140.61.HW8.Delahanty.Jeremy

Jeremy Delahanty

2024-04-17

```
# Install necessary library
library(devtools, quietly = TRUE)
devtools::install_github("bllfrg/SPH.140.615", quiet = TRUE)
library(SPH.140.615)
```

1

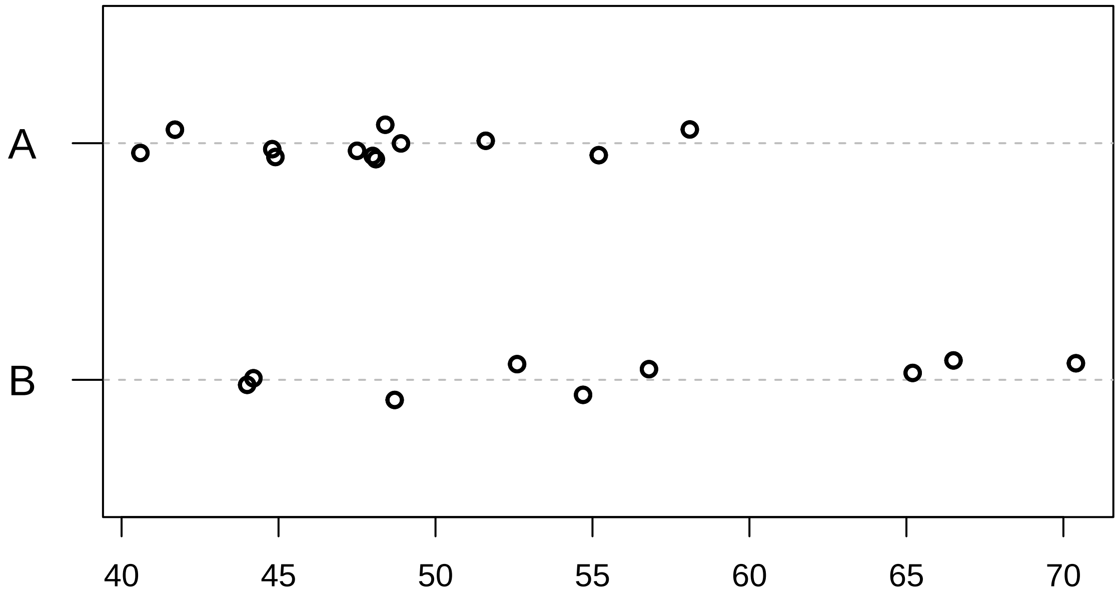
Consider the data on the treatment response of 12 mice from strain A and 9 from strain B.

a) Plot the data

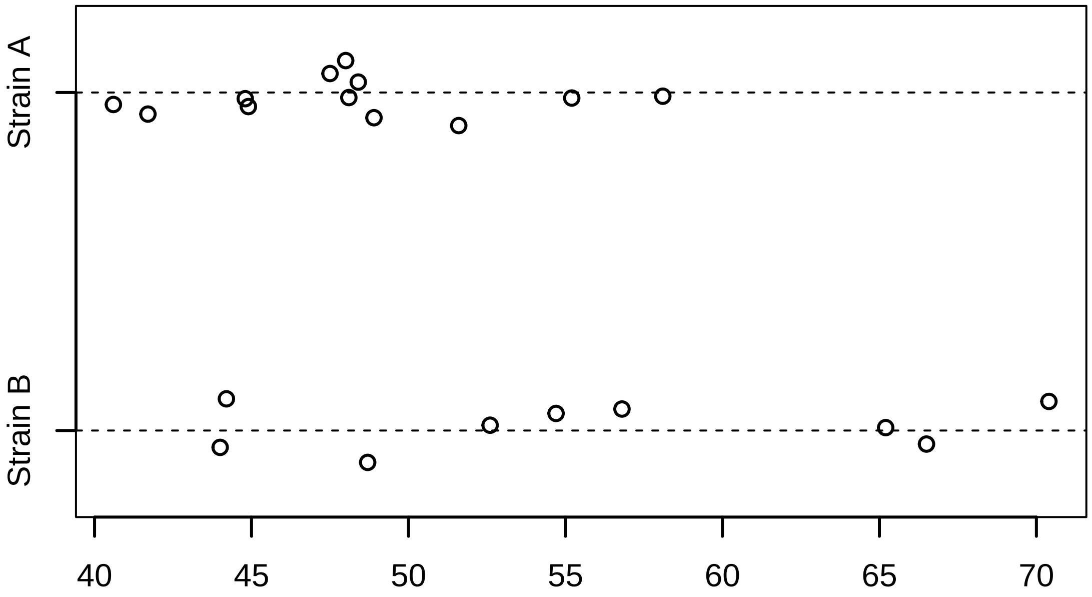
We can first perform plotting with our class `dot.plot` code which looks pretty similar to the strip chart. It's worth comparing the two since they look quite similar.

```
# Define the data arrays
strain_a <- c(55.2, 58.1, 41.7, 44.9, 44.8, 48.9, 47.5, 48.1, 48.4, 51.6, 40.6, 48.0)
strain_b <- c(48.7, 52.6, 65.2, 70.4, 44.2, 54.7, 44.0, 66.5, 56.8)

# Plot using class library
dot.plot(strain_a, strain_b, includeCI = FALSE)
```



```
# Compare to the stripchart function
stripchart(list(strain_b, strain_a), method = "jitter", vertical = FALSE,
           pch = 1, lwd = 1.5, group.names = c("Strain B", "Strain A"))
# Draw a horizontal line for each group
abline(h = 1:2, lty = 2)
```



b) Calculate the 95% CI for the population standard deviation in each group.

```
n_strain_a <- length(strain_a)
sd_strain_a <- sd(strain_a)
df_strain_a <- n_strain_a - 1

# Determine test statistic critical values using a chi-square distribution
# We are using an alpha of 0.95, so we want both tails to include 0.025
# of the distribution.
upper_critical <- qchisq(0.975, df_strain_a)
lower_critical <- qchisq(0.025, df_strain_a)

# Once critical values are found, we need to find the actual values of the
# bounds
upper_bound_a <- sd_strain_a*sqrt(df_strain_a/upper_critical)
lower_bound_a <- sd_strain_a*sqrt(df_strain_a/lower_critical)

# Repeat the procedure for Strain B
n_strain_b <- length(strain_b)
sd_strain_b <- sd(strain_b)
df_strain_b <- n_strain_b - 1
upper_critical <- qchisq(0.975, df_strain_b)
lower_critical <- qchisq(0.025, df_strain_b)
upper_bound_b <- sd_strain_b*sqrt(df_strain_b/upper_critical)
```

```r
lower_bound_b <- sd_strain_b*sqrt(df_strain_b/lower_critical)

# Display the results
cat("95% CI for Strain A:", upper_bound_a, lower_bound_a, "\n")
```

```
## 95% CI for Strain A: 3.587217 8.597825
```

```r
cat("95% CI for Strain B:", upper_bound_b, lower_bound_b)
```

```
## 95% CI for Strain B: 6.551324 18.58125
```

c) Calculate a 95% confidence interval for the ratio of the population standard deviations.

It was shown in class that we can use a ratio of the variances for both datasets to determine the CI for a ratio of population SDs by dividing this ratio from critical values of the F distribution for our dataset.

```r
# Calculate ratio of variances
var_ratio <- var(strain_a)/var(strain_b)

# Calculate critical values from F distribution with qf function
upper_critical <- qf(0.975, df_strain_a, df_strain_b)
lower_critical <- qf(0.025, df_strain_a, df_strain_b)

# Calculate CI bounds, make sure to do square root to get std. dev!
upper_bound_ratio <- sqrt(var_ratio/upper_critical)
lower_bound_ratio <- sqrt(var_ratio/lower_critical)

# Display the data
cat("95% CI for ratio of population SD:", upper_bound_ratio, lower_bound_ratio)
```

```
## 95% CI for ratio of population SD: 0.2534505 0.9993504
```

d) Confirm using var.test

```r
# Variance caluclation
confirm <- var.test(strain_a, strain_b)
confirm$conf.int
```

```
## [1] 0.06423715 0.99870113
## attr(,"conf.level")
## [1] 0.95
```

```r
# Standard Deviation Calculation: Do Square Root!
cat("Std Dev \n")
```

```
## Std Dev
```

```
sqrt(confirm$conf.int)
```

```
## [1] 0.2534505 0.9993504
## attr(,"conf.level")
## [1] 0.95
```

2

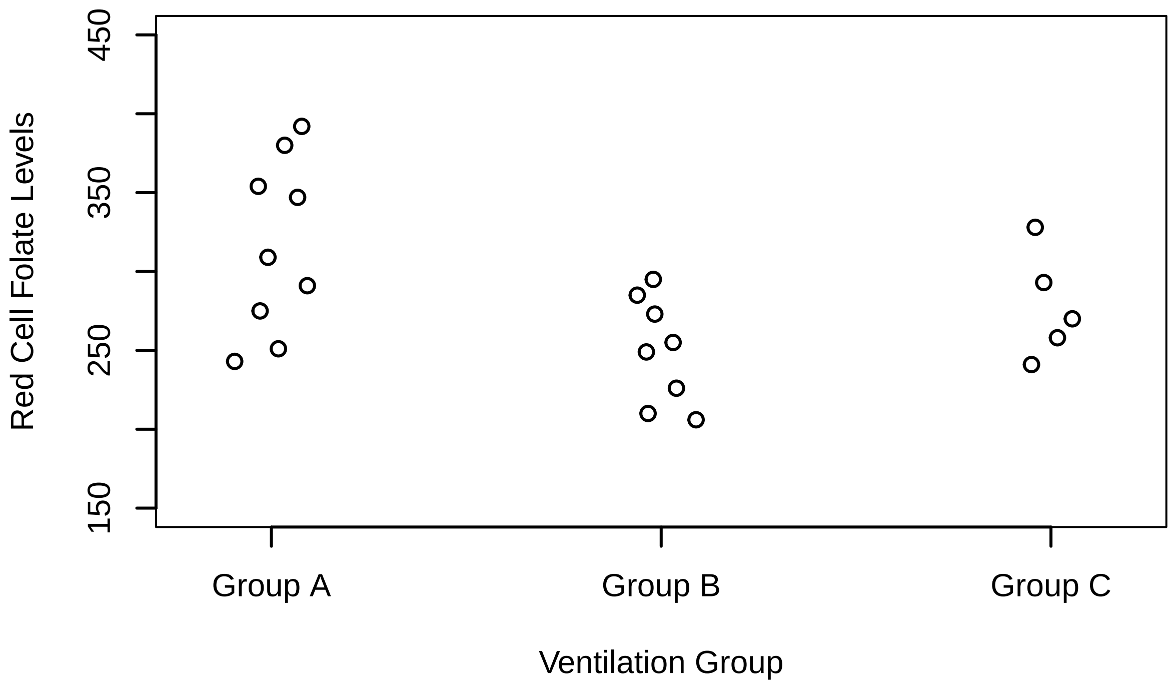
Cardiac bypass surger randomized to three different ventilation groups and their red cell folate levels are measured. Did the different ventilation methods change the mean counts?

a) Plot the data

```
# Define data vectors
group_a <- c(243,251,275,291,347,354,380,392,309)
group_b <- c(206,210,226,249,255,273,285,295)
group_c <- c(241,258,270,293,328)

# Group all treatments together into one object
cell_levels <- c(group_a, group_b, group_c)

#Plot using stripchart
stripchart(list(group_a, group_b, group_c),method='jitter', pch=1, lwd = 1.5, vertical=TRUE,
group.names = c("Group A", "Group B", "Group C"), xlab = "Ventilation Group",
ylab = "Red Cell Folate Levels", ylim=c(150, 450))
```



b) Derive the ANOVA table manually. Calculate a test statistic and look up the associated p-value.

For this question, I will use a significance level of $\alpha = 0.05$.

```
# Create labels for our dataset
# Use rep on the LETTERS object for letters A, B, C with number of labels
# equal to size of data set for each label
labels <- rep(LETTERS[1:3], c(9,8,5))

# Establish these values as factors
factors <- factor(labels, levels = c("A", "B", "C"))

# Calculate means
treatment_means <- tapply(cell_levels, labels, mean)
total_mean <- mean(cell_levels)
num_treatments <- 3
group_sizes <- tapply(cell_levels, factors, length)
total_subjects <- sum(group_sizes)

# Calculate the sum of squares with sum(total*(Yt - Y)^2)
sum_squares_btwn <- sum(group_sizes*(treatment_means - total_mean)^2)

# Calculate sum of squares within treatments
sum_squares_within <- sum((cell_levels - rep(treatment_means, group_sizes))^2)
```

```
# Calculate total sum of squares
sum_squares_total <- sum((cell_levels - total_mean)^2)

# Calculate degrees of freedom:
# Between treatments: num_treatments factors - 1
# Within treatments: N - k, N is total number treatments
# Total DF: Total size - 1
df_btwn <- num_treatments - 1
df_within <- total_subjects - num_treatments
df_total <- total_subjects - 1

# Calculate the mean squares
# Between treatments: sum_squares_btwn / df_btwn
# Within treatments: sum_squares_within / df_within
mean_square_btwn <- sum_squares_btwn / df_btwn
mean_square_within <- sum_squares_within / df_within

# Calculate test statistic and p-value
# Test statistic is found by the ratio of the mean of squares
# Can use pf with the test statistic to determine the p-value
test_statistic <- mean_square_btwn / mean_square_within
p_value <- pf(test_statistic, df_btwn, df_within, lower.tail = FALSE)

# Display the values
cat("Sum of squares between:", sum_squares_btwn, "\n",
    "Sum of squares within:", sum_squares_within, "\n",
    "Sum of squares total:", sum_squares_total, "\n",
    "Mean Square Between:", mean_square_btwn, "\n",
    "Mean Square Within:", mean_square_within, "\n",
    "Test Statistic from F Distribution:", test_statistic, "\n",
    "P-Value:", p_value)
```

```
## Sum of squares between: 18571.43
##  Sum of squares within: 36660.43
##  Sum of squares total: 55231.86
##  Mean Square Between: 9285.717
##  Mean Square Within: 1929.496
##  Test Statistic from F Distribution: 4.812508
##  P-Value: 0.02037393
```

c) Check with aov() to confirm

```
check <- aov(cell_levels~factors)
check
```

```
## Call:
##    aov(formula = cell_levels ~ factors)
##
## Terms:
##                   factors Residuals
## Sum of Squares  18571.43  36660.43
## Deg. of Freedom        2        19
##
## Residual standard error: 43.92603
## Estimated effects may be unbalanced
```

```
summary(check)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## factors      2  18571    9286   4.813 0.0204 *
## Residuals   19  36660    1929
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(check)
```

```
## Analysis of Variance Table
##
## Response: cell_levels
##           Df Sum Sq Mean Sq F value  Pr(>F)
## factors    2  18571  9285.7  4.8125 0.02037 *
## Residuals 19  36660  1929.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We performed our calculations correctly!

Our null hypothesis was that there is no difference between the ventilation methods influence upon red cell folate levels. The alternative is that there is a difference in the ventilation methods yielding different red cell folate levels.

With a p-value of 0.02037, which is less than our value of $\alpha = 0.05$, we can reject the null hypothesis and conclude that there are differences in ventilation methods influence upon red cell folate levels.

3

With a table of mean/SD values of ILEA scores with 50 pupils in each do…

a) State your null hypothesis if there are differences in test scores between the schools and explain why a random effects model is appropriate for this data set.

The random effects model is best for our data because the dataset was generated through random sampling of the population of students with the intention of generalizing our results from a small sample of individuals to the schools generally. There is variability introduced due to the sampling of students across different schools which are not the same. The goal is to determine if there is a difference in test results across schools. Therefore, the null hypothesis is that the variance between schools is zero. The alternative hypothesis is that the variance is not equal to zero which implies a differences between the schools.

b)

I will use a significance value of 0.05 for this question.

```
factors <- factor(1:7)
ilea_mean <- c(28.3, 21.1, 14.3, 16.3, 26.5, 20.3, 16.8)
ilea_sd <- c(13.4, 14.2, 11.8, 14.9, 13.4, 12.1, 13.9)

# Build a dataframe of this data
scores <- data.frame(
  School = factors,
  Mean=ilea_mean,
  SD=ilea_sd
  )

aov_result <- aov(ilea_mean ~ factors, data = scores)
anova(aov_result)
```

```
## Warning in anova.lm(aov_result): ANOVA F-tests on an essentially perfect fit
## are unreliable
```

```
## Analysis of Variance Table
##
## Response: ilea_mean
##           Df Sum Sq Mean Sq F value Pr(>F)
## factors    6 167.01  27.835     NaN    NaN
## Residuals  0   0.00     NaN
```

```
# Calculate relevant statistics from our dataset using the previous methods
# because we area not getting an interpretable F statistic or P-value.

# Number of factors
k <- 7
# Number of students per group
sample_size <- 50
# Total number of students
n <- 7*50

# Calculate degrees of freedom
df_btwn <- 7-1
df_within <- n - k
df_total <- n - 1

# Cacluate difference in means
overall_mean <- mean(ilea_mean)
diff_means <- ilea_mean - overall_mean

# Calculate sum of squares
sum_squares_btwn <- sum(sample_size * diff_means^2)
sum_squares_within <- sum((sample_size - 1)*ilea_sd^2)
sum_squares_total <- sum_squares_btwn + sum_squares_within

# Calculate group mean squares
mean_square_btwn <- sum_squares_btwn / df_btwn
mean_square_within <- sum_squares_within / df_within

# Calculate test statistic
```

```
test_statistic <- mean_square_btwn / mean_square_within

# Calculate the p-value
p_value <- pf(test_statistic, df_btwn, df_within, lower.tail = FALSE)

# Display the values
cat("Sum of squares between:", sum_squares_btwn, "\n",
    "Sum of squares within:", sum_squares_within, "\n",
    "Sum of squares total:", sum_squares_total, "\n",
    "Mean Square Between:", mean_square_btwn, "\n",
    "Mean Square Within:", mean_square_within, "\n",
    "Test Statistic from F Distribution:", test_statistic, "\n",
    "P-Value:", p_value)
```

```
## Sum of squares between: 8350.429
##  Sum of squares within: 61819.87
##  Sum of squares total: 70170.3
##  Mean Square Between: 1391.738
##  Mean Square Within: 180.2329
##  Test Statistic from F Distribution: 7.721889
##  P-Value: 8.44437e-08
```

It's possilbe to format this data nicely in a table! Thank you to my classmate for showing me how to do it.

```
anova_table <- data.frame(
  Source = c("Between Groups", "Within Groups", "Total"),
  DF = c(df_btwn, df_within, df_total),
  Sum_Sq = c(sum_squares_btwn, sum_squares_within, sum_squares_total),
  Mean_Sq = c(mean_square_btwn, mean_square_within, NA),
  F = c(test_statistic, NA, NA),
  P_Value = c(p_value, NA, NA)
)
print(anova_table)
```

```
##           Source  DF    Sum_Sq   Mean_Sq        F     P_Value
## 1 Between Groups   6  8350.429 1391.7381 7.721889 8.44437e-08
## 2  Within Groups 343 61819.870  180.2329       NA          NA
## 3          Total 349 70170.299        NA       NA          NA
```

Our p-value is substantially smaller than our value for α and we can thus reject the null hypothesis. We can conclude there there is variability between the schools test scores.

4

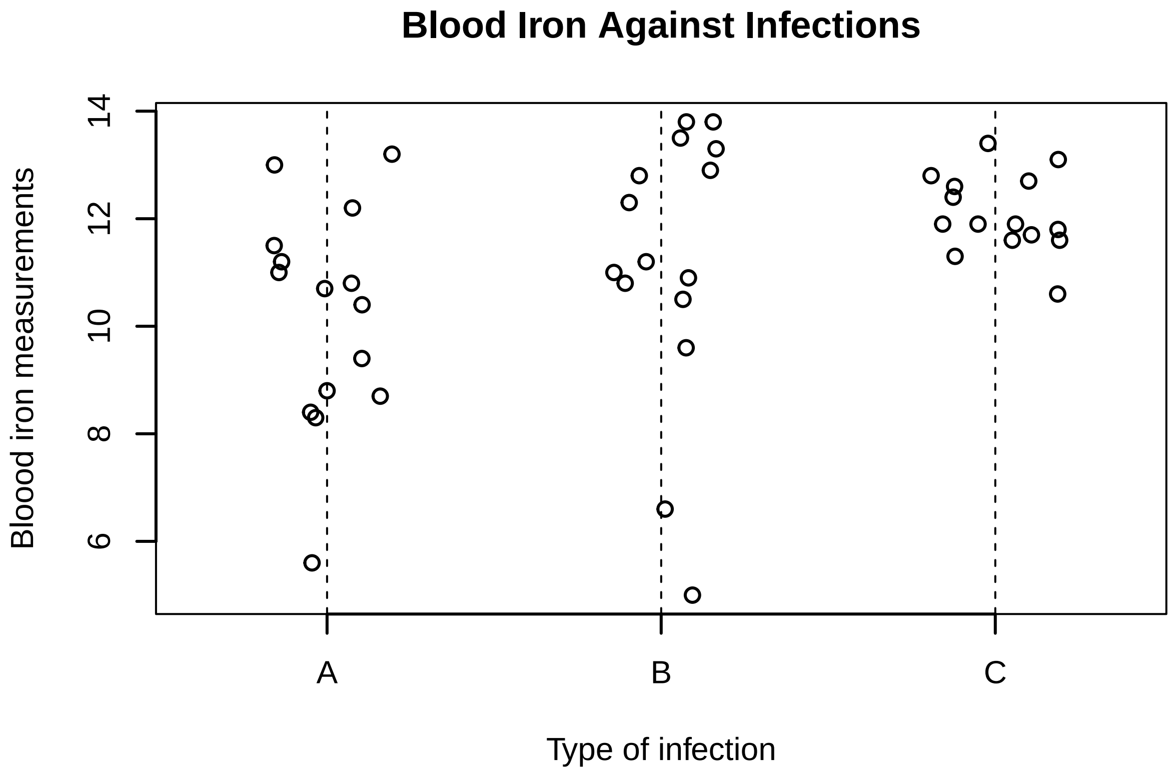
Using blood iron measurements from patients suffering from infections...

a) Graph the data in the groups

```
# Define Data Columns
A <- c(8.7,10.4,8.3,13.2,10.7,8.4,11.0,11.5,11.2,9.4,13.0,10.8,8.8,5.6,12.2)
B <- c(10.9,11.0,13.8,13.5,12.8,13.8,13.3,11.2,5.0,12.9,10.8,10.5,6.6,9.6,12.3)
C <- c(13.4,11.9,12.4,11.9,10.6,11.6,11.6,11.9,13.1,12.8,11.8,11.7,12.7,12.6,11.3)

# Place all data into one dataset
infections <- c(A, B, C)

# Create Strip Chart
stripchart(list(A,B,C),method='jitter',jitter=0.2, vertical=TRUE, pch=1,lwd=1.5,
           group.names=c("A","B","C"),ylab="Bloood iron measurements",
           xlab="Type of infection",main="Blood Iron Against Infections")
abline(v=1:3, lty=2)
```



b) Use the aov function to run an ANOVA

```
# Define factors
types <- rep(LETTERS[1:3], c(length(A), length(B), length(C)))
factors <- factor(types)

# Create a dataframe
data <- data.frame(
  Type = types,
  Blood_Iron = infections
)

aov_result <- aov(Blood_Iron ~ Type, data = data)
summary(aov_result)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Type          2  26.35  13.173   3.499 0.0393 *
## Residuals    42 158.13   3.765
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

c) What are the null and alternative hypotheses in this situation?

The null hypothesis in this analysis is that the blood iron levels are not different across infections. The alternative hypothesis is that there is a difference in blood iron levels across infections. With the significance value of 0.05 and a p-value of 0.0393, we can reject the null hypothesis that there is no difference in iron levels between infections and conclude there there is a difference in blood iron levels.