BIOS:4120 – Introduction to Biostatistics
Unit 3: Numerical Summary Measures

Knute D. Carter

Department of Biostatistics
The University of Iowa

September 2, 2025

Learning Objectives

At the end of this session, you should be able to:

- Demonstrate understanding of the Sigma summation notation.
- Calculate the population and sample mean and understand the difference between them.
- Find the median and mode of a data set.
- Describe the relationships between mean, median and mode for skewed data;
- Calculate the range, IQR, variance, standard deviation and coefficient of variation for either a population or sample; and
- Explain the property of robustness and identify robust and non-robust measures

Overview

- Measures of Central Tendency
  - Mean
  - Median
  - Mode
- Measures of Dispersion
  - Range and Interquartile Range
  - Variance
  - Standard Deviation

Measures of Central Tendency

Mean / Median / Mode

Measures of Central Tendency

The two most fundamental characteristics of a variable of any data set are:

1) the *center* of the data set; and
2) the *spread* of the data set.

Measures of Central Tendency

- Numbers designed to reflect the center of a data set are called *Measures of Central Tendency*.
- A *Statistic* is a numerical value describing a sample characteristic.
- A *Parameter* is a numerical value describing a population characteristic.

Summation Notation

- Consider a set of n observations denoted as

$$x_1, x_2, \ldots, x_n$$

- To represent the sum

$$x_1 + x_2 + \cdots + x_n$$

we often use an abbreviated *Sigma* (summation) notation:

$$\sum_{i=1}^{n} x_i \quad \text{or} \quad \textstyle\sum_{i=1}^{n} x_i \quad \text{or} \quad \sum x_i$$

Summation Notation Extensions

$$\sum_{i=5}^{7} y_i = y_5 + y_6 + y_7$$

$$\sum_{i=1}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2$$

$$\left(\sum_{k=1}^{4} x_k \right)^2 = (x_1 + x_2 + x_3 + x_4)^2$$

$$\sum_{i=1}^{n} (x_i - c)^2 = (x_1 - c)^2 + (x_2 - c)^2 + \cdots + (x_n - c)^2$$

Summation Notation Extensions

$$\sum_{k=1}^{n} k = 1 + 2 + 3 + \cdots + (n-1) + n = \frac{n(n+1)}{2}$$

$$\sum_{x=0}^{4} 2^x = 2^0 + 2^1 + 2^2 + 2^3 + 2^4 = 1 + 2 + 4 + 8 + 16$$

$$\sum_{i=1}^{5} 4x_i = 4x_1 + 4x_2 + 4x_3 + 4x_4 + 4x_5$$

$$\sum_{i=1}^{n} 7 = 7 + 7 + 7 + \cdots + 7 = 7n$$

Population Mean

- Consider a population of N observations denoted as

$$x_1, x_2, \ldots, x_N$$

- The *Population Mean* is denoted by μ, (called mu), and is given by

$$\begin{aligned} \mu &= \frac{x_1 + x_2 + \cdots + x_N}{N} \\ &= \frac{1}{N} \sum_{i=1}^{N} x_i \end{aligned}$$

Sample Mean

- Consider a sample of n observations denoted as

$$x_1, x_2, \ldots, x_n$$

- The *Sample Mean* is denoted by $\bar{x}$, (called x bar), and is given by

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

Example of Calculation of Sample Mean

- Sample data: `1, 2, 4, 7, 8, 8`.
- A sample of $n = 6$ numbers.
- The sample mean is:

$$\bar{x} = \frac{(1+2+4+7+8+8)}{6} = \frac{30}{6} = 5$$

Median

- The *Median* of a data set is the 50th percentile: a value which exceeds about half of the observations and is exceeded by about half.
- For an odd number of observations, the median is the middle observation when the data are arranged in (ascending) order.
- For an even number of observations, the median is the mean of the middle two observations when the data are arranged in order.
- Examples:

$$1, 2, 2, 3, 5, 7, 9, 10, 11 \Rightarrow \text{Median } = 5$$

$$1, 2, 2, 3, 5, 6, 7, 9, 10, 11 \Rightarrow \text{Median } = (5+6)/2 = 5.5$$

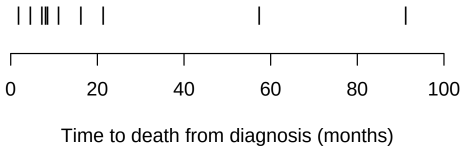
Comparison of Mean and Median

- The mean and the median are both measures of central tendency, but do they always give roughly the same impression of the center of a data set?

Comparison of Mean and Median, Example

- Suppose we had the following sample of survival times (months) of 10 people diagnosed with a fatal disease:

1.80	4.56	7.20	8.04	8.52
11.04	16.20	21.36	57.36	91.20



- Mean = 22.728 months
- Median = (8.52+11.04)/2 = 9.78 months
- Note that in this case the median better reflects the typical survival time.

Comparison of Mean and Median

- The mean is highly sensitive to outliers; the median is not.
- A measure which is not greatly influenced by outliers is called *Robust* or *Resistant*.

Mode

- The *Mode* of a data set is the most frequently occurring value in the data set.
- A data set may have more than one mode. A data set with two modes is said to be *Bimodal*.
- If each of the values in a data set is unique, the mode is said to be undefined.

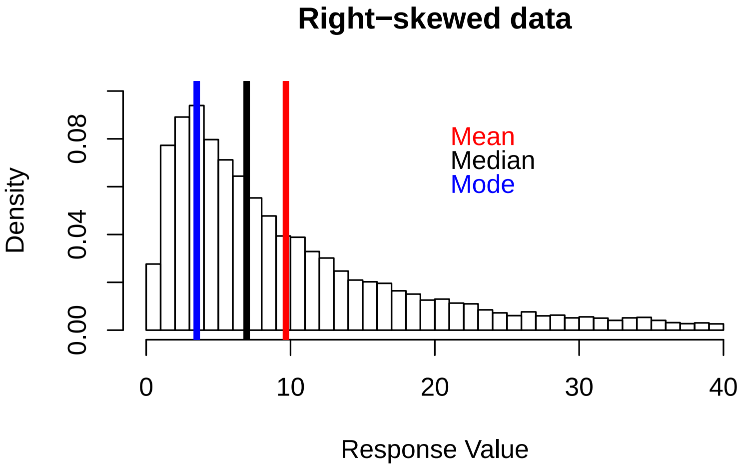
Mode

- The mode can be used as a descriptive measure for nominal or ordinal data.
- The mean and median *cannot* be used for nominal data, even if the categorical values are numerically coded.
- The mean and median can only be used for ordinal data if the categorical values are numerically coded and the coding is sensible.
  - For example, letter grades are often assigned grade points.
  - GPAs are then computed using these points.

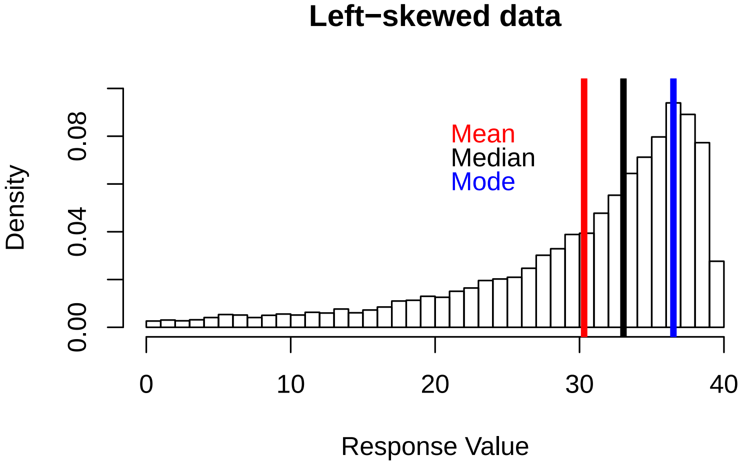
Histograms and Measures of Central Tendency

- The mean, median, and mode of a data set can be estimated based on a histogram constructed from the data set.
- The mode of the histogram is that point along the horizontal axis which corresponds to the histogram's peak (or the midpoint of the interval that features the tallest rectangle).
- The median on a histogram is that point along the horizontal axis which divides the total area of the histogram in half.
- The mean on a histogram is that point along the horizontal axis which corresponds to the histogram's center of gravity (i.e., balancing point).

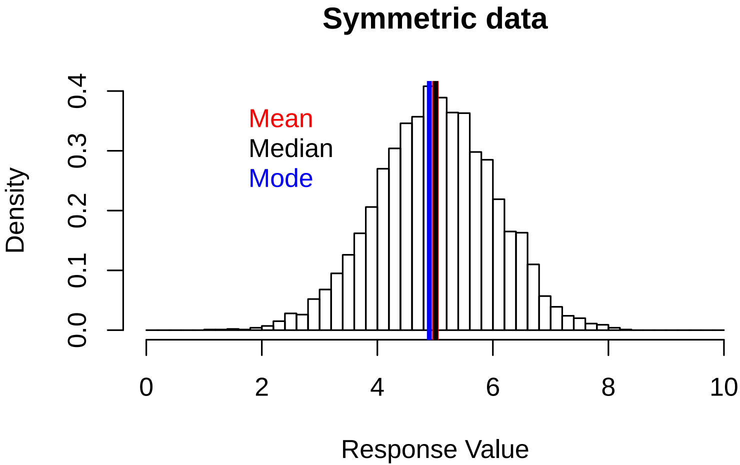
Histograms and Measures of Central Tendency



Histograms and Measures of Central Tendency



Histograms and Measures of Central Tendency



Some Questions

- If you add a constant to all the values in the sample (e.g., add '10' to each one), what will happen to the sample mean? to the median? to the mode?
- If you multiply each observation in the sample by some number (e.g., 100), what will happen to the sample mean? to the median? to the mode? (Note: This is what we often do when we change scale/units - e.g., going from feet to inches or vice versa).
- If the data are skewed right, will the mean or median be larger?
- If the data are skewed left, will the mean or median be larger?
- If the data are both unimodal and skewed right, is the median or mode larger?

Measures of Dispersion

Range and IQR / Variance / Standard Deviation

Measures of Dispersion

- Numbers designed to reflect the degree of spread or variability within a data set are called *Measures of Dispersion*.
- Even if two sets of data have the same mean, median, and mode, we don't know if they have the exact same shape.
- The spread of the data is another aspect to consider.

Range

- The *Range* of a data set is the difference between the largest value and the smallest value: i.e.,

$$\text{maximum} - \text{minimum}$$

- The range is not robust.
- If the largest or the smallest value in the data set is an outlier, the range may provide a misleading indication of spread.

Interquartile Range

Recall from last lecture that:

- the lower quartile, Q_1, is defined as the 25$^{\text{th}}$ percentile; and
- the upper quartile, Q_3, is defined as the 75$^{\text{th}}$ percentile.

Find Q_1 and Q_3 for the following survival data:

1.80	4.56	7.20	8.04	8.52
11.04	16.20	21.36	57.36	91.20

$Q_1 =$

$Q_3 =$

Interquartile Range

- The *Interquartile Range* (IQR) of a data set is the difference between the 75th percentile (third quartile, Q_3) and the 25th percentile (first quartile, Q_1): i.e., $Q_3 - Q_1$.
- The IQR is the length of the interval that captures the middle 50% of the observations.
- The IQR is robust (or resistant).
- The survival data:

```
 1.80   4.56   7.20   8.04   8.52
11.04  16.20  21.36  57.36  91.20
```

IQR $=$

Variance and Standard Deviation

- The *Variance* of a data set is essentially the average of the squared differences between each data value and the mean.
- The *Standard Deviation* (SD) is the square root of the variance.

Variance and Standard Deviation

- Consider a population of N observations denoted as

$$x_1, x_2, \ldots, x_N$$

- The *Population Variance* is denoted by σ^2, (sigma squared), and is given by

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

- The *Population Standard Deviation* is denoted by σ, and is given by $\sqrt{\sigma^2}$.

Variance and Standard Deviation

- Consider a sample of n observations denoted as

$$x_1, x_2, \ldots, x_n$$

- The *Sample Variance* is denoted by s^2, and is given by

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- The *Sample Standard Deviation* is denoted by s, and is given by $\sqrt{s^2}$.

Variance and Standard Deviation, Example

- Compute the sample mean, variance, and standard deviation of the following sample of diastolic blood pressure readings (in mm Hg): `65, 74, 82, 68, 78` $\Rightarrow \bar{x} = 73.4$ mm Hg.

x_i	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
65	-8.4	70.56
74	0.6	0.36
82	8.6	73.96
68	-5.4	29.16
78	4.6	21.16
Total	0.0	195.20

- $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{5-1} 195.2 = 48.8 \text{ mm}^2 \text{ Hg}$
- $s = \sqrt{48.8} = 6.9857$ mm Hg

Properties of the Standard Deviation

- The SD measures spread about the mean, and should only be used when the mean is chosen to reflect the center of the data set.
- SD $= 0$ only when all of the observations in the data set are the same.
- The SD is not robust.

Coefficient of Variation

- The *Coefficient of Variation* (CV) is the ratio of the standard deviation to the mean, multiplied by 100.
- For the population, the CV would be given by

$$CV = \frac{\sigma}{\mu} \times 100$$

- For the sample, the CV would be given by

$$CV = \frac{s}{\bar{x}} \times 100$$

- The coefficient of variation is a unitless quantity. It is therefore useful for comparing relative variation in different data sets.

Coefficient of Variation

- For example, one might wish to compare the relative variation of body weights in three species: mice, chimpanzees, and humans.
- The means and standard deviations of body weights would invariably be largest for humans, followed by chimpanzees, followed by mice.
- The SD's would be inappropriate for comparing relative variation.
- The CV's scale each SD by the corresponding mean, adjusting for the innate differences in body sizes among the three species.

Learning Objectives

At the end of this session, you should be able to:

- Demonstrate understanding of the Sigma summation notation.
- Calculate the population and sample mean and understand the difference between them.
- Find the median and mode of a data set.
- Describe the relationships between mean, median and mode for skewed data;
- Calculate the range, IQR, variance, standard deviation and coefficient of variation for either a population or sample; and
- Explain the property of robustness and identify robust and non-robust measures