TMFVC Assignment 11

Guangxin Zhao st194136

January 18, 2025

Exercise 11.1

a)

Code:

```
# Exercise 11.1

# a) Read and inspect the data
data <- read.csv("DatasaurusDozen.csv")

str(data)
print(unique(data$dataset))
```

Output:

```
'data.frame':   1846 obs. of  3 variables:
 $ dataset: chr  "dino" "dino" "dino" "dino" ...
 $ x      : num  55.4 51.5 46.2 42.8 40.8 ...
 $ y      : num  97.2 96 94.5 91.4 88.3 ...
 [1] "dino"       "away"       "h_lines"    "v_lines"    "x_shape"
 [6] "star"       "high_lines" "dots"       "circle"     "bullseye"
[11] "slant_up"   "slant_down" "wide_lines"
```

b)

Code:

```
# b) Compute the mean and standard deviation of the x and y values for each
↪  dataset
library(dplyr)

result <- data %>%
    group_by(dataset) %>%
    summarise(
        mean_x = mean(x),
        sd_x = sd(x),
        mean_y = mean(y),
        sd_y = sd(y))

print(result)
```

Output:

```
# A tibble: 13 × 5
   dataset    mean_x  sd_x mean_y  sd_y
   <chr>       <dbl> <dbl>  <dbl> <dbl>
 1 away         54.3  16.8   47.8  26.9
 2 bullseye     54.3  16.8   47.8  26.9
 3 circle       54.3  16.8   47.8  26.9
 4 dino         54.3  16.8   47.8  26.9
 5 dots         54.3  16.8   47.8  26.9
 6 h_lines      54.3  16.8   47.8  26.9
 7 high_lines   54.3  16.8   47.8  26.9
 8 slant_down   54.3  16.8   47.8  26.9
 9 slant_up     54.3  16.8   47.8  26.9
10 star         54.3  16.8   47.8  26.9
11 v_lines      54.3  16.8   47.8  26.9
12 wide_lines   54.3  16.8   47.8  26.9
13 x_shape      54.3  16.8   47.8  26.9
```

c)

Code:

```r
# c) Scatter plot of the data
library(ggplot2)

data %>%
    ggplot(aes(x = x, y = y, group = dataset)) +
    geom_point() +
    facet_wrap(~dataset)
```

Output:

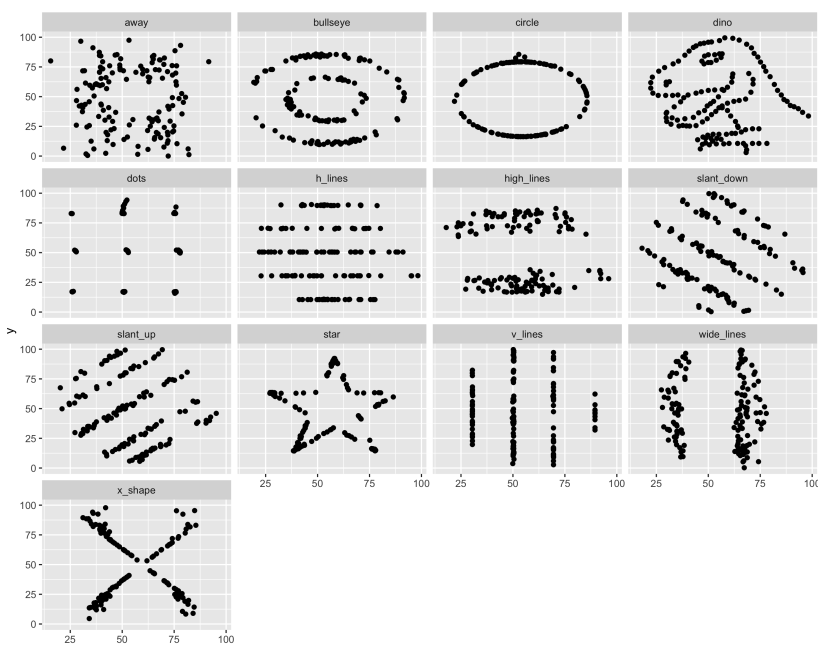

Figure 1: Scatter Plot of the Data

d)

The mean and standard deviation for the x and y variables for each group in the "dataset" variable are the same. However, as shown in Fig. 1, the datasets have different shapes and patterns. This demonstrates that summary statistics alone are insufficient to understand data distributions. Visualization is essential to better identify underlying structures.

Exercise 11.2

a)

1. The result is statistically significant.
2. We can reject H_0.

b)

Effect size gives an objective, standardized measure of the magnitude of the observed effect.

Two common measures for effect sizes are **Cohen's d** and **Pearson's r**.

c)

Two things one can do with power analysis is **a priori power analysis** and **sensitivity analysis**.

d)

- **Benefits**:
  1. Reduces the number of errors in the data.
  2. Improves the accuracy of statistical metrics such as mean, variance, and correlation.
  3. Reduces the risk of overfitting by allowing models to focus on the majority of data points.
- **Drawbacks**:
  1. Leads to subjective decisions when deleting valuable information.
  2. May introduce bias if outliers are removed arbitrarily or without proper justification.
  3. Potentially oversimplifies the data by removing valid complexities or underlying patterns.

Exercise 11.3

a)

I would perform **Shapiro-Wilk** since it is more sensitive and powerful to detect a significant effect on small sample sizes.

b)

I can then **visually** check the normality of the data (e.g., using a **histogram** or **Q-Q plot**). I can also apply **data transformation**, such as **log transformation** or **square-root transformation**.