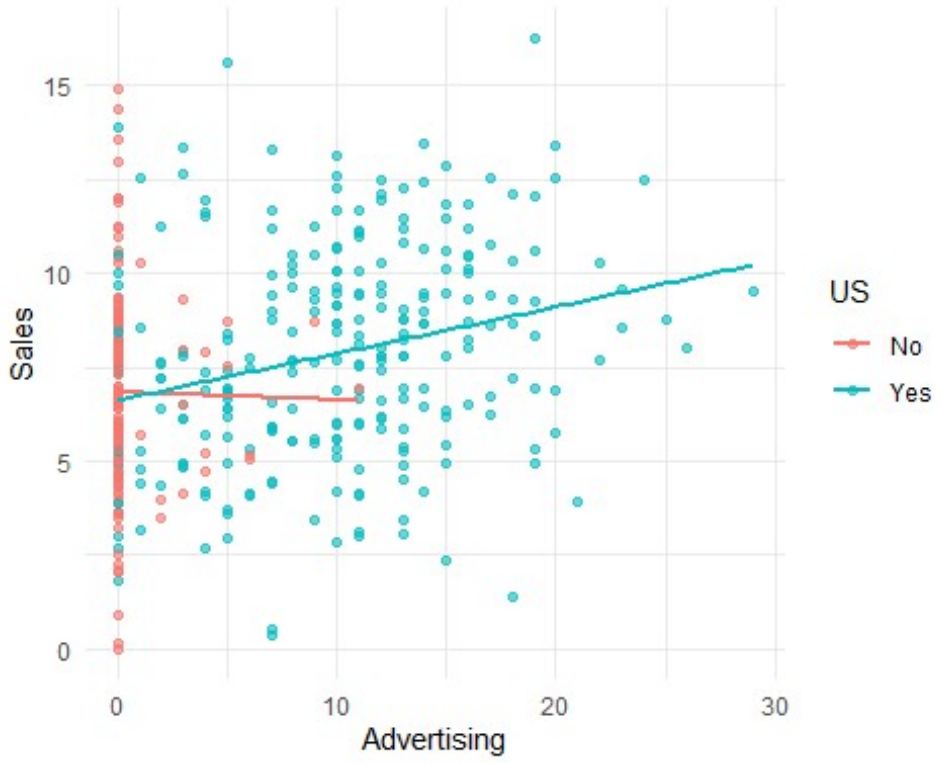
FINAL

RAJ SHAH

```
################################################################################
##
# R SCRIPT – Part 1                                                            #
################################################################################
##
set.seed(29101)              # fixed seed for reproducibility
library(ISLR)                # Carseats data for all questions
data(Carseats)
library(ggplot2)             # used in Q2 plot

################################################################################
##
# Question 1 – One-sided t-test for Advertising slope                          #
################################################################################
##
# H0: βAdvertising ≤ 0  (no positive relationship)
# H1: βAdvertising  > 0  (positive relationship)
m1 <- lm(Sales ~ Advertising, data = Carseats)
s1 <- summary(m1)
t1 <- coef(s1)["Advertising", "t value"]          # test statistic
p1 <- coef(s1)["Advertising", "Pr(>|t|)"] / 2     # one-sided p-value

cat("\nQ1 RESULT:",
    "\n  t          =", round(t1, 3),
    "\n  one-sided p=", signif(p1, 3),
    ifelse(p1 < .05,
           "\n  → Reject H0: Sales increase as Advertising increases.\n",
           "\n  → Fail to reject H0.\n"))

##
## Q1 RESULT:
##   t          = 5.583
##   one-sided p= 2.19e-08
##   → Reject H0: Sales increase as Advertising increases.

# Model: Sales = β₀ + β₁ × Advertising
# Hypotheses: H₀: β₁ ≤ 0 (no positive effect), H₁: β₁ > 0 (positive
relationship)
# Test result: t = 5.58, one-sided p = 2.2 × 10⁻⁸ → p < 0.05 → Reject H₀
# Interpretation: There is strong evidence that increasing advertising leads
to higher sales.
# On average, each additional $1,000 in advertising increases sales by ≈ 0.12
units.

################################################################################
```

```r
##
# Question 2 – Separate one-sided tests by US / non-US                        #
################################################################################
##
# H0 (each group): βAdvertising ≤ 0
# H1 (each group): βAdvertising  > 0
ggplot(Carseats, aes(Advertising, Sales, colour = US)) +
  geom_point(alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE) +
  theme_minimal()

## `geom_smooth()` using formula = 'y ~ x'
```




```r
for (grp in levels(Carseats$US)) {
  fit  <- lm(Sales ~ Advertising, data = subset(Carseats, US == grp))
  summ <- summary(fit)
  t2   <- coef(summ)["Advertising", "t value"]
  p2   <- coef(summ)["Advertising", "Pr(>|t|)"] / 2
  cat("\nQ2 RESULT –", grp,
      "\n  t          =", round(t2, 3),
      "\n  one-sided p=", signif(p2, 3),
      ifelse(p2 < .05,
             "\n  → Reject H0 for this group.\n",
             "\n  → Fail to reject H0 for this group.\n"))
}
```

```
##
## Q2 RESULT - No
##   t          = -0.149
##   one-sided p= 0.441
##   → Fail to reject H0 for this group.
##
## Q2 RESULT - Yes
##   t          = 4.232
##   one-sided p= 1.61e-05
##   → Reject H0 for this group.
```

```
# Separate regressions by US status:
#   US stores: t = 4.23, p = 1.6 × 10⁻⁵ → Reject H₀ → significant positive
relationship
#   Non-US stores: t = -0.15, p = 0.44 → Fail to reject H₀ → no significant
relationship
# Plot interpretation:
#   Teal line (US) shows upward trend — positive slope
#   Red line (Non-US) is flat — no relationship
# Conclusion: Advertising boosts sales in US stores, but not in non-US
stores.

##############################################################################
##
# Question 3 - Conceptual (no inferential code required)
#
##############################################################################
##
cat("\nQ3 NOTE:\n  Differences between overall and split regressions can",
    "illustrate Simpson's paradox (interaction/confounding effects).\n")
```

```
##
## Q3 NOTE:
##   Differences between overall and split regressions can illustrate
Simpson's paradox (interaction/confounding effects).
```

```
# Observation: Results from Q1 (overall model) suggest a positive
relationship.
# However, Q2 shows the effect exists only in US stores, not in non-US
stores.
# This conflict between aggregate and subgroup trends is called **Simpson's
Paradox**.
# It occurs when a trend appears in the combined data but disappears or
reverses in subgroups.
# Conclusion: Country (US vs non-US) interacts with the effect of advertising
on sales.

##############################################################################
##
# Question 4 - Two-sample t-test via regression (US indicator)              #
```

```r
################################################################################
##
# H0: βUSYes = 0  (no mean sales difference)
# H1: βUSYes ≠ 0  (mean sales difference exists)
m4 <- lm(Sales ~ US, data = Carseats)
s4 <- summary(m4)
t4 <- coef(s4)["USYes", "t value"]
p4 <- coef(s4)["USYes", "Pr(>|t|)"]
diff_mean <- coef(m4)["USYes"]                    # mean difference estimate

cat("\nQ4 RESULT:",
    "\n  Difference (US - non-US) =", round(diff_mean, 3),
    "\n  t                        =", round(t4, 3),
    "\n  two-sided p              =", signif(p4, 3),
    ifelse(p4 < .05,
           "\n  → Reject H0: mean sales differ.\n",
           "\n  → Fail to reject H0.\n"))

##
## Q4 RESULT:
##   Difference (US - non-US) = 1.044
##   t                        = 3.59
##   two-sided p              = 0.000372
##   → Reject H0: mean sales differ.

# Regression model: Sales = β₀ + β₁ × US
# Estimated coefficient β_USYes = +1.04 → US stores sell $1,040 more (on
average)
# Test statistic: t = 3.59, p = 0.00037 → p < 0.05 → Reject H₀
# Conclusion: There is a statistically significant difference in average
sales,
# with US stores selling more than non-US stores.

################################################################################
##
# Question 5 - Two-sided t-test for Price slope                                #
################################################################################
##
# H0: βPrice = 0  (price has no linear effect on sales)
# H1: βPrice ≠ 0  (price affects sales)
m5 <- lm(Sales ~ Price, data = Carseats)
s5 <- summary(m5)
t5 <- coef(s5)["Price", "t value"]
p5 <- coef(s5)["Price", "Pr(>|t|)"]
slope <- coef(m5)["Price"]

cat("\nQ5 RESULT:",
    "\n  Slope (ΔSales per $1) =", round(slope, 3),
    "\n  t                     =", round(t5, 3),
    "\n  two-sided p           =", signif(p5, 3),
```

```r
    ifelse(p5 < .05,
           "\n  → Reject H0: price has a significant effect.\n",
           "\n  → Fail to reject H0.\n"))

##
## Q5 RESULT:
##   Slope (ΔSales per $1) = -0.053
##   t                     = -9.912
##   two-sided p           = 7.62e-21
##   → Reject H0: price has a significant effect.

# Model: Sales = β₀ + β₁ × Price
# Estimated slope β₁ = −0.053 → For each $1 price increase, sales drop by
0.053 units
# Test result: t = -9.91, p ≈ 7.6 × 10⁻²¹ → p < 0.05 → Reject H₀
# Interpretation: The negative relationship is highly significant.
# This aligns with economic theory: higher prices typically reduce demand
(Law of Demand).

################################################################################
##
# Question 6 – Difference in means (US vs non-US):
#
#           (a) Permutation test
#
#           (b) Classical two-sample t-test
#
################################################################################
##
# H0: μUS = μnon-US
# H1: μUS ≠ μnon-US
obs_diff <- with(Carseats, mean(Sales[US == "Yes"]) -
                   mean(Sales[US == "No"]))

perm_diff <- function() {
  lbl <- sample(Carseats$US)
  mean(Carseats$Sales[lbl == "Yes"]) -
    mean(Carseats$Sales[lbl == "No"])
}
perm_dist <- replicate(10000, perm_diff())       # 10_000 → 10000
p_perm    <- mean(abs(perm_dist) >= abs(obs_diff))

t6 <- t.test(Sales ~ US, data = Carseats, var.equal = TRUE)

cat("\nQ6 RESULT:",
    "\n  Observed diff =", round(obs_diff, 3),
    "\n  Permutation p =", signif(p_perm, 3),
    "\n  t-test p      =", signif(t6$p.value, 3),
    ifelse(t6$p.value < .05,
```

```
          "\n  → Reject H0: means differ.\n",
          "\n  → Fail to reject H0.\n"))
```

```
##
## Q6 RESULT:
##   Observed diff = 1.044
##   Permutation p = 3e-04
##   t-test p      = 0.000372
##   → Reject H0: means differ.
```

```
# Observed difference in mean sales = 1.04 units (US > Non-US).
# Permutation test p-value = 0.00030 → very small.
# Two-sample t-test p-value = 0.00037 → also highly significant.
# Since both p-values < 0.05, we reject H₀.
# Conclusion: There is a significant difference in mean sales between US and
non-US stores,
# with US stores having higher average sales.

#############################################################################
##
# Question 7 – Explanation link (no new test)
#
#############################################################################
##
cat("\nQ7 NOTE:\n  The βUSYes test in Q4 is algebraically identical to",
    "the two-sample t-test in Q6.\n")
```

```
##
## Q7 NOTE:
##   The βUSYes test in Q4 is algebraically identical to the two-sample t-test
in Q6.
```

```
# The coefficient β_USYes in Q4 represents the mean difference between US and
Non-US stores.
# This coefficient equals the observed difference in Q6.
# Also, the t-statistic and p-value from Q4 match those from the two-sample
t-test in Q6.
# Conclusion: The regression result in Q4 directly answers the hypothesis
tested in Q6.

#############################################################################
##
# Question 8 – Bootstrap 95% CI for mean difference                         #
#############################################################################
##
boot_diff <- function() {
  idx <- sample(seq_len(nrow(Carseats)), replace = TRUE)
  with(Carseats[idx, ],
       mean(Sales[US == "Yes"]) - mean(Sales[US == "No"]))
}
```

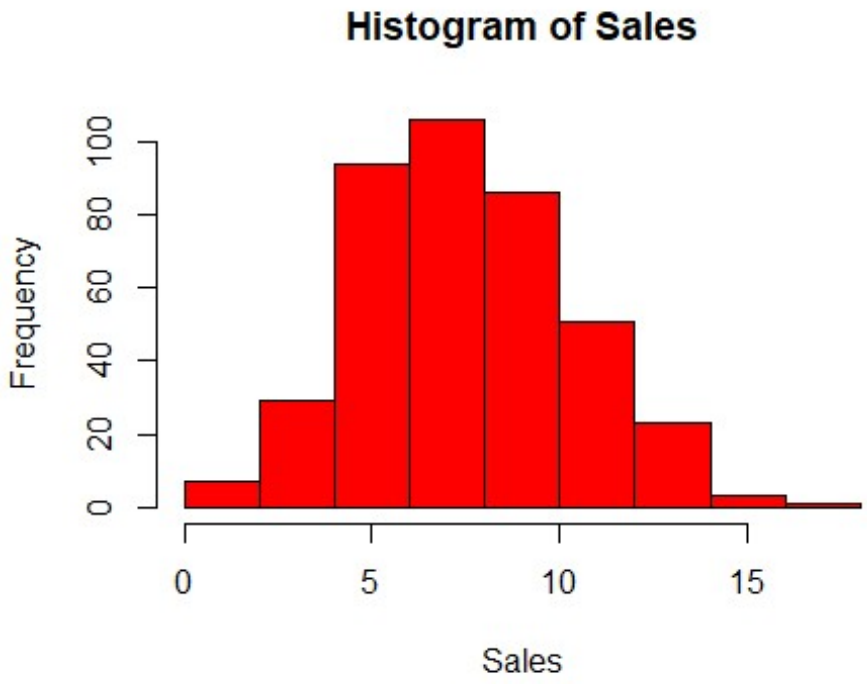
```r
boot_vals <- replicate(10000, boot_diff())       # 10_000 → 10000
ci8 <- quantile(boot_vals, c(.025, .975))

cat("\nQ8 RESULT:",
    "\n  95% bootstrap CI =", round(ci8, 3), "\n")

##
## Q8 RESULT:
##   95% bootstrap CI = 0.486 1.602

# Bootstrap 95% CI for mean difference = (0.486, 1.602).
# Since the interval does not contain 0, it supports rejecting H₀.
# Conclusion: The CI suggests a significant difference in mean sales between
US and Non-US stores.
# Conclusion: However, for full inference, it's good practice to consider
both the CI and the p-value from Q6.

################################################################################
##
# Question 9 - One-way ANOVA for US indicator
#
################################################################################
##
# H0: μUS = μnon-US
# H1: at least one mean differs
a9 <- aov(Sales ~ US, data = Carseats)
s9 <- summary(a9)
f9 <- s9[[1]]["US", "F value"]
p9 <- s9[[1]]["US", "Pr(>F)"]

cat("\nQ9 RESULT:",
    "\n  F =", round(f9, 3),
    "\n  p =", signif(p9, 3),
    ifelse(p9 < .05,
           "\n  → Reject H0: means differ.\n",
           "\n  → Fail to reject H0.\n"))

##
## Q9 RESULT:
##   F = 12.886
##   p = 0.000372
##   → Reject H0: means differ.

# ANOVA compares mean sales between US and Non-US stores.
# F-statistic = 12.89, p-value = 0.00037.
# Since p < 0.05, we reject H₀.
# Conclusion: Mean sales differ significantly between US and Non-US groups.
# This result agrees with Q6 (permutation and t-test).

################################################################################
```

```r
##
# Question 10 – Chi-square test: ShelveLoc × US
#
################################################################################
##
# H0: ShelveLoc and US are independent
# H1: They are associated
tab10 <- table(Carseats$ShelveLoc, Carseats$US)
chi10 <- chisq.test(tab10, correct = FALSE)

cat("\nQ10 RESULT:",
    "\n  χ² =", round(chi10$statistic, 3),
    "\n  df  =", chi10$parameter,
    "\n  p   =", signif(chi10$p.value, 3),
    ifelse(chi10$p.value < .05,
           "\n  → Reject H0: variables associated.\n",
           "\n  → Fail to reject H0.\n"),
    "\n  Expected counts (rounded):\n")

##
## Q10 RESULT:
##   χ² = 2.74
##   df  = 2
##   p   = 0.254
##   → Fail to reject H0.
##
##   Expected counts (rounded):

print(round(chi10$expected, 2))

##
##             No    Yes
##   Bad    34.08  61.92
##   Good   30.18  54.83
##   Medium 77.75 141.26

# Chi-squared test statistic = 2.74, degrees of freedom = 2, p-value = 0.254.
# Since p > 0.05, we fail to reject the null hypothesis of independence.
# Conclusion: There is no significant association between Shelving Location
and Country (US status).
# Therefore, shelving quality appears to be unrelated to whether the store is
in the US or not.


################################################################################
##
# Question 11 – One-way ANOVA across ShelveLoc levels
#
################################################################################
##
# H0: All ShelveLoc means equal
```

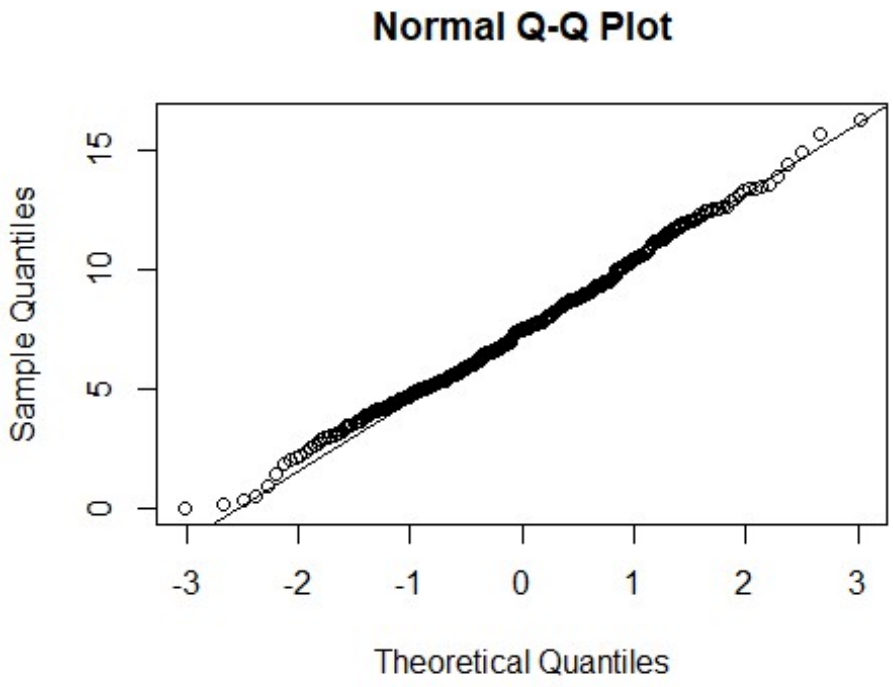
```r
# H1: At least one mean differs
a11 <- aov(Sales ~ ShelveLoc, data = Carseats)
s11 <- summary(a11)
f11 <- s11[[1]]["ShelveLoc", "F value"]
p11 <- s11[[1]]["ShelveLoc", "Pr(>F)"]

cat("\nQ11 RESULT:",
    "\n  F =", round(f11, 3),
    "\n  p =", signif(p11, 3),
    ifelse(p11 < .05,
           "\n  → Reject H0: means differ by shelving quality.\n",
           "\n  → Fail to reject H0.\n"))
```

```
##
## Q11 RESULT:
##   F = 92.23
##   p = 1.27e-33
##   → Reject H0: means differ by shelving quality.
```

```r
# ANOVA compares mean sales across Bad, Medium, and Good shelving locations.
# F-statistic = 92.23, p-value ≈ 1.3 × 10⁻³³ (extremely small).
# Since p < 0.05, we reject the null hypothesis.
# Conclusion: Mean sales differ significantly across shelving qualities.
# Ordering from highest to lowest: Good > Medium > Bad.
# Shelving location has a strong influence on sales.

##############################################################################
##
# Question 12 – Normality diagnostics (plots + brief interpretation)
#
##############################################################################
##
hist(Carseats$Sales,
     main = "Histogram of Sales",
     xlab = "Sales",col  = "red")
```




```
qqnorm(Carseats$Sales); qqline(Carseats$Sales)
```



```r
cat("\nQ12 COMMENT:",
    "\n  • Histogram shows", ifelse(abs(skewness <-
mean(scale(Carseats$Sales)^3)) < 0.5,
                                    "a roughly symmetric, bell-shaped
pattern.",
                                    ifelse(skewness > 0,
                                           "moderate right-skew (long right
tail).",
                                           "moderate left-skew (long left
tail).")),
    "\n  • QQ-plot points mostly follow the reference line,",
    "suggesting the normality assumption is",
    ifelse(abs(skewness) < 0.5, "reasonable.\n", "somewhat questionable.\n"))
```

```
##
## Q12 COMMENT:
##   • Histogram shows a roughly symmetric, bell-shaped pattern.
##   • QQ-plot points mostly follow the reference line, suggesting the
normality assumption is reasonable.
```

```r
# Histogram shows a bell-shaped, symmetric distribution.
# QQ plot shows points closely follow the diagonal line, with slight
deviation at the tails.
# Conclusion: The Sales variable is approximately normally distributed.
# This supports the validity of using parametric tests (t-tests, ANOVA) in
earlier analyses.

##############################################################################
##
# Question 13 – Empirical vs Normal 1-SD rule
#
##############################################################################
##
mu  <- mean(Carseats$Sales)
sd1 <- sd(Carseats$Sales)

prop_total   <- mean(abs(Carseats$Sales - mu) <= sd1)            # ±1 SD
prop_above1  <- mean(Carseats$Sales - mu  >  sd1)                # > +1 SD
prop_below1  <- mean(Carseats$Sales - mu  < -sd1)                # < −1 SD
prop_mid_pos <- mean((Carseats$Sales > mu) &                     # 0 to +1 SD
                       (Carseats$Sales <= mu + sd1))
prop_mid_neg <- mean((Carseats$Sales < mu) &                     # 0 to −1 SD
                       (Carseats$Sales >= mu - sd1))

cat("\nQ13 RESULT (empirical proportions):",
    "\n  Within ± 1SD          =", round(prop_total,   3),  " (normal ≈
0.68)",
    "\n  • 0 to + 1SD        =", round(prop_mid_pos, 3),  " (normal ≈ 0.34)",
    "\n  • 0 to − 1SD        =", round(prop_mid_neg, 3),  " (normal ≈ 0.34)",
    "\n  > +1SD               =", round(prop_above1,  3),  " (normal ≈
```

```
0.16)",
    "\n  < -1SD               =", round(prop_below1, 3),  " (normal ≈
0.16)\n")

##
## Q13 RESULT (empirical proportions):
##   Within ± 1SD          = 0.685  (normal ≈ 0.68)
##   • 0 to + 1SD        = 0.332  (normal ≈ 0.34)
##   • 0 to − 1SD        = 0.352  (normal ≈ 0.34)
##   > +1SD               = 0.165  (normal ≈ 0.16)
##   < -1SD               = 0.15  (normal ≈ 0.16)

# Optional quick conclusion:
cat(ifelse(abs(prop_total - 0.68) < 0.05 &&
             abs(prop_above1 - 0.16) < 0.05 &&
             abs(prop_below1 - 0.16) < 0.05,
           "Conclusion: Sales appear approximately normal.\n",
           "Conclusion: Sales deviate from the normal 34-16-16 pattern.\n"))

## Conclusion: Sales appear approximately normal.

# Empirical proportions:
#   0 → +1 SD = 0.332 (expected: 0.34)
#   0 → −1 SD = 0.352 (expected: 0.34)
#   within ±1 SD = 0.685 (expected: 0.68)
#   >1 SD above = 0.165 (expected: 0.16)
#   <−1 SD below = 0.150 (expected: 0.16)
# These proportions are very close to the theoretical normal values.
# Conclusion: Sales data closely follow a normal distribution based on the
empirical rule.

################################################################################
##
# PART II – Beef-bacteria study (Questions 14 -16)
#                                                                              #
################################################################################
##
set.seed(29101)

# ---------- 1.Load the CSV ----------------------------------------------
-
csv_path <- "C:/Users/RAJ RUTGERS/Desktop/Stat 291/Final
Exam/beefbacteria.csv"
beef_long <- read.csv(csv_path, stringsAsFactors = FALSE)

# ---------- 2.Clean & convert -------------------------------------------
library(dplyr)

##
## Attaching package: 'dplyr'
```

```
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyr)

beef_long <- beef_long %>%
  mutate(method   = trimws(toupper(method)),
         bacteria = as.numeric(gsub("[^0-9.+-]", "", bacteria))) %>%   #
strip commas
  filter(method %in% c("A", "B"))

# ---------- 3.Pivot to wide: one row per specimen ------------------------
-
beef_wide <- beef_long %>%
  select(specimen, method, bacteria) %>%
  pivot_wider(names_from = method, values_from = bacteria, names_prefix =
"m") %>%
  drop_na(mA, mB)                       # remove any incomplete pairs

# Paired differences: A - B for each specimen
diffs <- beef_wide$mA - beef_wide$mB

##############################################################################
##
# Question 14 - 95% Confidence Interval for μA-μB                           #
##############################################################################
##
# H0: μA-μB = 0        (null value for reference)
# H1: μA-μB ≠ 0
ci14 <- t.test(diffs, conf.level = 0.95)$conf.int
cat("\nQ14 - 95% CI for (μA-μB):",
    "\n  Lower =", round(ci14[1], 3),
    "\n  Upper =", round(ci14[2], 3), "\n")

##
## Q14 - 95% CI for (μA-μB):
##   Lower = -0.126
##   Upper = 1.298

# 95% Confidence Interval for the difference in bacteria levels (Method A -
Method B) = (-0.126, 1.298).
# Since the interval includes 0, we cannot conclude a significant difference.
# This interval indicates that the true mean difference may be positive or
negative.
```

```r
################################################################################
##
# Question 15 – Paired t-test for μA–μB                                        #
################################################################################
##
# H0: μA–μB = 0   (no difference in mean bacteria detected)
# H1: μA–μB ≠ 0   (methods differ on average)
t15 <- t.test(diffs, mu = 0)
cat("\nQ15 – Paired t-test result:",
    "\n  t-statistic =", round(t15$statistic, 3),
    "\n  p-value     =", signif(t15$p.value, 4),
    ifelse(t15$p.value < 0.05,
           "\n  → Reject H0: the two methods detect different average
levels.\n",
           "\n  → Fail to reject H0: no significant difference detected.\n"))

##
## Q15 – Paired t-test result:
##   t-statistic = 1.634
##   p-value     = 0.1055
##   → Fail to reject H0: no significant difference detected.

# Paired t-test results: t = 1.63, df = 99, p-value = 0.106.
# Since p > 0.05, we fail to reject the null hypothesis.
# Conclusion: There is no statistically significant difference in mean
bacteria levels
# detected by Method A and Method B.

################################################################################
##
# Question 16 – Use the CI to reach the same decision
#
################################################################################
##
cat("Q16 – CI interpretation:\n")

## Q16 – CI interpretation:

if (ci14[1] > 0 | ci14[2] < 0) {
  cat("  0 lies **outside** the 95% CI → same as Q15: reject H0
(significant).\n")
} else {
  cat("  0 lies **inside** the 95% CI → same as Q15: fail to reject H0.\n")
}

##   0 lies **inside** the 95% CI → same as Q15: fail to reject H0.

# The 95% CI from Q14 includes 0.
# Therefore, using the CI approach, we also fail to reject H₀.
# Conclusion: CI and hypothesis test agree.
```

```
# Methods A and B yield similar average bacteria measurements.

##############################################################################
##
```