STAT 2593

Lecture 003 - Measures of Location

Dylan Spicker

Measures of Location

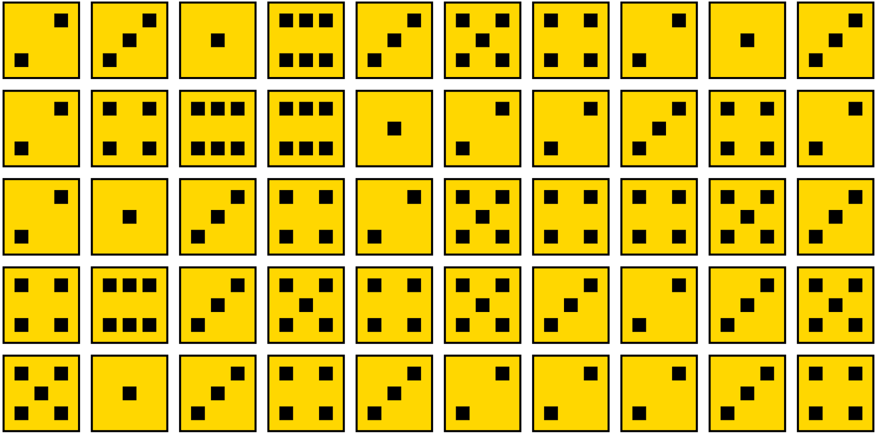
Learning Objectives

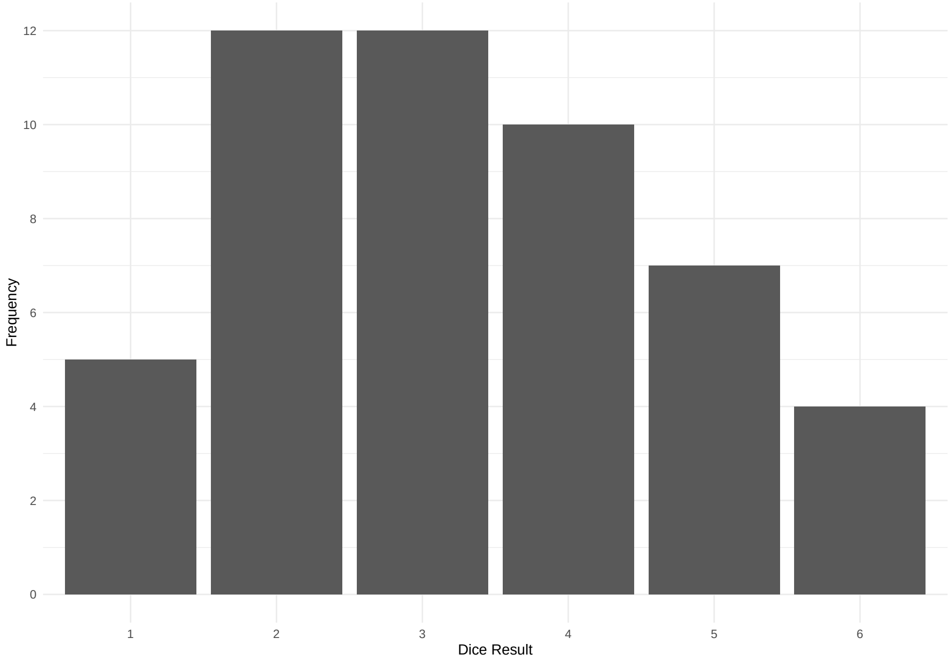
1. Understand and interpret the mean, median, and mode
2. Understand and interpret the sample proportion

Where is our Data?

Given a large dataset, how do we understand where observations typically fall?

Success: 50 of 50 (100%)




12
10
8
6
4
2
0
Frequency
1
2
3
4
5
6
Dice Result

Measures of Location

- **Sample mean** is the standard average of a distribution.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Measures of Location

- **Sample mean** is the standard average of a distribution.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- **Sample median** is the halfway point of a dataset, when the data are ordered.

$$\text{median} = \begin{cases} \left(\frac{n+1}{2}\right)^{\text{th}} \text{observation} & n \text{ is odd.} \\ \text{Mean of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observations} & n \text{ is even} \end{cases}$$

Measures of Location

- **Sample mean** is the standard average of a distribution.

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- **Sample median** is the halfway point of a dataset, when the data are ordered.

$$\text{median} = \begin{cases} \left(\frac{n+1}{2}\right)^{\text{th}} \text{observation} & n \text{ is odd.} \\ \text{Mean of } \left(\frac{n}{2}\right)^{\text{th}} \text{ and } \left(\frac{n}{2}+1\right)^{\text{th}} \text{ observations} & n \text{ is even} \end{cases}$$

- **Sample mode** is the most common (set of) observation(s).

Mean versus Median

- When data are approximately symmetric, the mean and median will be similar.

Mean versus Median

- When data are approximately symmetric, the mean and median will be similar.
- If data are skewed, the mean is *pulled* towards the long tail of the distribution.

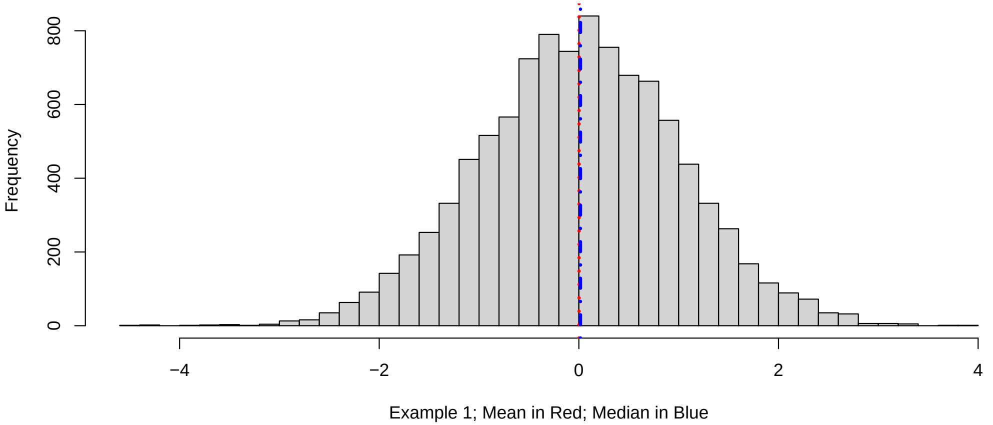
Mean versus Median

- When data are approximately symmetric, the mean and median will be similar.
- If data are skewed, the mean is *pulled* towards the long tail of the distribution.
    - In this way, the mean is more sensitive to *skewed* outliers than the median.

Mean versus Median

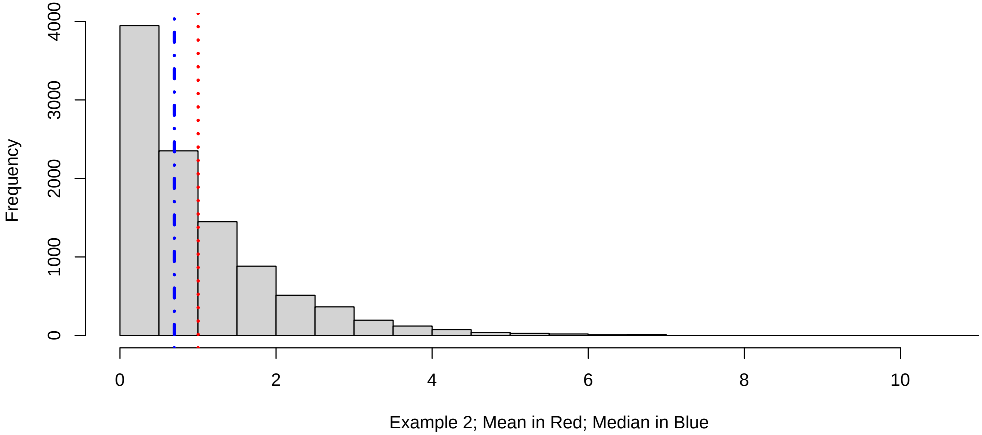
- When data are approximately symmetric, the mean and median will be similar.
- If data are skewed, the mean is *pulled* towards the long tail of the distribution.
  - In this way, the mean is more sensitive to *skewed* outliers than the median.
- We generally prefer the median if data are skewed, and the mean otherwise.

Examples

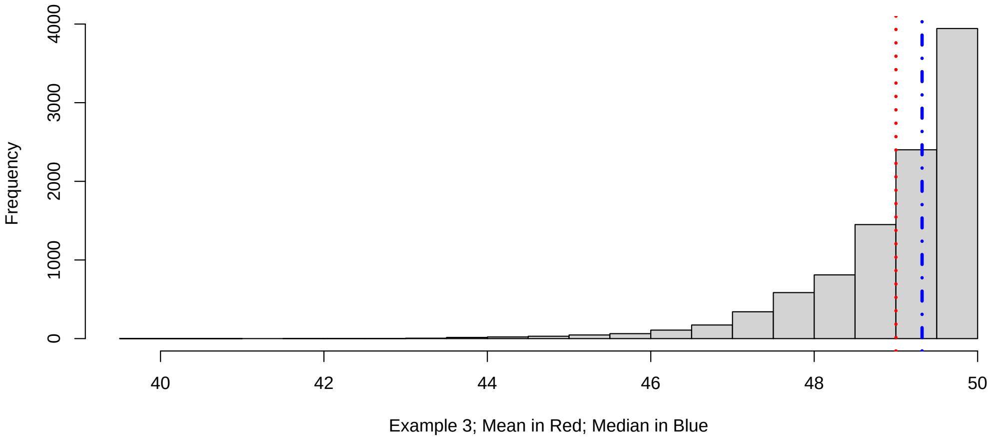


Frequency
0
200
400
600
800
−4
−2
0
2
4
Example 1; Mean in Red; Median in Blue

Examples

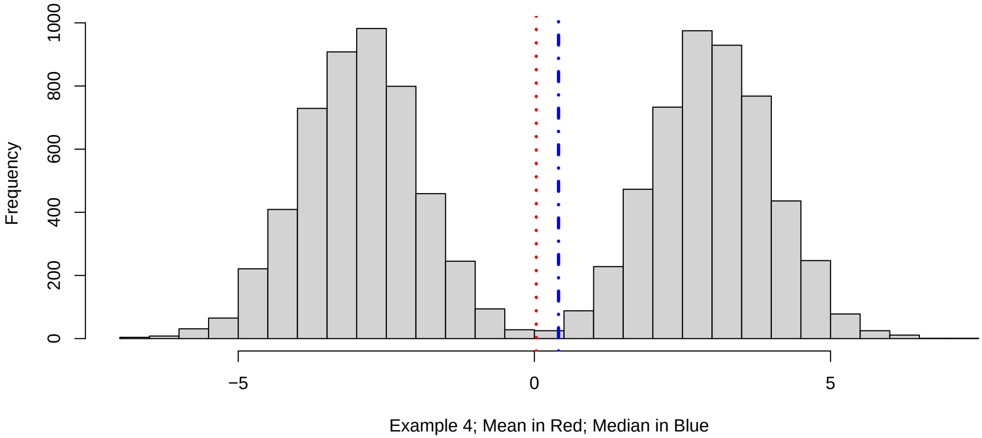

Frequency
0
1000
2000
3000
4000
0
2
4
6
8
10
Example 2; Mean in Red; Median in Blue

Examples



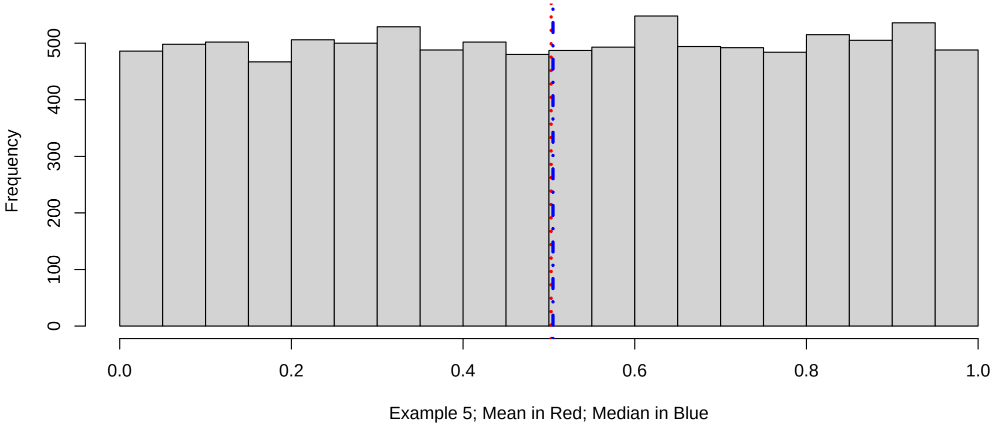
Frequency
0
1000
2000
3000
4000
40
42
44
46
48
50
Example 3; Mean in Red; Median in Blue

Examples

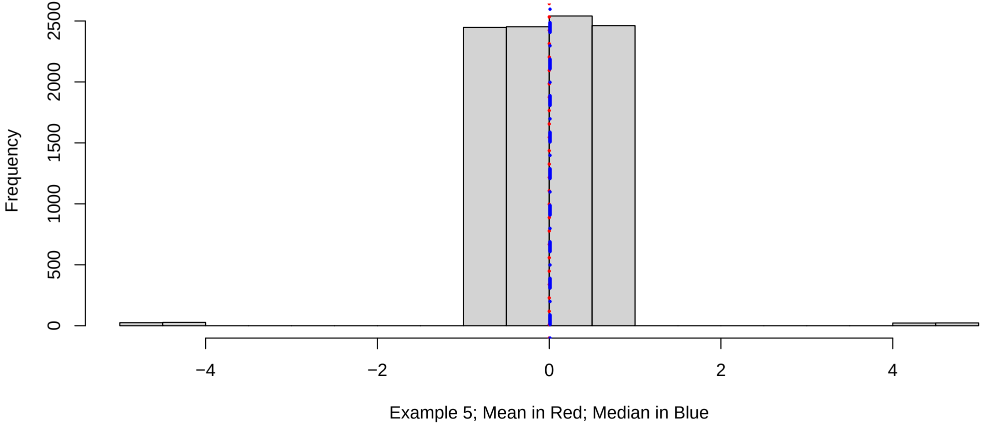


Frequency
0
200
400
600
800
1000
−5
0
5
Example 4; Mean in Red; Median in Blue

Examples


Frequency
0
100
200
300
400
500
0.0
0.2
0.4
0.6
0.8
1.0
Example 5; Mean in Red; Median in Blue

Examples



Frequency
0
500
1000
1500
2000
2500
−4
−2
0
2
4
Example 5; Mean in Red; Median in Blue

Sample Proportion

- For categorical data, we can consider the relative frequency of each category as the sample proportion.

Sample Proportion

- For categorical data, we can consider the relative frequency of each category as the sample proportion.
- Assume that there are categories, $c_1, c_2, \ldots, c_k$, and observations $x_1, x_2, \ldots, x_k$.

Sample Proportion

- For categorical data, we can consider the relative frequency of each category as the sample proportion.
- Assume that there are categories, $c_1, c_2, \ldots, c_k$, and observations $x_1, x_2, \ldots, x_k$.
- Define $z_{i,j} = I(x_i = c_j)$, where $I(\cdot)$ is an indicator function.

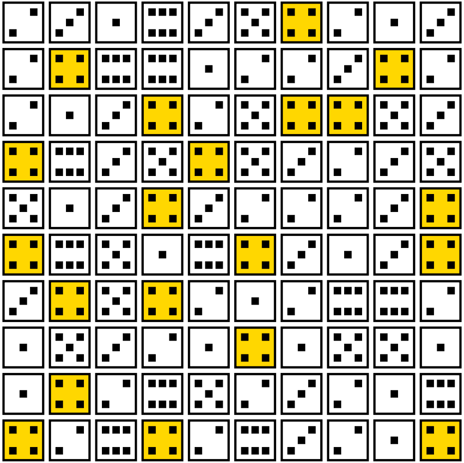
Sample Proportion

- For categorical data, we can consider the relative frequency of each category as the sample proportion.
- Assume that there are categories, $c_1, c_2, \ldots, c_k$, and observations $x_1, x_2, \ldots, x_k$.
- Define $z_{i,j} = I(x_i = c_j)$, where $I(\cdot)$ is an indicator function.
- Then, we can write the j-th sample proportion as

$$p_j = \bar{z}_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} z_{i,j}.$$

Example

Success: 20 of 100 (20%)



Summary

- Measures of location indicate what *usually* happens in a dataset
- The mean, median, and mode can be computed for quantitative variables
- The mean and median are most commonly used; the median is preferable for skewed data
- The sample proportion is used for indicating the location of a categorical variable