PS_ Descriptive Statistics for TAs

Pablo E. Gutiérrez-Fonseca

2023-08-13

R practice.

Install packages.

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.2

## Warning: package 'ggplot2' was built under R version 4.3.2

## Warning: package 'tibble' was built under R version 4.3.1

## Warning: package 'tidyr' was built under R version 4.3.1

## Warning: package 'readr' was built under R version 4.3.2

## Warning: package 'purrr' was built under R version 4.3.1

## Warning: package 'dplyr' was built under R version 4.3.1

## Warning: package 'stringr' was built under R version 4.3.1

## Warning: package 'forcats' was built under R version 4.3.2

## Warning: package 'lubridate' was built under R version 4.3.2

## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
library(lessR)
```

```
## Warning: package 'lessR' was built under R version 4.3.2
```

```
##
## lessR 4.3.0                         feedback: gerbing@pdx.edu
## --------------------------------------------------------------
## > d <- Read("")   Read text, Excel, SPSS, SAS, or R data file
##   d is default data frame, data= in analysis routines optional
##
## Learn about reading, writing, and manipulating data, graphics,
## testing means and proportions, regression, factor analysis,
## customization, and descriptive statistics from pivot tables
##   Enter:  browseVignettes("lessR")
##
## View changes in this and recent versions of lessR
##   Enter: news(package="lessR")
##
## Interactive data analysis
##   Enter: interact()
##
##
## Attaching package: 'lessR'
##
## The following objects are masked from 'package:dplyr':
##
##     recode, rename
```

```r
library(DT)
library(e1071)
```

```
##
## Attaching package: 'e1071'
##
## The following object is masked from 'package:lessR':
##
##     kurtosis
```

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
```

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

```r
round(mean(df_snow$Mean_Dec_May_snow_depth),2)
```

```
## [1] 50.8
```

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

```r
round(median(df_snow$Mean_Dec_May_snow_depth),2)
```

```
## [1] 51.03
```

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

```r
round(sd(df_snow$Mean_Dec_May_snow_depth),2)
```

```
## [1] 14.51
```

2.1

```r
round(var(df_snow$Mean_Dec_May_snow_depth),2)
```

```
## [1] 210.48
```

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

```r
# Calculate the IQR for the specified column
round(iqr_value <- IQR(df_snow$Mean_Dec_May_snow_depth),2)
```

```
## [1] 18.52
```

```r
# Print the IQR
print(iqr_value)
```

```
## [1] 18.52056
```

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

```
# 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 > uppe

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

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

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

5. 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)

```
# Calculate Pearson's skew
skew <- skew(df_snow$Mean_Dec_May_snow_depth)
# Print the skew rounded to 2 decimal places
print(round(skew, 2))
```

```
## [1] -0.03
```

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

```
# Calculate the skewness and standard error of skew
ses <- round(sqrt(6/length(df_snow$Mean_Dec_May_snow_depth)),2)
ses
```

```
## [1] 0.31
```

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

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

```r
kurtosis <- round(kurtosis(df_snow$Mean_Dec_May_snow_depth, type = 1) - 3, 2)
kurtosis
```

```
## [1] -2.66
```

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

```r
# Calculate standard error of kurtosis (SES)
ses_kurt <- round(sqrt(24/length(df_snow$Mean_Dec_May_snow_depth)),2)
ses_kurt
```

```
## [1] 0.63
```

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

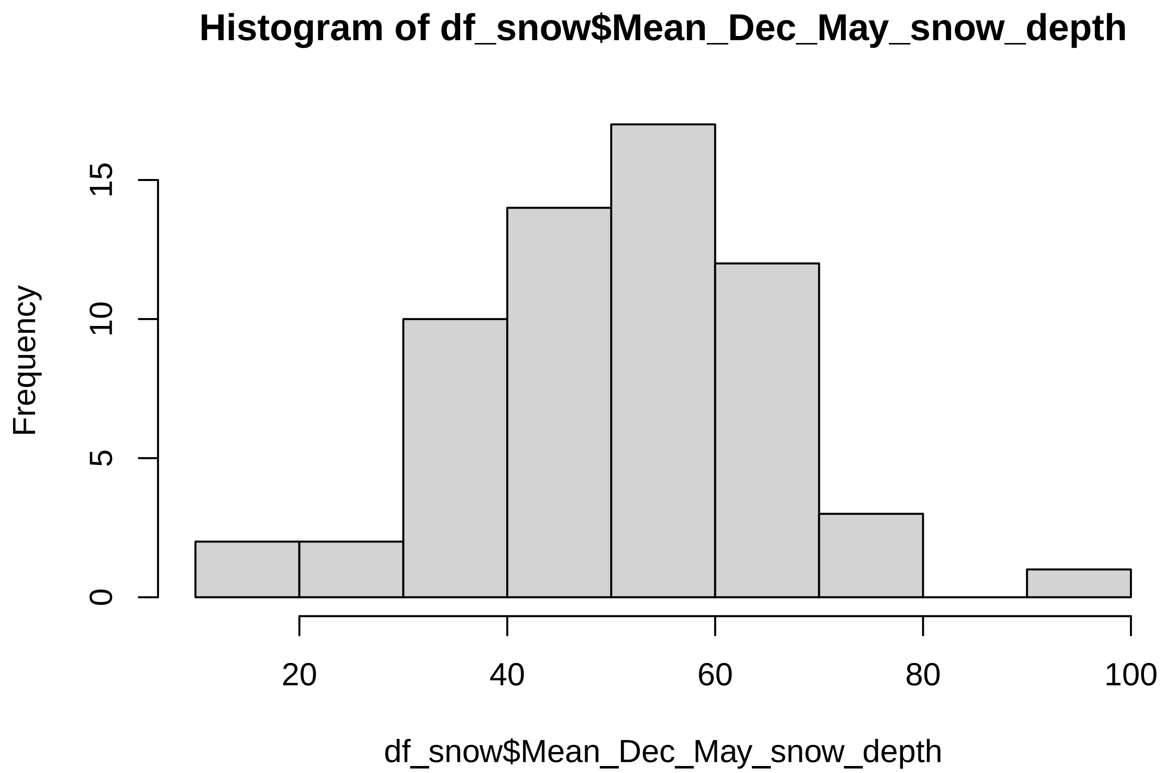
14. Based on your statistical summary values, do you think that your data is normally distributed? **No**

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

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

```
##
##  Shapiro-Wilk normality test
##
## data:  df_snow$Mean_Dec_May_snow_depth
## W = 0.99221, p-value = 0.9661
```

```r
hist(df_snow$Mean_Dec_May_snow_depth)
```


Histogram of df_snow$Mean_Dec_May_snow_depth
Frequency
0
5
10
15
20
40
60
80
100
df_snow$Mean_Dec_May_snow_depth