MATH 2310 Lab 3 - Numerical Summaries & Boxplots

Jakob Balkovec

2024-04-16

Table of Contents

Utility Functions

```
# {
# @brief Creates a summary table of numerical statistics for the given data frame.
#
# @param data_frame_sub: data.frame
#   - The data frame containing the numerical data.
# @return A summary table of mean, standard deviation, and five-number summary.
#
# Table Legend:
#  Mean    ...  -> Mean
#  SD      ...  -> Standard Deviation
#  Min     ...  -> Minimum
#  Q1      ...  -> 1st Quartile
#  Median  ...  -> Median
#  Q3      ...  -> 3rd Quartile
#  Max     ...  -> Maximum
#
# }
create_summary_table <- function(data_frame_sub) {
  mean <- mean(data_frame_sub)
  sd <- sd(data_frame_sub)

  fivenum <- fivenum(data_frame_sub)

  summary_table <- data.frame(
      Mean = round(mean, digits = 2),
      SD = round(sd, digits = 2),
      Min = round(fivenum[1], digits = 2),
      Q1 = round(fivenum[2], digits = 2),
      Median = round(fivenum[3], digits = 2),
      Q3 = round(fivenum[4], digits = 2),
      Max = round(fivenum[5], digits = 2)
  )

  knitr::kable(summary_table, format = "html") %>%
  kable_styling(bootstrap_options = "striped", full_width = TRUE) %>%
  add_header_above(c("Summary" = 2, "Five-Number Summary" = 5))
}
```

```
# {
# @brief Plots a comparative boxplot of rainfall data by treatment.
#
# @param data_frame_sub: data.frame
#   - The data frame containing the rainfall data.
#
# @param log: bool
#   - indicating whether to plot the log-transformed rainfall data.
# }
plot_rainfall_boxplot <- function(data_frame_sub, log) {

  if(!log) {
    y_label <- "Rainfall"
    title <- "Comparative Boxplot of Rainfall by Treatment"
  } else {
    y_label <- "Log_Rainfall"
    title <- "Comparative Boxplot of Log Rainfall by Treatment"
  }

  ggplot(data_frame_sub, aes(x = Treatment,
                             y = !!rlang::sym(y_label),
                             fill = Treatment)) +
    geom_boxplot(color = "black",
                 alpha = 0.7) +

    scale_fill_manual(values = c("#FF5733", "#33FF57"),
                      name = "Treatment") +

    labs(x = "Treatment",
         y = y_label,
         title = title) +

    theme_minimal() +
    theme(legend.position = "bottom",
          plot.title = element_text(face = "bold",
                                    size = 14,
                                    hjust = 0.5))
}
```

```
# Read the excel file and define the data as a variable
data_frame <- read_excel("clouds.xlsx")

# Subcategories
seeded <- data_frame$Rainfall[data_frame$Treatment == "Seeded"]
unseeded <- data_frame$Rainfall[data_frame$Treatment == "Unseeded"]

# Log Transformed Data
data_frame$Log_Rainfall <- log(data_frame$Rainfall)
```

Activity One

[Question A] *- Put together a summary of the cloud seeding data, including the mean, standard deviation, and five-number summary for each of the two groups (seeded and unseeded clouds).*

[Answer A] - Seeded

```
create_summary_table(seeded)
```

Summary		Five-Number Summary				
Mean	SD	Min	Q1	Median	Q3	Max
441.98	650.79	4.1	92.4	221.6	430	2745.6

[Answer A] - Unseeded

```
create_summary_table(unseeded)
```

Summary		Five-Number Summary				
Mean	SD	Min	Q1	Median	Q3	Max
164.59	278.43	1	24.4	44.2	163	1202.6

[Question A] *Of these numerical summaries, which are the most appropriate to use in describing this data? Why?*

I think it's challenging to determine which summary is the most appropriate since all of them provide valuable insights into the provided data.

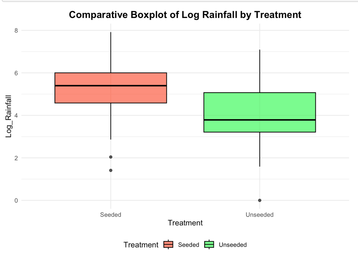
- **Mean**: It gives us the average rainfall across both treatment groups, which is useful but may be skewed by outliers (as evident in the dataset).
- **Standard Deviation**: The standard deviation seems quite high right off the bat. This suggests considerable variability in the dataset and a wide range of values from the average.
- **Five-Number Summary**: This one might seem the most useful as it provides a decent amount of information about the data. The **median** (Q2) gives us an idea of the middle point of the dataset, while the quartiles (**Q1** and **Q3**) provide insights into the spread of the data. The **minimum** and **maximum** represent the range of rainfall observations, indicating potential outliers at both extremes.

To conclude, I believe the **Five-Number Summary** is the most appropriate summary to describe the clouds dataset.

[Question B] *- Construct comparative boxplots for the cloud seeding data.*

[Answer B] - Boxplot

```
# Create a boxplot
plot_rainfall_boxplot(data_frame, log=TRUE)
```



[Answer B]

- **Unseeded Clouds:**
  - **Summary:** The unseeded clouds have a mean rainfall of approximately **164.59 mm** with a standard deviation of **278.43 mm**.
  - **Five-Number Summary:** The data range from a minimum of 1 mm to a maximum of **1202.6 mm**. The median rainfall (Q2) is **44.2 mm**, and the interquartile range (IQR) is from **24.4 mm** (Q1) to **163 mm** (Q3).
- **Seeded Clouds:**
  - **Summary:** The seeded clouds show a higher mean rainfall of around **441.98 mm** but with a larger standard deviation of **650.79 mm**.
  - **Five-Number Summary:** The seeded cloud data range from a minimum of **4.1 mm** to a maximum of **2745.6 mm**. The median rainfall (Q2) is **221.6 mm**, and the IQR extends from **92.4 mm** (Q1) to **430 mm** (Q3).
- **Observations:**
  - **Effectiveness:** The mean rainfall for seeded clouds is noticeably higher than that for unseeded clouds, suggesting that cloud seeding increases rainfall. However, the large standard deviation for seeded clouds indicates a high variability in the observed rainfall data, which may imply inconsistent effectiveness or some other influencing factors.
  - **Outliers:** Seeded clouds exhibit more extreme values, with a maximum rainfall almost double that of unseeded clouds. This suggests that while cloud seeding may increase overall rainfall, it also has the potential to produce some more extreme weather events.
  - **Interpretation:** While the mean rainfall for seeded clouds appears higher, further analysis would be needed to determine the significance of this difference and the reliability of cloud seeding as a "rainfall increase approach/method".

Activity Two

[Question A] *Put together a summary of the log transformed cloud seeding data, including the mean, standard deviation, and five-number summary for each of the two groups (seeded and unseeded clouds).*

```
log_seeded <- data_frame$Log_Rainfall[data_frame$Treatment == "Seeded"]
log_unseeded <- data_frame$Log_Rainfall[data_frame$Treatment == "Unseeded"]
```

[Answer A] - Log Seeded

```
create_summary_table(log_seeded)
```

Summary		Five-Number Summary				
Mean	SD	Min	Q1	Median	Q3	Max
5.13	1.6	1.41	4.53	5.4	6.06	7.92

[Answer A] - Log Unseeded

```
create_summary_table(log_unseeded)
```

Summary		Five-Number Summary				
Mean	SD	Min	Q1	Median	Q3	Max
3.99	1.64	0	3.19	3.79	5.09	7.09

[Question A] *Of these numerical summaries, which are the most appropriate to use in describing this data? Why?*

[Answer A]

- **Mean**: The mean of the log-transformed data gives us a measure of central tendency, indicating the average value of the data points on the log scale.
- **Standard Deviation**: The standard deviation of the log-transformed data measures the spread or variability of the data points around the mean on the log scale. It tells us how much the data deviates from the mean value.
- **Five-Number Summary**: This one might not be the most useful as we can't rely on the minimum or the maximum. That doesn't necessarily mean the other values in the Five-Number Summary are useless. We can focus on **Q1**, **Median (Q2)**, and **Q3**. These quartiles provide us with information about the spread and distribution of the data across different percentiles on a log scale.

Again, to conclude, I believe a combination of the **Mean**, **SD**, and the **Five-Number Summary (excluding Min and Max)** is the most appropriate summary to describe the clouds dataset.

Jakob Balkovec