Random Variable and its measurement

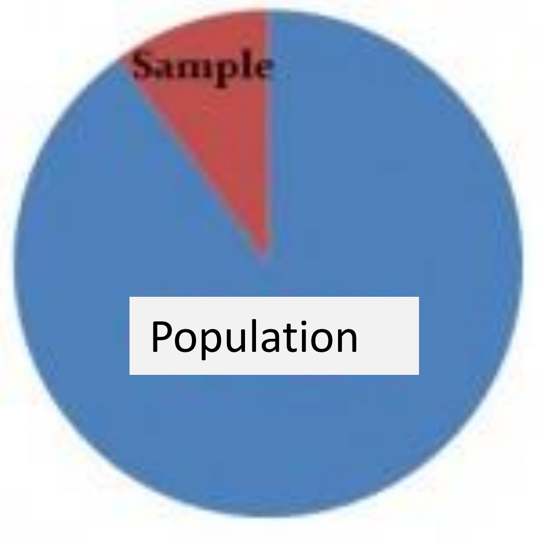
Outline

- Variable and its types
- Percentile
- Mean or expected value
- Variance
- Other measurement of random variables

Populations and Samples

- The study of statistics revolves around the study of data sets

- Two important types of data sets - **populations** and **samples**.



- A population includes all of the elements from a set of data.
- A sample consists one or more observations drawn from the population.

Variable

- In statistics, a **variable** has two defining characteristics:

✓A variable is an attribute that describes a person, place, thing, or idea.

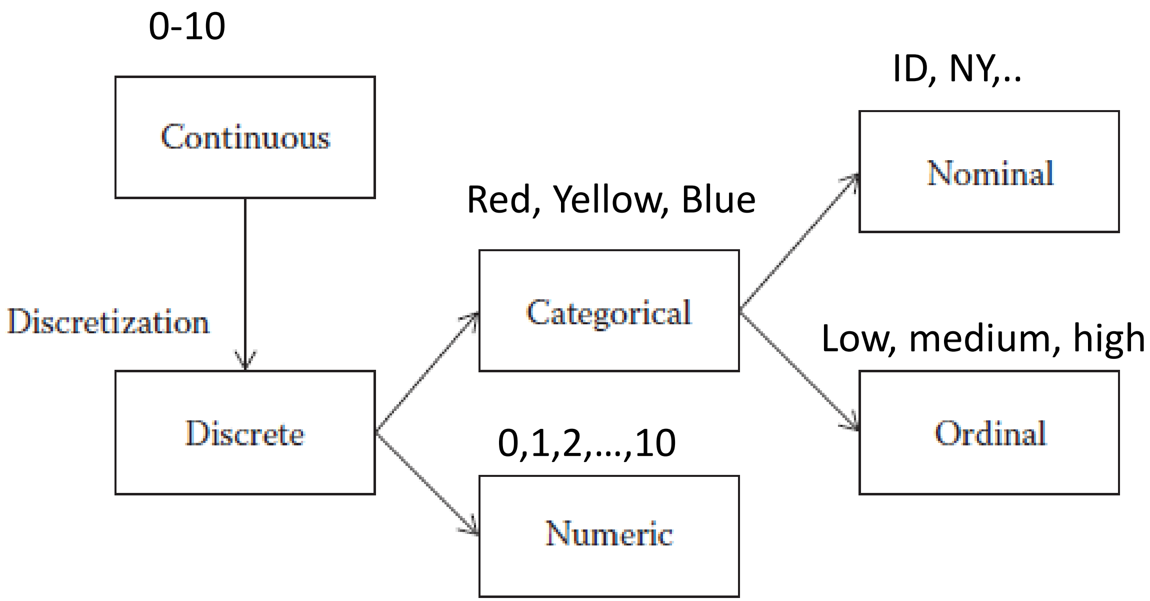
✓The value of the variable can "vary" from one entity to another.

- For example, suppose we let the variable *x* represent the color of a person's hair. The variable *x* could have the value of "blond" for one person, and "brunette" for another.

Types of Variables

A **variable:**

represents a characteristic of an object or a system that we intend to measure or to assign values, and of course, that varies



0-10

Continuous

Discretization

Discrete

Red, Yellow, Blue

Categorical

0,1,2,…,10

Numeric

ID, NY,..

Nominal

Low, medium, high

Ordinal

FIGURE 1.1 Simple classification of variables according to the nature of the values.

Independent vs. Dependent Variables

- The independent forces or drives the dependent, as in cause–effect, or factor–consequence.
- In some cases, this distinction is clear and, in other cases, it is an arbitrary choice to establish the predictor of a dependent variable Y based on the independent variable X.
- As the popular cautionary statement warns us, we could draw wrong conclusions about cause and effect when the quantitative method employed can only tell about the existence of a relationship.

- **Qualitative variable** – a variable that differs in kind.
- **Quantitative variable** – a variable that differs in amount.

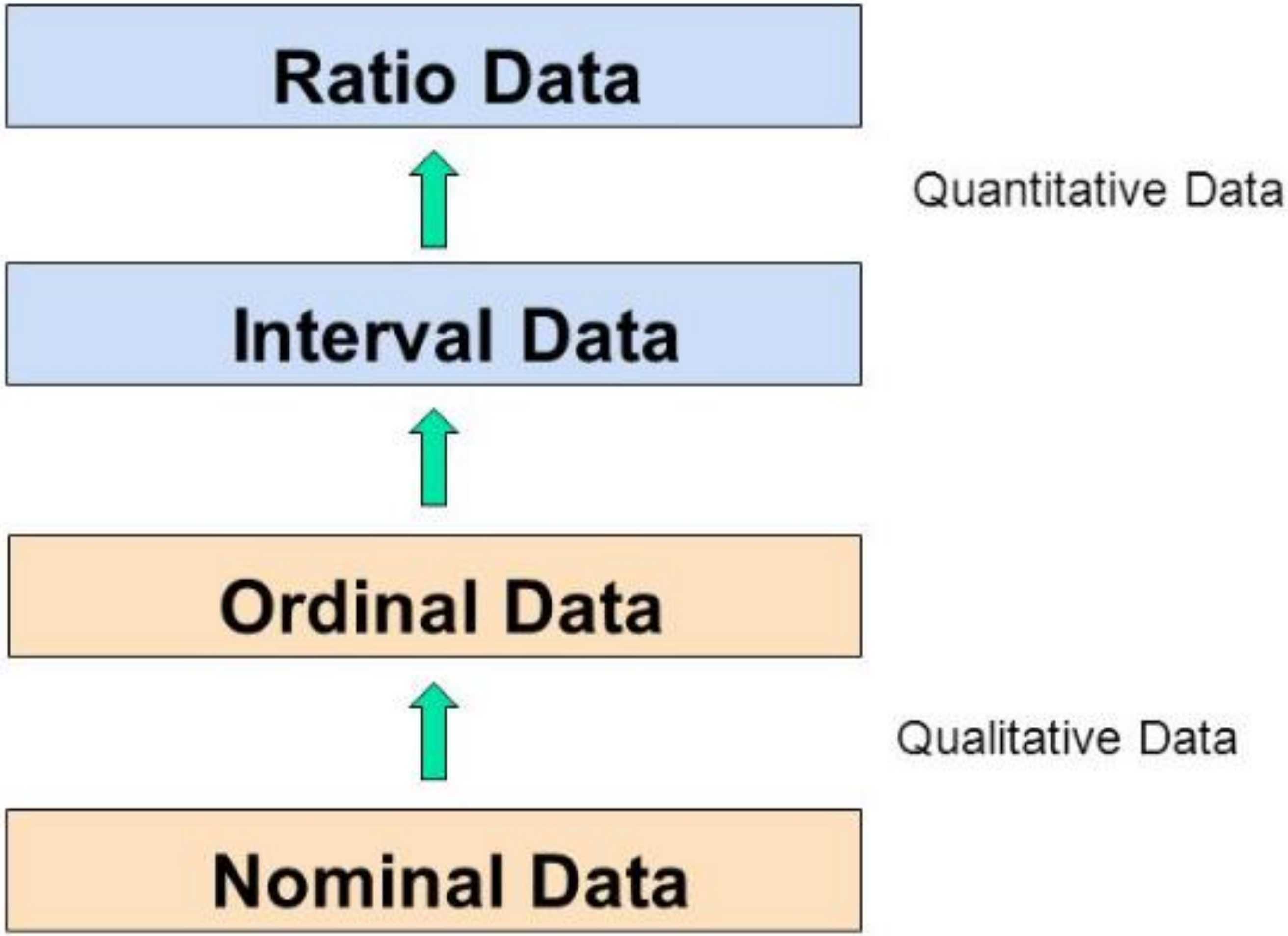
Random Variable

- When the value of a variable is the outcome of a statistical experiment , that variable is a **random variable**.
- Just like variables from a data set, random variables are described by measures of central tendency (like the mean) and measures of variability (like variance).

Scales of Measurement

- Used to classify variables into 4 different levels, or scales, depending on their mathematical properties.
- The type of scale determines the type of statistical operations that can be used to measure the variable.
- Each subsequent scale adds additional information to the previous scale.

Types of Scales of Measurements



Ratio Data

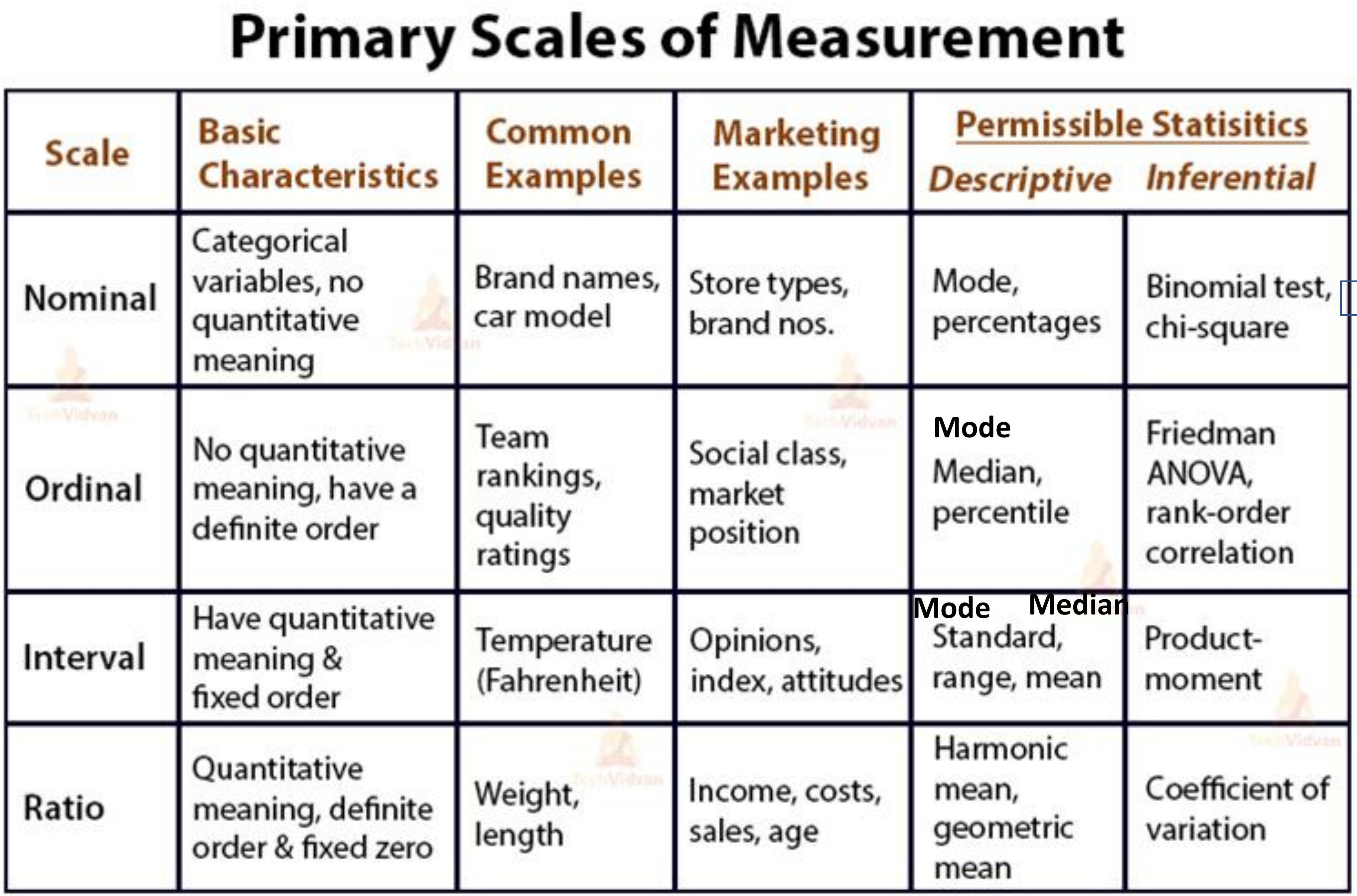
Quantitative Data

Interval Data

Ordinal Data

Qualitative Data

Nominal Data



Primary Scales of Measurement

Scale	Basic Characteristics	Common Examples	Marketing Examples	Permissible Statisitics	
				Descriptive	*Inferential*
Nominal	Categorical variables, no quantitative meaning	Brand names, car model	Store types, brand nos.	Mode, percentages	Binomial test, chi-square
Ordinal	No quantitative meaning, have a definite order	Team rankings, quality ratings	Social class, market position	**Mode** Median, percentile	Friedman ANOVA, rank-order correlation
Interval	Have quantitative meaning & fixed order	Temperature (Fahrenheit)	Opinions, index, attitudes	**Mode Median** Standard, range, mean	Product-moment
Ratio	Quantitative meaning, definite order & fixed zero	Weight, length	Income, costs, sales, age	Harmonic mean, geometric mean	Coefficient of variation

Graphs

Bar
Pie

Bar
Pie
Stem and leaf

Bar
Pie
Stem and leaf
Box plot
Histogram

Histogram
Box plot

Measures of Central Tendency

Learning Objectives

- Compute mean
- Compute median
- Compute mode

Central tendency

- single values that describe the most typical or representative score in an entire distribution.
- the mode, the median, and the mean are the most common.

The Mode

Mode (MO)

- score value with the highest frequency
- arrange the scores in descending order and look for the score with the greatest frequency.

For Example,

In the set of scores below:

73, 73, 72, 70, 68, 68, 68, 68, 59, 59, 59, 55

the Mode is 68 and is written as follows:

MO = 68

In a grouped frequency distribution, the mode would be the **<u>mid-point</u>** of the class interval with the greatest frequency.

For Example,
Mode of a Grouped Frequency Distribution

Notice that the mode is not a frequency, but rather the value that occurs most often.

Class Intervals	Mid-Point	*f*
36 – 38	37	8
33 – 35	34	11
30 – 32	31	18
27 – 29	28	26
24 – 26	25	32
21 – 23	22	20
18 – 20	19	16
15 – 17	16	12
12 – 14	13	7
9 – 11	10	3

The class interval with the most frequently occurring scores is 24 – 26, with a frequency count of 32. The mid-point of that class interval is 25. Thus, MO = 25.

Mode

- ➢ quick and easy, but
- ➢ ignores all other scores
- ➢ not reliable

Consider the following distribution of scores on a 20-point quiz:

8, 8, 8, 11,11, 12, 13, 15, 15, 17, 18, 20, 20, 20, 20.

MO = 20

However, if one student scored 8, rather than 20, the mode would then become 8

The Median

Median (Mdn)

- middle point in a distribution; half of the scores are above this point and half are below it.

Counting Method

- Used for a short list of scores.
- The counting procedure used will differ depending on whether you have an odd or even number of scores.

For an odd number of scores:

- Arrange the scores in descending order from high to low.
- The median will be the score that has an equal number of scores above and below as determined by:

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

For Example,

For the following distribution of an odd number of scores:

26, 25, 24, 20, 18, 17, 17, 15, 12

$$\frac{9+1}{2} = 5$$

Looking for the 5th score. Thus, the median is 18 (i.e., Mdn = 18).

Note that 5 is not the median, but rather the location of the median.

For an **even** number of scores:

- Arrange the scores in descending order from high to low.
- Divide the distribution in half and draw a line between the two scores that separate the distribution into two halves.
- Add the two middle scores that surround the halfway point and divide by 2; resulting value is the median.

For Example,

For the following distribution of an **even** number of scores:

$$92,\ 91,\ 90,\ 90,\ 87,\ \underbrace{82, \mid 77,}_{\textbf{Middle Scores}}\ 75,\ 75,\ 70,\ 68,\ 60$$

$$Mdn = \frac{82 + 77}{2} = 79.5$$

The expected value or mean

- The expected value or mean is a **population** concept. To calculate it we need its density or mass function. The mean is not the same as the **statistic** known as the **"sample mean"** .
- The **sample mean** is the arithmetic average of *n* data values x_i comprising a sample.

The Mean

Mean

- sum total of the scores divided by number of scores.
- *N* (population); *n* (sample)
- μ (population); *M* or $\overline{X}$ (sample)

For a population,

$$\mu = \frac{\sum X}{N}$$

For a sample,

$$\overline{X} = \frac{\sum X}{n}$$

Calculations are the same; only the symbols are different.

For Example,

The mean for the following set of scores from a population:

78, 63, 42, 98, 87, 52, 72, 64, 75, 89

$$\mu = \frac{\sum X}{N} = \frac{720}{10} = 72$$

The mean for the following set of scores from a sample:

3, 8, 6, 9, 10, 17, 5, 8, 1

$$\overline{X} = \frac{\sum X}{n} = \frac{67}{9} = 7.44$$

When to Use Which Measure of Central Tendency

Scale of measurement and shape of the distribution need to be considered.

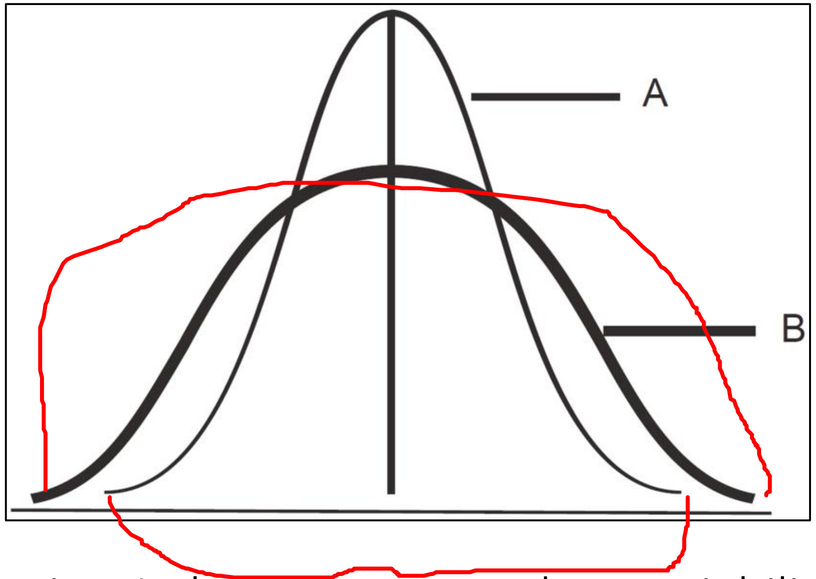
Scale of Measurement

- The **mode** *can* be used for all scales, but is the ***only*** measure of central tendency that can be used for nominal variables.
- The **median** can be used for all scales, except nominal.
- The **mean** can only be used for interval and ratio data.

Variability

Variability

- how much spread the scores have



- Distribution A shows less spread, or variability.

- Distribution B shows a greater amount of spread, or variability.

Three common measures of variability are the **range**, the **interquartile range**, and the **standard deviation**.

Range

The **range (***R***)**

$$R = X_{UL\text{-}High} – X_{LL\text{-}Low}$$

For Example,

For the following set of scores: 3, 3, 5, 7, 7, 7, 8, 9

$R = 9 – 3 = 6$

Definitions of Percentile

- Definition 1 : Using the 65th percentile as an example, the 65th percentile can be defined as the lowest score that is greater than 65% of the scores. This is the way we defined it above and we will call this "Definition 1."
- Definition 2 : The 65th percentile can also be defined as the smallest score that is greater than or equal to 65% of the scores.

- Definition 3: A weighted average of the percentiles computed according to the first two definitions. This third definition handles rounding more gracefully than the other two and has the advantage that it allows the *median* to be defined conveniently as the 50th percentile.

An example of Percentile Calculation

1. Compute the rank (R) of the 25th percentile. This is done using the following formula:

$$R = P/100 \times (N + 1)$$

where P is the desired percentile (25 in this case) and N is the number of numbers (8 in this case). Therefore,

$$R = 25/100 \times (8 + 1) = 9/4 = 2.25.$$

2. If R is an integer, the Pth percentile is the number with rank R. When R is not an integer, we compute the Pth percentile by interpolation as follows:

2.1 Define IR as the integer portion of R (the number to the left of the decimal point). For this example, IR = 2.

2.2 Define FR as the fractional portion of R. For this example, FR = 0.25.

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

Table 1. Test Scores.

Number	Rank
3	1
5	2
7	3
8	4
9	5
11	6
13	7
15	8

- 3. Find the scores with Rank I_R and with Rank $I_R + 1$. For this example, this means the score with Rank 2 and the score with Rank3. *The scores are 5 and 7.*

- 4. Interpolate by multiplying the difference between the scores by F_R and add the result to the lower score. For these data, this is $\underline{(0.25)(7 - 5)} + \underline{5} = \boxed{5.5}$.

Therefore, the 25th percentile is 5.5.

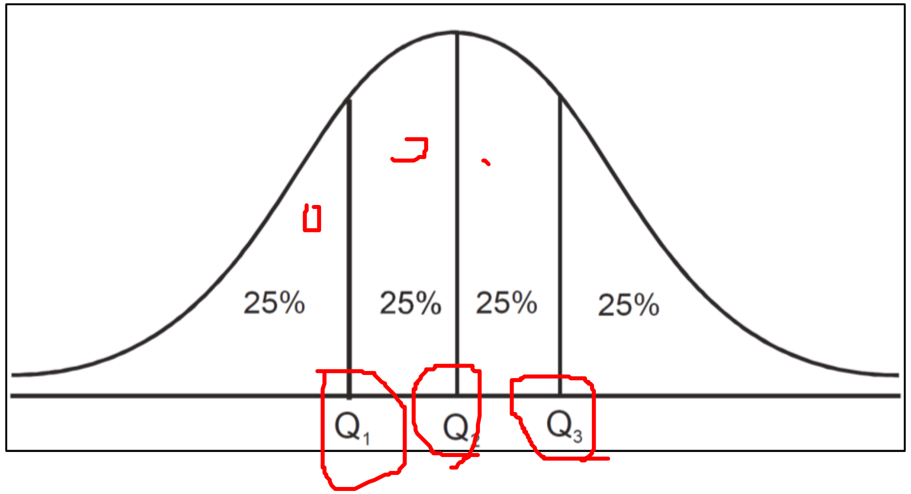
Definition 1: 7 (25th percentile)

Definition 2: 5 (25th percentile)

Interquartile Range

Interquartile range (*IQR)*

- Describes the range of scores from the middle 50% of the distribution.
- Will divide the distribution into four equal parts, producing three **quartiles**.



Q_1 = the point at or below which 25% of the scores lie
Q_2 = the point at or below which 50% of the scores lie
Q_3 = the point at or below which 75% of the scores lie

Steps for Determining the Interquartile Range

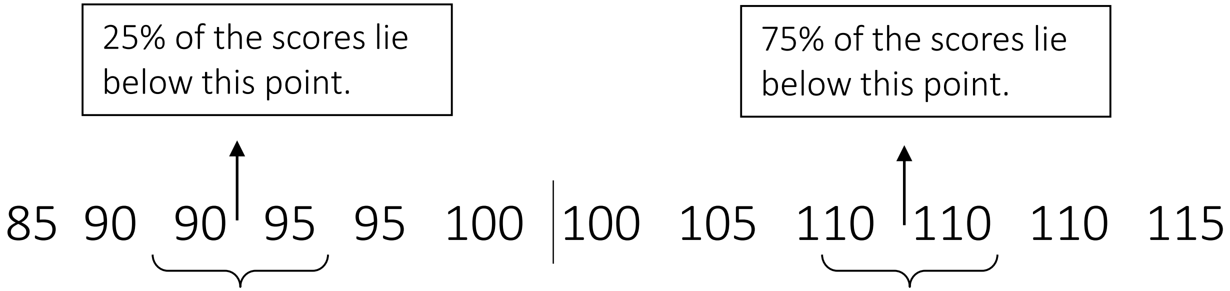
1. Arrange scores in ascending order from low to high.
2. Divide the distribution of scores into four equal parts.
3. Find the points below which 25% of the scores and 75% of the scores lie.
4. Identify the two scores that “bracket” these points.
5. Determine the means of each of these two pairs of scores to determine Q_1 and Q_3.
6. Subtract Q_1 from Q_3.

$$IQR = Q_3 – Q_1$$

For Example,

Compute the interquartile range for the following scores:

85, 115, 90, 90, 105, 100, 110, 110, 95, 110, 95, 100



25% of the scores lie below this point.

75% of the scores lie below this point.

85 90 90 95 95 100 | 100 105 110 110 110 115

$$Q_1 = \frac{90 + 95}{2} = 92.50$$

$$Q_3 = \frac{110 + 110}{2} = 110$$

IQR = *Q*$_3$ – *Q*$_1$
= 110 – 92.50 = 17.50

Second Central Moment or Variance

- The second **central** (i.e., with respect to the mean) moment is the **variance** or the expected value of the square of the difference with respect to the mean

$$\sigma_X^2 = E[\underline{(X-\mu_X)^2}]$$

- If *X* is discrete,

$$\sigma_X^2 = E[(X-\mu_X)^2] = \sum_i \underline{(x_i-\mu_X)^2 p(x_i)}$$

- If *X* is continuous,

$$\sigma_X^2 = E[(X-\mu_X)^2] = \int_{-\infty}^{+\infty} \underline{(x-\mu_X)^2 p(x)dx}$$

- The variance or second central moment is not the same as the **statistic** known as the **“sample variance”**
- The sample variance is the variability measured relative to the arithmetic average of *n* data values x_i comprising a sample, $\bar{X}$ is the mean of the sample

$$\mathrm{var}(X) = s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2$$

- This is the average of the square of the deviations from the sample mean. Alternatively,

$$\mathrm{var}(X) = s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{X})^2$$

where n − 1 is used to account for the fact that the sample mean was already estimated from the n values. We write s_X^2 to denote the sample variance to distinguish from the variance σ_X^2. This equation can be converted in a more practical one by using Equation 3.13 and doing algebra to obtain

$$\mathrm{var}(X) = \boxed{s_X^2} = \underline{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right]}$$

Example: Suppose we get 6 heads in 10 tosses of a coin. The sum of squares is 6 and the sum of the x_i is also 6, then the sample variance is

$$s_X^2 = \frac{1}{9}\times\left[6-\frac{1}{10}\times 6^2\right] = \frac{1}{9}\times[6-3.6] = 0.26$$

The sample standard deviation is $s_X = \sqrt{.26} = 0.509$.

Note that the population variance is 0.25 and standard deviation 0.5 according to calculation in previous exercise. Therefore, the statistic sample variance has overestimated the variance.

The Standard Deviation

- involves first calculating the variance.
- The standard deviation is the square root of the variance.

Definitional formulas

- are written the way that statistics are defined.
- involve more computations but facilitate understanding.

Computational formulas

- are easier to use with a calculator.
- and lead you to the same conclusions.

Definitional Formula for Population Standard Deviation

$$\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$$

- Based on deviation scores ($X - \mu$).
- When the deviation scores are summed, the result will be 0. Square the deviation scores to get a non-zero value.
- After we divide by *N*, we then have to return to our original, unsquared, unit of measurement which is why we get the square root.
- The variance (σ^2) is the value under the square root (average of the squared deviations).
- The value in the numerator is called the sum of squares.

Formula Guide for Definitional Formula:

1. Using the appropriate symbols, create 4 columns as follows:
   - X
   - μ
   - $(X - \mu)$
   - $(X - \mu)^2$
2. List the raw scores under *X*.
3. Calculate the mean (μ) for the second column.
4. Subtract the mean from each raw score to find the deviation score.
5. Square each deviation score and then sum the squared deviations. This value is the numerator in the standard deviation formula. It is also the sum of squares $(SS = \sum(X - \mu)^2)$.
6. Divide *SS* by *N*. This value is the variance, symbolized by σ^2 $\left(\sigma^2 = \frac{SS}{N}\right)$.
7. Obtain the square root of the variance to arrive at the standard deviation. Thus, the following equations are equivalent.

$$\sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}} \qquad \text{and} \qquad \sigma = \sqrt{\frac{SS}{N}}$$

For Example,

X	μ	$(X-\mu)$	$(X-\mu)^2$
17	21.4	-4.4	19.36
24	21.4	2.6	6.76
22	21.4	0.6	.36
26	21.4	4.6	21.16
18	21.4	-3.4	11.56

$\Sigma X = 107$ 0 $\Sigma(X-\mu)^2 = 59.2$

Notice that the sum of the deviation scores equals zero [*(X − μ) = 0*].

$$\sigma = \frac{\sum(X-\mu)^2}{N} = \sqrt{\frac{59.2}{5}} = \sqrt{11.84} = 3.44$$

$SS = 59.2$
$\sigma^2 = 11.84$
$\sigma = 3.44$

Computational Formula for Population Standard Deviation

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma \mathrm{X})^2}{N}}{N}}$$

where: ΣX^2 = sum of the squared raw scores
$(\Sigma X)^2$ = square of the sum of the raw scores

- easier to use
- also called raw score formula

Formula Guide for Computational Formula:

1. Create two columns, X and X^2, and list the raw scores under X.
2. Square the individual raw scores and place these values in the X^2 column.
3. Sum the X column to obtain $\sum X$.
4. Sum the X^2 column to obtain $\sum X^2$.
5. Place these values into the formula along with the appropriate N.
6. Square the sum of the raw scores and divide the result by N to determine $\frac{\Sigma X^2}{N}$.
7. Subtract this result from $\sum X^2$. This value is SS.
8. Divide SS by N. This value is the variance (σ^2).
9. Find the square root of the variance to obtain the standard deviation (σ).

For Example,

X	X^2
17	289
24	576
22	484
26	676
18	324

$\Sigma X = 107$ $\Sigma X^2 = 2349$

$$\sigma = \sqrt{\frac{\Sigma X^2 - \frac{(\Sigma X)^2}{N}}{N}}$$

$$= \sqrt{\frac{2349 - \frac{(107)^2}{5}}{5}}$$

$$= \sqrt{\frac{2349 - 2289.8}{5}}$$

$$= \sqrt{\frac{59.20}{5}}$$

$$= \sqrt{11.84}$$

$$= 3.44$$

$SS = 59.2$
$\sigma^2 = 11.84$
$\sigma = 3.44$