

What is Statistics?

Tony W (FE UNY)

Bahan diadopsi dari Douglas A. Lind, William G. Marchal,
Samuel A. Wathen, Basic Statistics For Business and
Economics, McGraw Hill





Uses of Statistics

- Statistics is one of the tools used to make decisions in business
- We apply statistical concepts in our lives
- As a student of business or economics, basic knowledge and skills to organize, analyze, and transform data and to present the information.



Why Study Statistics?

1. Numerical information is everywhere
2. Statistical techniques are used to make decisions that affect our daily lives
3. The knowledge of statistical methods will help you understand how decisions are made and give you a better understanding of how they affect you.

No matter what line of work you select, you will find yourself faced with decisions where an understanding of data analysis is helpful.



Who Uses Statistics?

Statistical techniques are used extensively by marketing, accounting, quality control, consumers, professional sports people, hospital administrators, educators, politicians, physicians, etc...

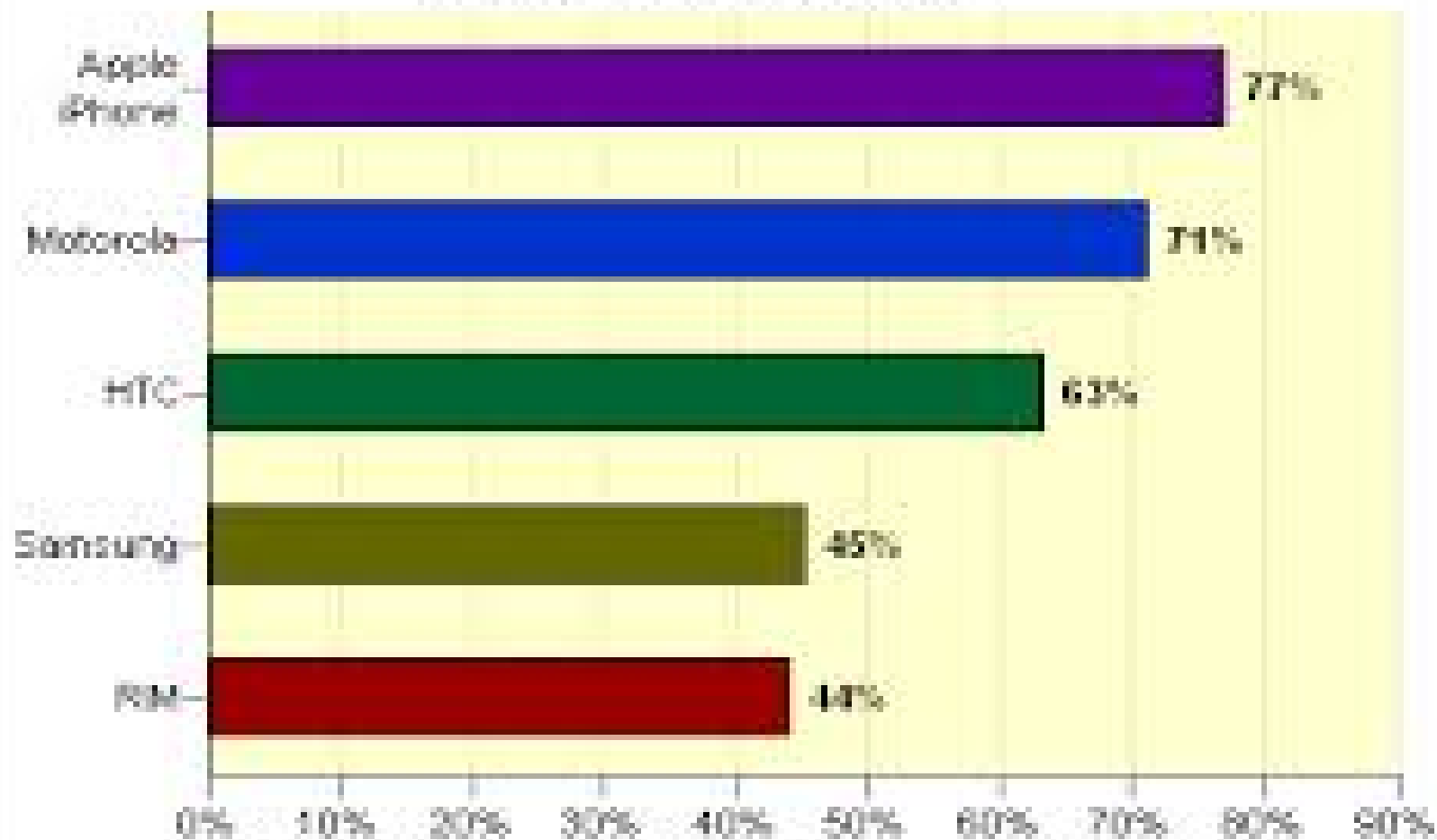
Arsenal	Vs	Liverpool
62.4	Possession	37.6
11	Total Shots	9
5	On Target	5
598	Total Passes	368
466	Accurate Passes	255
78%	Pass Completion %	69%
15	Total Tackles	22
11	Tackles Won	10
10	Corners	3
47	Possession Duels Won	46
13	Aerial Duels Won	10
12	Fouls Won	12

Player	Accurate Passes	Total Passes	Pass Completion %
Reina	17	37	46
Jame Carragher	16	19	84
Martin Skrtel	14	17	82
John Flanagan	23	37	62
Fabio Aurelio	4	4	100
Soto Kyrgiakos	4	5	80
Jack Robinson	12	17	71
Lucas Leiva	31	39	79
Raul Meireles	27	41	66
Jay Sparring	25	30	83
Dirk Kuyt	34	49	69
Jonjo Shelvey	9	12	75
Luis Suarez	34	45	76
Andy Carroll	5	16	31

Satisfaction Rating - By Manufacturer of Smart Phone Purchased Past 6 Months

Nov 2010

Customers Who Say They Are Very Satisfied With the Smart Phone They Purchased in the Past 6 Months

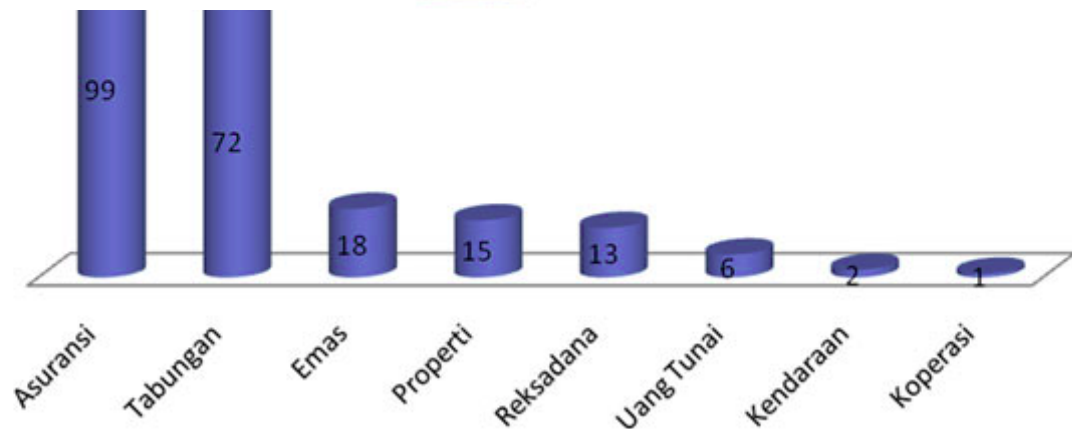
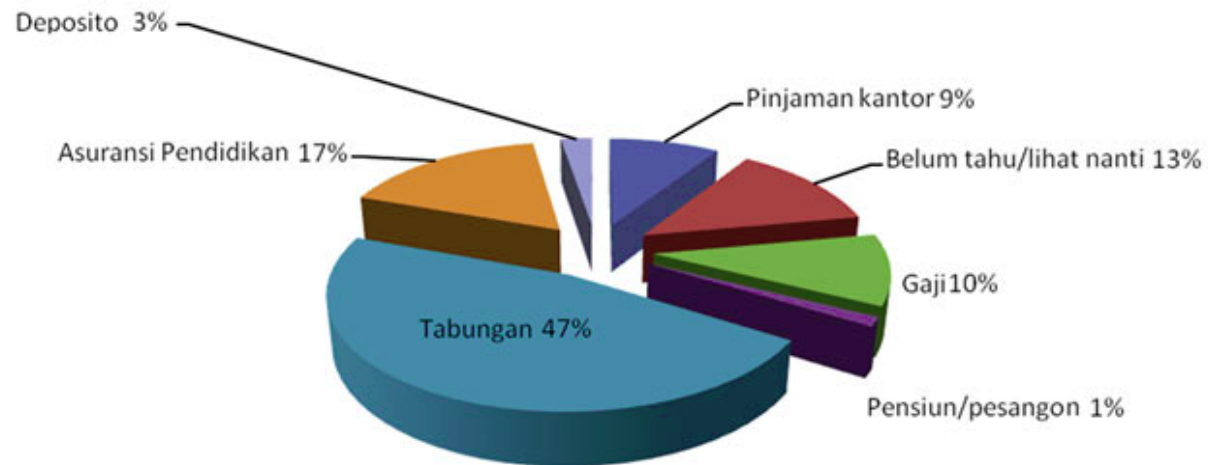


Example:

Grafik 1 : Anda Sudah Menyiapkan Dana Pendidikan Anak Sampai Lulus S1?



Grafik 3 : Bagaimana Kalau Belum Menyiapkan Dananya?



Types of Statistics – Descriptive Statistics and Inferential Statistics

Descriptive Statistics - methods of organizing, summarizing, and presenting data in an informative way.

EXAMPLE 1: The United States government reports the population of the United States was 179,323,000 in 1960; 203,302,000 in 1970; 226,542,000 in 1980; 248,709,000 in 1990, and 265,000,000 in 2000.

EXAMPLE 2: According to the *Bureau of Labor Statistics*, the average hourly earnings of production workers was \$17.90 for April 2008.

Types of Statistics – Descriptive Statistics and Inferential Statistics

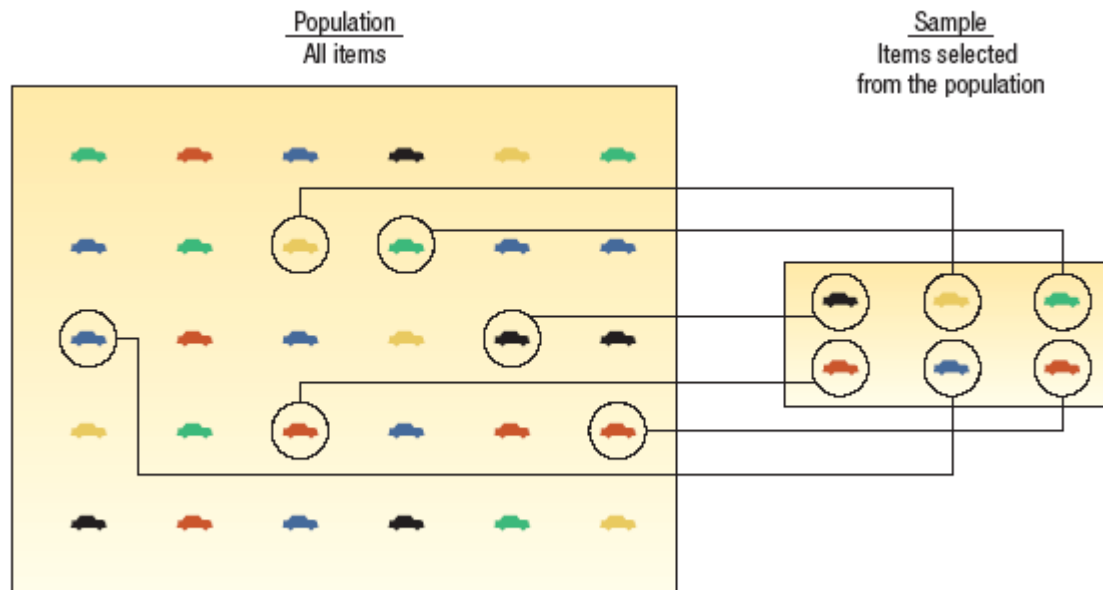
Inferential Statistics: A decision, estimate, prediction, or generalization about a population, based on a sample.


Note: In statistics the word *population* and *sample* have a broader meaning. A population or sample may consist of ***individuals*** or ***objects***

Population versus Sample

A **population** is a **collection** of **all** possible individuals, objects, or measurements of interest.

A **sample** is a **portion**, or **part**, of the population of interest





Why take a sample instead of studying every member of the population?

1. Prohibitive cost of census
2. Destruction of item being studied may be required
3. Not possible to test or inspect all members of a population being studied



Usefulness of a Sample in Learning about a Population

Using a sample to learn something about a population is done extensively in business, agriculture, politics, and government.

EXAMPLE: Television networks constantly monitor the popularity of their programs by hiring Nielsen and other organizations to sample the preferences of TV viewers.

Types of Variables

A. Qualitative or Attribute variable - the characteristic being studied is *nonnumeric*.

EXAMPLES: Gender, religious affiliation, type of automobile owned, state of birth, eye color are examples.

B. Quantitative variable - information is reported *numerically*.

EXAMPLES: balance in your checking account, minutes remaining in class, or number of children in a family.

Quantitative Variables - Classifications

Quantitative variables can be classified as either *discrete* or *continuous*.

A. Discrete variables: can only *assume certain values* and there are *usually “gaps”* between values.

EXAMPLE: the number of bedrooms in a house, or the number of hammers sold at the local Home Depot (1,2,3,...,etc).

B. Continuous variable can *assume any value* within a specified range.

EXAMPLE: The pressure in a tire, the weight of a pork chop, or the height of students in a class.

Summary of Types of Variables

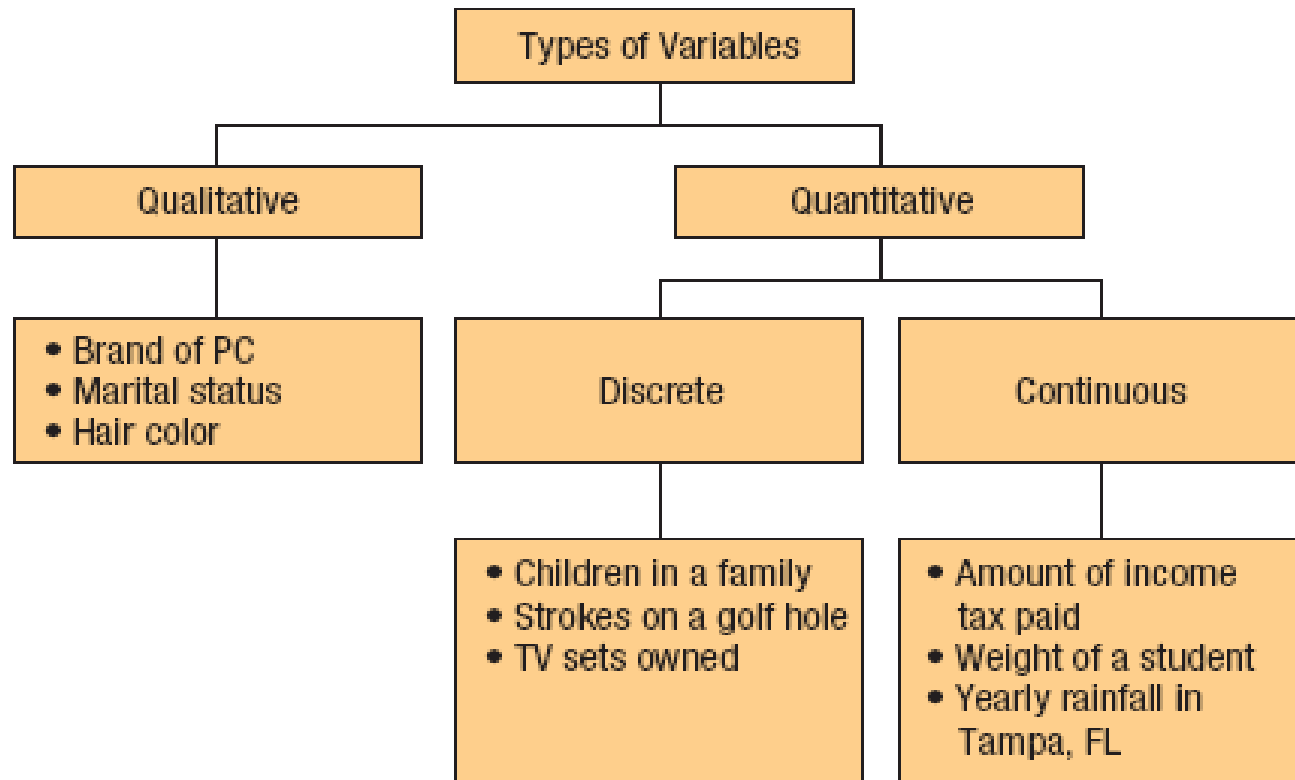


CHART 1–2 Summary of the Types of Variables

Four Levels of Measurement

Nominal level - data that is classified into categories and cannot be arranged in any particular order.

EXAMPLES eye color, gender, religious affiliation.

Interval level - similar to the ordinal level, with the additional property that meaningful amounts of differences between data values can be determined. There is no natural zero point.

EXAMPLE Temperature on the Fahrenheit scale.

Ordinal level - data arranged in some order, but the differences between data values cannot be determined or are meaningless.

EXAMPLE During a taste test of 4 soft drinks, Mellow Yellow was ranked number 1, Sprite number 2, Seven-up number 3, and Orange Crush number 4.

Ratio level - the interval level with an inherent zero starting point. Differences and ratios are meaningful for this level of measurement.

EXAMPLES Monthly income of surgeons, or distance traveled by manufacturer's representatives per month.

Nominal-Level Data

Properties:

1. Observations of a qualitative variable can only be ***classified*** and ***counted***.
2. There is ***no particular order*** to the labels.



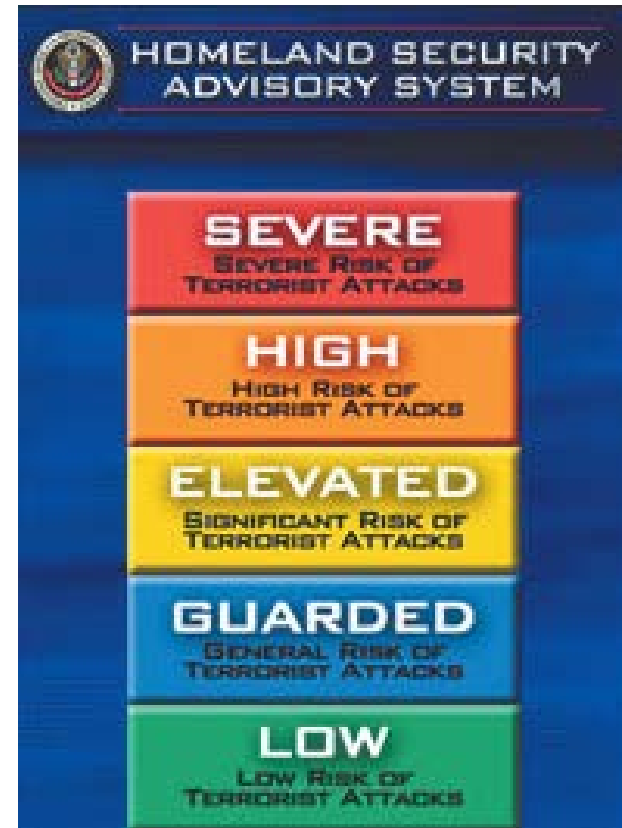
By Anne R. Carey and Chad Palmer, USA TODAY
Source: hudson-index.com



Ordinal-Level Data

Properties:

1. Data classifications are represented by sets of labels or names (high, medium, low) that have *relative values*.
2. Because of the relative values, the *data classified can be ranked or ordered*.



Interval-Level Data

Properties:

1. Data classifications are ordered according to the amount of the characteristic they possess.
2. Equal differences in the characteristic are represented by equal differences in the measurements.

Example: Women's dress sizes listed on the table.

Size	Bust (in)	Waist (in)	Hips (in)
8	32	24	35
10	34	26	37
12	36	28	39
14	38	30	41
16	40	32	43
18	42	34	45
20	44	36	47
22	46	38	49
24	48	40	51
26	50	42	53
28	52	44	55

Ratio-Level Data

- Practically all quantitative data is recorded on the ratio level of measurement.
- Ratio level is the “highest” level of measurement.

Properties:

1. Data classifications are **ordered** according to the amount of the characteristics they possess.
2. Equal differences in the characteristic are represented by equal differences in the numbers assigned to the classifications.
3. The zero point is the absence of the characteristic and the ratio between two numbers is meaningful.



Why Know the Level of Measurement of a Data?

- The level of measurement of the data dictates the calculations that can be done to summarize and present the data.
- To determine the statistical tests that should be performed on the data

Summary of the Characteristics for Levels of Measurement

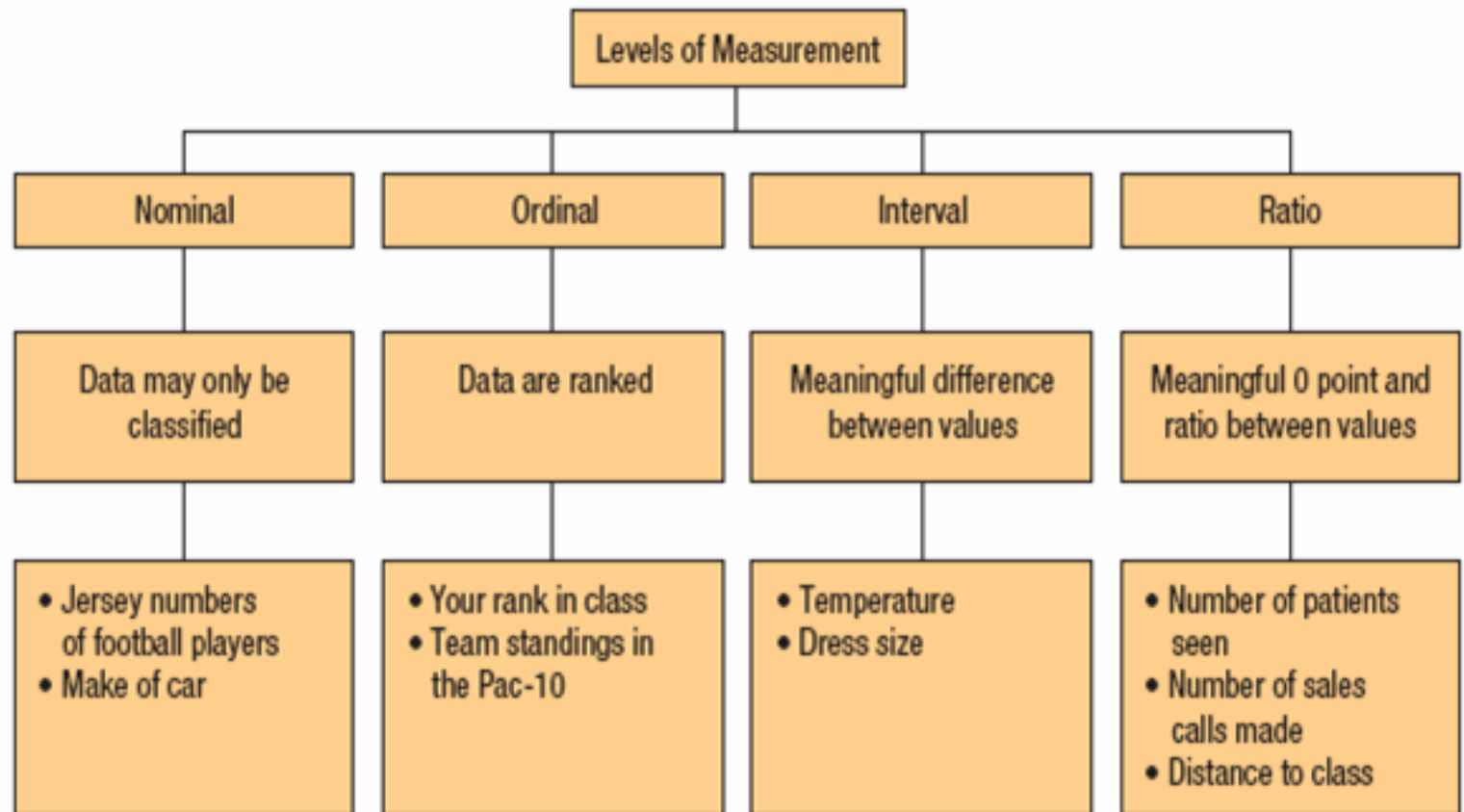


CHART 1-3 Summary of the Characteristics for Levels of Measurement



Descriptive Statistics

Measures of Central Tendency: Ungrouped Data

- Measures of central tendency yield information about “particular places or locations in a group of numbers.”
- Common Measures of Location
 - Mode
 - Median
 - Mean
 - Percentiles
 - Quartiles

Mode

- **Mode - the most frequently occurring value in a data set**
 - **Applicable to all levels of data measurement (nominal, ordinal, interval, and ratio)**
 - **Can be used to determine what categories occur most frequently**
 - **Sometimes, no mode exists (no duplicates)**
- **Bimodal – In a tie for the most frequently occurring value, two modes are listed**
- **Multimodal -- Data sets that contain more than two modes**

Median

- **Median - middle value in an ordered array of numbers.**
 - **Half the data are above it, half the data are below it**
 - **Mathematically, it's the $(n+1)/2^{\text{th}}$ ordered observation**
 - **For an array with an odd number of terms, the median is the middle number**
 - **$n=11 \Rightarrow (n+1)/2^{\text{th}} = 12/2^{\text{th}} = 6^{\text{th}}$ ordered observation**
 - **For an array with an even number of terms the median is the average of the middle two numbers**
 - **$n=10 \Rightarrow (n+1)/2^{\text{th}} = 11/2^{\text{th}} = 5.5^{\text{th}} = \text{average of } 5^{\text{th}} \text{ and } 6^{\text{th}}$ ordered observation**

Arithmetic Mean

- **Mean is the average of a group of numbers**
- **Applicable for interval and ratio data**
- **Not applicable for nominal or ordinal data**
- **Affected by each value in the data set, including extreme values**
- **Computed by summing all values in the data set and dividing the sum by the number of values in the data set**

Population Mean

For ungrouped data, the **population mean** is the sum of all the population values divided by the total number of population values:

POPULATION MEAN

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

[3-1]

where:

μ represents the population mean. It is the Greek lowercase letter “mu.”

N is the number of values in the population.

X represents any particular value.

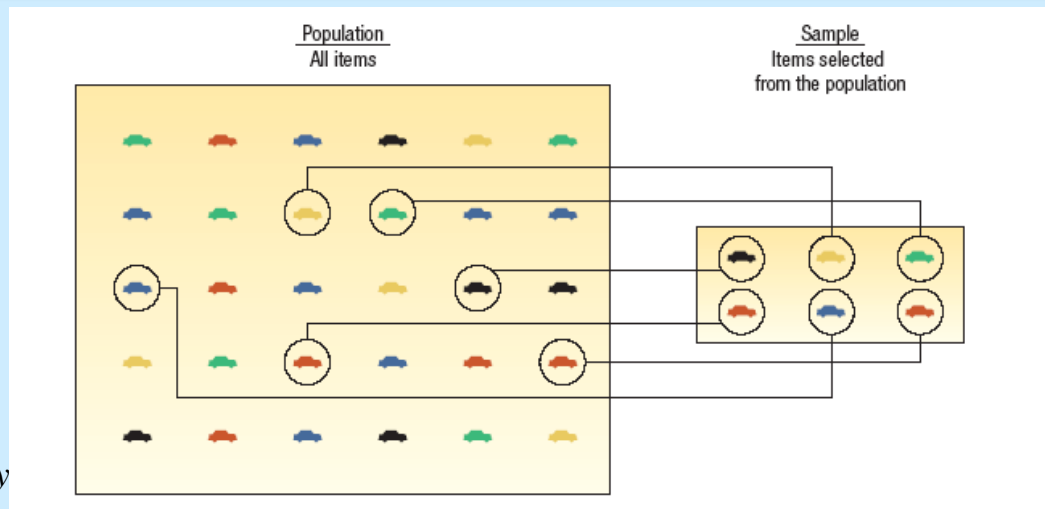
\sum is the Greek capital letter “sigma” and indicates the operation of adding.

$\sum X$ is the sum of the X values in the population.

Parameter Versus Statistics

PARAMETER A measurable characteristic of a *population*.

STATISTIC A measurable characteristic of a *sample*.



Sample Mean

- For ungrouped data, the sample mean is the sum of all the sample values divided by the number of sample values:

SAMPLE MEAN

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

[3-2]

where:

\bar{X} is the sample mean. It is read “X bar.”

n is the number of values in the sample.

The Geometric Mean

GEOMETRIC MEAN

$$GM = \sqrt[n]{(X_1)(X_2) \cdots (X_n)}$$

[3-4]

- Useful in finding the average change of percentages, ratios, indexes, or growth rates over time.
- It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the GDP, which compound or build on each other.
- The geometric mean will always be less than or equal to the arithmetic mean.
- The formula for the geometric mean is written:

EXAMPLE:

The return on investment earned by Atkins Construction Company for four successive years was: 30 percent, 20 percent, -40 percent, and 200 percent. What is the geometric mean rate of return on investment?

$$GM = \sqrt[n]{(X_1)(X_2) \cdots (X_n)} = \sqrt[4]{(1.3)(1.2)(0.6)(3.0)} = \sqrt[4]{2.808} = 1.294$$

Percentiles

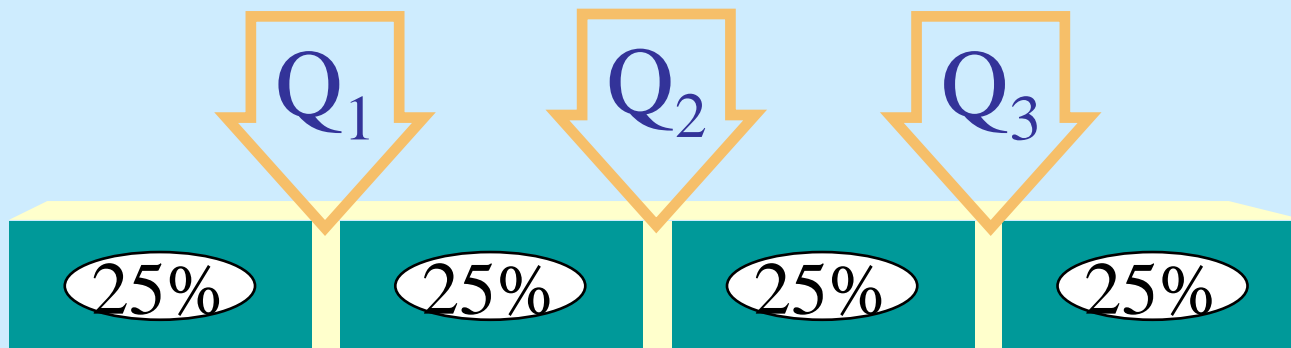
- **Percentile - measures of central tendency that divide a group of data into 100 parts**
- **At least $n\%$ of the data lie at or below the n^{th} percentile, and at most $(100 - n)\%$ of the data lie above the n^{th} percentile**
- **Example: 90th percentile indicates that at 90% of the data are equal to or less than it, and 10% of the data lie above it**

Calculating Percentiles

- **To calculate the p^{th} percentile,**
 - **Order the data**
 - **Calculate $i = N (p/100)$**
 - **Determine the percentile**
 - **If i is a whole number, then use the average of the i^{th} and $(i+1)^{\text{th}}$ ordered observation**
 - **Otherwise, round i up to the next highest whole number**

Quartiles

- Quartile - measures of central tendency that divide a group of data into four subgroups
- Q1: 25% of the data set is below the first quartile
- Q2: 50% of the data set is below the second quartile
- Q3: 75% of the data set is below the third quartile



Which Measure Do I Use?

- **Which measure of central tendency is most appropriate?**
 - In general, the mean is preferred, since it has nice mathematical properties (in particular, see chapter 7)
 - The median and quartiles, are resistant to outliers
- **Consider the following three datasets**
 - 1, 2, 3 (median=2, mean=2)
 - 1, 2, 6 (median=2, mean=3)
 - 1, 2, 30 (median=2, mean=11)
 - All have median=2, but the mean is sensitive to the outliers
- **In general, if there are outliers, the median is preferred to the mean**

Measures of Variability: Ungrouped Data

- **Measures of Variability - tools that describe the spread or the dispersion of a set of data.**
 - **Provides more meaningful data when used**
 - **with measures of central tendency**
 - **in comparison to other groups**

Measures of Spread or Dispersion: Ungrouped Data

- **Common Measures of Variability**
 - **Range**
 - **Inter-quartile Range**
 - **Mean Absolute Deviation**
 - **Variance and Standard Deviation**
 - **Coefficient of Variation**

Range

- **The difference between the largest and the smallest values in a set of data**
 - **Advantage – easy to compute**
 - **Disadvantage – is affected by extreme values**

Interquartile Range

- **Interquartile Range - range of values between the first and third quartiles**
- **Range of the “middle half”; middle 50%**
 - **Useful when researchers are interested in the middle 50%, and not the extremes**

$$\textit{Interquartile Range} = Q_3 - Q_1$$

- **Example: For the cars in service data, the IQR is $204,000 - 9,000 = 195,000$**

Deviations from the mean

- Useful for interval or ratio level data
- An examination of deviation from the mean can reveal information about the variability of the data
 - Deviations are used mostly as a tool to compute other measures of variability
- However, the sum of deviations from the arithmetic mean is always zero:
$$\text{Sum } (X - \mu) = 0$$
- There are two ways to solve this conundrum...

Mean Absolute Deviation (MAD)

- One solution is to take the absolute value of each deviation around the mean. This is called the Mean Absolute Deviation

\underline{X}	$\underline{X-\mu}$	$\underline{ X-\mu }$
5	-8	8
9	-4	4
16	3	3
17	4	4
18	5	5

$$MAD = \frac{\sum |X - \mu|}{n} = \frac{24}{5} = 8.4$$

- Note that while the MAD is intuitively simple, it is rarely used in practice

Sample Variance

- Another solution is to take the Sum of Squared Deviations (SSD) about the mean
- Sample Variance - average of the squared deviations from the arithmetic mean
- Sample Variance – denoted by s^2

<u>X</u>	<u>X-Xbar</u>	<u>(X-Xbar)²</u>
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444

$$s^2 = \frac{\sum (X - \mu)^2}{n - 1} = \frac{663,886}{3} = 221,289$$

Sample Standard Deviation

- **Sample standard deviation is the square root of the sample variance**
- **Same units as original data**

$$s = \sqrt{s^2} = \sqrt{221,289} = 470.4$$

Demonstration Problem 3.6

The effectiveness of district attorneys can be measured by several variables, including the number of convictions per month, the number of cases handled per month, and the total number of years of conviction per month. A researcher uses a sample of five district attorneys in a city and determines the total number of years of conviction that each attorney won against defendants during the past month, as reported in the first column in the following tabulations. Compute the mean absolute deviation, the variance, and the standard deviation for these figures.

Demonstration Problem 3.6

Solution

The researcher computes the mean absolute deviation, the variance, and the standard deviation for these data in the following manner.

<u>X</u>	<u>X-Xbar</u>	<u> X-Xbar </u>	<u>(X-Xbar)²</u>	
55	-41	41	1,681	
100	4	4	16	
125	29	29	841	
140	44	44	1,936	
60	-36	36	1,296	
SUM:	480	0	154	5,770

$$MAD = 154 / 5 = 30.8$$

$$s^2 = 5,770 / 4 = 1,443$$

$$s = \sqrt{1,443} = 38$$

Z Scores

- **Z score – represents the number of Std Dev a value (x) is above or below the mean of a set of numbers**
- **Z score allows translation of a value's raw distance from the mean into units of std dev**
- **$Z = (x - \mu) / \sigma$**

Coefficient of Variation

- **Coefficient of Variation (CV) – measures the volatility of a value (perhaps a stock portfolio), relative to its mean. It's the ratio of the standard deviation to the mean, expressed as a percentage**
- **Useful when comparing Std Dev computed from data with different means**
- **Measurement of relative dispersion**

$$C.V. = \frac{\sigma}{\mu} (100)$$

Coefficient of Variation

Consider two different populations

$$\mu_1 = 29$$

$$\sigma_1 = 4.6$$

$$\begin{aligned} C.V._1 &= \frac{\sigma_1}{\mu_1} (100) \\ &= \frac{4.6}{29} (100) \\ &= 15.86 \end{aligned}$$

$$\mu_2 = 84$$

$$\sigma_2 = 10$$

$$\begin{aligned} C.V._2 &= \frac{\sigma_2}{\mu_2} (100) \\ &= \frac{10}{84} (100) \\ &= 11.90 \end{aligned}$$

Since $15.86 > 11.90$, the first population is more variable, relative to its mean, than the second population

Calculation of Grouped Mean

Sometimes data are already grouped, and you are interested in calculating summary statistics

Interval	Frequency (f)	Midpoint (M)	f*M
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	<u>1</u>	75	<u>75</u>
	50		2150

$$\mu = \frac{\sum f * M}{\sum f} = \frac{2150}{50} = 43.0$$

Median of Grouped Data - Example

<u>Class Interval</u>	<u>Frequency</u>	<u>Cumulative Frequency</u>
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	<u>1</u>	50
	N = 50	

$$\begin{aligned}Md &= L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W) \\ &= 40 + \frac{\frac{50}{2} - 24}{11}(10) \\ &= 40.909\end{aligned}$$

Mode of Grouped Data

- Midpoint of the modal class
- Modal class has the greatest frequency

<u>Class Interval</u>	<u>Frequency</u>
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$\text{Mode} = \frac{30 + 40}{2} = 35$$

Variance and Standard Deviation of Grouped Data

$$S^2 = \frac{\sum f (M - \bar{X})^2}{n - 1}$$
$$S = \sqrt{S^2}$$

Population Variance and Standard Deviation of Grouped Data

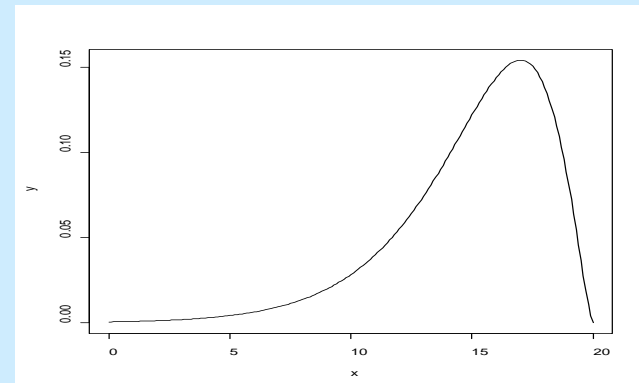
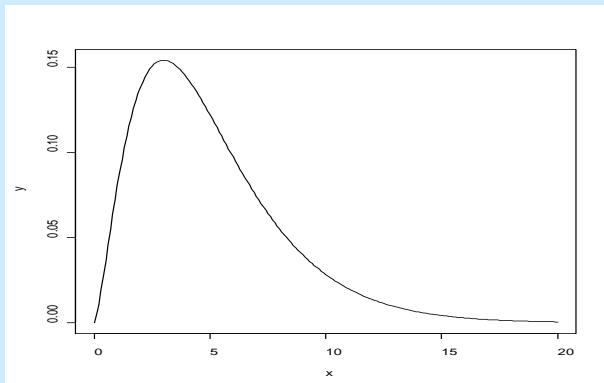
<i>Class Interval</i>	<i>f</i>	<i>M</i>	<i>fM</i>	<i>M - μ</i>	$(M - \mu)^2$	$f(M - \mu)^2$
20-under 30	6	25	150	-18	324	1944
30-under 40	18	35	630	-8	64	1152
40-under 50	11	45	495	2	4	44
50-under 60	11	55	605	12	144	1584
60-under 70	3	65	195	22	484	1452
70-under 80	1	75	<u>75</u>	32	1024	<u>1024</u>
	50		2150			7200

$$\sigma^2 = \frac{\sum f(M - \mu)^2}{N} = \frac{7200}{50} = 144$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{144} = 12$$

Measures of Shape

- **Symmetrical** – the right half is a mirror image of the left half
- **Skewed** – shows that the distribution lacks symmetry; used to denote the data is sparse at one end, and piled at the other end
 - **Absence of symmetry**
 - **Extreme values or “tail” in one side of a distribution**
 - **Positively- or right-skewed vs. negatively- or left-skewed**



Coefficient of Skewness

- **Coefficient of Skewness (S_k)** - compares the mean and median in light of the magnitude to the standard deviation; M_d is the median; S_k is coefficient of skewness; σ is the Std Dev

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

Coefficient of Skewness

- Summary measure for skewness

$$S_k = \frac{3(\mu - M_d)}{\sigma}$$

- If $S_k < 0$, the distribution is negatively skewed (skewed to the left).
- If $S_k = 0$, the distribution is symmetric (not skewed).
If S_k is close to 0, it's almost symmetric
- If $S_k > 0$, the distribution is positively skewed (skewed to the right).