

Assignment 4

STAT3373

Isaac Shoebottom

Oct 16th, 2025

```
library(tidyverse)
library(knitr)
```

Question 1

a)

```
# Create the dataset
data <- tibble(
  Farm = factor(1:4),
  Fert1 = c(48, 45, 52, 44),
  Fert2 = c(55, 50, 58, 49),
  Fert3 = c(52, 49, 55, 47)
)

# Convert to long format
long_data <- data %>%
  pivot_longer(
    cols = starts_with("Fert"),
    names_to = "Fertilizer",
    values_to = "Yield"
  ) %>%
  mutate(Fertilizer = factor(Fertilizer))

kable(long_data, caption = "Yield Data (Bushels per Acre)")
```

Table 1: Yield Data (Bushels per Acre)

Farm	Fertilizer	Yield
1	Fert1	48
1	Fert2	55
1	Fert3	52
2	Fert1	45
2	Fert2	50
2	Fert3	49
3	Fert1	52
3	Fert2	58
3	Fert3	55
4	Fert1	44

Farm	Fertilizer	Yield
4	Fert2	49
4	Fert3	47

b)

Model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

```
anova_model <- aov(Yield ~ Fertilizer + Farm, data = long_data)
anova_table <- summary(anova_model)
anova_table
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## Fertilizer  2  67.17   33.58    93.0 3.05e-05 ***
## Farm       3 127.33   42.44   117.5 1.02e-05 ***
## Residuals  6   2.17    0.36
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusions:

- Fertilizer effect is significant ($p < 0.05$)
- Farm (block) effect is also significant

c)

```
tukey_results <- TukeyHSD(anova_model, "Fertilizer")
tukey_results

##      Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = Yield ~ Fertilizer + Farm, data = long_data)
##
## $Fertilizer
##           diff          lwr          upr          p adj
## Fert2-Fert1  5.75  4.446234  7.0537659  0.0000248
## Fert3-Fert1  3.50  2.196234  4.8037659  0.0004243
## Fert3-Fert2 -2.25 -3.553766 -0.9462341  0.0044225
```

Results:

- Fertilizer 2 produces the highest yields
- All fertilizer pairs differ significantly
- Ordering of mean yields: Fert 2 > Fert 3 > Fert 1

Final Conclusion ($\alpha = 0.05$)

- There is strong statistical evidence that fertilizer type affects yield.
- Blocking by farm was appropriate and reduced error variability.
- Fertilizer 2 is the most effective option based on yield.

Question 2

a)

```
drug_data <- data.frame(
  patient = factor(rep(1:5, each = 3)),
  drug = factor(rep(c("A", "B", "C"), times = 5)),
  response_time = c(
    12, 10, 15, # Patient 1
    14, 11, 16, # Patient 2
    10, 8, 13,  # Patient 3
    13, 10, 14, # Patient 4
    11, 9, 14   # Patient 5
  )
)

kable(drug_data, caption = "Drug Trial Response Times (seconds)")
```

Table 2: Drug Trial Response Times (seconds)

patient	drug	response_time
1	A	12
1	B	10
1	C	15
2	A	14
2	B	11
2	C	16
3	A	10
3	B	8
3	C	13
4	A	13
4	B	10
4	C	14
5	A	11
5	B	9
5	C	14

b)

Model:

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}$$

```
anova_model <- aov(response_time ~ drug + patient, data = drug_data)
summary(anova_model)
```

```
##           Df Sum Sq Mean Sq F value    Pr(>F)
## drug        2   57.60   28.800   132.92 7.28e-07 ***
## patient     4   18.67    4.667    21.54 0.000243 ***
## Residuals   8    1.73    0.217
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

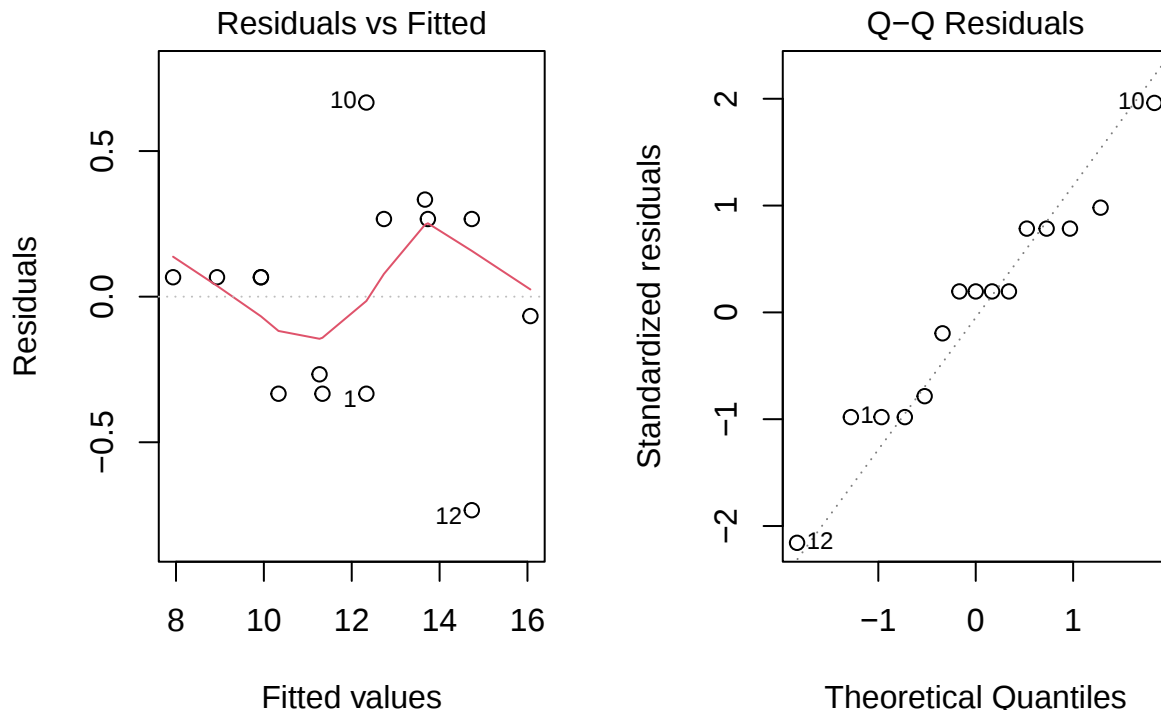
Decision (alpha = 0.05):

- Drug effect is significant
- Patient (block) effect is significant

c)

Residual Diagnostics

```
par(mfrow = c(1, 2))
plot(anova_model, which = 1) # Residuals vs Fitted
plot(anova_model, which = 2) # Normal Q-Q
```



```
par(mfrow = c(1, 1))
```

Formal Tests

```
# Normality of residuals
```

```
shapiro.test(residuals(anova_model))
```

```
##
```

```
## Shapiro-Wilk normality test
```

```
##
```

```
## data: residuals(anova_model)
```

```
## W = 0.95664, p-value = 0.6342
```

```
# Homogeneity of variance
```

```
bartlett.test(response_time ~ drug, data = drug_data)
```

```
##
```

```
## Bartlett test of homogeneity of variances
```

```
##
```

```
## data: response_time by drug
```

```
## Bartlett's K-squared = 0.54312, df = 2, p-value = 0.7622
```

Results:

- Residuals are approximately normally distributed
- Variances across drug groups are homogeneous

d)

Multiple Comparisons

```
tukey_results <- TukeyHSD(anova_model, "drug")
tukey_results

##    Tukey multiple comparisons of means
##      95% family-wise confidence level
##
## Fit: aov(formula = response_time ~ drug + patient, data = drug_data)
##
## $drug
##      diff      lwr      upr    p adj
## B-A -2.4 -3.241209 -1.558791 9.92e-05
## C-A  2.4  1.558791  3.241209 9.92e-05
## C-B  4.8  3.958791  5.641209 5.00e-07
```

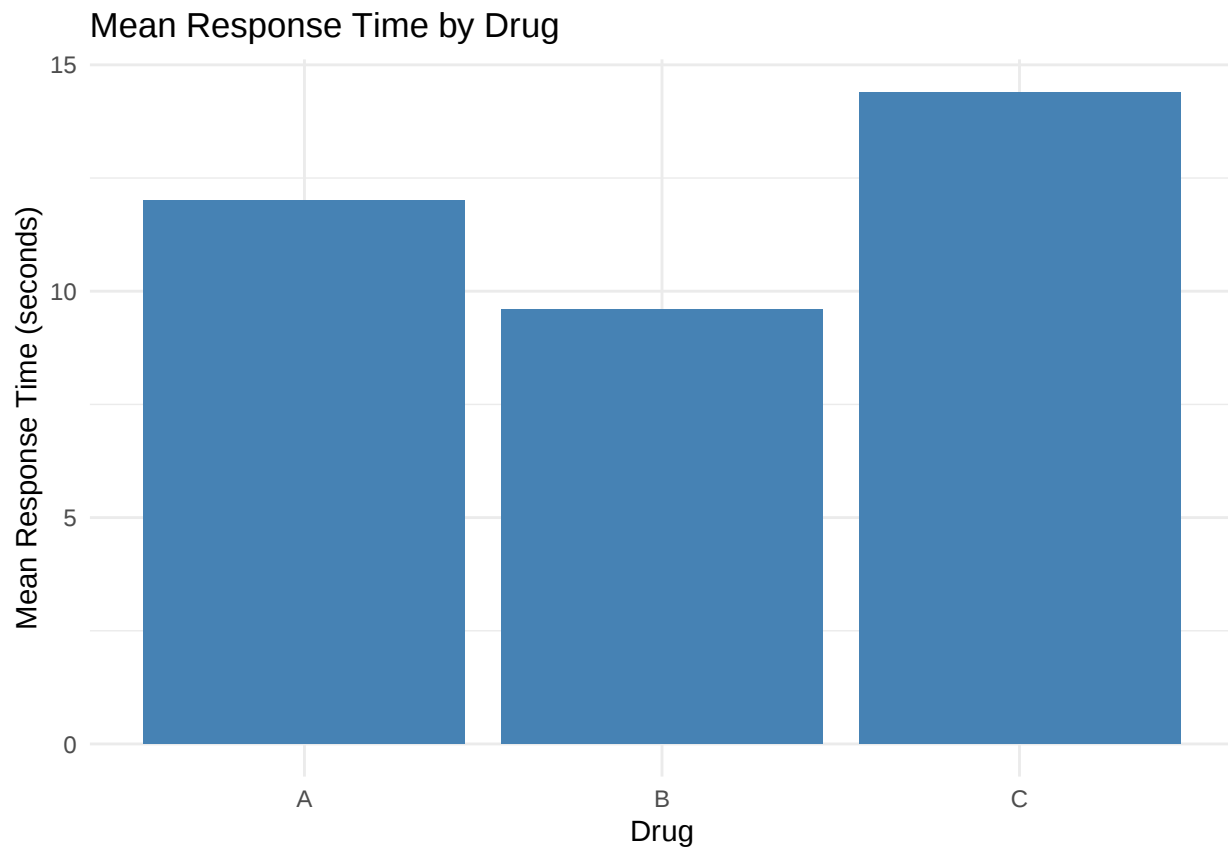
Results:

- All drug pairs differ significantly
- Ordering of mean response times: Drug B < Drug A < Drug C

e)

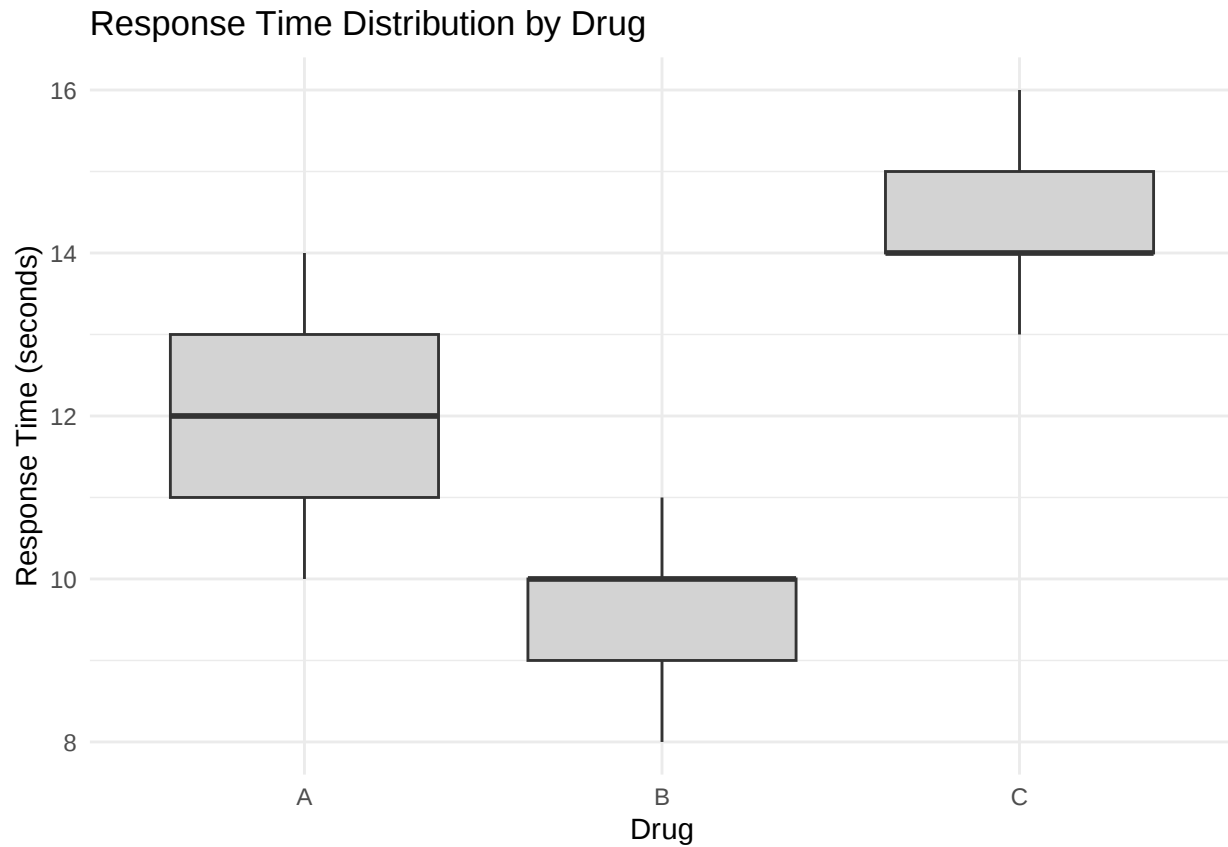
Mean Response Times by Drug

```
drug_data %>%
  group_by(drug) %>%
  summarise(mean_time = mean(response_time)) %>%
  ggplot(aes(x = drug, y = mean_time)) +
  geom_col(fill = "steelblue") +
  labs(
    title = "Mean Response Time by Drug",
    x = "Drug",
    y = "Mean Response Time (seconds)"
  ) +
  theme_minimal()
```



Boxplot by Drug

```
ggplot(drug_data, aes(x = drug, y = response_time)) +  
  geom_boxplot(fill = "lightgray") +  
  labs(  
    title = "Response Time Distribution by Drug",  
    x = "Drug",  
    y = "Response Time (seconds)"  
  ) +  
  theme_minimal()
```



f)

Conclusion:

At the 5% significance level, there is strong evidence that drug formulation affects patient response time. Blocking by patient was effective and significantly reduced unexplained variability. Post-hoc analysis using Tukey's HSD showed that all three drugs differ significantly, with Drug B producing the fastest (best) response times, followed by Drug A, and then Drug C.