally distributed, have constant variance, and are independent. Examination of residual plots can help validate these assumptions. If the residuals violate these assumptions, it may indicate that Body Length is not an appropriate covariate.

**Practicality and Interpretability:** A suitable covariate should be practically measurable and interpretable in the context of the study. Body Length is often easily measured and interpretable in studies involving organisms.

Statistical Significance: The ANCOVA output indicates that 'Body Length' is statistically significant (p < 0.05) in predicting 'Yield (mg)' after adjusting for 'Expression'. This suggests that changes in 'Body Length' are associated with changes in 'Yield (mg)'.

Adjusted R-squared Value: The adjusted R-squared value of the model including 'Body Length' as a covariate is relatively high (0.8195), suggesting that 'Body Length' explains a substantial portion of the variability in 'Yield (mg)'.

Based on the ANCOVA results, Body Length was found to have a statistically significant and positive linear relationship with Venom Yield. Additionally, there was no significant interaction between Body Length and gene Expression concerning Venom Yield, indicating consistent relationships across different levels of Expression.

Therefore, considering these factors, Body Length appears to be a suitable covariate for the ANCOVA in this study. It demonstrates a plausible theoretical link to Venom Yield, exhibits a significant linear relationship, and does not violate assumptions of homogeneity of regression slopes. However, it's essential to acknowledge that the appropriateness of a covariate is context-specific and should be evaluated based on the study's objectives and empirical findings.

2(e)

the choice between performing an ANOVA (b) and ANCOVA (c) depends on the research question, the variables' relationships, and the study's objectives:

1. ANOVA (b):
   * Advantages: ANOVA helps analyze whether there are significant differences in mean venom yield among different levels of the 'Body Class' and 'Expression' factors. It examines group differences without considering other potential influential variables (covariates).
   * Suitability: ANOVA is appropriate when there is no need to control for or consider the influence of other continuous variables on the outcome variable (venom yield). It assumes that the effect of the factors (e.g., 'Body Class' and 'Expression') on the outcome is not confounded by other variables.
   * Limitations: ANOVA does not account for or control the potential impact of covariates like 'Body Length'. If 'Body Length' has a significant influence on venom yield and is related to the categorical predictors ('Body Class' and 'Expression'), ANOVA might not fully account for this relationship.
2. ANCOVA (c):
   * Advantages: ANCOVA extends ANOVA by including one or more continuous variables (covariates) alongside categorical predictors. It enables the assessment of group differences while controlling for the effects of continuous variables.
   * Suitability: ANCOVA is suitable when there's a need to control for the influence of continuous variables ('Body Length') on the outcome ('Yield (mg)') while examining group differences associated with categorical predictors ('Body Class' and 'Expression').
   * Limitations: ANCOVA assumes a linear relationship between the covariate(s) and the outcome variable. If this assumption is violated, the results might be misleading. Additionally, ANCOVA assumes that the covariate(s) do not interact with the categorical predictors, which should be assessed and considered in the analysis.

The better approach depends on whether 'Body Length' is considered an influential variable that impacts venom yield and if it is related to the categorical predictors. If 'Body Length' is deemed an essential variable influencing venom yield and is related to the categorical predictors, ANCOVA would be a better choice to control for its influence. If 'Body Length' is not considered influential or is independent of the categorical predictors, ANOVA might be more appropriate for exploring group differences in venom yield.

3(a)

A computer code with blue text

Description automatically generated

A graph of a function

Description automatically generated

A graph of a function

Description automatically generated

In survival analysis, the Survival Function (S(t)) and Failure Probability Density Function (f(t)) are essential in understanding the probability of survival and the distribution of failure events over time.

1. **Survival Function S(t):**
   * The plot of the Survival Function S(t) shows the probability of survival beyond each time point along the x-axis (Time in years).
   * Initially, at time t = 0, the probability of survival is 1 (S(0) = 1), indicating that all subjects or units are alive at the starting time point.
   * As time progresses, the Survival Function decreases, indicating the declining probability of survival. The downward slope of the curve signifies the cumulative proportion of subjects experiencing the event of interest (failure) as time passes.
2. **Failure Probability Density Function f(t):**
   * The plot of the Failure Probability Density Function f(t) displays the density of failure events occurring at different time points along the x-axis (Time in years).
   * The function illustrates the likelihood of the event (failure) occurring at each specific time point.
   * Peaks or higher values on the graph indicate time periods when failure events are more likely to happen, showing periods of increased risk or susceptibility to the event.

In summary, these plots are used to analyze and visualize survival data, providing insights into how survival probabilities change over time and when failure events are more prevalent. They are fundamental tools for assessing and modeling survival probabilities and understanding the dynamics of events in survival analysis.

Top of Form

Bottom of Form

3(b)

> # Given data

> time\_intervals <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9)

> ninint <- 100

> nlost <- c(1, 0, 3, 4, 11, 8, 8, 15, 17, 10)

> nevent <- c(0, 0, 2, 0, 0, 5, 3, 4, 2, 1)

>

> # Generating the life table

> my\_table <- lifetab(time\_intervals, ninint, nlost, nevent)

Warning messages:

1: In c(diff(-1 \* Sj), NA)/diff(tis) :

longer object length is not a multiple of shorter object length

2: In nevent/diff(tis) :

longer object length is not a multiple of shorter object length

3: In Sj \* qj/diff(tis) :

longer object length is not a multiple of shorter object length

4: In hmj \* diff(tis) :

longer object length is not a multiple of shorter object length

>

> print(my\_table)

nsubs nlost nrisk nevent surv pdf hazard se.surv se.pdf se.hazard

0-1 100 1 99.5 0 1.0000000 0.00000000 0.00000000 0.00000000 NaN NaN

1-2 99 0 99.0 0 1.0000000 0.00000000 0.00000000 0.00000000 NaN NaN

2-3 99 3 97.5 2 1.0000000 0.02051282 0.02072539 0.00000000 0.01435522 0.01465428

3-4 94 4 92.0 0 0.9794872 0.00000000 0.00000000 0.01435522 NaN NaN

4-5 90 11 84.5 0 0.9794872 0.00000000 0.00000000 0.01435522 NaN NaN

5-6 79 8 75.0 5 0.9794872 0.06529915 0.06896552 0.01435522 0.02822868 0.03082397

6-7 66 8 62.0 3 0.9141880 0.04423490 0.04958678 0.03123227 0.02495929 0.02862014

7-8 55 15 47.5 4 0.8699531 0.07325921 0.08791209 0.03878172 0.03520520 0.04391356

8-9 36 17 27.5 2 0.7966939 0.05794138 0.07547170 0.04990109 0.03961934 0.05332854

9-1 17 10 12.0 1 0.7387525 NA NA 0.06080799 NA NA

>

> # Extracting S, f, and h

> S <- my\_table[, 5]

> f <- my\_table[, 6]

> h <- my\_table[, 7]

>

> # Adjusted time intervals

> t <- 0.5 + c(0:9)

>

> # Setting up a single plot area to combine all plots

> par(mfrow = c(3, 1)) # 3 rows, 1 column

>

> # Plotting all functions in one figure

> plot(t, S, type = 'l', col = 'blue', xlab = 'Time (years)', ylab = 'Survival Probability', main = 'Survival Function S(t)')

>

> plot(t, f, type = 'l', col = 'green', xlab = 'Time (years)', ylab = 'Failure Probability Density', main = 'Failure Density Function f(t)')

>

> plot(t, h, type = 'l', col = 'red', xlab = 'Time (years)', ylab = 'Hazard Function', main = 'Hazard Function h(t)')

>

A screenshot of a computer

Description automatically generated

Life table:

A number of numbers on a white background

Description automatically generated

A screenshot of a graph

Description automatically generated

Interpretation :

The Survival function suggest the units have reasonable good life span with survival probability that decreases gradually over time.

The failure density appears to have peaks and troughs, indicating that there are specific times when the units are more likely to fail and other times when failures are less frequent

The hazard function curve is not constant, indicating that the risk of failure is not uniform

Showing some variability due to aging process of components used in AC .

Conclusion on whether model in part(a) is accurate or not:

The model provided in (a) is a basic theoretical representation of survival analysis based on a hazard function, survival function, and failure probability density function. However, without specific context or real data to compare these theoretical functions against observed survival data, it's difficult to evaluate the accuracy or applicability of this model in a real-world scenario.

Here are some considerations to assess the model:

1. Theoretical Soundness: The model is built upon fundamental concepts of survival analysis, including the hazard function, survival function, and failure density function. The functions are derived mathematically and follow the standard definitions used in survival analysis.
2. Hazard Function: The hazard function 1−exp(−t) used in this model is a simple form that indicates an increasing hazard over time. Depending on the context and characteristics of the data, this may or may not accurately represent the true hazard pattern. Real hazard functions can be more complex and may vary for different populations or conditions.
3. Survival Function and Failure Density: The survival function and failure density are calculated based on the hazard function and integral computations. These functions describe the cumulative survival probability and failure density over time. Their accuracy depends on the accuracy of the hazard function and the integration method used.
4. Validation against Data: To assess the accuracy of the model, it needs to be validated against empirical data from the specific population or context it aims to represent. Real-life survival data should be used to fit the model and check how well it predicts observed survival patterns.
5. Assumptions: The model assumes a specific form for the hazard function, which might not hold true for all scenarios. Survival analysis often involves making assumptions about the underlying hazard function, and the appropriateness of these assumptions should be evaluated based on the data.

In summary, while the provided code demonstrates fundamental concepts of survival analysis by defining and plotting theoretical survival and failure density functions based on a hazard function, its accuracy and applicability in real scenarios can only be assessed by comparing these functions against observed survival data from the relevant population or study context. Without such validation against empirical data, it's challenging to ascertain the model's accuracy or usefulness for real-world predictions or analysis.

Top of Form

Bottom of Form