# Statistical Methods for the Biological, Environmental, and Health Sciences

STAT 007

# Describing, Exploring, and Comparing Data

Chapter 3

# Measure of Center

Section 3-1

- In this section we will:
  - Discuss different measurements of center of the data such as mean, median, mode, and midrange.

- In this section we focus in obtaining values that measure the **center** of a data set.
- A **measure of center** is a value at the center or middle of a data set.
- There are different approaches for measuring the center of the data set.
- As measures of center of data we will discuss: the *mean*, the *median*, the *mode*, and the *midrange*.
- All these measures of center, when computed for a sample from the population, are statistics.
- Critical Thinking: Before computing a measure of center, always ask yourself whether it makes sense to do that. The answer will depend on the type of data you are summarizing.

# Measures of Center: Mean

- **Mean**: is the measure of center found by adding all of the data values and dividing the total by the number of data values.
- Mean computed for a sample from the population is denoted $\overline{x}$ (pronounced x-bar) also called sample mean. $\overline{x} = \frac{x_1+x_2,+\ldots+x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$,
- Mean computed for the entire population is denoted $\mu$ (Greek letter mu) also called population mean. $\mu = \frac{x_1+x_2,+\ldots+x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$
- Properties: (i) Sample means drawn from the same population tend to vary less than other measurements of center. (ii) The mean of a data set uses every data value. (iii) One extreme value can change the value of the mean substantially.

**Example**

Consider the following data of IQ scores:
`96; 87; 101; 103; 127; 96; 88; 85; 97; 124`
The sample mean is

$$\overline{x} = \frac{96 + 87 + 101 + 103 + 127 + 96 + 88 + 85 + 97 + 124}{10} = 100.4$$

# Measures of Center: Median

- **Median**: is the measure of center that is the *middle* value when the original data values are arranged in order of increasing (or decreasing) magnitude.
- Median computed for a sample from the population is denoted $\tilde{x}$ (pronounced x-tilde), $M$ or Med.
- Median computed for the entire population is called population median.
- Properties: (i) The median does not directly use every data value. (ii) The median does not change by large amounts when we include just few extreme values

**Example**

Consider the following data of IQ scores: `96; 87; 101; 103; 127; 96; 88; 85; 97; 124`
Sorted values: `85; 87; 88; 96; 96; 97; 101; 103; 124; 127`
The sample median is

$$\tilde{x} = \frac{96 + 97}{2} = 96.5.$$

# Measures of Center: Mode

- **Mode**: is the value(s) that occurs with the greatest frequency.
- Only measure of center that can be used for categorical data. Not so used for quantitative data.
- Properties: (i) Can be computed for categorical data. (ii) A data set can have *no mode* (no data value is repeated), *one mode*, be *bimodal* (two modes), or *multimodal* (more than two modes).

**Example**

Consider the following data of IQ scores: 96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Sorted values: 85; 87; 88; 96; 96; 97; 101; 103; 124; 127
The sample mode is 96.

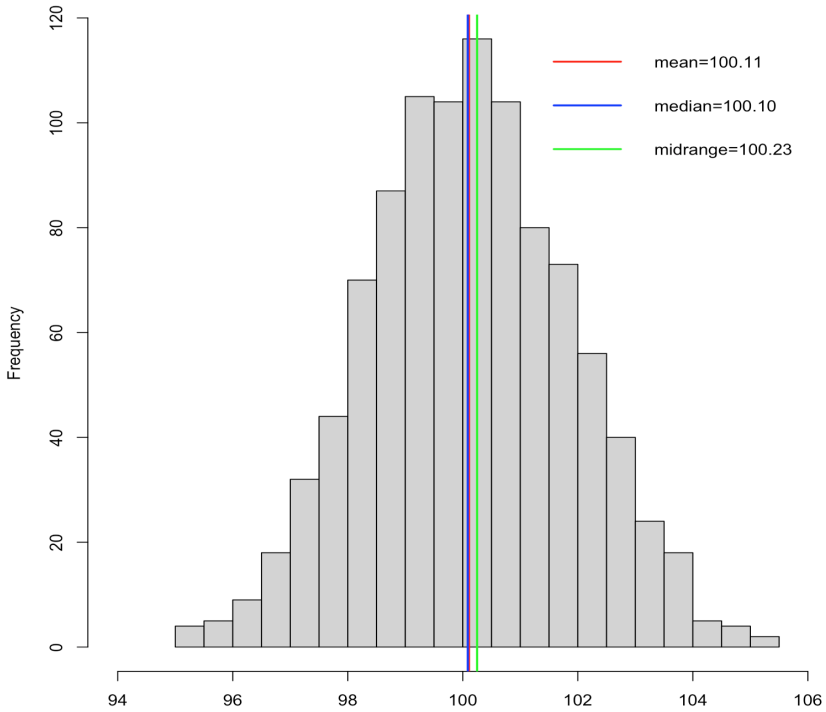
# Measures of Center: Midrange

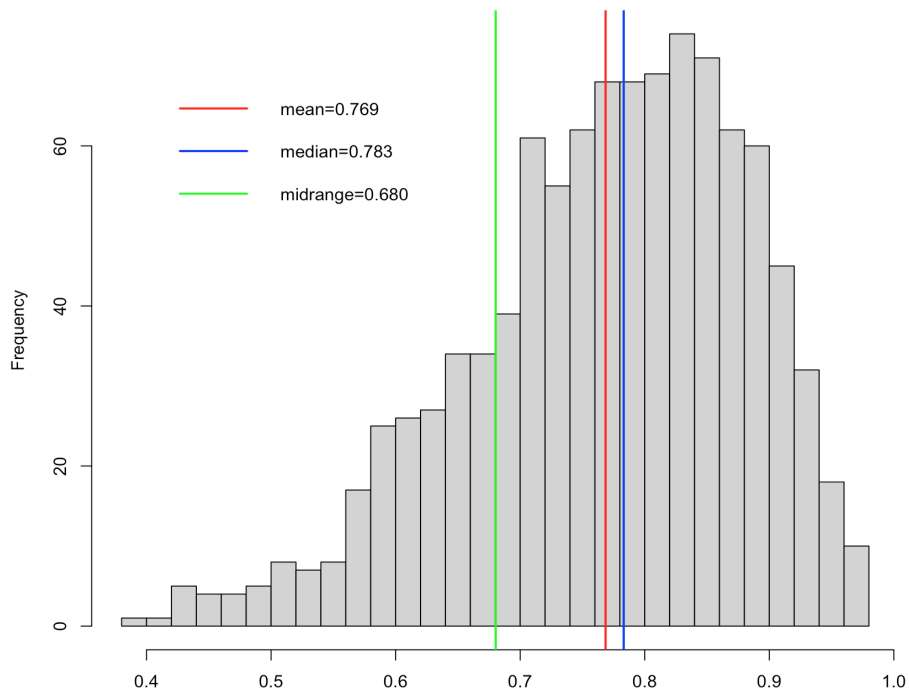
- **Midrange**: is the measure of center that is the value midway between the maximum and minimum values in the original data set.
- $\textit{Midrange} = \frac{\text{maximum data value} + \text{minimum data value}}{2}$.
- Properties: (i) Very sensitive to extreme values. (ii) Very easy to compute. (iii) Reinforces the fact that there are many measures of center.

**Example**

Consider the following data of IQ scores: 96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Sorted values: 85; 87; 88; 96; 96; 97; 101; 103; 124; 127
The sample midrange is $\frac{85+127}{2} = 106$.

# Measures of Center



# Practice

Look at the exercises at the end of Section 3-1 in page 85.

Specially, look at exercises:
3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 15, 16, 17, 18, 19, 20.

# Measures of Variation

Section 3-2

- In this section we will:
  - Discuss different measurements of variation of the data such as range, standard deviation and variance.
  - Introduce a rule of thumb to identify significantly low or large values in a data set.
  - Discuss how to compare variation in different data sets.

- In this section we focus in obtaining values that measure the **variation** of a data set.
- As measures of variation of data we will discuss: the *range*, the *standard deviation*, and the *variance*.
- All these measures of variation, when computed for a sample from the population, are statistics.

# Measures of Variation: Range

- **Range**: is the difference between the maximum data value and minimum data value.
- *Range* = maximum data value − minimum data value.
- Properties: (i) Very sensitive to extreme data values. (ii) Because its value is only based on the minimum and maximum, it does not reflect the variation among all data values.

**Example**

Consider the following data of IQ scores:
`96; 87; 101; 103; 127; 96; 88; 85; 97; 124`
The range is $127 - 85 = 42$.

# Measures of Variation: Standard Deviation

- **Standard Deviation**: is a measure of how much data values deviate away from the mean.
- Standard deviation computed for a sample from the population is denoted $s$ and is also called sample standard deviation. $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$.
- Standard deviation computed for the entire population is denoted $\sigma$ (Greek letter sigma) and is also called population standard deviation. $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$.
- Properties: (i) $s$ measures how much data values deviate from the mean. (ii) $s$ is always positive and equal to zero when all data values are exactly the same. (iii) Larger values of $s$ indicate greater amounts of variation. (iv) Is sensible to outliers. (v) The units of $s$ are the same as the units of the original data values. (vi) $s$ does not center around the population standard deviation (biased estimator).

# Measures of Variation: Standard Deviation

- **Range Rule of Thumb for Identifying Significant Values**: is a crude but simple tool for understanding and interpreting standard deviation.
- Is based on the principle that the vast majority ( 95%) of the sample values lie within 2 standard deviations of the mean.

| | based on parameters | based on statistics |
|---|---|---|
| Significantly low | lower than $\mu - 2\sigma$ | lower than $\overline{x} - 2s$ |
| Significantly high | higher than $\mu + 2\sigma$ | higher than $\overline{x} + 2s$ |
| Not Significant | between $(\mu \mp 2\sigma)$ | between $(\overline{x} \mp 2s)$ |

**Example**

Consider the following data of IQ scores:
96; 87; 101; 103; 127; 96; 88; 85; 97; 124. The sample standard deviation is

$$s = \sqrt{\frac{(96 - 100.4)^2 + (87 - 100.4)^2 + (101 - 100.4)^2 + \ldots + (97 - 100.4)^2 + (124 - 100.4)^2}{10 - 1}} = 14.50$$

Note that $\overline{x} - 2s = 71.40$ and that $\overline{x} + 2s = 129.401$. Discuss whether there are significantly low or high values.

# Measures of Variation: Variance

- **Variance**: is a measure of variation equal to the square of the standard deviation.
- Variance computed for a sample from the population is denoted $s^2 =$ square of $s$.
- Variance computed from the entire population is denoted $\sigma^2 =$ square of $\sigma$.
- Properties: (i) The units of variance are the squares of the units of the original data values. (ii) Is sensible to outliers. (iii) is always positive and equal to zero when all data values are exactly the same. (iv) The variance centers around the population variance (unbiased estimator).

# Comparing Variation in Different Samples

- When comparing the standard deviation from two samples, the sample means should be similar and the units of measurement should be the same.
- In contrary case, use the coefficient of variation.
- The **Coefficient of Variation** for a set of nonnegative sample data describes the standard deviation relative to mean.
- The coefficient of variation is denoted $CV$ and is given by $CV = \frac{s}{\bar{x}} * 100\%$.
- Higher values of $CV$ for different samples indicate higher degrees of variation in the samples.

# Practice

Look at the exercises at the end of Section 3-2 in page 97.

Specially, look at exercises:
2, 3, 4, 5-16, 17-20.

# Measures of Relative Standing

Section 3-3

- In this section we will:
  - Introduce a standardized score to compare data sets and a rule of thumb to identify significantly low or large values in a data set.
  - Define Outliers and resistant statistics.

- Measures of relative standing describe the location of data values *relative* to other values within the same data set.
- Here we discuss *z*Scores as a measure of relative standing.
- A *z*Score for a value $x$ is found by converting that value to a standardized scale.
- *z*Scores allow us to compare values from different data sets.
- Other measures of position for comparing values within the same data set or between different data sets are quartiles.
- Quartiles are also used to identify potential outlier values.

# *z*Scores

- ***z*Scores** (or standard score or standard value): is the number of standard deviations that a given value $x$ is above or below the mean.
- The *z*Score can be computed for a sample: $z = \frac{x-\bar{x}}{s}$ or a population: $z = \frac{x-\mu}{\sigma}$.
- Properties: (i) A *z*score is the number of standard deviations that a given value $x$ is above or below the mean; (ii) *z*Scores are expressed as numbers with no units of measurement; (iii) A data value is *significantly low* if its $z$ score is less than or equal to $-2$ or the value is *significantly high* if its $z$ score is greater than or equal to $+2$ (iv) If an individual data value is less than the mean, its corresponding $z$ score is a negative number.

**Example**

Consider the following data of IQ scores:
96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Compute the *z*Scores of each data value and determine whether there are significantly low or high values.
-0.30 -0.92 0.04 0.18 1.83 -0.30 -0.86 -1.06 -0.23 1.63

# Outliers

- **Outlier**: is a sample value that lie very far away from the vast majority of the other values in a set of data.
- Outliers can strongly affect the values of some important statistics, such a the mean and standard deviation.
- A statistic is **resistant** if the presence of extreme values or outliers does not cause it to change very much.
- $Q_1$, $Q_2$, and $Q_3$ are the first, second and third **quartiles** of the data set:
  $Q_1$: divides the first 25% of the *sorted* data from the top 75%.
  $Q_2$: divides the first 50% of the *sorted* data from the top 50%.
  $Q_3$: divides the first 75% of the *sorted* data from the top 25%.
- A more specific criteria for identifying outliers:
  values above $Q_3$ by an amount greater $1.5(Q_3 - Q_1)$
  values below $Q_1$ by an amount greater $1.5(Q_3 - Q_1)$

# Outliers

- Computation of quartiles:
  $Q_1$: from the sorted data values, is the one in position $L = \frac{25}{100} * n$.
  $Q_2$: from the sorted data values, is the one in position $L = \frac{50}{100} * n$.
  $Q_3$: from the sorted data values, is the one in position $L = \frac{75}{100} * n$.
  If $L$ is a whole number, the quartile of interest is $\frac{x_L + x_{L+1}}{2}$.
  If $L$ is not a whole number, round $L$ to the next whole number and the quartile is $x_L$.

**Example**

Consider the following data of IQ scores: 96; 87; 101; 103; 127; 96; 88; 85; 97; 124
Note that the ordered data set is: 85; 87; 88; 96; 96; 97; 101; 103; 124; 127;
Are 85 and 127 outliers?

Note that $\frac{25}{100} * 10 = 2.5$, $\frac{50}{100} * 10 = 5$, and $\frac{75}{100} * 10 = 7.5$.

So, $Q_1 = x_3 = 88$, $Q_2 = \frac{x_5 + x_6}{2} = \frac{96+97}{2} = 96.5$, and $Q_3 = x_8 = 103$.

So, outliers are values smaller than $Q_1 - 1.5 * (Q_3 - Q_1) = 88 - 1.5 * 15 = 65.5$ and greater than $Q_3 + 1.5 * (Q_3 - Q_1) = 103 + 1.5 * 15 = 125.5$.

# Practice

Look at the exercises at the end of Section 3-3 in page 112

Specially, look at exercises:
1, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14, 15, 16