Displaying and Summarizing Data II

Sean Hellingman ©

Introduction to Statistical Data Analysis (ADSC1000)

shellingman@tru.ca

Fall 2024

THOMPSON RIVERS UNIVERSITY

Topics

Introduction

- Visual summaries are useful tools for learning about samples.
- Data analysis usually requires more formal calculations.
- **Remember:** we are still not trying to gain any insights about the population.
- Goal is to generate several numbers that characterize some of the most important features of the data.
  - Measures of location (mean & median).
  - Measures of variability (sample variance & sample standard deviation).

Measures of Location

Mean

- The most familiar and useful measure of the center is the mean, or arithmetic average of the set.
- The **sample mean** $\bar{x}$ of the observations x_1, x_2, ... x_n is given by

$$\bar{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1}$$

- In R: `mean(vector)`

Example 1

- Assume we have a sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Calculate the sample mean of the test scores.
  - Confirm your findings using the `mean()` function in R.

Question

- Can you think of any potential problems with using the mean as a measure of center?

Median

- The **median** or middle, is the middle value when the observations are ordered smallest to largest.
- The median is denoted by $\tilde{x}$.

$$\tilde{x} = \begin{cases} (\frac{n+1}{2})^{th} \text{ Ordered value }, \text{ if } n \text{ is odd} \\ \text{The average of } (\frac{n}{2})^{th} \text{and } (\frac{n}{2}+1)^{th} \text{ ordered values, if } n \text{ is even} \end{cases} \tag{2}$$

- In R: `median(vector)`

Example 2

- Assume we have a sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Calculate the sample median of the test scores.
  - Confirm your findings using the `median()` function in R.

Question

- Can you think of any potential problems with using the median as a measure of center?

Graphics of Measures of Center

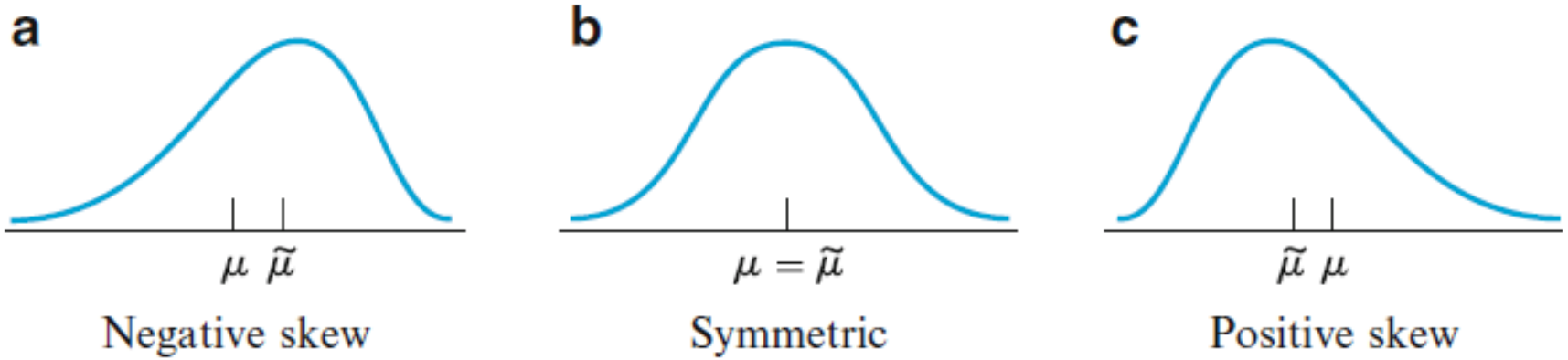


Figure: *source:* (2)

Example 3

- Assume we have a sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Generate a histogram with the mean and median lines included (see provided example for R code).
  - Comment on the shape of the distribution.

Quantiles

- The median simply divides the ordered data into two parts of equal size
  - We can divide the ordered data into more than two equal parts.
- **Quartiles** divide the ordered data into four equal parts
  - Observations above the third quartile represent the upper quarter of the data.
  - The second quartile is identical to the median.
  - Observations below the first quartile represent the lowest quarter of the data.
- We can also divide the data using **percentiles**
  - For example: the 99th percentile separates the highest 1% from the lowest 99%.

Quantiles in R

- The default in R is to calculate the quartiles: `quantile(vector)`
- You can also specify the divisions: `quantile(vector, probs = seq(0, 1, 1/4))`

Trimmed Mean

- The mean can be greatly influenced by outliers.
  - Income.
  - Number of goals scored per game.

- A **trimmed mean** is a mean that is calculated after removing a percentage of observations from each end of the ordered data.
  - For example: A 10% trimmed mean is computed by eliminating the smallest 10% and the largest 10% of the sample and then averaging what remains.

- In R: `mean(vector, trim=0.1)`

Example 4

- Assume we have the same sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Calculate the quartiles and tertiles (3^{rd}s).
  - Calculate the mean and 10% trimmed mean.
  - What do you notice?

Sample Proportions

- Sample proportions are a natural way to examine categorical data.
- The numerical summaries account for the individual frequencies and the relative frequencies.
  - What brand of laptops students own.
  - Own a laptop or not.
- A sample proportion reflects how many observations fall into a certain category relative to the other possible categories.

Explanation

- Assume a sample $n = 10$ is drawn from a dichotomous variable (two categories) that takes the outcomes A or B:
  - Let x denote the number in the sample falling in category A.
  - The **sample proportion** for category A is x/n.
  - Then the sample proportion for category B is 1-x/n.
- Denote a response that falls in category A by a 1 and a response that falls in category B by a 0.
- The sample 1, 0, 1, 1, 1, 1, 0, 0, 1, 1 would yield the sample proportion for A:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1 + 0 + \ldots + 1 + 1}{n} = \frac{7}{10} = \frac{x}{n}.$$

Sample Proportions in R

- One way to do this in R:

  `sum(as.numeric(vector == A))/length(vector)`

- Note: *use* `nrow(data.frame$variable.name)` *instead of* `length(vector)` *if data come from a data frame.*

Example 5

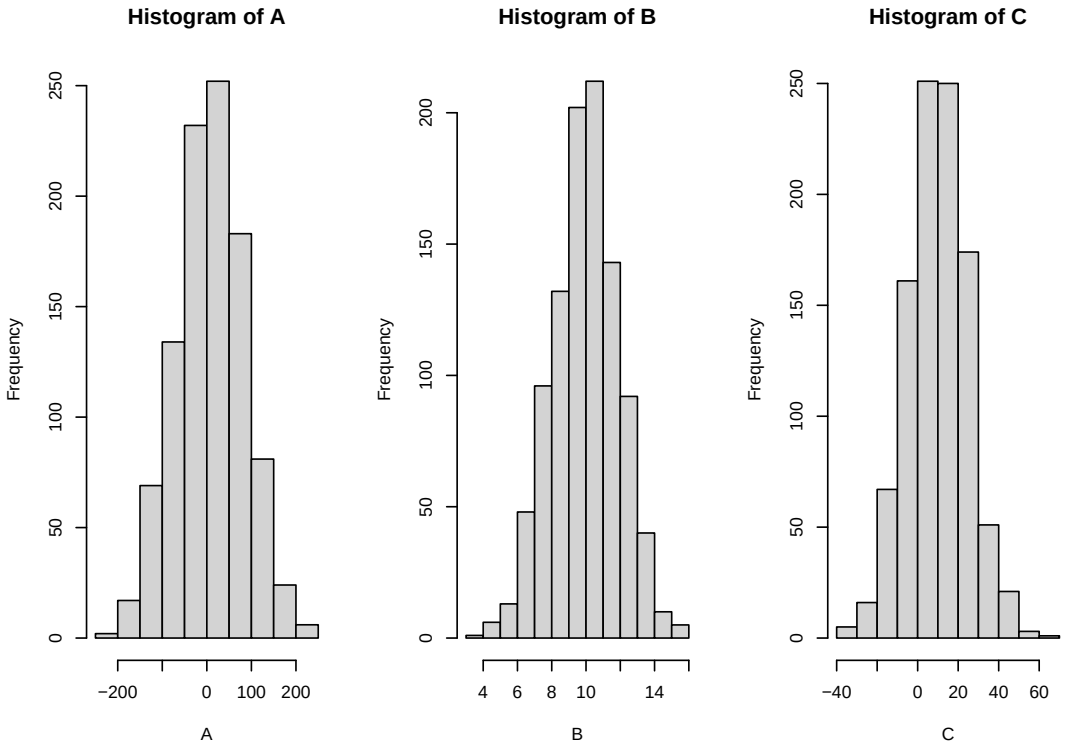
- Assume we have a sample of the favourite colour of ten ($n = 10$) students: red, red, blue, green, green, orange, red, orange, yellow, green.
    - Calculate the sample proportion for students who has a favourite colour of red.
    - Confirm your answer using R.

Measures of Variability

Variability

- Examining measures of center only give partial information about the data.

- Different populations or samples may have the same measures of center but differ in other ways.



Question

- Can you think of any ways to measure variability in a sample?

Some Measures of Variability

- **Range** is the difference between the largest and smallest sample values.
- Main measures of variability involve deviations from the mean (x_1 - $\bar{x}$, x_2 - $\bar{x}$, ..., x_n - $\bar{x}$)
  - Sum of deviations is not informative $\sum(x_i - \bar{x}) = 0$
  - Average absolute deviations $\sum |x_i - \bar{x}|$ (Theoretical difficulties).
  - **We will focus on squared deviations.**

Sample Variance and Standard Deviation

- The **sample variance** s^2 is calculated by:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{S_{xx}}{n-1}. \tag{3}$$

- The **sample standard deviation** s is calculated by:

$$s = \sqrt{s^2}. \tag{4}$$

- Note: We divide the sum of squared deviations by $n-1$ instead of n.

Sample Variance and Standard Deviation in R

- Sample variance in R: `var(vector)`
- Sample standard deviation in R: `sd(vector)`

Example 6

- Assume we have the same sample of fifteen ($n = 15$) test scores [0-100]: 93, 84, 86, 78, 95, 81, 72, 93, 84, 78, 45, 71, 78, 95, 88.
  - Using R, calculate the sample variance and sample standard deviation.
  - Generate a boxplot of the sample test scores.

Population Variance and Standard Deviation

- The **population variance** σ^2 is calculated by:

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{N}, \tag{5}$$

- The **population standard deviation** σ is calculated by:

$$\sigma = \sqrt{\sigma^2}. \tag{6}$$

- Note: We divide the sum of squared deviations by the population size N.
- Note: μ is the population mean.

Notation

Notation	Description
$\bar{x}$	Sample mean
$\tilde{x}$	Sample median
μ	Population mean
s^2	Sample variance
s	Sample standard deviation
σ^2	Population variance
σ	Population standard deviation

Exercise 1

- The 2022/23 Premier League points totals $n = 20$ are 67, 62, 61, 60, 59, 89, 84, 75, 71, 36, 34, 31, 25, 52, 45, 44, 41, 40, 39, 38.
- Calculate the mean and median for this data.
- Repeat this step using R.
- Create a histogram of the points and add the mean and median to the plot.
- By hand, calculate the variance in the 2022/23 Premier League points totals.

Exercise 2

- The 2022/23 La Liga points totals $n = 20$ are 88, 37, 25, 50, 49, 49, 49, 64, 60, 53, 51, 43, 42, 42, 42, 41, 40, 78, 77, 71.
- Calculate the 5% trimmed mean and the Quartiles for this data.
- Repeat this step using R.
- Compare the boxplots of the Premier League points (Excercise 1) and La Liga points.
- Which appears to have more variability?

Exercise 3

- Using the *iris* dataset in R:
  - Calculate the sample mean and median for the *Sepal.Length* and the *Petal.Length*.
  - Calculate the sample variance and standard deviation for the *Sepal.Length* and the *Petal.Length*.
  - Which measure has more variability?

Exercise 4

- Using the *iris* dataset in R:
  - Calculate the sample means of *Sepal.Length* for the three different *Species*.
  - Calculate the sample variances of *Sepal.Length* for the three different *Species*.
  - Comment on any potential differences you might uncover.

Exercise 5

- Load the *ToothGrowth* dataset in base R using `data("ToothGrowth")`.
- Use `?ToothGrowth` to familiarise yourself with the data.

Exercise 6

- Using the *ToothGrowth* dataset in R:
  - Using only the sample mean and median calculations comment on any potential skew of the *len* variable.
  - Verify your findings by generating a histogram and a boxplot.

Exercise 7

- Using the *ToothGrowth* dataset in R:
  - Calculate the 95th and 5th percentiles for the *len* variable.
  - Calculate the range for the *len* variable.
  - Compare the mean and 10% trimmed mean of the *len* variable.
  - Comment on any differences.

Exercise 8

- Using the *ToothGrowth* dataset in R:
  - Calculate the proportion of guinea pigs that received the supplement type VC in the sample.
  - Calculate the proportion of guinea pigs that received the supplement type OJ AND a dose size of 2 mg/day in the sample.

Exercise 9

- The three measures of center introduced in this section are the *mean*, *median*, and *trimmed mean*. Two additional measures of center that are occasionally used are the *midrange*, which is the average of the smallest and largest observations, and the *midfourth*, which is the average of the two fourths. Which of these five measures of center are resistant to the effects of outliers and which are not? Explain your reasoning.

References & Resources

1. Kohl, K., (2022). *Introduction to statistical data analysis with R (Second Edition)* Retrieved from https://github.com/stamats/ISDR/blob/main/IntroductionToStatistical DataAnalysisWithR_ed2.pdf.

2. Devore, J. L., Berk, K. N., & Carlton, M. A. (2012). *Modern mathematical statistics with applications (Second Edition).* New York: Springer.

- https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/quantile
- https://en.wikipedia.org/wiki/Quantile