PS_ Descriptive Statistics for TAs

Pablo E. Gutiérrez-Fonseca

2023-08-13

R practice.

Install packages.

```
library(ggplot2)
library(moments)
```

Load the water pollution data into R.

```
##       Year      Mean_Dec_May_snow_depth
##  Min.   :1954   Min.   :13.84
##  1st Qu.:1969   1st Qu.:41.87
##  Median :1984   Median :51.03
##  Mean   :1984   Mean   :50.80
##  3rd Qu.:1999   3rd Qu.:60.39
##  Max.   :2014   Max.   :92.60
```

```
##      Type              Density          Strength
##  Length:20          Min.   :0.8700   Min.   :390.0
##  Class :character   1st Qu.:0.9375   1st Qu.:482.5
##  Mode  :character   Median :0.9800   Median :555.0
##                     Mean   :0.9740   Mean   :541.5
##                     3rd Qu.:1.0200   3rd Qu.:602.5
##                     Max.   :1.0400   Max.   :650.0
```

1. the **MEAN Yearly Mean Snow Depth**.

```
mean(df_snow$Mean_Dec_May_snow_depth)
```

```
## [1] 50.79623
```

2. the **MEDIAN Yearly Mean Snow Depth**.

```
median(df_snow$Mean_Dec_May_snow_depth)
```

```
## [1] 51.02649
```

3. If you round **Yearly Mean Snow Depth** to the nearest integer (whole number with no decimals), **what is the MODE**?

R does not have a standard in-built function to calculate mode. So we create a user function to calculate mode of a data set in R. This function takes the vector as input and gives the mode value as output.

```
# Round the continuous column "value" to 1 decimal place
df_snow$Mean_Dec_May_snow_depth <- round(df_snow$Mean_Dec_May_snow_depth, 0)

calculate_mode <- function(x) {
  uniq_x <- unique(x)
  freq <- tabulate(match(x, uniq_x))
  uniq_x[which.max(freq)]
}

# Calculate the mode of the rounded column
mode_value <- calculate_mode(df_snow$Mean_Dec_May_snow_depth)
# Print the mode
print(paste("Mode:", mode_value))
```

```
## [1] "Mode: 43"
```

4. Still in Excel, now calculate the **STANDARD DEVIATION** for the Yearly Mean Snow Depth.

```
sd(df_snow$Mean_Dec_May_snow_depth)
```

```
## [1] 14.54088
```

5. What is the **INTER-QUARTILE RANGE** for **Yearly Mean snow Depth**?

```r
# Calculate the IQR for the specified column
iqr_value <- IQR(df_snow$Mean_Dec_May_snow_depth)

# Print the IQR
print(iqr_value)
```

```
## [1] 18
```

6. Using the Interquartile range technique, how many **OUTLIER** years are there in your **Yearly Mean Snow Depth Data**?

```r
# Calculate the quartiles and IQR
q1 <- quantile(df_snow$Mean_Dec_May_snow_depth, 0.25, na.rm = TRUE)
q3 <- quantile(df_snow$Mean_Dec_May_snow_depth, 0.75, na.rm = TRUE)
iqr <- q3 - q1

# Define the lower and upper bounds for outliers
lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr

# Identify outlier years
outlier_years <- df_snow$Mean_Dec_May_snow_depth < lower_bound | df_snow$Mean_Dec_May_snow_depth > upper

# Count the number of outlier years
num_outliers <- sum(outlier_years)

# Print the number of outlier years
print(num_outliers)
```

```
## [1] 2
```

8. Now calculate **Pearson's skew for the Yearly Mean Snow Depth Data**. Enter your answer rounded to 2 decimal places. (do this in R; you've already proved your excel skills above)

```r
# Calculate Pearson's skew
skew <- skewness(df_snow$Mean_Dec_May_snow_depth, na.rm = TRUE)

# Print the skew rounded to 2 decimal places
print(round(skew, 2))
```

```
## [1] -0.01
```

9. Calculate the **standard error of skew (ses)** for this data so we can determine the significance of our skew.

```r
# Calculate the skewness and standard error of skew
n <- length(df_snow$Mean_Dec_May_snow_depth)
se_skew <- sqrt(6 * n * (n - 1) / ((n - 2) * (n + 1) * (n + 3)))

# Print the skew and standard error of skew
cat("Skewness:", skew)
```

```
## Skewness: -0.01381543
```

```r
cat("Standard Error of Skewness:", se_skew)
```

```
## Standard Error of Skewness: 0.3062699
```

10. Based on the ses technique, is this **skew significant**?
    No

11. Now determine the **kurtosis for the Yearly Mean Snow Depth data**.

```r
# Calculate the kurtosis
kurt <- kurtosis(df_snow$Mean_Dec_May_snow_depth, na.rm = TRUE)

# Print the kurtosis
cat("Kurtosis :", kurt)
```

```
## Kurtosis : 3.349084
```

12. Calculate the **standard error of kurtosis (sek)** for this data so we can determine the significance of our skew.

```r
# Calculate the kurtosis and sample size
kurt <- kurtosis(df_snow$Mean_Dec_May_snow_depth, na.rm = TRUE)
n <- length(df_snow$Mean_Dec_May_snow_depth)

# Calculate the standard error of kurtosis
sek <- sqrt((24 * n * (n - 2) * (n - 3)) / ((n + 1) * (n + 1) * (n + 3) * (n + 5))) * (1 - (6 / (n + 1)

# Print the kurtosis and standard error of kurtosis
cat("Kurtosis:", kurt)
```

```
## Kurtosis: 3.349084
```

```r
cat("Standard Error of Kurtosis:", sek)
```

```
## Standard Error of Kurtosis: 0.4930314
```

13. Based on the sek technique, is this **kurtosis significant**? **No**

14. Based on your statistical summary values, do you think that your data is normally distributed? **No**

15. Is your data normally distributed? How can you tell? Be sure to report your certainty of this conclusion.

```
shapiro.test(df_snow$Mean_Dec_May_snow_depth)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_snow$Mean_Dec_May_snow_depth
## W = 0.99172, p-value = 0.9548
```

16. Based on these results, which measure of central tendency would be best describe central tendency for this dataset? **Mean**

17. Now use the **Boardstrength data** down-loadable from this blackboard folder. This datafile contains measurements of board strength for several different types of wood. Your task is to describe this data (**Strength variable**) as though you were presenting it in a paper, poster or presentation.

```
# Calculate mean and standard deviation
mean_strength <- mean(df_trees$Strength, na.rm = TRUE)
sd_strength <- sd(df_trees$Strength, na.rm = TRUE)

# Calculate quartiles (25th, 50th, 75th percentiles)
quartiles <- quantile(df_trees$Strength, probs = c(0.25, 0.5, 0.75), na.rm = TRUE)

# Test for normality using Shapiro-Wilk test
normality_test <- shapiro.test(df_trees$Strength)

# Calculate median
median_strength <- median(df_trees$Strength, na.rm = TRUE)

# Calculate mode using a custom function
calculate_mode <- function(x) {
  uniq_x <- unique(x)
  freq <- tabulate(match(x, uniq_x))
  uniq_x[which.max(freq)]
}
mode_strength <- calculate_mode(df_trees$Strength)

# Identify outlier years using IQR technique
q1 <- quantile(df_trees$Strength, 0.25, na.rm = TRUE)
q3 <- quantile(df_trees$Strength, 0.75, na.rm = TRUE)
iqr <- q3 - q1
lower_bound <- q1 - 1.5 * iqr
upper_bound <- q3 + 1.5 * iqr
outlier_years <- df_trees$Strength < lower_bound | df_trees$Strength > upper_bound
num_outliers <- sum(outlier_years)
```

```r
# Print the results
print(paste("Mean:", mean_strength))
```

```
## [1] "Mean: 541.5"
```

```r
print(paste("Standard Deviation:", sd_strength))
```

```
## [1] "Standard Deviation: 73.8615347511103"
```

```r
print("Quartiles:")
```

```
## [1] "Quartiles:"
```

```r
print(quartiles)
```

```
##   25%   50%   75%
## 482.5 555.0 602.5
```

```r
print("Shapiro-Wilk Test for Normality:")
```

```
## [1] "Shapiro-Wilk Test for Normality:"
```

```r
print(normality_test)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  df_trees$Strength
## W = 0.94921, p-value = 0.3553
```

```r
print(paste("Median:", median_strength))
```

```
## [1] "Median: 555"
```

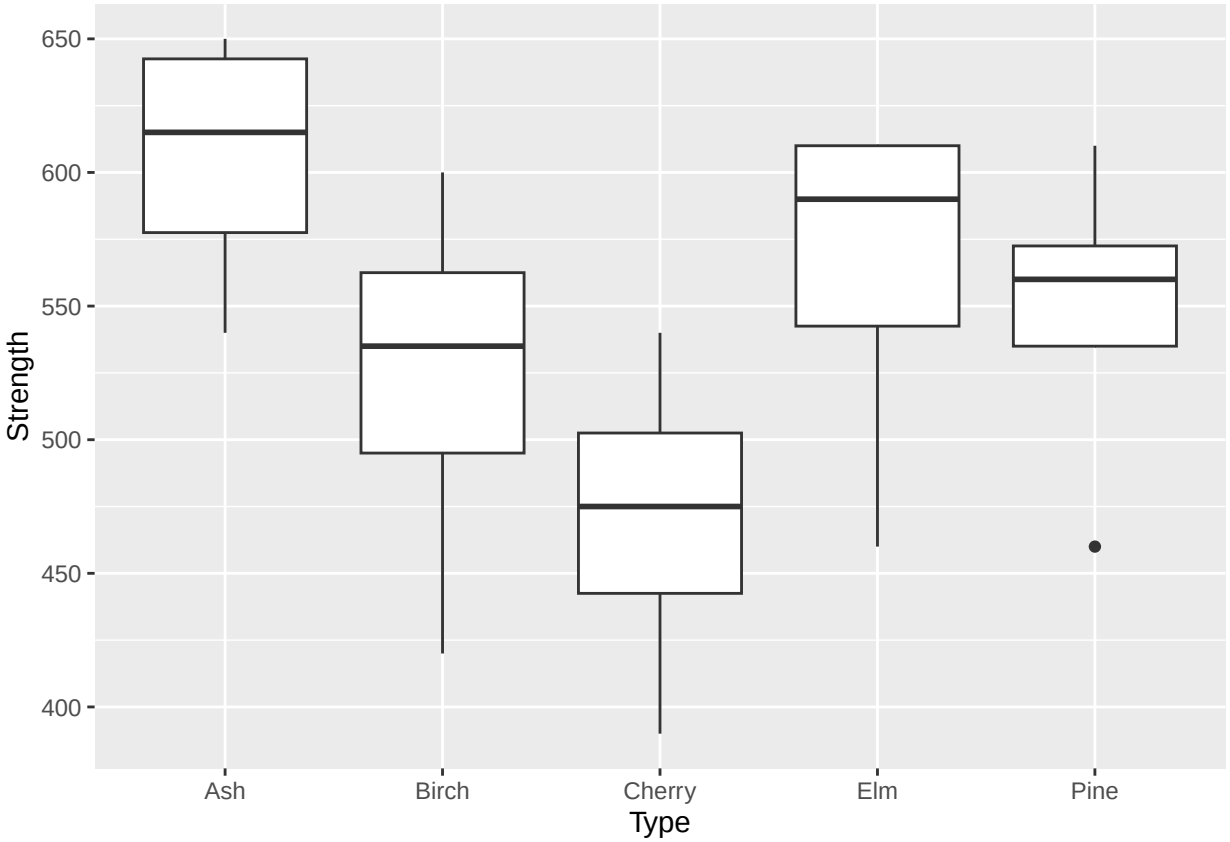
```r
print(paste("Mode:", mode_strength))
```

```
## [1] "Mode: 460"
```

```r
print(paste("Number of Outliers:", num_outliers))
```

```
## [1] "Number of Outliers: 0"
```

```
ggplot(df_trees, aes(x=Type, y=Strength)) +
  geom_boxplot()
```



18. If you wanted to highlight the differences between species, and describe strength for each of them, how would your table look different? Consider that the actual statistical analysis to test for differences among species is run on the full sample of 20 observations, so metrics like kurtosis, skew and goodness of fit tests would be calculated on the full data set.