Lesson 003
Measures of Location

Friday, September 15

Survey Feedback

- **Note:** Questions mentioning "standard deviation", "variation", or "boxplots" will be covered next lesson.
- **Note:** Problem 32 had a missing histogram in the problem set.

- Comparing means and medians.
- How do we calculate the mean of categorical data?
- Questions 29, 31, and 32 in the problem set

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.

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

- **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).

Example

Calculate the Mean, Median, and Mode of the Following Data

28	93	31	1
14	7	67	36
21	41	1	30

What is the mean of: 5, 5, 10, 20?

10	0%
7.5	0%
5	0%
40	0%

What is the median of: 5, 5, 10, 20?

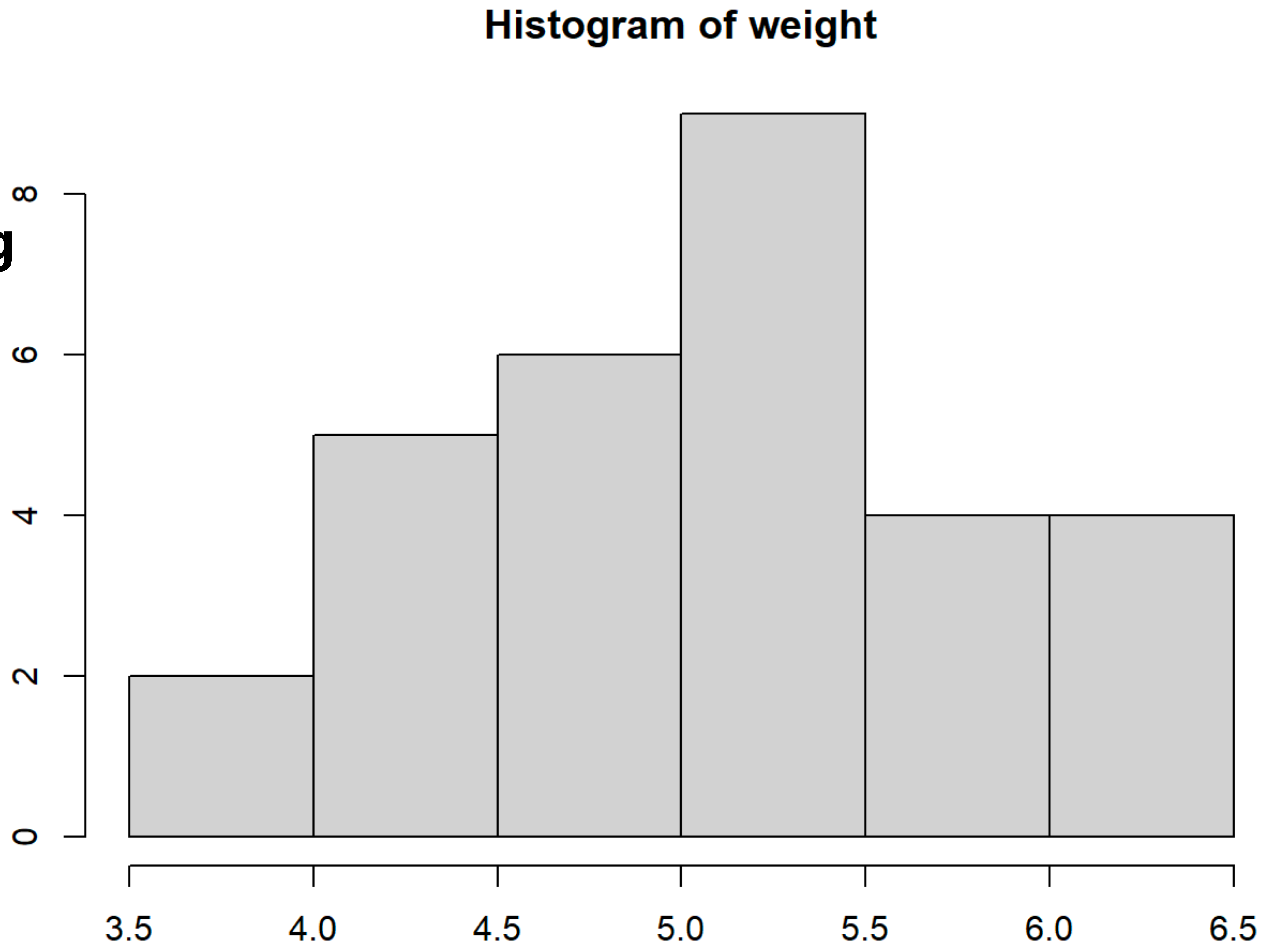
10	0%
7.5	0%
5	0%
40	0%

One batch from a manufacturing process had 15 items, with a mean weight of 10kg. A second batch had 10 items, with a mean weight of 9kg. What is the total weight of the two batches?

150kg	0%
240kg	0%
90kg	0%
9.6kg	0%

Example

Find the (approximate) mean from the following histogram



Example

Find the (approximate) mean from the following histogram

3.75	2	**5.25**	9
4.25	5	**5.75**	4
4.75	6	**6.25**	4

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.

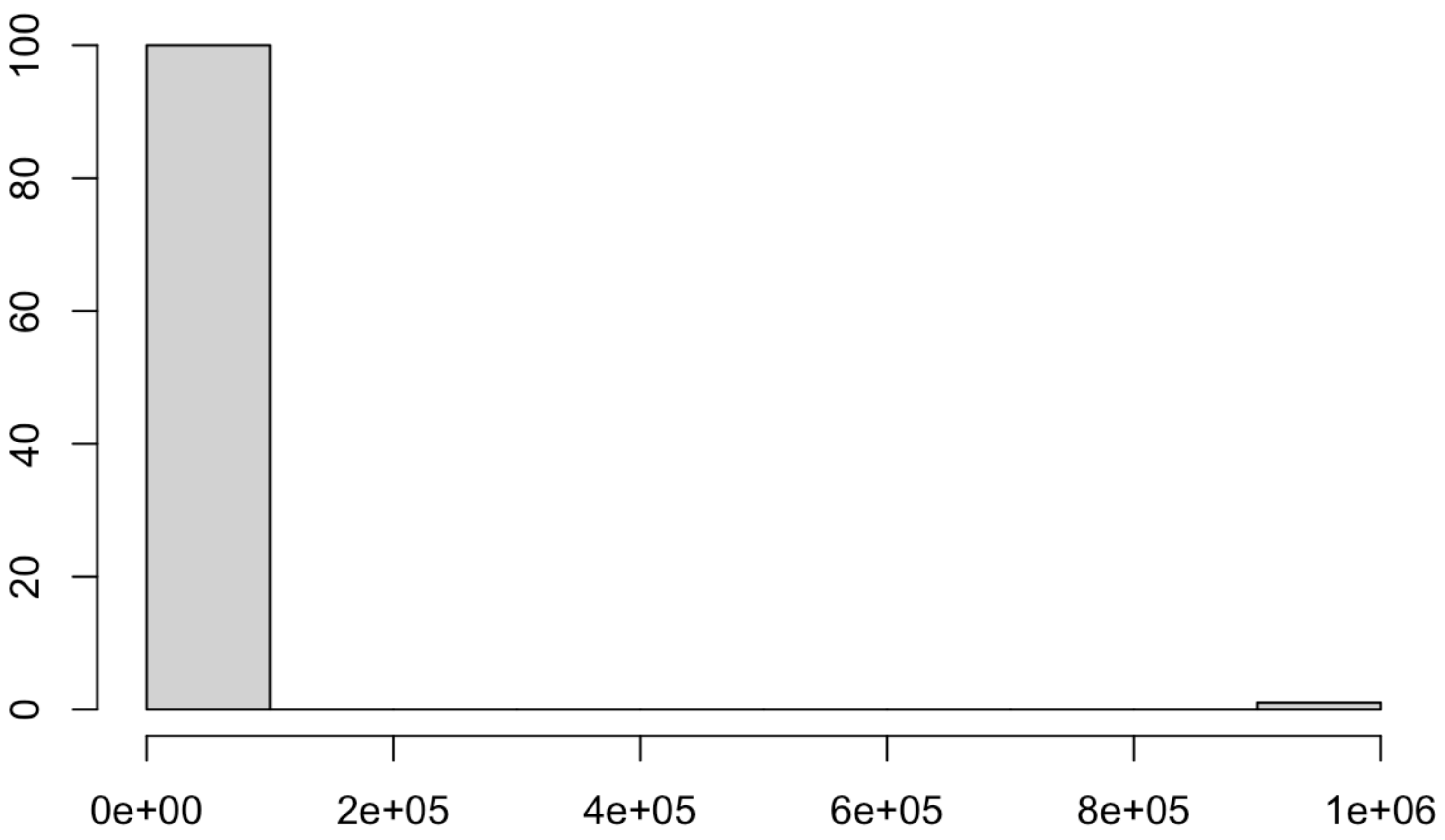
Mean versus Median

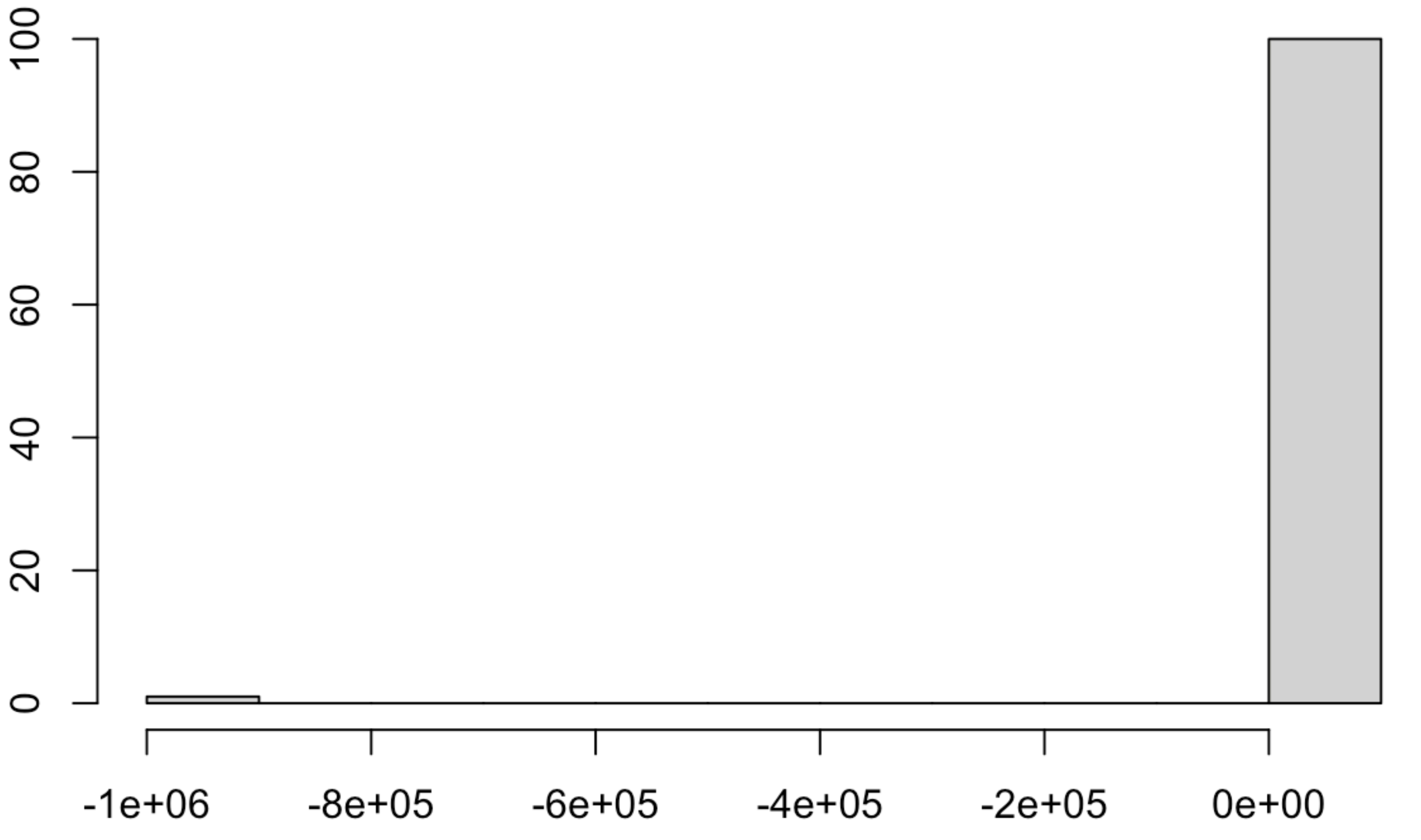
- When data are approximately symmetric, the mean and median will be similar.
- If data are skewed, the mean is *pulled* towards the long tail of the distribution.
  - In this way, the mean is more sensitive to *skewed* outliers than the median.
- We generally prefer the median if data are skewed, and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

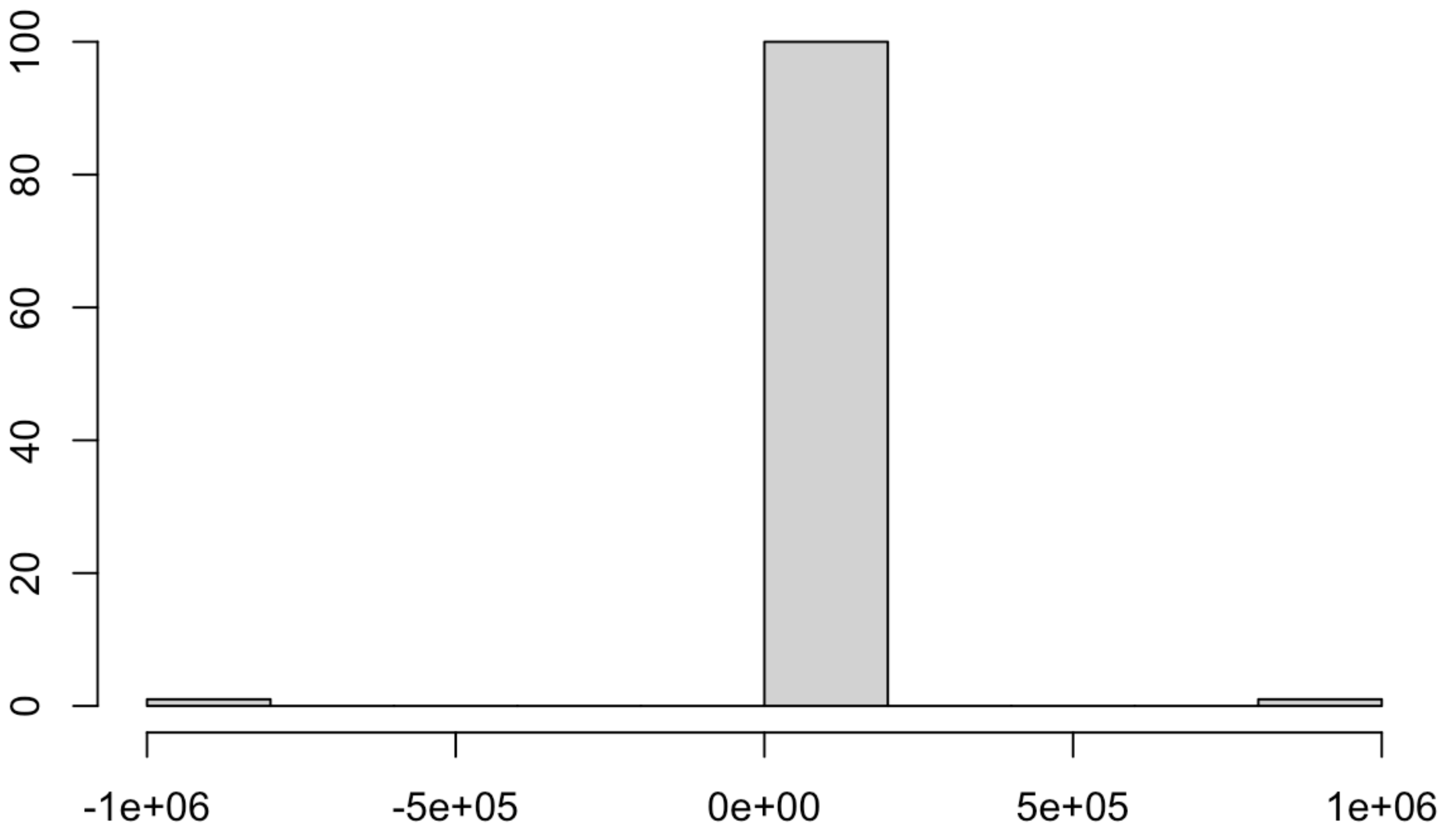
Example

Consider the mean and median of the following three sets of numbers (without calculation)

1, 1000000	1, -1000000	1, 1000000, -1000000







If you wanted a measure of location for household income, would you prefer the mean or median?

It does not matter since there will be no meaningful difference between them. 0%

The mean since the data are likely to be right skewed. 0%

The median since the data are likely to be right skewed. 0%

The median since the data are likely to have outliers. 0%

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

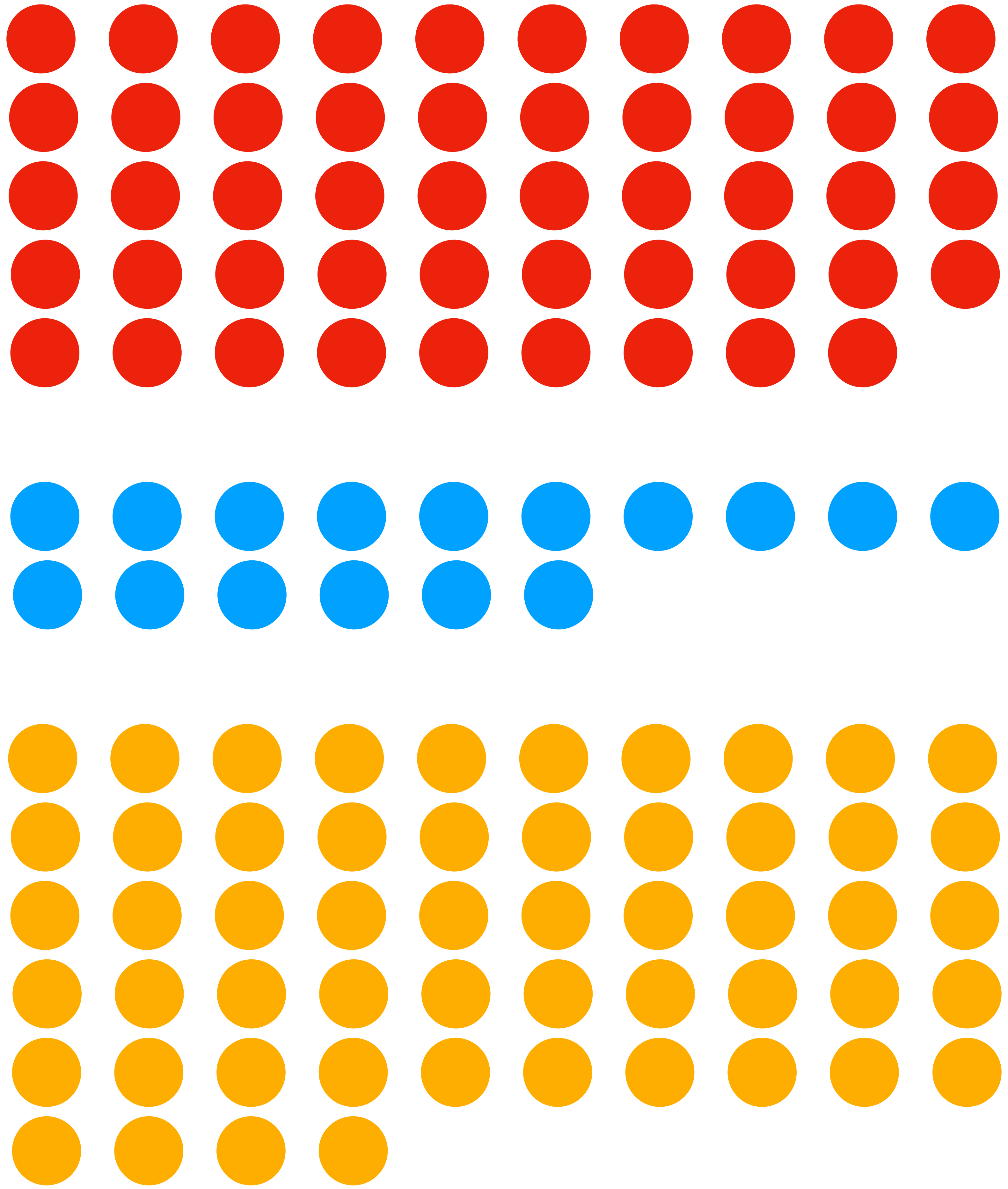
- For categorical data, we can consider the relative frequency of each category as the sample proportion.
- Assume that there are categories, $c_1, c_2, \ldots, c_k$, and observations $x_1, x_2, \ldots, x_k$.
- Define $z_{i,j} = I(x_i = c_j)$, where $I(\cdot)$ is an indicator function.
- Then, we can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

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

Example

Compute the Sample Proportion of 'Red'

	Count
Red	49
Blue	16
Yellow	54
	119



1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1

0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0

0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0
0 0 0 0

$$49 / 119 = 0.41176$$

In the data from Spotify, 550 of the 953 songs were in a major key, the rest were in a minor key. What proportion of songs were in a minor key?

550/953 = 0.5771 — 0%

403/550 = 0.7327 — 0%

403/953 = 0.4229 — 0%

550/403 = 1.3648 — 0%