Laboratory 2 Skills

Steven Roberts

2025-01-18

Exercise 1

Sperm whales (Physeter macrocephalus) are among the deepest ocean divers among mammals. A team of marine mammal researchers attached time-depth recorders to sperm whales in the Pacific and Atlantic oceans to record individual diving depth for dives lasting longer than 30 minutes. The results of the study are presented below, showing dive depth in meters.

Pacific sperm whales: 420, 729, 442, 529, 484, 720, 453, 1002, 561, 980, 806, 263, 897, 652, 575, 346, 794, 553, 410, 417, 830, 1050, 1196, 687, 629, 1125, 496, 674, 998, 62, 735, 628, 1082, 1014, 732, 474, 111, 716, 567, 646, 286, 531, 550, 1346, 1401, 949, 644, 787, 929, 756, 763, 1035, 631, 1235, 395, 302, 804, 388, 574, 568, 491, 1389, 1125

Atlantic sperm whales: 614, 283, 415, 756, 288, 435, 473, 487, 512, 552, 795, 544, 472, 731, 290, 772, 674, 387, 670, 271, 648, 323, 344, 399, 812, 616, 396, 598, 1082, 594, 363, 332, 457, 456, 125, 566, 884, 801, 567, 442, 722, 374, 485, 370, 343, 658, 425, 429, 901, 489, 480, 431, 608, 576, 348, 331, 432, 538

Copy and paste each dataset into R. Don't forget to define the datasets with a name using the c() function, and remember: capitalization, spacing, and parentheses are important when using datasets and functions in R!

1. Use R to estimate the maximum, minimum, mean, median, variance, and standard deviation for each dataset using the max(), min(), mean(), median(), var(), and sd() functions.

```
pacific_sperm_whales <- c(420, 729, 442, 529, 484, 720, 453, 1002, 561, 980, 806, 263, 897,
                          652, 575, 346, 794, 553, 410, 417, 830, 1050, 1196, 687, 629,
                          1125, 496, 674, 998, 62, 735, 628, 1082, 1014, 732, 474, 111,
                          716, 567, 646, 286, 531, 550, 1346, 1401, 949, 644, 787, 929,
                          756, 763, 1035, 631, 1235, 395, 302, 804, 388, 574, 568, 491,
                          1389, 1125)

atlantic_sperm_whales <- c(614, 283, 415, 756, 288, 435, 473, 487, 512, 552, 795, 544, 472,
                           731, 290, 772, 674, 387, 670, 271, 648, 323, 344, 399, 812,
                           616, 396, 598, 1082, 594, 363, 332, 457, 456, 125, 566, 884,
                           801, 567, 442, 722, 374, 485, 370, 343, 658, 425, 429, 901,
                           489, 480, 431, 608, 576, 348, 331, 432, 538)


pacific_max <- max(pacific_sperm_whales)
pacific_min <- min(pacific_sperm_whales)
pacific_mean <- mean(pacific_sperm_whales)
pacific_median <- median(pacific_sperm_whales)
pacific_variance <- var(pacific_sperm_whales)
```

```r
pacific_sd <- sd(pacific_sperm_whales)


atlantic_max <- max(atlantic_sperm_whales)
atlantic_min <- min(atlantic_sperm_whales)
atlantic_mean <- mean(atlantic_sperm_whales)
atlantic_median <- median(atlantic_sperm_whales)
atlantic_variance <- var(atlantic_sperm_whales)
atlantic_sd <- sd(atlantic_sperm_whales)

# Create a data frame for the statistics
whale_statistics <- data.frame(
  Statistic = c("Maximum", "Minimum", "Mean", "Median", "Variance", "Standard Deviation"),
  Pacific = c(
    max(pacific_sperm_whales),
    min(pacific_sperm_whales),
    mean(pacific_sperm_whales),
    median(pacific_sperm_whales),
    var(pacific_sperm_whales),
    sd(pacific_sperm_whales)
  ),
  Atlantic = c(
    max(atlantic_sperm_whales),
    min(atlantic_sperm_whales),
    mean(atlantic_sperm_whales),
    median(atlantic_sperm_whales),
    var(atlantic_sperm_whales),
    sd(atlantic_sperm_whales)
  )
)

# Display the data frame
whale_statistics
```

```
##            Statistic    Pacific   Atlantic
## 1            Maximum  1401.0000  1082.0000
## 2            Minimum    62.0000   125.0000
## 3               Mean   704.1905   520.1034
## 4             Median   652.0000   482.5000
## 5           Variance 90138.3502 34033.9540
## 6 Standard Deviation   300.2305   184.4829
```

2. Calculate the range for each dataset.

```r
pacific_range <- max(pacific_sperm_whales) - min(pacific_sperm_whales)
atlantic_range <- max(atlantic_sperm_whales) - min(atlantic_sperm_whales)


data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  Range = c(pacific_range, atlantic_range)
)
```

```
##                 Dataset Range
```

```
## 1  Pacific Sperm Whales  1339
## 2 Atlantic Sperm Whales   957
```

3. Calculate the coefficient of variation for each dataset, and then interpret the results.

```r
pacific_cv <- (sd(pacific_sperm_whales) / mean(pacific_sperm_whales)) * 100
atlantic_cv <- (sd(atlantic_sperm_whales) / mean(atlantic_sperm_whales)) * 100


cv_results <- data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  Coefficient_of_Variation = c(pacific_cv, atlantic_cv)
)

cv_results
```

```
##                 Dataset Coefficient_of_Variation
## 1  Pacific Sperm Whales                 42.63484
## 2 Atlantic Sperm Whales                 35.47043
```

- The coefficient of variation (CV) is a measure of relative variability. It expresses the standard deviation as a percentage of the mean, allowing you to compare the variability of datasets regardless of their scale.
- A higher CV indicates greater relative variability, while a lower CV suggests the data is more consistent relative to the mean.
- Compare the CV values of the Pacific and Atlantic sperm whale datasets to see which has more relative variability in its values. For example:/ If the Pacific dataset has a higher CV, it suggests greater variability in the Pacific sperm whale data relative to its mean.
  If the Atlantic dataset has a higher CV, it suggests the same for the Atlantic sperm whale data.

4. Assume that the data from both datasets are normally distributed. Compute the extent of diving depths (i.e., from x to y meters) that correspond to approximately 95% of the diving depths for each dataset. Remember, this is the empirical rule.

```r
# Compute 95% diving depth range for Pacific sperm whales
pacific_mean <- mean(pacific_sperm_whales)
pacific_sd <- sd(pacific_sperm_whales)
pacific_lower <- pacific_mean - 2 * pacific_sd
pacific_upper <- pacific_mean + 2 * pacific_sd

# Compute 95% diving depth range for Atlantic sperm whales
atlantic_mean <- mean(atlantic_sperm_whales)
atlantic_sd <- sd(atlantic_sperm_whales)
atlantic_lower <- atlantic_mean - 2 * atlantic_sd
atlantic_upper <- atlantic_mean + 2 * atlantic_sd

# Display results
depth_range <- data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  Lower_Bound = c(pacific_lower, atlantic_lower),
  Upper_Bound = c(pacific_upper, atlantic_upper)
```

```
)

depth_range
```

```
##                 Dataset Lower_Bound Upper_Bound
## 1  Pacific Sperm Whales    103.7295   1304.6515
## 2 Atlantic Sperm Whales    151.1376    889.0693
```

Interpretation of 95% Diving Depths

Assuming the diving depths of both Pacific and Atlantic sperm whales are normally distributed, we used the empirical rule to estimate the range within which 95% of the diving depths fall. This rule states that:

$$\text{Range for } 95\% = \text{Mean} \pm 2 \times \text{Standard Deviation}$$

For each dataset:
- Pacific Sperm Whales:
- The diving depths from the lower bound to the upper bound encompass approximately 95% of the dives.
- Dives outside this range are considered uncommon, representing the extremes (2.5% on each tail) of the normal distribution.
- Atlantic Sperm Whales:
- Similarly, the diving depths from the calculated lower bound to upper bound represent 95% of the dives.
- Any dives outside this range are rare and represent outlier events.

5. Calculate the lower (Q1), and upper (Q3) quartiles of the distribution for each dataset. You can do this by using the quantile() command.

```
# Calculate quartiles for Pacific sperm whales
pacific_quartiles <- quantile(pacific_sperm_whales, probs = c(0.25, 0.75))

# Calculate quartiles for Atlantic sperm whales
atlantic_quartiles <- quantile(atlantic_sperm_whales, probs = c(0.25, 0.75))

# Create a data frame to display the results
quartile_results <- data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  Q1 = c(pacific_quartiles[1], atlantic_quartiles[1]),
  Q3 = c(pacific_quartiles[2], atlantic_quartiles[2])
)

quartile_results
```

```
##                 Dataset     Q1    Q3
## 1  Pacific Sperm Whales 493.50 913.0
## 2 Atlantic Sperm Whales 389.25 615.5
```

6. Calculate the interquartile range (IQR) for each dataset.

```
# Calculate the IQR for Pacific sperm whales
pacific_iqr <- IQR(pacific_sperm_whales)

# Calculate the IQR for Atlantic sperm whales
atlantic_iqr <- IQR(atlantic_sperm_whales)
```

```r
# Create a data frame to display the results
iqr_results <- data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  IQR = c(pacific_iqr, atlantic_iqr)
)

iqr_results
```

```
##                 Dataset    IQR
## 1  Pacific Sperm Whales 419.50
## 2 Atlantic Sperm Whales 226.25
```

7. Now compute the dive depth at the 95th and 5th percentile for each dataset. Note: The empirical rule tells us something subtly different from percentiles. Use quantile(data, probs = c(.05,.95)) to calculate percentiles. Note that we can also use this function to compute deciles.

```r
# Calculate the 5th and 95th percentiles for Pacific sperm whales
pacific_percentiles <- quantile(pacific_sperm_whales, probs = c(0.05, 0.95))

# Calculate the 5th and 95th percentiles for Atlantic sperm whales
atlantic_percentiles <- quantile(atlantic_sperm_whales, probs = c(0.05, 0.95))

# Create a data frame to display the results
percentile_results <- data.frame(
  Dataset = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  P5 = c(pacific_percentiles[1], atlantic_percentiles[1]),
  P95 = c(pacific_percentiles[2], atlantic_percentiles[2])
)

percentile_results
```

```
##                 Dataset     P5    P95
## 1  Pacific Sperm Whales 287.60 1231.1
## 2 Atlantic Sperm Whales 287.25  822.8
```

8. Calculate the z-scores for a Pacific sperm whale that dove to a depth of 949 meters and an Atlantic sperm whale that dove to 538 meters, and then compare the results.

```r
# Calculate mean and standard deviation for both datasets
pacific_mean <- mean(pacific_sperm_whales)
pacific_sd <- sd(pacific_sperm_whales)

atlantic_mean <- mean(atlantic_sperm_whales)
atlantic_sd <- sd(atlantic_sperm_whales)

# Calculate z-scores
pacific_z <- (949 - pacific_mean) / pacific_sd
atlantic_z <- (538 - atlantic_mean) / atlantic_sd

# Create a data frame to display the results
z_score_results <- data.frame(
```

```
  Dataset = c("Pacific Sperm Whale", "Atlantic Sperm Whale"),
  Depth = c(949, 538),
  Z_Score = c(pacific_z, atlantic_z)
)

z_score_results
```

```
##                Dataset Depth    Z_Score
## 1  Pacific Sperm Whale   949 0.81540526
## 2 Atlantic Sperm Whale   538 0.09700925
```

9. Determine if there are any outliers. In this case we want to look for z-scores less than -3 or greater than 3. Report the z-score as well as the depth associate with that z-score for any outliers.

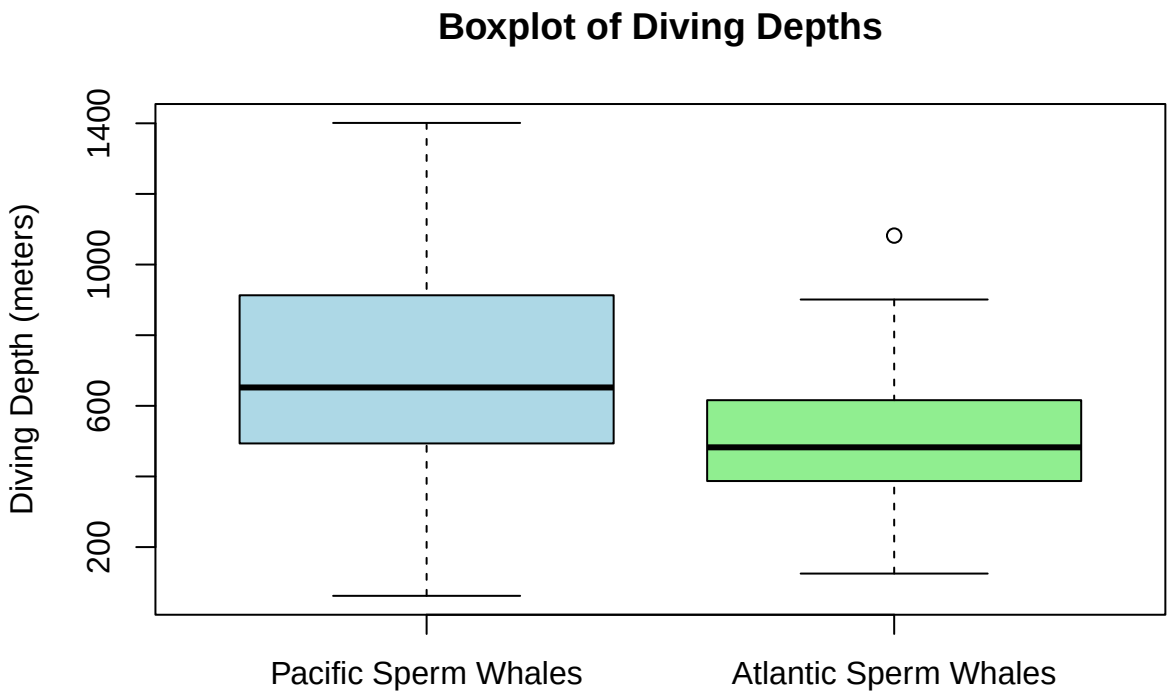
```
# Calculate z-scores for Pacific sperm whales
pacific_z_scores <- (pacific_sperm_whales - pacific_mean) / pacific_sd

# Identify outliers in the Pacific dataset
pacific_outliers <- pacific_sperm_whales[abs(pacific_z_scores) > 3]
pacific_outlier_z_scores <- pacific_z_scores[abs(pacific_z_scores) > 3]

# Calculate z-scores for Atlantic sperm whales
atlantic_z_scores <- (atlantic_sperm_whales - atlantic_mean) / atlantic_sd

# Identify outliers in the Atlantic dataset
atlantic_outliers <- atlantic_sperm_whales[abs(atlantic_z_scores) > 3]
atlantic_outlier_z_scores <- atlantic_z_scores[abs(atlantic_z_scores) > 3]

# Combine results into a data frame
outlier_results <- data.frame(
  Dataset = c(
    rep("Pacific Sperm Whales", length(pacific_outliers)),
    rep("Atlantic Sperm Whales", length(atlantic_outliers))
  ),
  Depth = c(pacific_outliers, atlantic_outliers),
  Z_Score = c(pacific_outlier_z_scores, atlantic_outlier_z_scores)
)

outlier_results
```

```
##                 Dataset Depth  Z_Score
## 1 Atlantic Sperm Whales  1082 3.045791
```

10. Use the boxplot() command to create a box-whisker plot for both the Atlantic and Pacific sperm whale data side-by-side and copy and paste it below. Hint: The names argument in the plot() function lets you label the x-axis ticks.

```
# Create a side-by-side box-whisker plot
boxplot(
  pacific_sperm_whales, atlantic_sperm_whales,
  names = c("Pacific Sperm Whales", "Atlantic Sperm Whales"),
  main = "Boxplot of Diving Depths",
```

```
  ylab = "Diving Depth (meters)",
  col = c("lightblue", "lightgreen") # Optional colors for the boxplots
)
```



Exercise 2

The gender and age of students from a past QSCI381 class were surveyed. The results of this survey can be found in the "lab2data.csv" file. REMEMBER: capitalization, spacing, and parentheses are important when using datasets and functions in R!

Include all code for each question, plus output from R or graphs, and answer all sub-questions.

1. Read in the "lab2data.csv" file using the read.csv() command and name it data. Look at the first six rows of data.

```
lab2data <- read.csv("http://gannet.fish.washington.edu/seashell/snaps/lab2data.csv")

lab2data
```

```
##       ID Gender Age HomeTown
## 1   BA_1   Male  22  Seattle
## 2   BA_2 Female  28   Tacoma
## 3   BA_3 Female  24   Tacoma
## 4   BA_4 Female  28  Seattle
## 5   BA_5   Male  20   Tacoma
```

```
## 6   BA_6   Male  22   Tacoma
## 7   BA_7 Female  24   Tacoma
## 8   BA_8 Female  25  Seattle
## 9   BA_9 Female  19  Seattle
## 10 BA_10 Female  26   Tacoma
## 11 BA_11   Male  19   Tacoma
## 12 BA_12   Male  24  Seattle
## 13 BA_13 Female  19  Seattle
## 14 BA_14 Female  23   Tacoma
## 15 BA_15 Female  23   Tacoma
## 16 BA_16   Male   1   Tacoma
## 17 BA_17 Female  20   Tacoma
## 18 BA_18 Female  24  Seattle
## 19 BA_19 Female  22   Tacoma
## 20 BA_20 Female  21   Tacoma
## 21 BA_21 Female  19   Tacoma
## 22 BA_22 Female  27   Tacoma
## 23 BA_23 Female  23  Seattle
## 24 BA_24 Female  19  Seattle
## 25 BA_25 Female  27   Tacoma
## 26 BA_26 Female  18  Seattle
## 27 BA_27   Male  19  Seattle
## 28 BA_28 Female  20   Tacoma
## 29 BA_29   Male  22  Seattle
## 30 BA_30 Female  24  Seattle
```

2. Change the Age column values of data from numeric to character ("under21" , "over21") values using the ifelse() function. data$AgeFactor <- ifelse(data$AgeFactor < 21, "under21", "over21")

```r
# Add the AgeFactor column
lab2data$AgeFactor <- ifelse(lab2data$Age < 21, "under21", "over21")

# Display the updated dataset
head(lab2data)
```

```
##     ID Gender Age HomeTown AgeFactor
## 1 BA_1   Male  22  Seattle    over21
## 2 BA_2 Female  28   Tacoma    over21
## 3 BA_3 Female  24   Tacoma    over21
## 4 BA_4 Female  28  Seattle    over21
## 5 BA_5   Male  20   Tacoma   under21
## 6 BA_6   Male  22   Tacoma    over21
```

3. Create a contingency table for Gender and Age from data and name it conting. Add the row and column totals to complete the contingency table.

```r
# Create a contingency table for Gender and Age
conting <- table(lab2data$Gender, lab2data$Age)

# Add row and column totals to the contingency table
conting_with_totals <- addmargins(conting)
```

```r
# Display the contingency table with totals
print(conting_with_totals)
```

```
##
##           1 18 19 20 21 22 23 24 25 26 27 28 Sum
##   Female  0  1  4  2  1  1  3  4  1  1  2  2  22
##   Male    1  0  2  1  0  3  0  1  0  0  0  0   8
##   Sum     1  1  6  3  1  4  3  5  1  1  2  2  30
```

4. What is the probability of being female? Of being female and over 21? Of being female OR over 21?

a) P(Female)

```r
# Total number of females
num_females <- sum(lab2data$Gender == "Female")

# Total number of individuals
total_individuals <- nrow(lab2data)

# Probability of being female
P_Female <- num_females / total_individuals
print(P_Female)
```

```
## [1] 0.7333333
```

b) P(Female AND over21)

```r
# Number of females over 21
num_female_over21 <- sum(lab2data$Gender == "Female" & lab2data$Age > 21)

# Probability of being female and over 21
P_Female_Over21 <- num_female_over21 / total_individuals
print(P_Female_Over21)
```

```
## [1] 0.4666667
```

c) P(Female OR over21)

```r
# Total number of individuals over 21
num_over21 <- sum(lab2data$Age > 21)

# Probability of being over 21
P_Over21 <- num_over21 / total_individuals

# Probability of being female OR over 21
P_Female_Or_Over21 <- P_Female + P_Over21 - P_Female_Over21
print(P_Female_Or_Over21)
```

```
## [1] 0.8666667
```

5. What is the probability of being male given one is over 21? What is the probability of being male given one is not over 21? Why do these two probabilities not sum to one?

a) P(Male|over21)

```
# Count males over 21
num_male_over21 <- sum(lab2data$Gender == "Male" & lab2data$Age > 21)

# Count total individuals over 21
num_over21 <- sum(lab2data$Age > 21)

# Probability of being male given over 21
P_Male_Given_Over21 <- num_male_over21 / num_over21
print(P_Male_Given_Over21)
```

```
## [1] 0.2222222
```

b) P(Male|under21)

```
# Count males not over 21
num_male_not_over21 <- sum(lab2data$Gender == "Male" & lab2data$Age <= 21)

# Count total individuals not over 21
num_not_over21 <- sum(lab2data$Age <= 21)

# Probability of being male given not over 21
P_Male_Given_NotOver21 <- num_male_not_over21 / num_not_over21
print(P_Male_Given_NotOver21)
```

```
## [1] 0.3333333
```

c) Why do these two probabilities not sum to one?

The probabilities P(Male | Over 21) and P(Male | Not Over 21) represent conditional probabilities, and they are calculated with respect to different subsets of the population. • P(Male | Over 21) only considers individuals over 21. • P(Male | Not Over 21) only considers individuals not over 21.

The denominators are different because they represent different subsets of the data. As such, these two probabilities are not complementary and do not sum to 1.