DESCRIPTIVE STATISTICS

MAT 152 – Statistical Methods I

Lecture 3

Instructor: Dustin Roten

Fall 2020

A Primer on Summation Notation Σ

The "big sigma" typically means "sum" in mathematics.

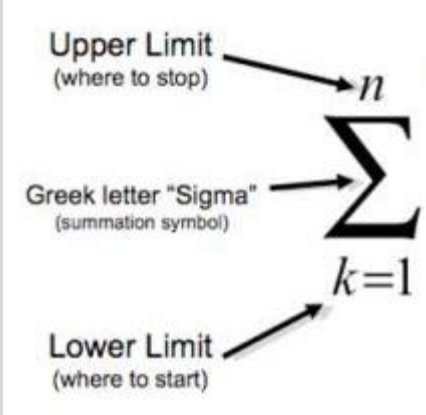
The bottom and top indices tell you where to start and stop.

$$\sum_{\text{start}}^{\text{stop}} \text{argument}$$

$$\sum_{i=2}^{7} i = 2 + 3 + 4 + 5 + 6 + 7 = 27$$

$$\sum_{a=1}^{3} (a+1)^2 = (1+1)^2 + (2+1)^2 + (3+1)^2 = 2^2 + 3^2 + 4^2 = 4 + 9 + 16 = 29$$

$$3 \cdot \sum_{p=1}^{4} \frac{1}{p} = 3 \cdot \left(\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}\right) = 3 \cdot \left(\frac{25}{12}\right) = \frac{75}{12}$$



A Primer on Summation Notation Σ

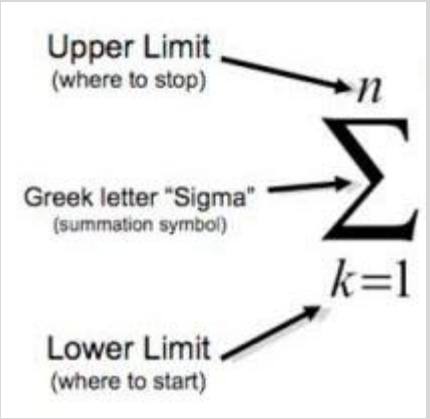
Consider a set of three numbers: $X = \{x_1, x_2, x_3\}$. Here, X is the name of the set and the values are x_k (k is the index).

Let's introduce some numeric values: $X = \{4, 5, 7\}$.
$x_1 = 4$
$x_2 = 5$
$x_3 = 7$

If we want to add up all values, they could be listed out: $x_1 + x_2 + x_3 = 4 + 5 + 7 = 16$.

Or, **summation notation** can be used:

$$\sum_{k=1}^{3} x_k = x_1 + x_2 + x_3$$



Measures of the Center of the Data

The **center** of a data set is also a way of describing location.

Two widely used ways used measures of the "center" of the data are the **mean** (average) and the **median.**

Median

The middle value of the ordered data

$\frac{n+1}{2}$

Mean

The notation that represents the **sample mean** is $\bar{x}$.

The notation that represents the **population mean** is μ.

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

Note: A variable's mean value in a sample will be similar to the population's mean value if the sampling method used was truly random.

Mode

The most frequent value. If two values have the same frequency, the data is said to be **bimodal**.

Calculating the median, mean, and mode

Consider the small **ordered** dataset below.

1, 1, 2, 3, 5

What is the **median** (center value)?

A simple trick for finding the median value is: $\frac{n+1}{2}$. *(n* is the number of values in the dataset.*)*

$\frac{5+1}{2} = \frac{6}{2} = 3$. The third value in the ordered dataset is 2.

What is the **mean**?

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5}(x_1 + x_2 + x_3 + x_4 + x_5) = \frac{1}{5}(1 + 1 + 2 + 3 + 5) = \frac{1}{5}(12) = \frac{12}{5} = 2.4$$

What is the **mode** (most frequent values)?

1

Calculating the Mean from Frequency Tables

Suppose thirty randomly selected students were asked the number of movies they watched the previous week. The results are shown in the relative frequency table below.

# of movies	Relative Frequency
0	$\frac{5}{30}$
1	$\frac{15}{30}$
2	$\frac{6}{30}$
3	$\frac{3}{30}$
4	$\frac{1}{30}$

Determine the mean, median, and mode of the data.

Median: 1

Mode: 1

The Mean can be determined using the frequencies.

$$\bar{x} = \frac{1}{30}\sum_{i=1}^{30} x_i = \frac{1}{30}\left(5(0) + 15(1) + 6(2) + 3(3) + 1(4)\right) \approx 1.33$$

Calculating the Mean from Grouped Frequency Tables

When only grouped data are available, individual data values are not known.

An exact mean cannot be determined.

The mean can be approximated using the information given.

The midpoint of each interval can be found. This value is assumed for all data falling in the interval.

$$\text{midpoint} = \frac{\text{lower value} + \text{upper value}}{2}$$

$$\bar{x}_{mid} = \frac{1}{n}\sum_{i=1}^{n} f_i m_i$$

Calculating the Mean from Grouped Frequency Tables

Midpoint	Grade Interval	Number of Students
53.25	50–56.5	1
59.5	56.5–62.5	0
65.5	62.5–68.5	4
71.5	68.5–74.5	4
77.5	74.5–80.5	2
83.5	80.5–86.5	3
89.5	86.5–92.5	4
95.5	92.5–98.5	1

$$53.25(1) + 59.5(0) + 65.5(4) + 71.5(4) + 77.5(2) + 83.5(3) + 89.5(4) + 95.5(1) = 1460.25$$

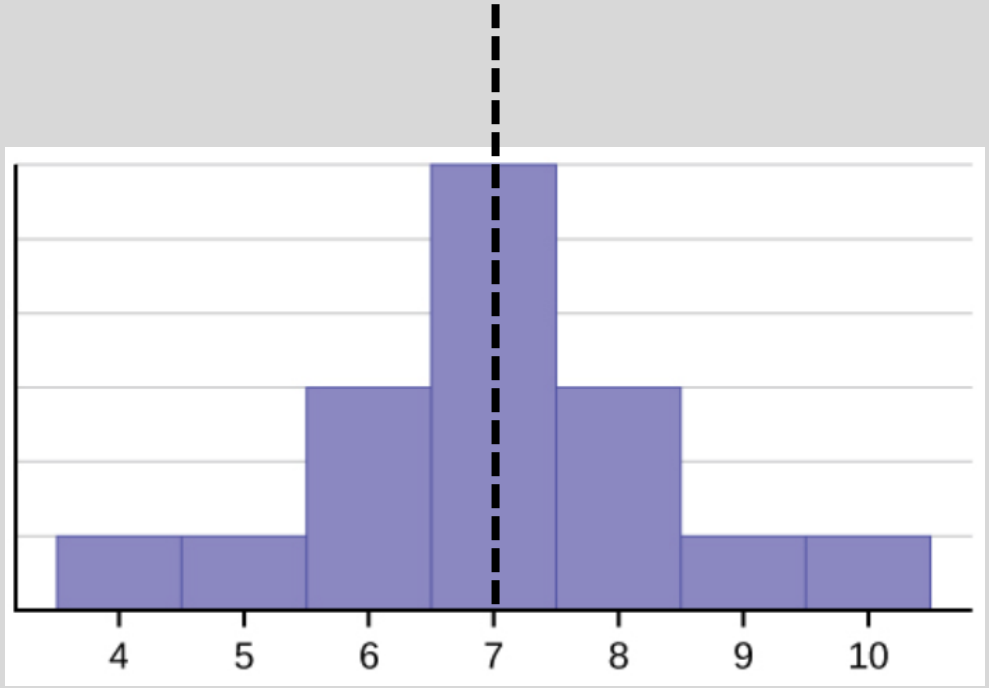
$$\mu = \frac{\sum fm}{\sum f} = \frac{1460.25}{19} = 76.86$$

Each term is the midpoint multiplied by the frequency.

Skewness (Symmetrical Data)

A distribution is **symmetrical** if the upper half of the data has the opposite distribution characteristics as the lower half of the data.

In a perfectly symmetrical distribution, the **mean** and the **median** are the same.

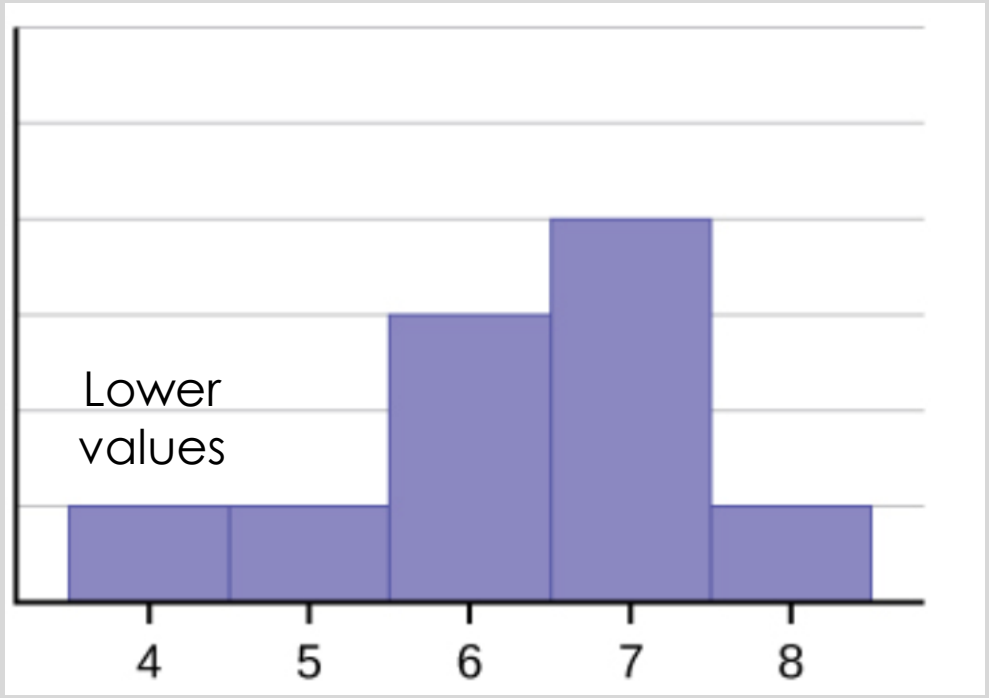


Skewness (Left-Skewed Data)

If there are a handful of values that are considerably smaller than the other data, the distribution is **skewed to the left.**

The **mean** is **sensitive** to outlier values. So, the **mean** is "dragged" down by the lower values in a **left-skewed** dataset. The **median** is unchanged.

$\text{mean} < \text{median} < \text{mode}$

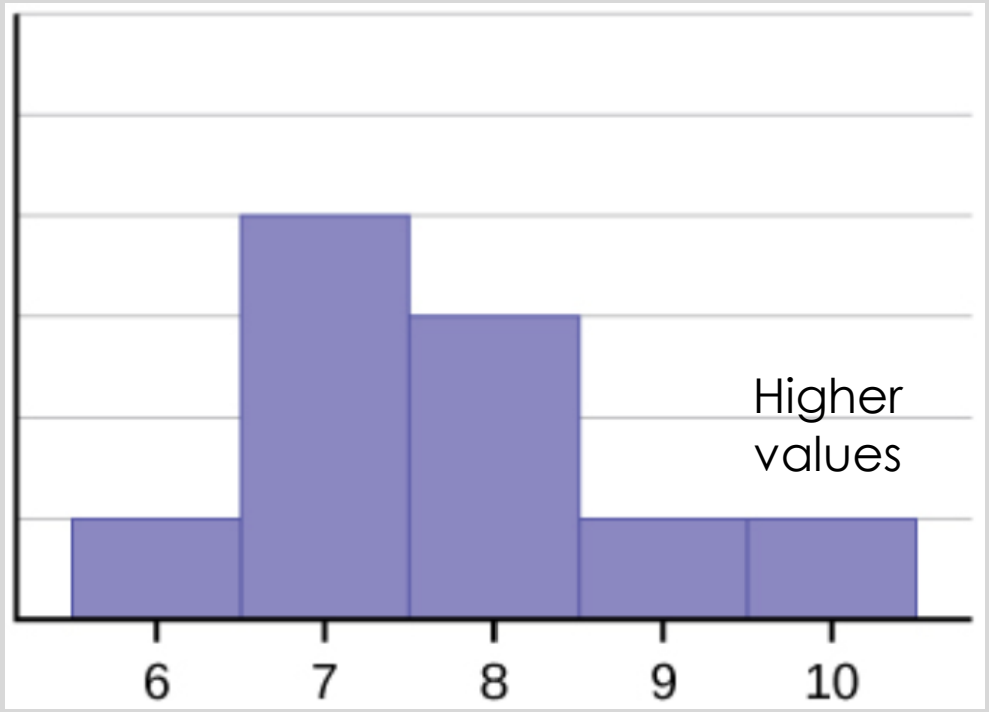


Skewness (Right-Skewed Data)

If there are a handful of values that are considerably larger than the other data, the distribution is **skewed to the right.**

The **mean** is **sensitive** to outlier values. So, the **mean** is "dragged" up by the higher values in a **right-skewed** dataset. The **median** is unchanged.

$\text{mode} < \text{median} < \text{mean}$



A Quick Review

- The **center** of a data set is also a way of describing location.
- Common measures of the center of the data are **mean**, **median**, and **mode**.
- The mean of a dataset can be calculated from its frequency table.
- The mean of a dataset can be *approximated* from its grouped frequency table.
- If a dataset is **symmetrical**, $\text{Mode} \approx \text{Median} \approx \text{Mean}$.
- The **mean** is sensitive to outlier values, the **median** is not.
- If a dataset is **left-skewed**, $\text{Mean} < \text{Median} < \text{Mode}$.
- If a dataset is **right-skewed**, $\text{Mode} < \text{Median} < \text{Mean}$.