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. Variations in Median and Spread:Higher acidity level observed in Sydney location , lower acidity level observed in New York
2. 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

1. Spread :Interquartile range highest in Sydney ,followed by Birmingham suggesting higher variability in acidity levels
2. Outliers: The outliers tending to the highest acidity values are observed in Paris, while tending to the minimum acidity values are observed in Birmingham.
3. Comparison of Acidity Levels Between Locations:Sydney has consistently higher acidity values , while New York has consistently lower acidity values.

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:**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.

Anova analysis

> grp = factor(vinegar\_data$Site)

> grp

[1] London London London London London London London Birmingham Birmingham Birmingham

[11] Birmingham Birmingham Birmingham Birmingham Sydney Sydney Sydney Sydney Sydney Sydney

[21] Sydney New York New York New York New York New York New York New York New York Paris

[31] Paris Paris Paris Paris Paris Paris

Levels: Birmingham London New York Paris Sydney

> y = vinegar\_data$pH

> y

[1] 3.96 4.26 2.11 3.47 3.79 2.99 2.08 3.01 2.50 5.03 3.80 5.25 4.95 4.05 6.68 4.55 5.60 4.36 6.82 5.56 4.41 3.77 3.38

[24] 4.36 3.19 1.89 4.30 2.57 1.86 3.51 5.61 5.15 4.16 3.55 3.86 4.30

> drug.lm = lm(y ~ grp)

> anova(drug.lm)

Analysis of Variance Table

Response: y

Df Sum Sq Mean Sq F value Pr(>F)

grp 4 24.571 6.1427 6.682 0.0005335 \*\*\*

Residuals 31 28.498 0.9193

---

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

>

> oneway.test(y ~ grp, var.equal = TRUE)

One-way analysis of means

data: y and grp

F = 6.682, num df = 4, denom df = 31, p-value = 0.0005335

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 )

**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

data: acidityLevels

Test Statistic = 0.28038, p-value = 0.8884

**Visual Inspection of Variance Homogeneity:**

> anova\_model <- aov(vinegar\_data$pH ~ vinegar\_data$Site, data = vinegar\_data)

> residuals <- residuals(anova\_model)

> print(residuals)

1 2 3 4 5 6 7

0.722857143 1.022857143 -1.127142857 0.232857143 0.552857143 -0.247142857 -1.157142857

8 9 10 11 12 13 14

-1.074285714 -1.584285714 0.945714286 -0.284285714 1.165714286 0.865714286 -0.034285714

15 16 17 18 19 20 21

1.254285714 -0.875714286 0.174285714 -1.065714286 1.394285714 0.134285714 -1.015714286

22 23 24 25 26 27 28

0.605000000 0.215000000 1.195000000 0.025000000 -1.275000000 1.135000000 -0.595000000

29 30 31 32 33 34 35

-1.305000000 -0.795714286 1.304285714 0.844285714 -0.145714286 -0.755714286 -0.445714286

36

-0.005714286

> print(anova\_model)

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

# Plot residuals against fitted values or group labels

plot(fitted(anova\_model), residuals, xlab = "Fitted Values", ylab = "Residuals", main = "Residuals vs. Fitted")

A graph with numbers and lines

Description automatically generated with medium confidence

**Interpretation**:

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)

Tukey multiple comparisons of means

95% family-wise confidence level

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

$`vinegar\_data$Site`

diff lwr upr p adj

London-Birmingham -0.84714286 -2.3307257 0.6364400 0.4765222

New York-Birmingham -0.91928571 -2.3557587 0.5171872 0.3632112

Paris-Birmingham 0.22142857 -1.2621543 1.7050114 0.9923747

Sydney-Birmingham 1.34142857 -0.1421543 2.8250114 0.0916315

New York-London -0.07214286 -1.5086158 1.3643301 0.9998934

Paris-London 1.06857143 -0.4150114 2.5521543 0.2519132

Sydney-London 2.18857143 0.7049886 3.6721543 0.0015038

Paris-New York 1.14071429 -0.2957587 2.5771872 0.1723996

Sydney-New York 2.26071429 0.8242413 3.6971872 0.0006834

Sydney-Paris 1.12000000 -0.3635829 2.6035829 0.2118486

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

1 2 3 4 5 6 7 8

0.722857143 1.022857143 -1.127142857 0.232857143 0.552857143 -0.247142857 -1.157142857 -1.074285714

9 10 11 12 13 14 15 16

-1.584285714 0.945714286 -0.284285714 1.165714286 0.865714286 -0.034285714 1.254285714 -0.875714286

17 18 19 20 21 22 23 24

0.174285714 -1.065714286 1.394285714 0.134285714 -1.015714286 0.605000000 0.215000000 1.195000000

25 26 27 28 29 30 31 32

0.025000000 -1.275000000 1.135000000 -0.595000000 -1.305000000 -0.795714286 1.304285714 0.844285714

33 34 35 36

-0.145714286 -0.755714286 -0.445714286 -0.005714286

**Key Findings**:

Some pairwise comparisons, such as Sydney-London and Sydney-New York, show substantial estimated mean differences with adjusted p-values (< 0.05), suggesting significant differences in pH levels between these pairs of locations.

Conversely, comparisons like New York-Birmingham and Paris-Birmingham have wide confidence intervals that include zero, indicating non-significant differences in mean pH levels between these locations.

**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.

1(e)

### Holm's Method:

# Assuming 'p\_values' contains the p-values obtained from the Tukey's test

p\_values <- c(0.4765222, 0.3632112, 0.9923747, 0.0916315, 0.9998934, 0.2519132, 0.0015038, 0.1723996, 0.0006834, 0.2118486)

# Apply Holm's correction

holm\_corrected <- p.adjust(p\_values, method = "holm")

# Display the corrected p-values

holm\_corrected

> holm\_corrected

[1] 1.0000000 1.0000000 1.0000000 0.7330520 1.0000000 1.0000000 0.0135342 1.0000000 0.0068340 1.0000000

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:

# Apply Bonferroni correction

bonferroni\_corrected <- p.adjust(p\_values, method = "bonferroni")

# Display the corrected p-values

bonferroni\_corrected

> bonferroni\_corrected

[1] 1.000000 1.000000 1.000000 0.916315 1.000000 1.000000 0.015038 1.000000 0.006834 1.000000

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.

1(f)

The experiment assessing balsamic vinegar acidity across factory locations reveals limited significant differences after rigorous corrections for multiple comparisons. To improve:

1. Increase Sample Size: Expanding the sample size may enhance statistical power, enabling better detection of subtle differences among locations.
2. Comprehensive Factors: Consider examining additional variables or factors influencing acidity (e.g., production methods, storage conditions) for a more nuanced understanding.
3. Replication: Replicate the study over time to validate findings and account for seasonal or temporal variations in acidity levels.
4. Diverse Metrics: Incorporate diverse acidity metrics beyond pH levels for a comprehensive assessment of vinegar quality.
5. Control Variables: Ensure rigorous control over environmental variables across factory locations to minimize external influences on acidity.
6. Collaborative Studies: Collaborate with experts in the field to introduce varied perspectives and refine experimental design for more robust conclusions.

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

Interpretation:

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’

2(b)

model <- aov(`Yield (mg)` ~ `Body Class` \* `Expression`, data = VenomYield)

# Perform the two-way ANOVA

summary(model)

> model <- aov(`Yield (mg)` ~ `Body Class` \* `Expression`, data = VenomYield)

> summary(model)

Df Sum Sq Mean Sq F value Pr(>F)

`Body Class` 1 68190 68190 49.160 3.17e-08 \*\*\*

Expression 1 53329 53329 38.446 3.75e-07 \*\*\*

`Body Class`:Expression 1 976 976 0.704 0.407

Residuals 36 49936 1387

---

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

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.

'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.

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.

2 (c )

# Running ANCOVA

ancova\_result <- lm(`Yield (mg)` ~ `Expression` \* `Body Length (cm)`, data = VenomYield)

summary(ancova\_result)

> summary(ancova\_result)

Call:

lm(formula = `Yield (mg)` ~ Expression \* `Body Length (cm)`,

data = VenomYield)

Residuals:

Min 1Q Median 3Q Max

-69.880 -15.209 -7.919 14.999 56.419

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 99.338 16.136 6.156 4.30e-07 \*\*\*

Expressionlow -53.046 23.213 -2.285 0.0283 \*

`Body Length (cm)` 40.678 4.908 8.287 7.28e-10 \*\*\*

Expressionlow:`Body Length (cm)` -7.467 6.990 -1.068 0.2925

---

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

Residual standard error: 28.25 on 36 degrees of freedom

Multiple R-squared: 0.8334, Adjusted R-squared: 0.8195

F-statistic: 60.01 on 3 and 36 DF, p-value: 4.406e-14

Here is the interpretation of the ANCOVA results:

Expression (High/Low):

The coefficient for 'Expressionlow' (-53.046) indicates that, on average, spiders with 'low' expression have 53.046 mg lower venom yield compared to those with 'high' expression, holding 'Body Length' constant.

This effect is statistically significant as indicated by the p-value (0.0283 < 0.05).

Body Length (cm):

The coefficient for 'Body Length (cm)' (40.678) signifies that for every 1 cm increase in body length, there is an average increase of 40.678 mg in venom yield, adjusting for 'Expression'.

This effect is statistically significant with a very low p-value (7.28e-10).

Interaction Term: Expressionlow x Body Length (cm):

The interaction term 'Expressionlow:Body Length (cm)' (-7.467) shows the change in the slope of 'Body Length' concerning 'Expression'.

The interaction effect is not statistically significant (p-value = 0.2925 > 0.05), indicating that the relationship between 'Body Length' and 'Yield (mg)' does not differ significantly between 'Expression' levels.

Overall Model:

The overall model is statistically significant (p-value: 4.406e-14), suggesting that the combined effects of 'Expression', 'Body Length', and their interaction can predict 'Yield (mg)' significantly well.

The Adjusted R-squared value (0.8195) indicates that around 81.95% of the variability in 'Yield (mg)' can be explained by 'Expression', 'Body Length', and their interaction in this model.

This ANCOVA indicates that both 'Expression' and 'Body Length' have significant effects on 'Yield (mg)' in funnel-web spiders. 'Expression' affects venom yield, and for every additional centimeter in 'Body Length', there is an increase in venom yield, while the interaction between 'Expression' and 'Body Length' does not significantly affect venom yield.

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 Justification: There exists a plausible theoretical relationship between body size and venom yield in spiders. Larger spiders potentially have larger venom glands or can produce more venom due to increased metabolic activity.

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)'.

2(e)

he 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)

# Define the hazard function

hazard <- function(t) {

return(1 - exp(-t))

}

# Survival function

survival\_function <- function(t) {

return(exp(-integrate(hazard, lower = 0, upper = t)$value))

}

# Failure probability density function

failure\_density <- function(t) {

return(hazard(t) \* survival\_function(t))

}

# Create a sequence of time points

time\_points <- seq(0, 10, by = 0.1)

# Calculate S(t) and f(t) for the time points

survival\_values <- sapply(time\_points, survival\_function)

failure\_density\_values <- sapply(time\_points, failure\_density)

# Plotting the survival function S(t)

plot(time\_points, survival\_values, type = 'l', xlab = 'Time (years)', ylab = 'Survival Probability', main = 'Survival Function S(t)')

# Plotting the failure probability density function f(t)

plot(time\_points, failure\_density\_values, type = 'l', xlab = 'Time (years)', ylab = 'Failure Density', main = 'Failure Probability Density Function f(t)')

A graph of a function

Description automatically generated

A graph of a function

Description automatically generated

3(b)

# Failure and censoring data

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

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

# Calculate the total number of units (population)

total\_units <- 100

# Calculate the remaining units at risk (accounts for censoring)

units\_at\_risk <- total\_units - cumsum(censoring)

# Calculate the survival probability at each time point

survival\_prob <- cumprod(1 - failures / units\_at\_risk)

# Calculate the hazard function

hazard <- failures / units\_at\_risk

# Calculate the failure probability density function

failure\_density <- c(0, diff(failures) / units\_at\_risk[-length(units\_at\_risk)])

# Plot the survival function S(t)

plot(failures, survival\_prob, type = "s", xlab = "Time (years)", ylab = "Survival Probability", main = "Survival Function S(t)")

# Plot the hazard function h(t)

plot(failures, hazard, type = "s", xlab = "Time (years)", ylab = "Hazard Rate", main = "Hazard Function h(t)")

# Plot the failure probability density function f(t)

plot(failures, failure\_density, type = "s", xlab = "Time (years)", ylab = "Failure Density", main = "Failure Probability Density Function f(t)")

A graph with a number of numbers

Description automatically generated with medium confidence

A graph with lines on it

Description automatically generated

The model in part (a) suggests a hazard function of which is a common form used to model increasing hazard rates over time. This hazard function implies that the risk of failure increases with time but eventually levels off asymptotically.

By examining the derived survival function *S*(*t*) and failure probability density function *f*(*t*) from this hazard function, one can assess its accuracy by comparing the resulting functions to expected behaviours for survival and failure rates.

**Survival Function :** The Survival function decreases over time but never reaches zero, suggesting decreasing probability of survival as time progresses

**Failure Probability Density Function :** The function indicates an increasing trend of density of failures as time progresses