My reservoir is like the distance between Aden and Amman in al-Balqa

Al-Balqa Applied University

Faculty of Medicine

Epidemiology and Biostatistics

الوبائيات والإحصاء الحيوي (31505204)

Lecture 5+6+7+8+9

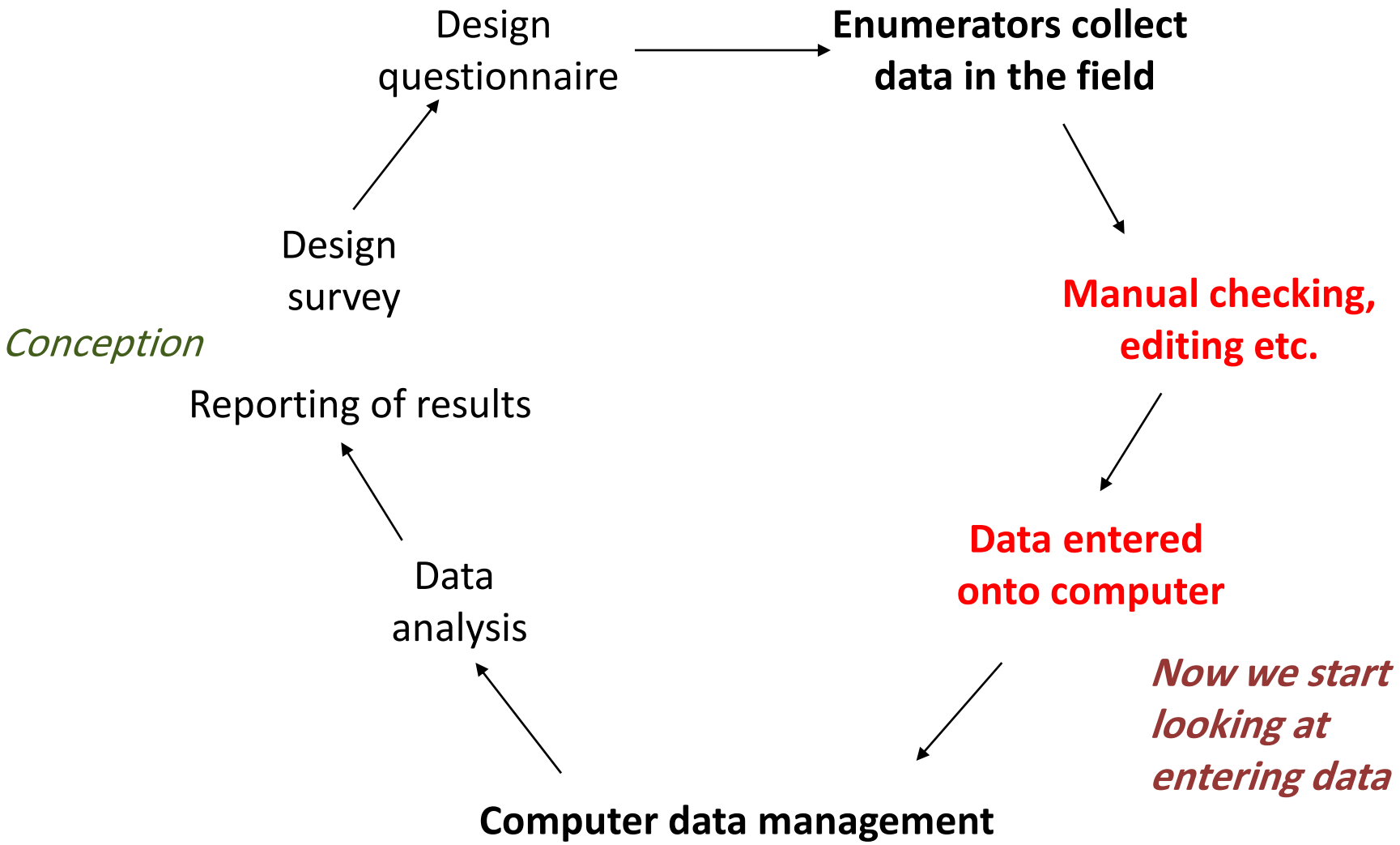
Descriptive statistics

23+24+25+26+30-6-2019

Presentation outline 23+24+25+26+30-6-2019

Lectures	Date
Data, Variables, Types of Measurement Scales.	23-6-2019
Frequency Distribution, Measure of Central Tendency.	24-6-2019
Measure of Dispersion.	25-6-2019
SPSS and Data Entry.	26-6-2019
Ways to Display Data ,Graphics and Interpret in a Public Health Context.	30-6-2019

Data Management Cycle



Data Preparation

- **Data:** The simple concept of data is 0,1, it has no actual meaning (group of numbers or coding). This data when processed gives meaningful " information".

- In a research , the data is collected using questionnaire. Ask participants to fill it up, or you make direct observation and collect data.

- The questionnaire that you collect from consist of questions , and each question asks about something, e.g. age.

Data Preparation ……

- When you applied to the university , you were asked about your name, gender, nationality , etc. , each of them represents a variable.

- **Each question** has got **an answer**; **the answer** maybe :

1. Open ended : It is the most used on researches.
2. Coding : The options come in the form of :

   (A.xxx B. yyy C. zzz) or (1.Xxx 2. Yyy 3.Zzz).

- **Example**: What is your gender? (it is a **question** which means **a variable**) 1. Male 2. Female

  (*The answer is the coding for this question*).

Types of Data

- **The data subdivided into:**

1. **Quantitative Data (Numerical):** They are variables that **can be measured**, counted & have a numeric meaning such as ; age , weight , height.

2. **Qualitative Data (Categorical):** Information which **can not be expressed as a number**. It is something that you **cannot count** , assign a numeric value to it such as: gender , residency , nationality, etc.

Types of Variables

- **Quantitative data**: can be **discrete** taking only certain values or **continuous**, taking any value.

A. **Discrete variable (count data)** that have only certain fixed values and no intermediate values possible *(number of students in a class room*).

A. **Continuous variable (real-values)** where between any two points. There are at least theoretically infinite number of values (weight , height, etc.).

- **Example:** The number of times a patient is admitted to a hospital is **discrete** (*a patient cannot be admitted 0.8 times*), while a patient's weight is a **continuous** (a patient's weight could take any value within a range).

Types of Variables …

- **Qualitative data (Categorical)**: can be **nominal variable Or ordinal variables.**

A. **Nominal variable (not ordered)/ Name only:** The variables are divided into a number of named categories that **cannot be ordered** one above the other. It has no ordinary sense. No ordering of the categories (*The answer is determined*).

- **Example:** a patient's ethnicity , gender, eye color, names , marital status, blood groups, etc.

❖ **Binary variable**: A variable has two answer options.

- **Example:** yes or no questions, gender questions.

➢ This type of variable makes the data analysis easier.

Types of Variables …

B. Ordinal variable (ordered): The variables are divided into number of name categories that **can ordered from lowest to highest or vice versa.** Has sense of ordering. Categories can be ordered.

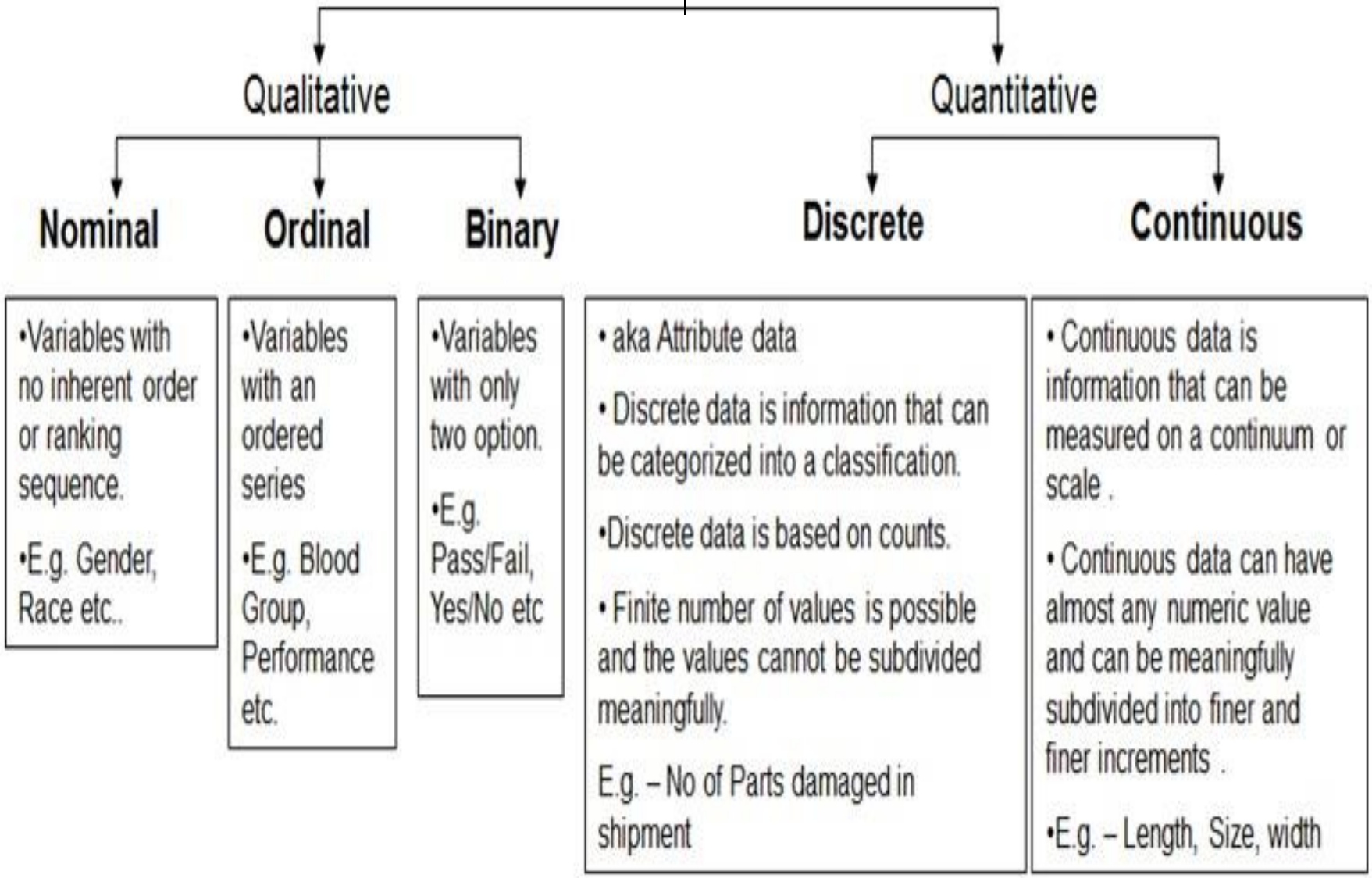
- **Example:** ***Response to treatment*** **,** Educational level (high school, university degree, college degree) can be organized in to an ascending or descending order.

Types of Variables …

- **One variable** could be quantitative or qualitative according to how it's presented.

- So, just saying blood pressure as a number means we classify it as **quantitative variable.** While saying types of blood pressure (explain high, normal, and low) then it's going to turn **a qualitative.**

Types of Variables

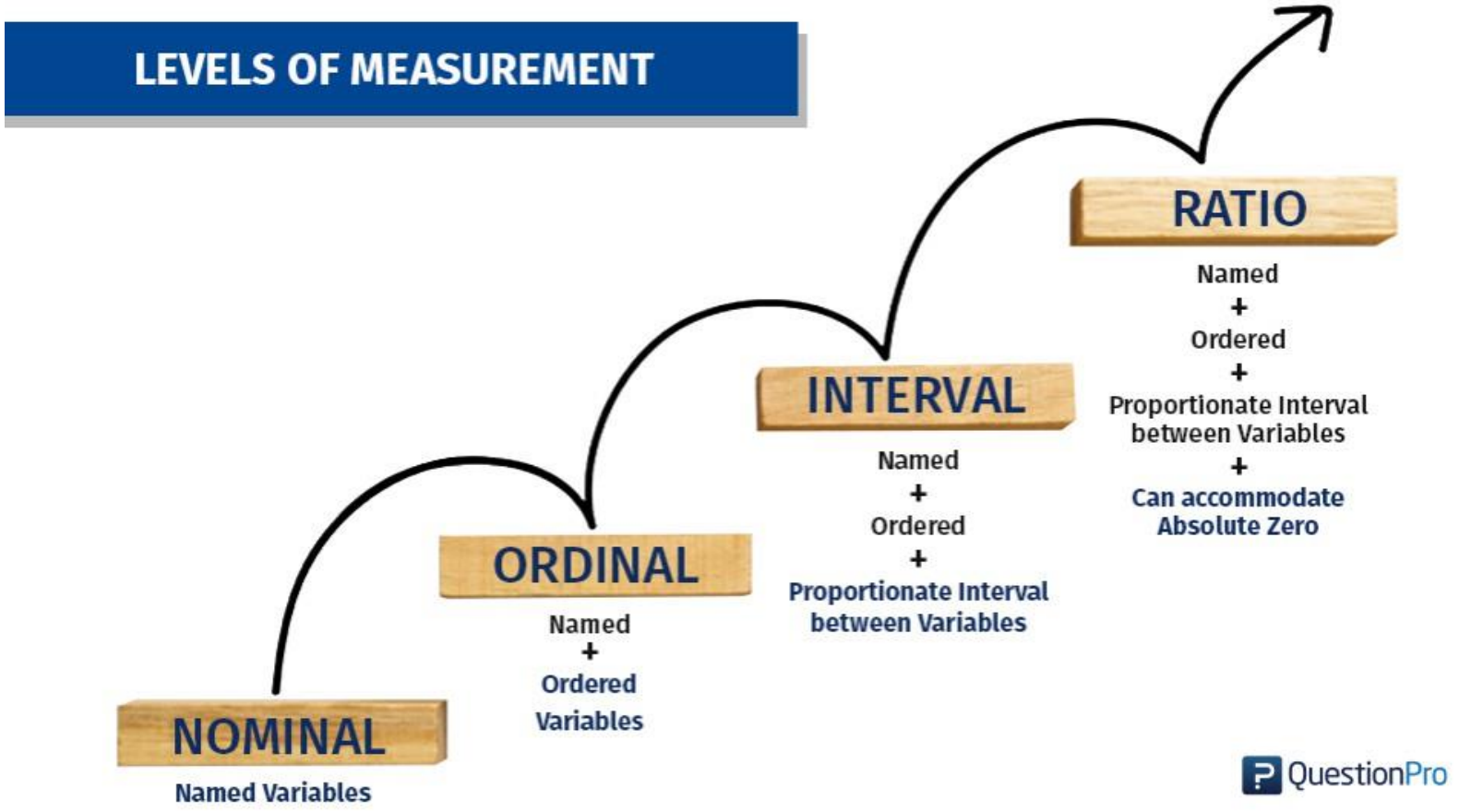


Types of Measurement Scales

1. **Nominal Scale (المقياس الأسمي / التنصيفي) ***
2. **Ordinal Scale) (المقياس الترتيبي **
3. **Interval Scale (مقياس الفتره) ***
4. **Ratio Scale (المقياس النسبي) ****

- The four scale types are ordered in that all later scales have all the properties of earlier scales—**plus** additional properties.

Level of Measurement



Nominal Scale

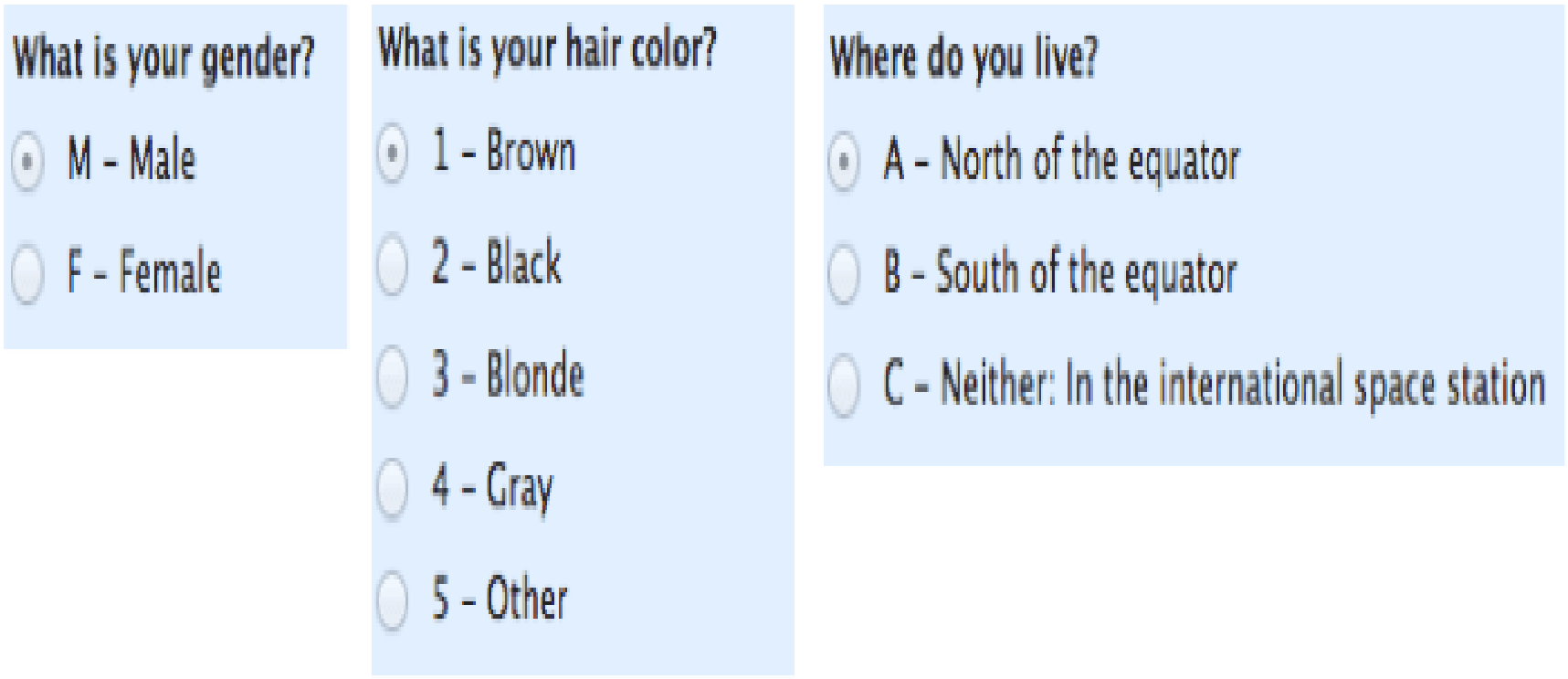
- Not really a ‘scale’ because it does not scale objects along any dimension.
- It simply labels objects.

Example: Gender is a nominal scale

Male = 1

Female = 2

Nominal Scale….



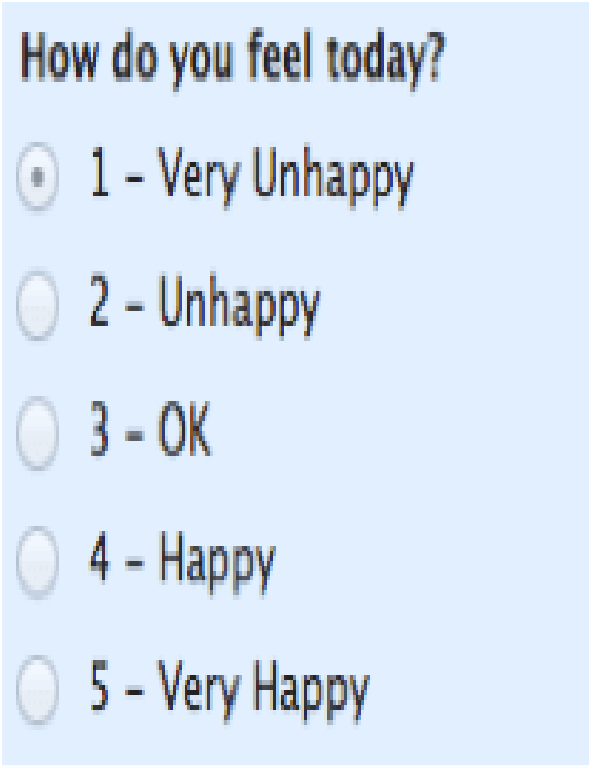
Ordinal Scale

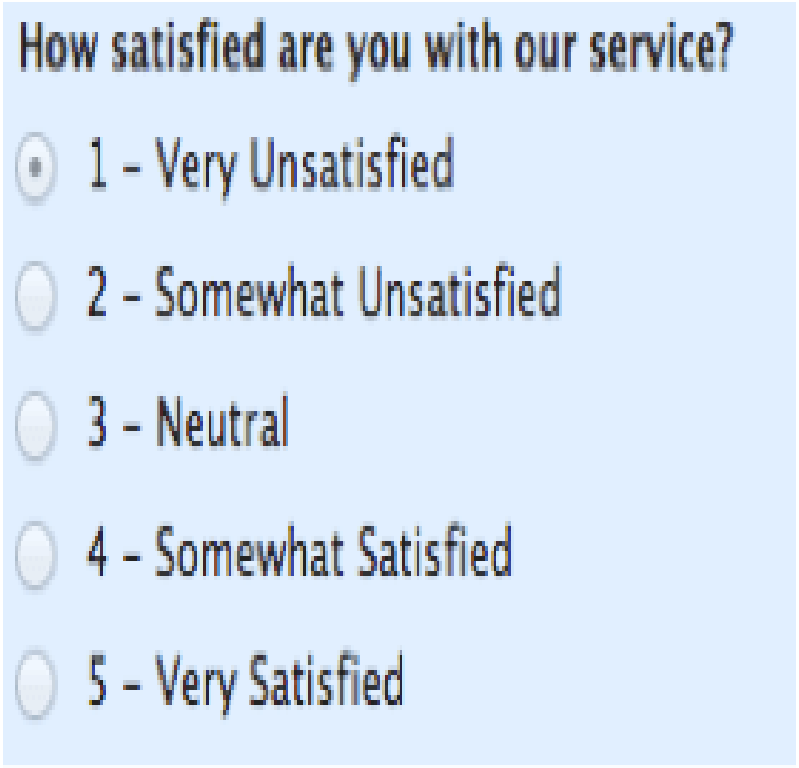
- **Ordinal Scale:** Nominal categories with implied order- Low, medium, high.

- Numbers are used to place objects in order.

- **But**, there is no information regarding the differences (intervals) between points on the scale.

Ordinal Scale ….

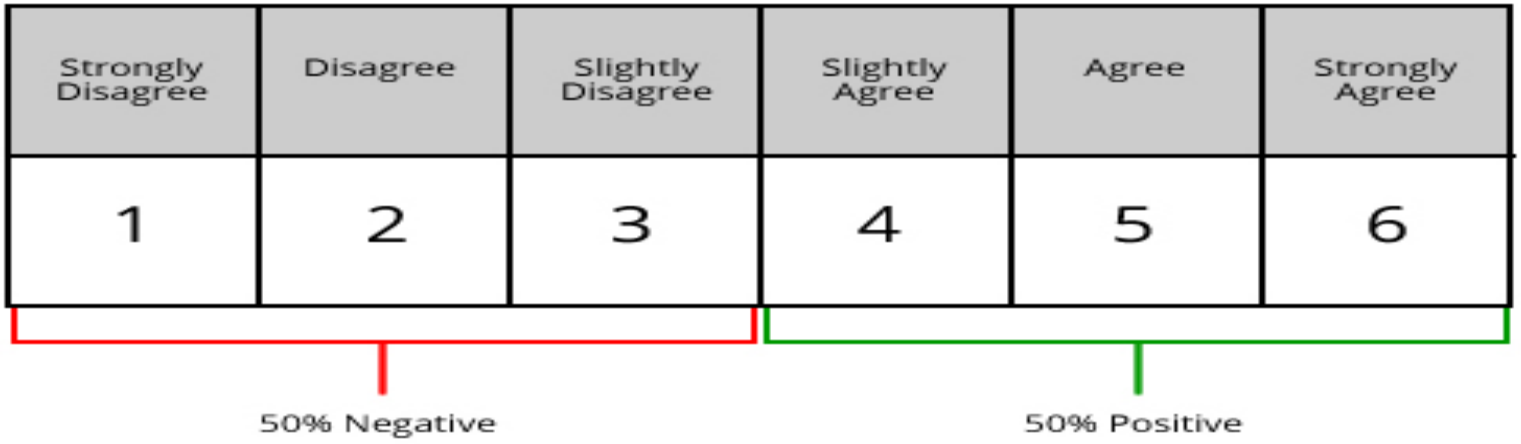




Likert Scale

Question: Compared to others, what is your satisfaction rating of the National Practitioner Data Bank?

1	2	3	4	5
Very Satisfied	Somewhat Satisfied	Neutral	Somewhat Dissatisfied	Very Dissatisfied



Interval Scale

- **Interval scale (**Numeric scales)**:** An interval scale is a scale on which equal intervals between objects, represent equal differences.
- The interval differences are meaningful. But, we can't defend ratio relationships.

- Differences *can* be compared; no true zero. **Ratios *cannot* be compared.**
- **Example: Temperature in Celsius.**

The difference between 10 and 20 degrees is the same as between 80 and 90 degrees but, we can't say that 80 degrees is twice as hot as 40 degrees.

Interval Scale….

- **Interval scales** are nice because the realm of statistical analysis on these data sets opens up. For example, *central tendency* can be measured by mode, median, or mean; standard deviation can also be calculated.

Ratio Scale

Ratio scale: Order and distance implied. Differences *can* be compared; has a true zero. **Ratios *can* be compared**.

Examples: Height, weight, blood pressure

- Ratios are meaningful.
- We can say that 20 seconds is twice as long as 10 seconds.

Summary of data-types and scales

Provides:	Nominal	Ordinal	Interval	Ratio
The "order" of values is known		✔	✔	✔
"Counts," aka "Frequency of Distribution"	✔	✔	✔	✔
Mode	✔	✔	✔	✔
Median		✔	✔	✔
Mean			✔	✔
Can quantify the difference between each value			✔	✔
Can add or subtract values			✔	✔
Can multiple and divide values				✔
Has "true zero"				✔

Two types of variables

1. **Dependent variables**: The variable that is used to describe or measure the problem under study. It is the **center of the study**.

2. **Independent variables**: The variable that are used to describe or measure the factors that are assumed to cause or at least to influence the problem.

- **Example:** If we are studying the blood pressure on a group of people take into consideration their age and environment, so the **center of study** is Hypertension, other variable are called **independent**.

- Whether a variable is dependent or independent is determined by the statement of the problem and study objectives.

Broad Categories of Statistics

- Statistics can broadly be split into two categories ***Descriptive Statistics and Inferential Statistics:***

1. **Descriptive statistics** deals with the meaningful presentation of data such that its characteristics can be effectively observed.

2. **Inferential statistics** on other hand, deals with drawing inferences and taking decision by studying a subset or sample from the population.

Descriptive Biostatistics

- The best way to work with data is to **summarize and organize** them.

- Numbers that have **not been summarized and organized are called raw data.**

Definition

- **Data is any type of information.**
- **Raw data** is a data collected as they receive.
- **Organize data** is data organized either in ascending, descending or in a grouped data.

Descriptive Statistics

1. **Frequency Distribution.**

2. **Measure of Central Tendency.**

3. **Measure of Dispersion.**

Frequency Distribution

- **Two simple ways of representing data**, which are: Tables , Graphics.

- **A frequency distribution** is a tabular summary of data showing the frequency (or number) of items in each of several non-overlapping classes.

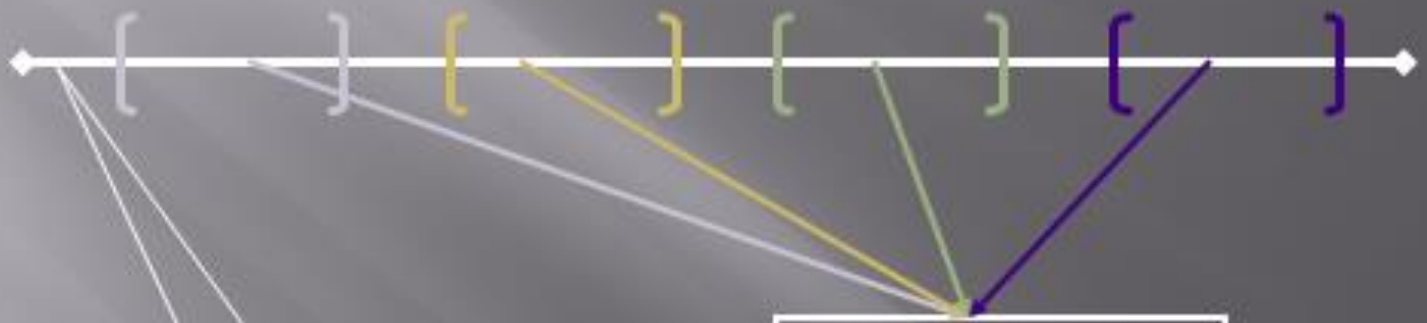
- The **objective** is to provide insights about the data that cannot be quickly obtained by looking only at the original data.

Frequency Distribution…

Patient Name	Patient Age	Patient Name	Patient Age	Patient Name	Patient Age
1	50	51	30	101	50
2	20	52	66	102	47
3	13	53	28	103	25
…	…	…	…	…	…
50	70	100	35	150	19

Frequency Distribution

- Data Grouping



Class Intervals

Data Line

Range = L - S

$$\text{Class Interval Width (w)} = \frac{R}{k}$$

No. of class intervals

$k.= 1+ 3.322(\log n)$

Frequency Distribution ….

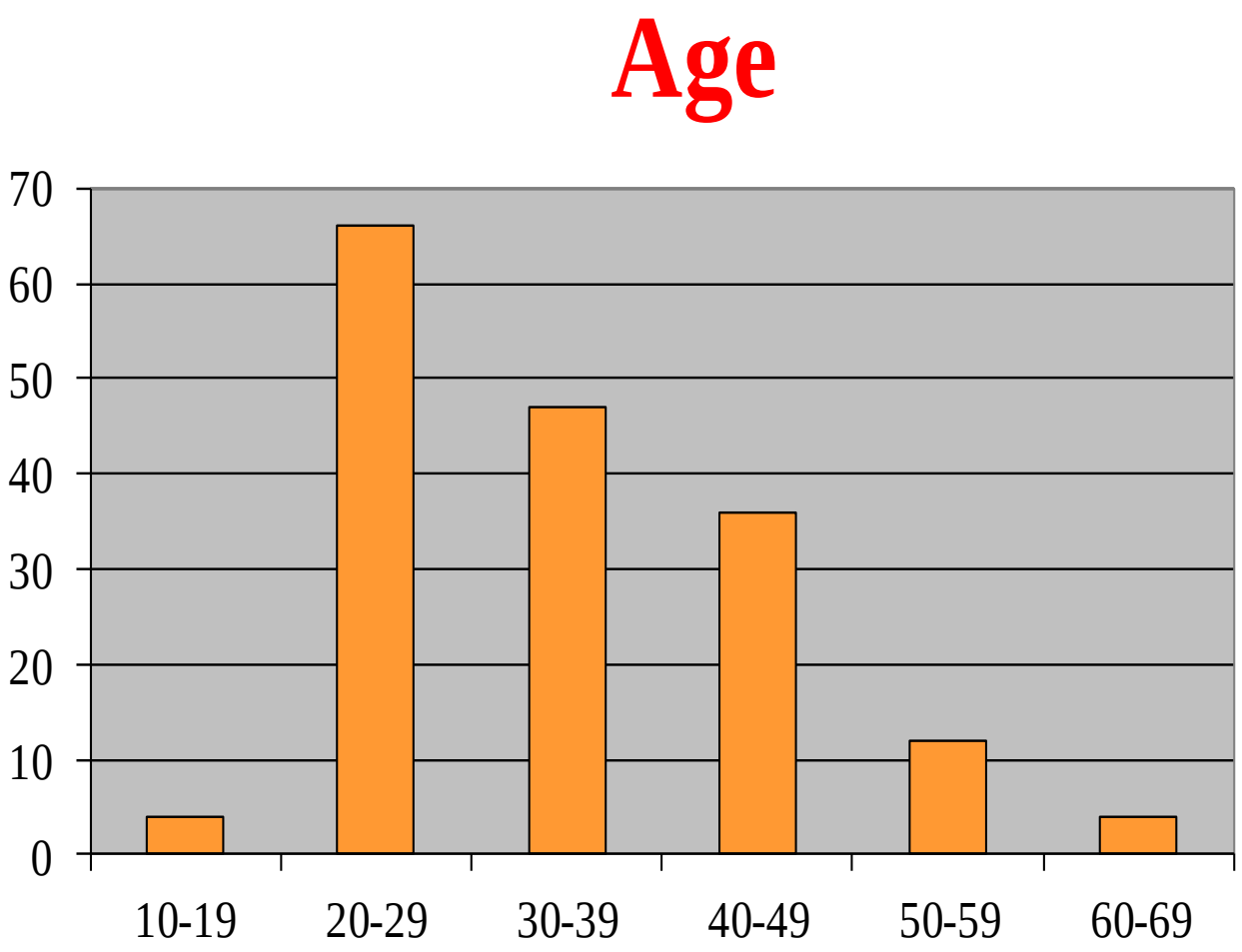
Frequency Distribution	
class intervals	Frequency
10-19	4
20-29	66
30-39	47
40-49	36
50-59	12
60-69	4
Total	169

- Freq. Dist. Is a table shows the way in which the variable values are distributed among the specified **class intervals**.

Frequency Distribution…

Frequency, Cumulative Frequency, Relative Frequency, and Cumulative Relative Frequency Distribution				
class intervals	Frequency	Cumu. F.	Relative F	C. R. F.
10-19	4	4	.0237	.0237
20-29	66	70	.3905	.4142
30-39	47	117	.2781	.6923
40-49	36	153	.2130	.9053
50-59	12	165	.0710	.9763
60-69	4	169	.0237	1.00
Total	169			

Frequency Distribution…



Measures of Location (Center)

- It is a property of the data that they **tend to be clustered about a center point.**

- Measures of ***central tendency*** *(i.e., central location)* ***help find the approximate center of the dataset.***

- Researchers usually **do not use** the term average, because there are **three alternative types of average:**

➢ These include the ***mean, the median, and the mode***.

Commonly Used Symbols

For a Sample

$\bar{x}$ sample mean

s^2 sample variance

s sample standard deviation

For a Population

μ population mean

σ^2 population variance

σ population standard deviation

Measures of central tendency

1. **Mean** (الوسط الحسابي)

2. **Median** (الوسيط)

3. **Mode** (المنوال)

Mean

- It is the arithmetic mean and is also known as **average**.
- It is calculated by totaling the results of all observations and dividing by the total number of observations.

- **Example:** height of 6 girls are as follow:

➢ 160, 168, 160, 177, 155, 169 cm

➢ Total is 989

➢ **Mean** is 989/6=165cm

Why do we measure the mean?

- In order to be able to use **only one number** for representing the data.

- If we have many data, the **mean** can give us an indication about the data using one value.

- If the data are from a sample , the mean is denoted by $\bar{x}$.

The Mean

- The sample mean is the sum of all the observations (ΣX_i) divided by the number of observations (n):

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{where } \Sigma X_i = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$$

- **Example.** 1, 2, 2, 4, 5, 10. Calculate the mean.
  Note: n = 6 (six observations)

$$\Sigma X_i = 1 + 2 + 2 + 4 + 5 + 10 = 24$$
$$\bar{X} = 24 / 6 = 4.0$$

7

- If the data are from a population , the mean is denoted by **μ.**

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ = population mean

$\sum$ = summation sign

x_i = value of element i of the sample

N = population size

Notes on Sample Mean

- Also called *sample average or arithmetic mean*

- Mean for the **sample** = $\bar{X}$.
- Mean for **population** = ***mew (μ).***

- Uniqueness: For a given set of data there is one **and only one mean.**

- **N** =Population, **n** = sample.

- **Simplicity**: The mean is easy to calculate.

- ***Sensitive or highly affected by extreme values.***

The Mean

Example.

For the data: 1, 1, 1, 1, 51. Calculate the mean.
Note: n = 5 (five observations)

$\Sigma X_i = 1 + 1 + 1 + 1 + 51 = 55$

$\bar{X} = 55 / 5 = 11.0$

- Here we see that the mean is affected by extreme values.

The Median

- The median is the **middle value of the** ***ordered data.***

- It is the value divides the distribution into **equal halves.**
- **50%** of the data lies **above** the median.
- **50%** of the data lies **below** the median.

- **To get the median**, there is **two steps**:

1. **Rearrange** the data into an ***ordered array (in ascending or descending order).***

- List all observation from **lowest value to the highest value.**

2. **Determine the position** of the median by using this equation **(n+1)/2**

(*where n= the number of values*).

The Median….

- **Example 1:** **Odd number of values*
- Weights of 7 boys are as follow:

➢ 47,42,44,40,43,72,41 kg.

1. **Sort first** 40,41,42,**43**,44,47,72 kg.
2. The **position of the median** is (7+1)/2=4 (*n is odd*).

(the 4th one, which is 43).

- Median is **less or not affected** by extreme values.
- In this case , it is the best measure of central tendency.

- If **n** is **odd**, the median is the middle observation of the ordered array. If n is **even**, it is midway **between the *two central observations.***

The Median….

- **Example 2:** **Even number of values*
- Find the median of the following data set::

➢ 12 18 16 21 10 13 17 19

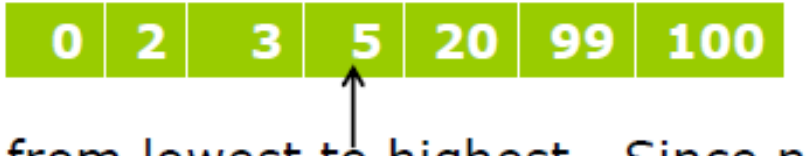
1. **Sort first** 10,12,13,**16,17**,18,19,21
2. The **position of the median** is 4.5^{th} value, so we take the 4^{th} and 5^{th} values (16+17)/2=16.5.

- For **odd number** of data we have one median.

- For **even number** of data we have 2 median , then we take their average.

The Median

Example:



Note: Data has been ordered from lowest to highest. Since n is odd (n=7), the median is the (n+1)/2 ordered observation, or the 4th observation.

Answer: The median is 5.

The mean and the median are unique for a given set of data. There will be exactly one mean and one median.

Unlike the mean, the median is not affected by extreme values.

Q: What happens to the median if we change the 100 to 5,000? Not a thing, the median will still be 5. Five is still the middle value of the data set.

The Median

Example:

Note: Data has been ordered from lowest to highest. Since n is even (n=6), the median is the (n+1)/2 ordered observation, or the 3.5^{th} observation, *i.e.*, the average of observation 3 and observation 4.

Answer: The median is 35.

The Mode

- **The Mode** is the **value of the data that occurs with the greatest frequency**. It is the most frequently occurring value in a set of observation.
- Its useful for categorized data.

- **Example:** Weight of 7 girls are as follow:

➢ 47,44,44,40,43,72,44 kg

➢ Sort first 40, 43,44,44,44,47,72 kg

➢ **The Mode** is 44

- The mode is **not** affected by extreme values.

The Mode…

- **Unstable index**: values of modes tend to fluctuate from one sample to another drawn from the same population.

- **Example 1:** 1, 1, 1, 2, 3, 4, 5

➢ The mode is 1 since it occurs three times. The other values each appear only once in the data set.

➢

- **Example 2:** 5, 5, 5, 6, 8, 10, 10, 10.

➢ The mode is: 5, 10.

➢ **There are two modes. This is a *bi-modal dataset.***

The Mode

- **The mode** is the value that occurs most often in a data set.

unimodal	A data set that has **only one value that occurs** with the greatest frequency .
bimodal	A data set has **two values that occur** with the same greatest frequency ,both values are considered to be the mode and the data set.
multimodal	A data set has **more than two values that occur** with the same greatest frequency ,each value is used as the mode, and the data set.
No mode	Each value occurs only once .

The Mode…

- The mode is different from the mean and the median in that **those measures always exist and are always unique.** For any numeric data set there will be one mean and one median.

- ***The mode may not exist.***
  - Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 0
  - Here you have 10 observations and they are all different.

- ***The mode may not be unique.***
  - Data: 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7
  - Mode = 1, 2, 3, 4, 5, and 6. There are ***six modes.***

Comparison of the Mode, the Median, and the Mean

- In a normal distribution, *the mode , the median, and the mean have the same value.*

- The **mean** is the widely reported index of central tendency for variables measured on an interval and ratio scale.

- The mean takes each and every score into account.

- It also the most stable index of central tendency and thus yields the most reliable estimate of the central tendency of the population.

Comparison of the Mode, the Median, and the Mean

- The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.

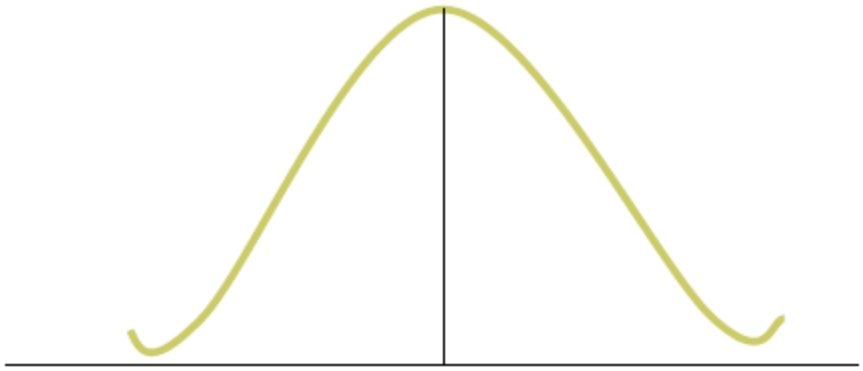
- For the variables that **positively skewed** (like income), the mean is higher than the mode or the median.

- For **negatively skewed** variables (like age at death) the mean is lower.
- When there are extreme values in the distribution (even if it is approximately normal), researchers sometimes report means that have been adjusted for **outliers.**

- To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.

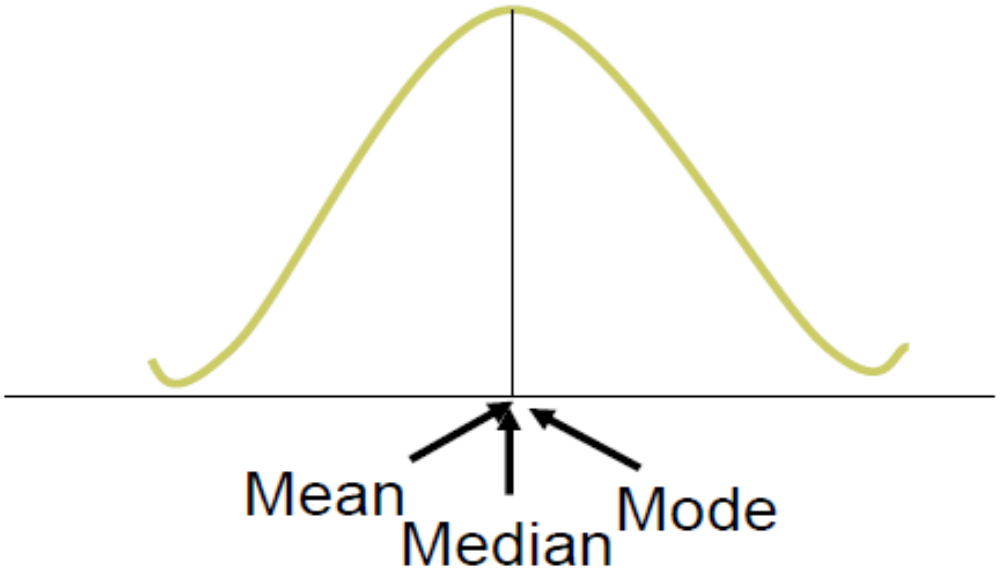
Distribution Characteristics

- **Mode: Peak(s)**
- **Median: Equal areas point**
- **Mean: Balancing point**



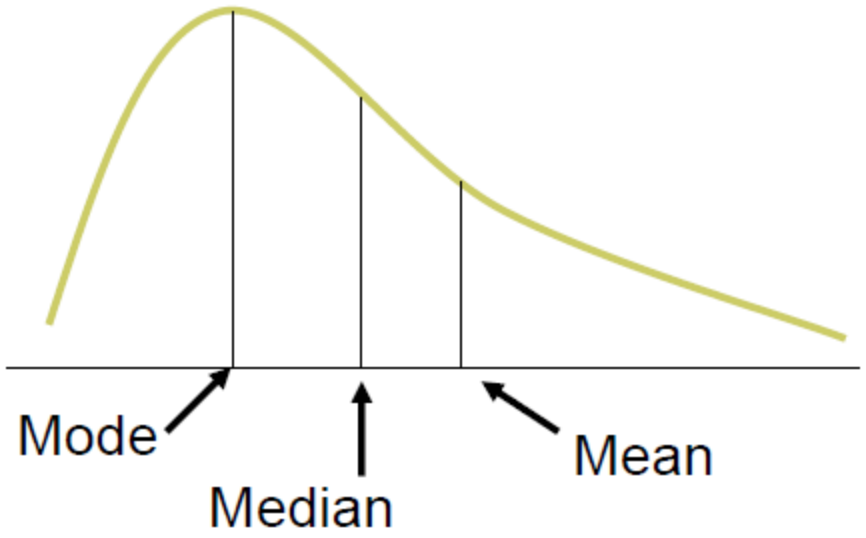
Shapes of Distributions

- **Symmetric (Right and left sides are mirror images)**
- Left tail looks like right tail
- **Mean = Median = Mode**



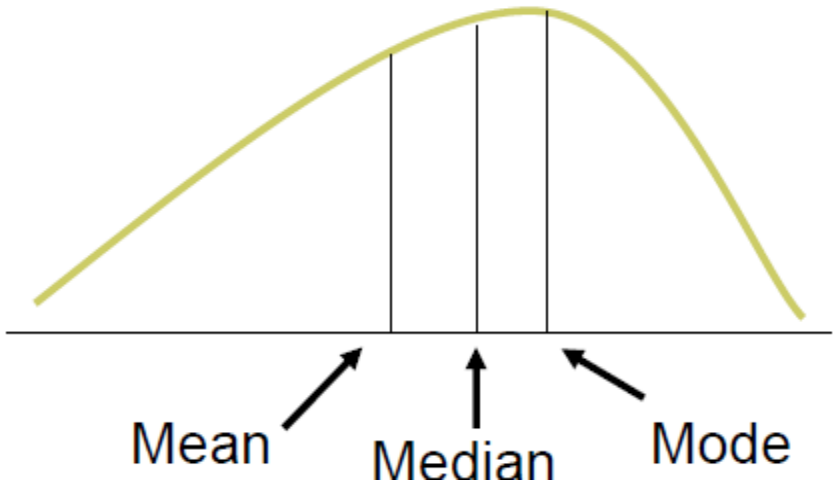
Shapes of Distributions

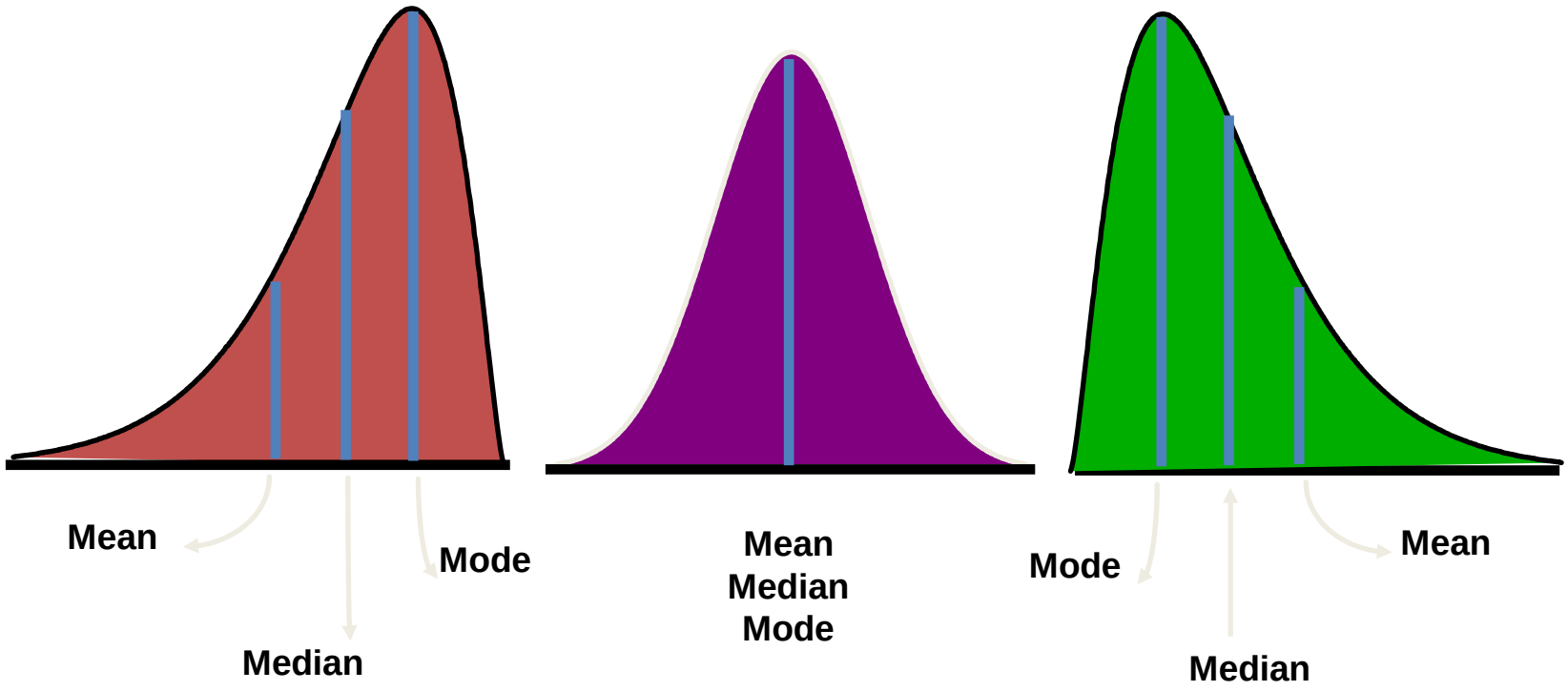
- **Right skewed (positively skewed)**
  - Long right tail
  - Mean > Median



Shapes of Distributions

- **Left skewed (negatively skewed)**
  - Long left tail
  - Mean < Median





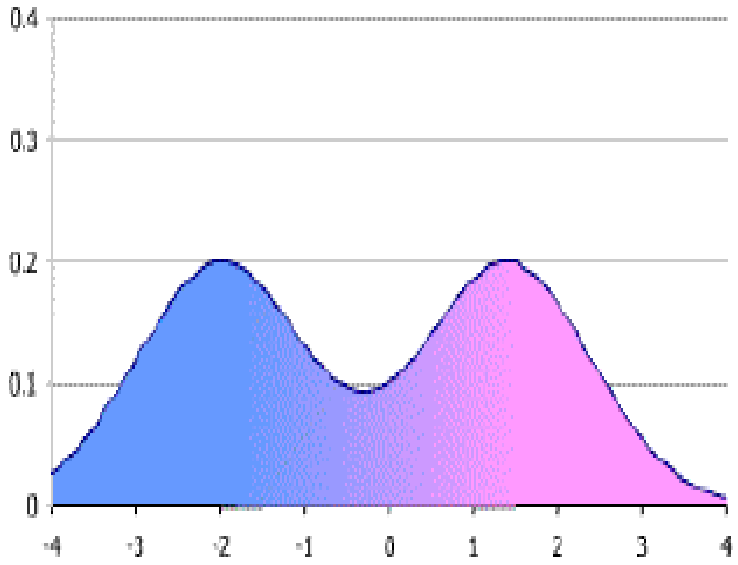
Negatively Skewed

Symmetric (Not Skewed)

Positively Skewed

Modality

- A distribution with two or more peaks called multimodal distribution.



Quantiles

- Measures of **non-central location** used to summarize a set of data.

- Examples of commonly used **quantiles:**
  - **Quartiles**
  - **Quintiles**
  - **Deciles**
  - **Percentiles**

Quartiles

- ***Quartiles split a set of ordered data into four parts.***

- Imagine cutting a chocolate bar into four equal pieces… How many cuts would you make? (yes, 3!).
- **Q1 is the First Quartile**

➢ 25% of the observations are smaller than Q1 and 75% of the observations are larger.

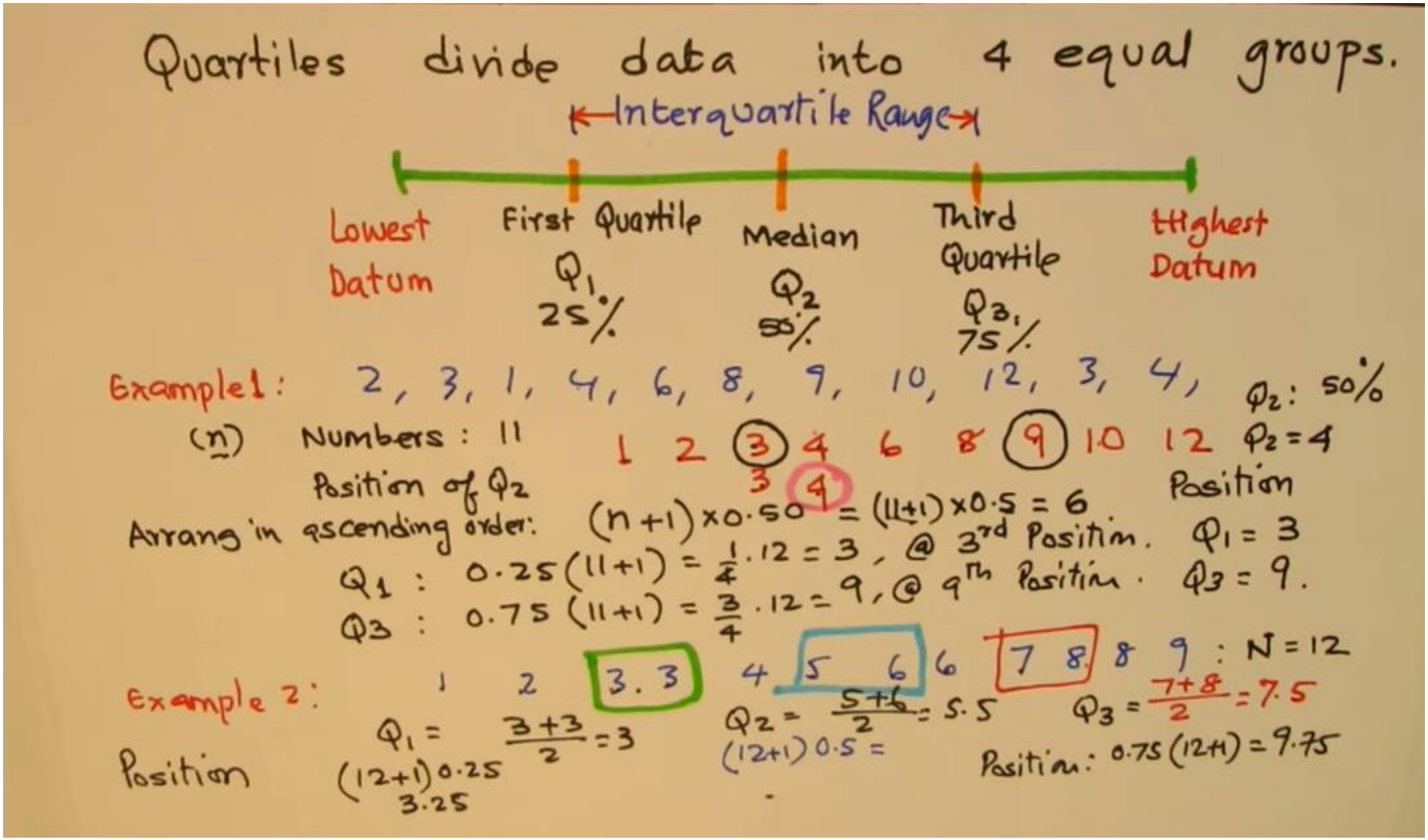
- **Q2 is the Second Quartile**

➢ **50%** of the observations are smaller than Q2 and 50% of the observations are larger. Same as the **Median**. It is also the **50th percentile**.

- **Q3 is the Third Quartile**

➢ 75% of the observations are smaller than Q3and 25% of the observations are larger.

Quartiles


Quartiles divide data into 4 equal groups.
Interquartile Range
Lowest Datum
First Quartile
Median
Third Quartile
Highest Datum
Q1 25%
Q2 50%
Q3 75%
Example 1: 2, 3, 1, 4, 6, 8, 9, 10, 12, 3, 4,
Q2: 50%
(n) Numbers : 11
1 2 3 4 6 8 9 10 12
Q2 = 4
Position of Q2
3 4
Position
Arrang in ascending order:
(n+1) x 0.50 = (11+1) x 0.5 = 6
Q1 = 3
Q1 : 0.25(11+1) = 1/4 . 12 = 3, @ 3rd Positim.
Q3 = 9.
Q3 : 0.75 (11+1) = 3/4 . 12 = 9, @ 9th Positim.
Example 2:
1 2 3.3 4 5 6 6 7 8 8 9 : N = 12
Q1 = (3+3)/2 = 3
Position
(12+1) 0.25
3.25
Q2 = (5+6)/2 = 5.5
(12+1) 0.5 =
Q3 = (7+8)/2 = 7.5
Positim: 0.75 (12+1) = 9.75

Exercise

Computer Sales (n = 12 salespeople)

Original Data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6

Compute the mean, median, mode, quartiles.

First order the data:

0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12

$\Sigma X_i = 76$

$\overline{X} = 76 / 12 = 6.33$ computers sold

Median = 6.5 computers

Mode = 10 computers

$Q_1 = 3.5$ computers, $Q_3 = 9.5$ computers

Other Quantiles

- Similar to what we just learned about quartiles, where 3 **quartiles** split the data into 4 equal parts,
- There are **9 deciles** dividing the distribution into 10 equal portions (tenths).
- There are **four quintiles** dividing the population into 5 equal portions and 99 percentiles.
- In all these cases, the convention is the same.

- The point, be it a quartile, decile, or percentile, takes the value of one of the observations **or** it has a value halfway between two adjacent observations.

Exercise

Data (n=16):

1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 5, 5, 6, 7, 8, 10

Compute the mean, median, mode, quartiles.

Answer.

1 1 2 2 | 2 2 3 3 | 4 4 5 5 | 6 7 8 10

Mean = 65/16 = 4.06

Median = 3.5

Mode = 2

$Q_1 = 2$

Q_2 = Median = 3.5

$Q_3 = 5.5$

Exercise: # absences

❑ Data – number of absences (n=13) :

- 0, 5, 3, 2, 1, 2, 4, 3, 1, 0, 0, 6, 12

➢ **Compute** the mean, median, mode, quartiles.

Answer. First order the data:

- 0, 0, 0, 1, 1, 2, 2 , 3, 3, 4, 5, 6, 12
- **Mean** = 39/13 = 3.0 absences
- **Median** = 2 absences
- **Mode** = 0 absences
- **Q1** = .5 absences
- **Q3** = 4.5 absences

Exercise: Reading Level

Data: Reading Levels of 16 eighth graders.
5, 6, 6, 6, 5, 8, 7, 7, 7, 8, 10, 9, 9, 9, 9, 9

Answer. First, order the data:

5 5 6 6 | 6 7 7 7 | 8 8 9 9 | 9 9 9 10

Sum=120.

Mean= 120/16 = 7.5 This is the average reading level of the 16 students.

Median = Q_2 = 7.5

Q_1 = 6, Q_3 = 9

Mode = 9

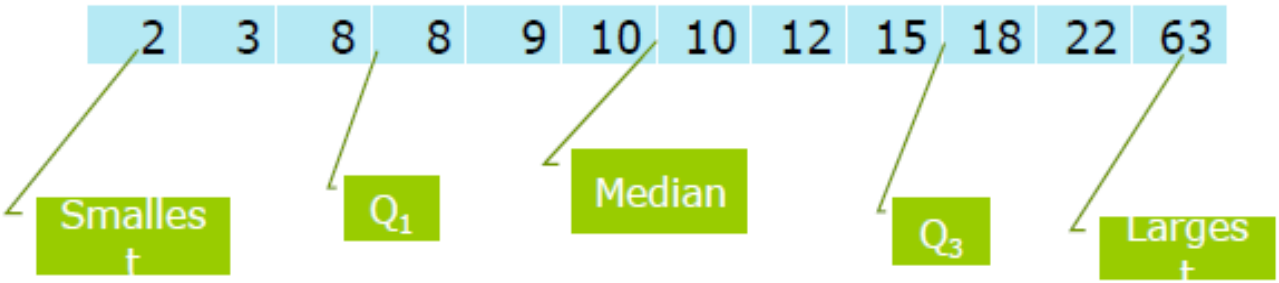
Five Number Summary

- When examining a distribution for shape, sometime the five number summary is useful:

  Smallest| Q1 | Median | Q3 | Largest

- Example:

$\overline{X} = 15$



5-number summary: 2 | 8 | 10 | 16.5 | 63

This data is right-skewed.

In right-skewed distributions, the distance from Q_3 to $X_{largest}$ (16.5 to 63) is significantly greater than the distance from $X_{smallest}$ to Q_1 (2 to 8).

Methods of description (Descriptive statistics)

- **Two most common methods of description:**

1. Measures of **location** (Central Tendency).

2. Measures of **spread** (variation or dispersion).

Measures of Dispersion (Variability)

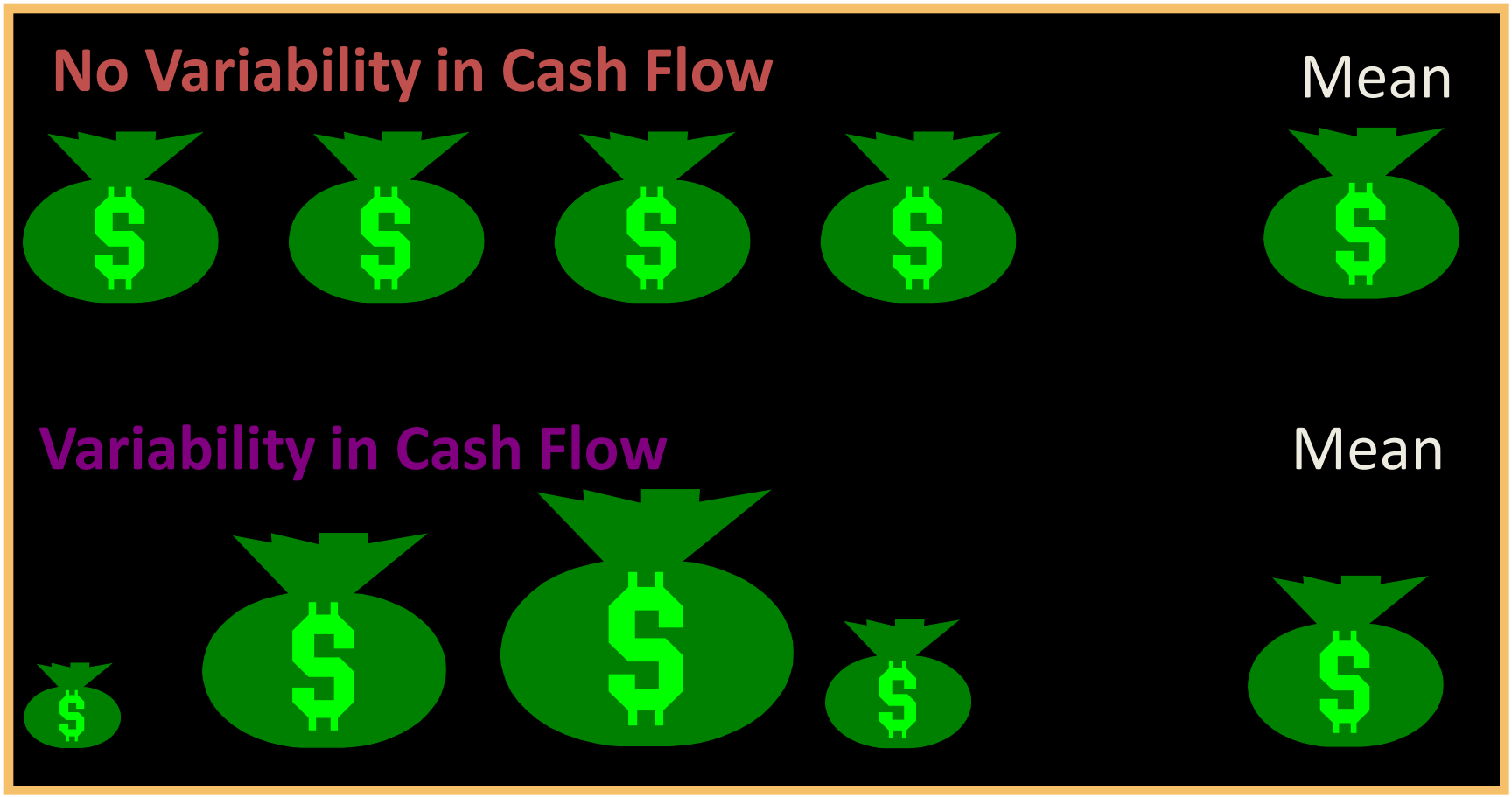
- A **measure** of **variability** is a summary statistic that represents the amount of **dispersion** in a dataset.
- It refers to **how spread out the scores are.**

- In other words, how similar or different participants are from one another on the variable. **It is either homogeneous or heterogeneous sample.**

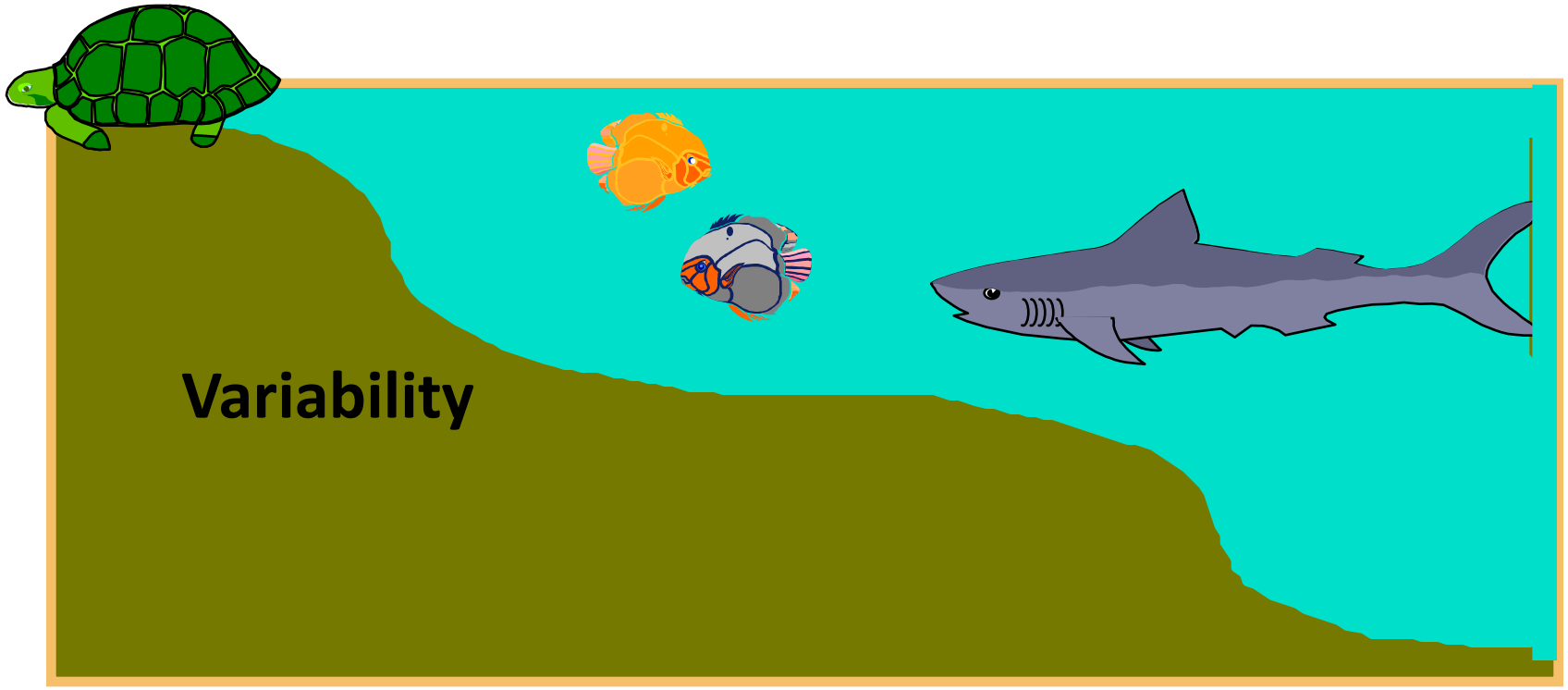
➢ A **low dispersion** indicates that the data points tend to be clustered tightly around the center.

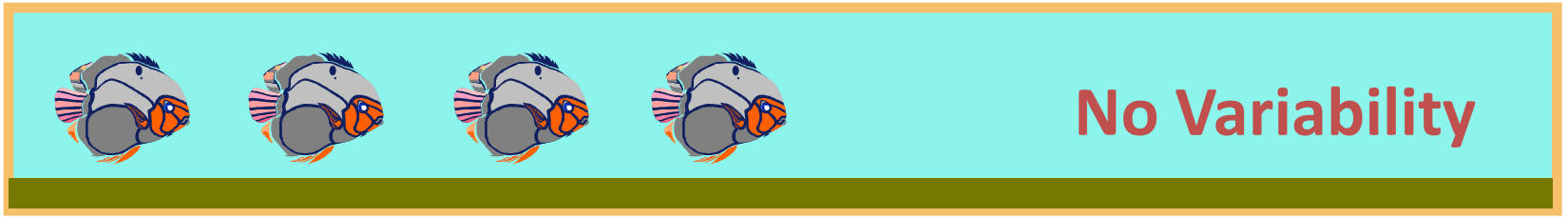
➢ **No Variability –- No Dispersion.**

Variability



Variability

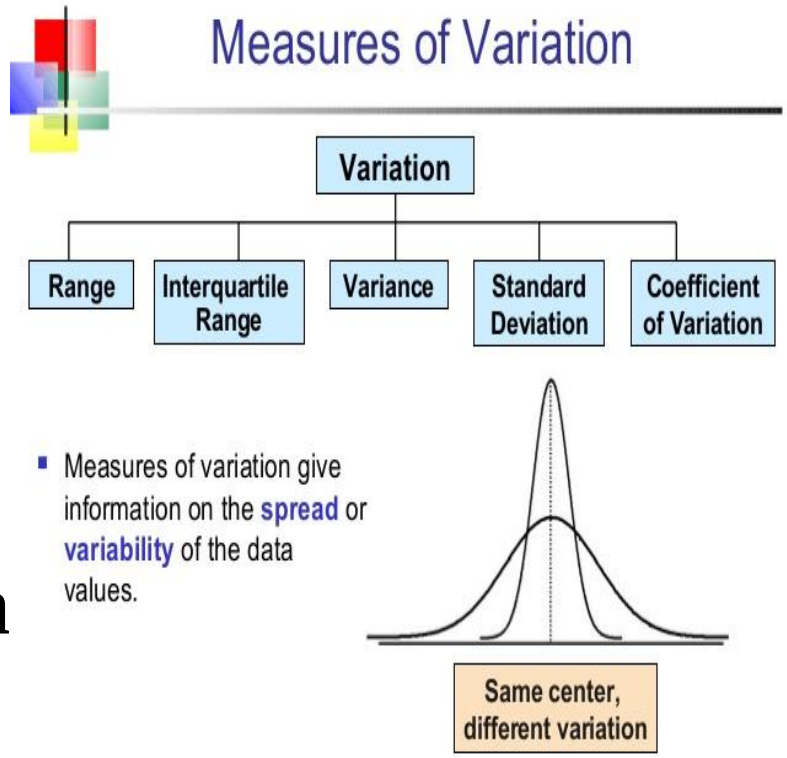




Measures of Dispersion

❑ We will study these **five measures of dispersion:**

- **Range**
- **Interquartile Range**
- **Variance**
- **Standard Deviation**
- **Coefficient of Variation**



Why do we need to look at measures of dispersion?

- **Why is it important to measure the spread of data?**

- A measure of spread gives us an idea of how well the mean.
- **For example**, represents the data. If the spread of values in the data set is **large**, the **mean** is **not as representative** of the data as if the spread of data is small.
- This is because a **large spread** indicates that there are probably **large differences** between individual scores.

Example / Why is it Important? ..

- **Example:** You want to choose the best brand of paint for your house. You are interested in how long the paint lasts before it fades and you must repaint. The choices are narrowed down to 2 different paints. The results are shown in the chart. Which paint would you choose?

The chart indicates the number of months a paint lasts before fading.

Paint A	Paint B
10	35
60	45
50	30
30	35
40	40
20	25
210	**210**

Does the Average Help?

- ***Paint A:*** Avg = 210/6 = 35 months

- ***Paint B:*** Avg = 210/6 = 35 months

- They both last 35 months before fading.
- No help in deciding which to buy.

Consider the Spread

- ***Paint A:*** Spread = 60 – 10 = 50 months

- ***Paint B:*** Spread = 45 – 25 = 20 months

- Paint B has a smaller ***variance*** which means that it performs more consistently.

➢ Choose paint B.

The Range

- The range of a data set is the **difference** between the **highest score** and the **lowest score** in the distribution.

- It is the simplest measure of variability.

- It provides a quick summary of a distribution's variability.

- It also provides useful information about a distribution when there are **extreme values.**

The Range …

Range = $X_{largest} - X_{smallest}$

- **Example:**

➢ **1**, 2, 3, 4, 5, 8, 9, 21, 25, **30**

➢ Answer: Range = 30 – 1 = 29.

- **Pros**:

➢ **Easy to calculate**

- **Cons:**

➢ Value of range is only determined by **two values.**

➢ One problem with the range is **that it is influenced by extreme values at either end.**

The Range …

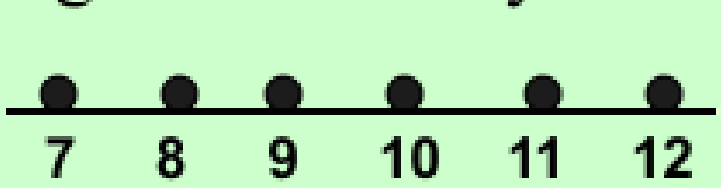
- If we have two extreme values **one very low** and the **other very high** then the **range** is **useless.**

- **Example:** If we have a set of marks for some students , one student got **zero** and another one got **100**, however the other students marks were between **60-70.**

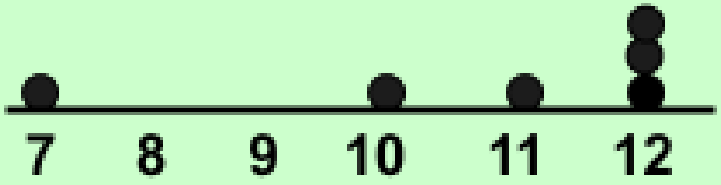
- **Theoretically,** the range should be **10**; however the **actual range** is **100** because of the two extreme values, so , in this case it is useless.

The Range …

- Ignores the way in which data are distributed



Range = 12 - 7 = 5



Range = 12 - 7 = 5

- Sensitive to outliers

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**5**

Range = 5 - 1 = 4

1,1,1,1,1,1,1,1,1,1,1,2,2,2,2,2,2,2,2,3,3,3,3,4,**120**

Range = 120 - 1 = 119

Standard Deviation

- The **standard deviation** is the **square root of the variance**.
- **For sample** $s = \sqrt{s^2}$

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

- It is measured in the **same units** as the data , making it more easily comparable , than the variance, to the mean.

- **Example:** If you are measuring **weight**, then the **SD** for the data would be in **Kg**.

➢ Standard deviation takes the units of measure of the data that it represents.

- Standard deviation tells us about how the data is distributed about the mean value.

Steps for Calculating Standard Deviation

1. Calculate the difference between each value and the mean.
2. Square each difference.
3. Add the squared differences.
4. Divide this total by n-1 to get the sample variance.
5. Take the square root of the sample variance to get the sample standard deviation.

Sample Data (X_i) : 10 12 14 15 17 18 18 24

$n = 8$ Mean $= \overline{X} = 16$

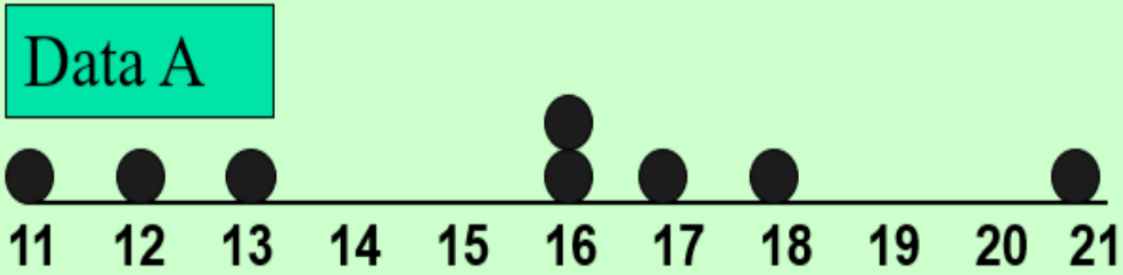
$$S = \sqrt{\frac{(10-\overline{X})^2 + (12-\overline{X})^2 + (14-\overline{X})^2 + \cdots + (24-\overline{X})^2}{n-1}}$$

$$= \sqrt{\frac{(10-16)^2 + (12-16)^2 + (14-16)^2 + \cdots + (24-16)^2}{8-1}}$$

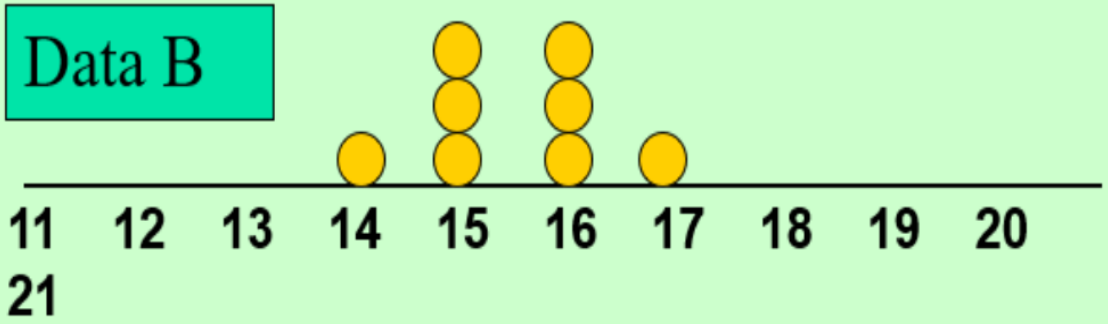
$$= \sqrt{\frac{130}{7}} = 4.3095$$

→ A measure of the "average" scatter around the mean

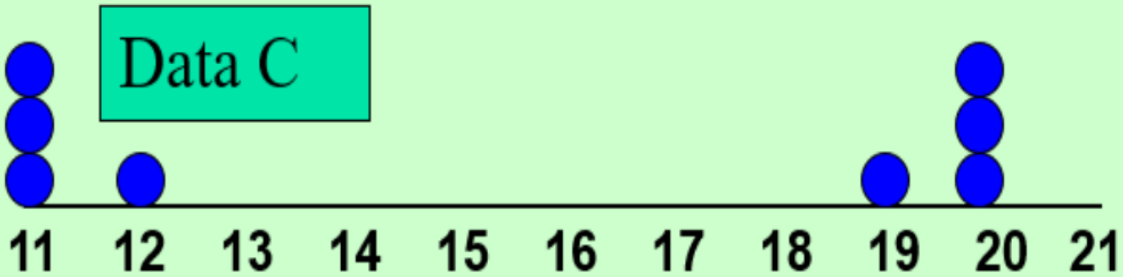
Comparing Standard Deviations



Mean = 15.5
S = 3.338

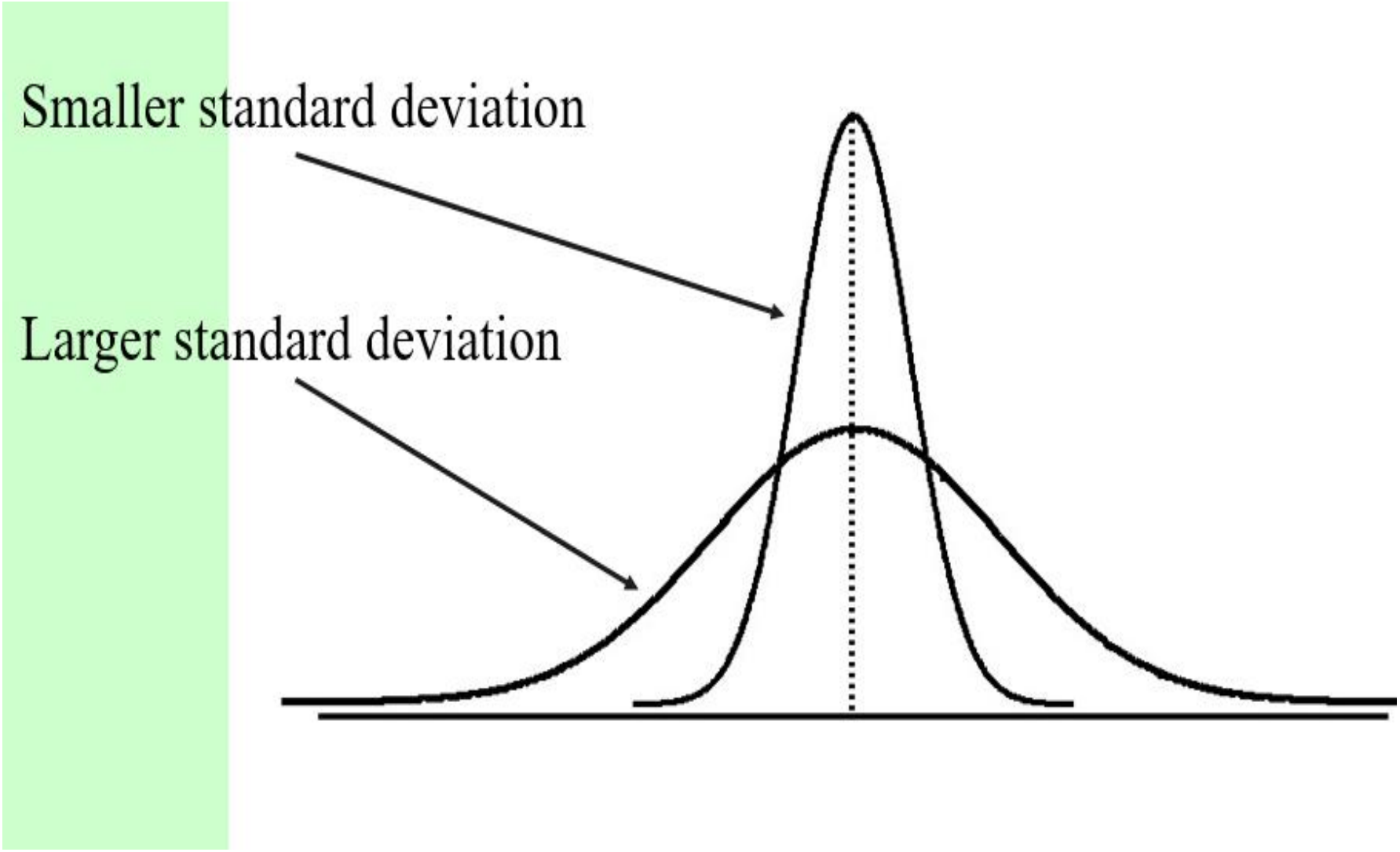


Mean = 15.5
S = 0.926



Mean = 15.5
S = 4.570

Comparing Standard Deviations



Standard Deviation ..

- Instead, we use:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- This is the "definitional formula" for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the "least squares property."
- Note we divide by (n−1), not n. This will be referred to as a loss of one degree of freedom.

Standard Deviation …

- **Standard Deviation** is a standardized measure of dispersion of the data around the mean, mathematically the standard deviation is the square root of the variance.
  - Interval, and ratio data.

Body Temperature	
Patient Name	Temp.
001	**37**
002	37
003	38
004	38.5
005	38.5
Mean	**37.8**

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

$$s = \sqrt{0.575} = 0.758$$

Standard Deviation..

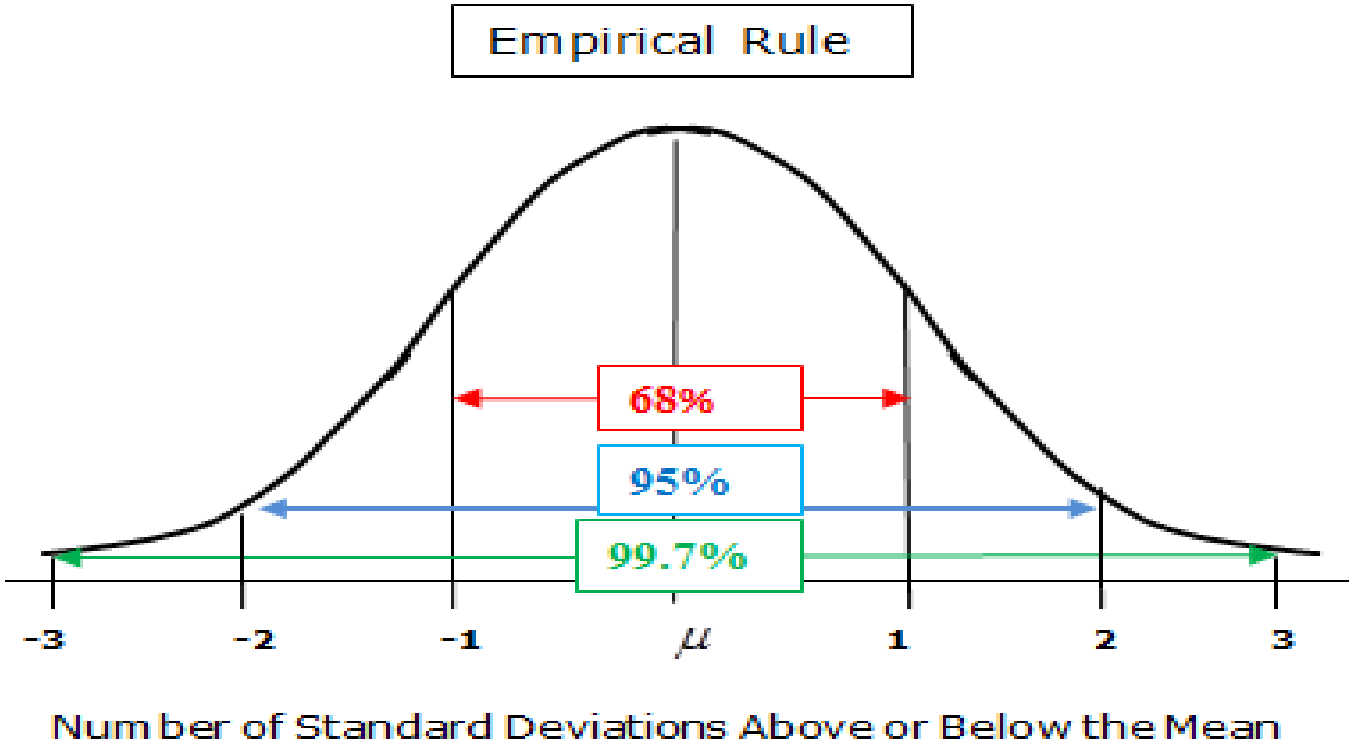
- The **smaller the standard deviation**, **the better is the mean as the summary of a typical score**.

- **Example 1:** 10 people weighted 150 kg, the SD would be zero, and the mean of 150 would communicate perfectly accurate information about all the participants wt.

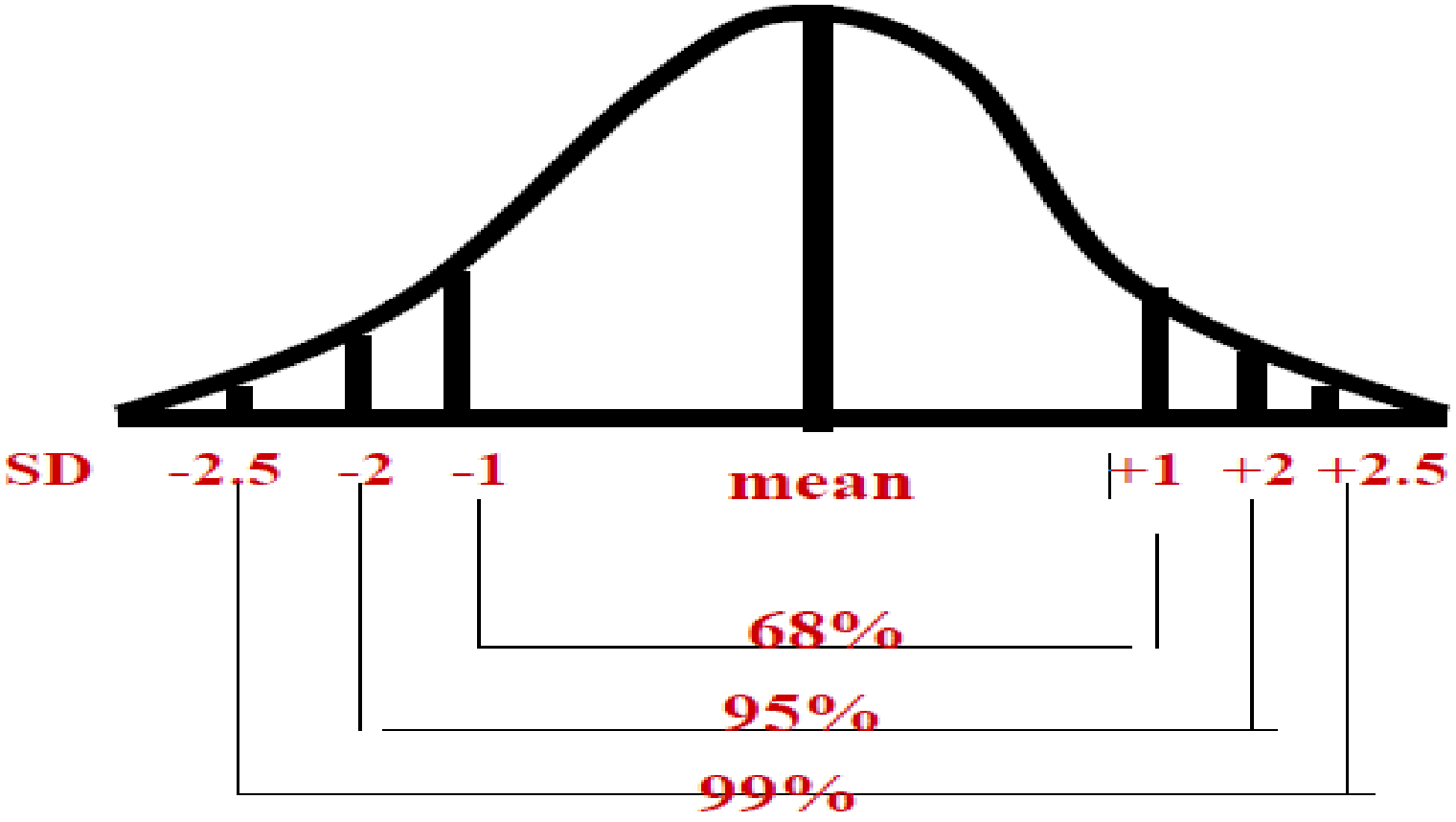
- **Example 2 :** Would be a heterogeneous sample 5 people 100 kg and another five people 200 kg. The mean still 150, but the SD would be 52.7.

Standard Deviation…

- In **normal distribution** there are **3 SDs above** the mean and **3 SDs below** the mean.



Relationship between SD and frequency distribution



Standard Deviation..

Example. Two data sets, X and Y. Which of the two data sets has greater variability? Calculate the standard deviation for each.

We note that both sets of data have the same mean:

$\overline{X} = 3$

$\overline{Y} = 3$

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

(continued...)

Standard Deviation …

$$S_X = \sqrt{\frac{10}{4}} = 1.58$$

X	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		$\Sigma=0$	10

$$S_Y = \sqrt{\frac{80}{4}} = = 4.47$$

Y	$\bar{Y}$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		$\Sigma=0$	80

[Check these results with your calculator.]

Standard Deviation: N vs. (n-1)

Note that $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(X_i-\mu)^2}{N}}$ and $s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}}$

- You divide by N only when you have taken a census and therefore know the population mean. This is rarely the case.
- Normally, we work with a sample and calculate sample measures, like the sample mean and the sample standard deviation:
- The reason we divide by n–1 instead of n is to assure that *s* is an *unbiased* estimator of σ.
  - We have taken a shortcut: in the second formula we are using the sample mean, $\overline{X}$, a statistic, in lieu of μ, a population parameter. Without a correction, this formula would have a tendency to understate the true standard deviation. We divide by n–1, which increases *s*. This makes it an *unbiased estimator* of σ.
  - We will refer to this as "losing one degree of freedom" (to be explained more fully later on in the course).

Variance

- **The variance** is the average is the squared differences between each data value and the mean.

Average (approximately) of squared deviations of values from the mean

- Sample variance:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$

Where $\overline{X}$ = arithmetic mean

n = sample size

X_i = i^{th} value of the variable X

Variance…

- **Variance** is a **measure of dispersion of the data around the mean, mathematically the variance is the average squared deviation from the mean.**
  - Interval, and ratio data.

Body Temperature	
Patient Name	Temp.
001	**37**
002	37
003	38
004	38.5
005	38.5
Mean	**37.8**

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

$$s^2 = \frac{(37-37.8)^2 + (37-37.8)^2 + (38-37.8)^2 + (38.5-37.8)^2 + (38.5-37.8)^2}{5-1}$$

$$s^2 = \frac{2.3}{4} = 0.575$$

Variance…

- It is **no actual unit** because of the square.

- **(n-1)**: is an indirect measure of variability , its called the **degree of freedom.**

- **Why do we square the result?**

➢ If we calculate the total differences between the values and the mean it will be **zero** ; so we square it to get the actual values.

Range, variance, and standard deviation.

- The **more** the data are spread out, the **greater** the range, variance, and standard deviation.

- The **less** the data are spread out, the **smaller** the range, variance, and standard deviation.

- If the values are **all the same** (no variation), all these measures will be zero.

Coefficient of Variation (C.V)

- **The coefficient of variation** is a **measure of comparing of two dispersions or more,** mathematically the standard deviation is divided by the mean.
- It is **relative variability** around the mean.
- This can be used to compare two distributions directly to see which has more dispersion because it **does not depend on units of the distribution.**

$$C.V. = \frac{s}{\bar{x}} * (100)$$

	Sample 1	Sample 2
Age	**25 years**	**11 Years**
Mean Weight	145	80
SD	10	10
C.V.	**6.9**	**12.5**

Properties of the Coefficient of Variation

1. Useful for **comparing the variability** of two or more variables (compare between different things).

2. It is **independent of unit** of measurement.

3. Measures the **relative variation**.

4. Always in **percentage (%).**

Coefficient of Variation..

- Stock A:
  - Average price last year = $50
  - Standard deviation = $5

$$CV_A = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$50} \cdot 100\% = 10\%$$

- Stock B:
  - Average price last year = $100
  - Standard deviation = $5

$$CV_B = \left(\frac{S}{\bar{X}}\right) \cdot 100\% = \frac{\$5}{\$100} \cdot 100\% = 5\%$$

Both stocks have the same standard deviation, but stock B is less variable relative to its price

Example: Stock Prices

Which stock is more volatile?

Closing prices over the last 8 months:

$$CV_A = \frac{\$1.62}{\$1.70} \times 100\% = 95.3\%$$

$$CV_B = \frac{\$11.33}{\$188.88} \times 100\% = 6.0\%$$

	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

Answer: The standard deviation of B is higher than for A, but A is more volatile:

Inter-Quartile Range-IQR

- **The interquartile rang** of a data set is the difference between the **third quartile** and the **first quartile.**

- The Interquartile range (IQR) is the score at the **75th percentile or 3rd quartile (Q3) minus the score at the 25th percentile or first quartile (Q1).** Are the most used to define **outliers.**

- ***It is not sensitive to extreme values.***

Inter-Quartile Range-IQR..

- IQR = Q3 – Q1
- **Example:**

➢ (n = 15): **0, 0, 2, 3, 4, 7, 9, 12, 17, 18, 20, 22, 45, 56, 98**

✓ Q1 = 3, Q3 = 22

✓ **IQR** = 22 – 3 = 19 (Range = 98).

- This is basically the range of the central 50% of the observations in the distribution.

- **Problem:** The Interquartile range does not take into account the variability of the total data (only the central 50%). We are "throwing out" half of the data.

Exercise: Test Scores

Data (n=10): 0, 0, 40, 50, 50, 60, 70, 90, 100, 100

Compute the mean, median, mode, quartiles (Q1, Q2, Q3), range, interquartile range, variance, standard deviation, and coefficient of variation. We shall refer to all these as the descriptive (or summary) statistics for a set of data.

Answer. First order the data:

0, 0, 40, 50, 50 ¦ 60, 70, 90, 100, 100

- Mean: $\Sigma X_i = 560$ and $n = 10$, so $\overline{X} = 560/10 = 56$.
  Median = $Q_2 = 55$
- $Q_1 = 40$; $Q_3 = 90$ (Note: Excel gives these as $Q_1 = 42.5$, $Q_3 = 85$.)
- Mode = 0, 50, 100
  Range = 100 – 0 = 100
- IQR = 90 – 40 = 50
- $s^2 = 11{,}840/9 = 1315.5$
- $s = \sqrt{1315.5} = 36.27$
- CV = (36.27/56) x 100% = 64.8%

Relative Standing

- It provides information about the **position of an individual score value within a distribution scores.**

- **Two types**:

➢ Percentile Ranks.

➢ Standard Scores

Percentile Ranks

- It is the **percentage of scores** in the distribution that **fall at or below a given value.**

- P = Number of scores less than a given score divided by total number scores X 100.

- **Example:**

✓ Suppose you received a score of 90 on a test given to a class of 50 people.
✓ Of your classmates, 40 had scores lower than 90.
✓ P = 40/50 X 100 = 80.
✓ YOU achieved a higher score than 80% of the people who took the test, which also means that almost 20% who took the test did better than you.

- Percentiles are symbolized by the **letter P**, with a subscript indicating the percentage below the score value.

- Hence, P60 refers to the 60th percentile and stands for the score below which 60% of values fall.

Percentile Ranks

- The statement P40= 55 means that 40% of the values in the distribution fall below the score 55.

- The are several interpercentile measures of variability. The most common being the Interquartile range (IQR).

Standard Scores

- There are scores that are expressed in **terms of their relative distance from the mean**.

- It provides information **not only about rank but also distance between scores.**

- **It often called Z-score.**

Z Score

- Is a standard score that indicates how many SDs from the mean a particular values lies.

- Z = Score of value – mean of scores divided by standard deviation.

Standard Normal Scores

- How many standard deviations away from the mean are you?

Standard Score (Z) = $\frac{\textit{Observation} - \textit{mean}}{\textit{Standard deviation}}$

"Z" is normal with mean 0 and standard deviation of 1.

Standard Normal Scores

- Example: Male Blood Pressure, mean = 125, s = 14 mmHg
  - BP = 167 mmHg
  - BP = 97

$$Z = \frac{167 - 125}{14} = 3.0$$

$$Z = \frac{97 - 125}{14} = -2.0$$

What is the Usefulness of a Standard Normal Score?

- It tells you **how many SDs** (s) an observation is from the mean.
- Thus, it is a way of quickly **assessing how "unusual" an observation** is.

Example: Suppose the mean BP is 125 mmHg, and standard deviation = 14 mmHg

- Is 167 mmHg an unusually high measure?
- If we know Z = 3.0, does that help us?

Standardizing Data: Z-Scores

To compute the Z-scores:

$$Z = \frac{X - \bar{X}}{s}$$

Example.

Data: 0, 2, 4, 6, 8, 10

$\bar{X} = 30/6 = 5$; s = 3.74

x	→	z
0	$\frac{0-5}{3.74}$	-1.34
2	$\frac{2-5}{3.74}$	-.80
4	$\frac{4-5}{3.74}$	-.27
6	$\frac{6-5}{3.74}$	.27
8	$\frac{8-5}{3.74}$	.80
10	$\frac{10-5}{3.74}$	1.34

Quantitative Analysis Software

- SAS (http://www.sas.com)
- SPSS (http://www-01.ibm.com/software/analytics/spss/)
- STATA (http://www.stata.com/)
- Microsoft Excel (!)
- Many others

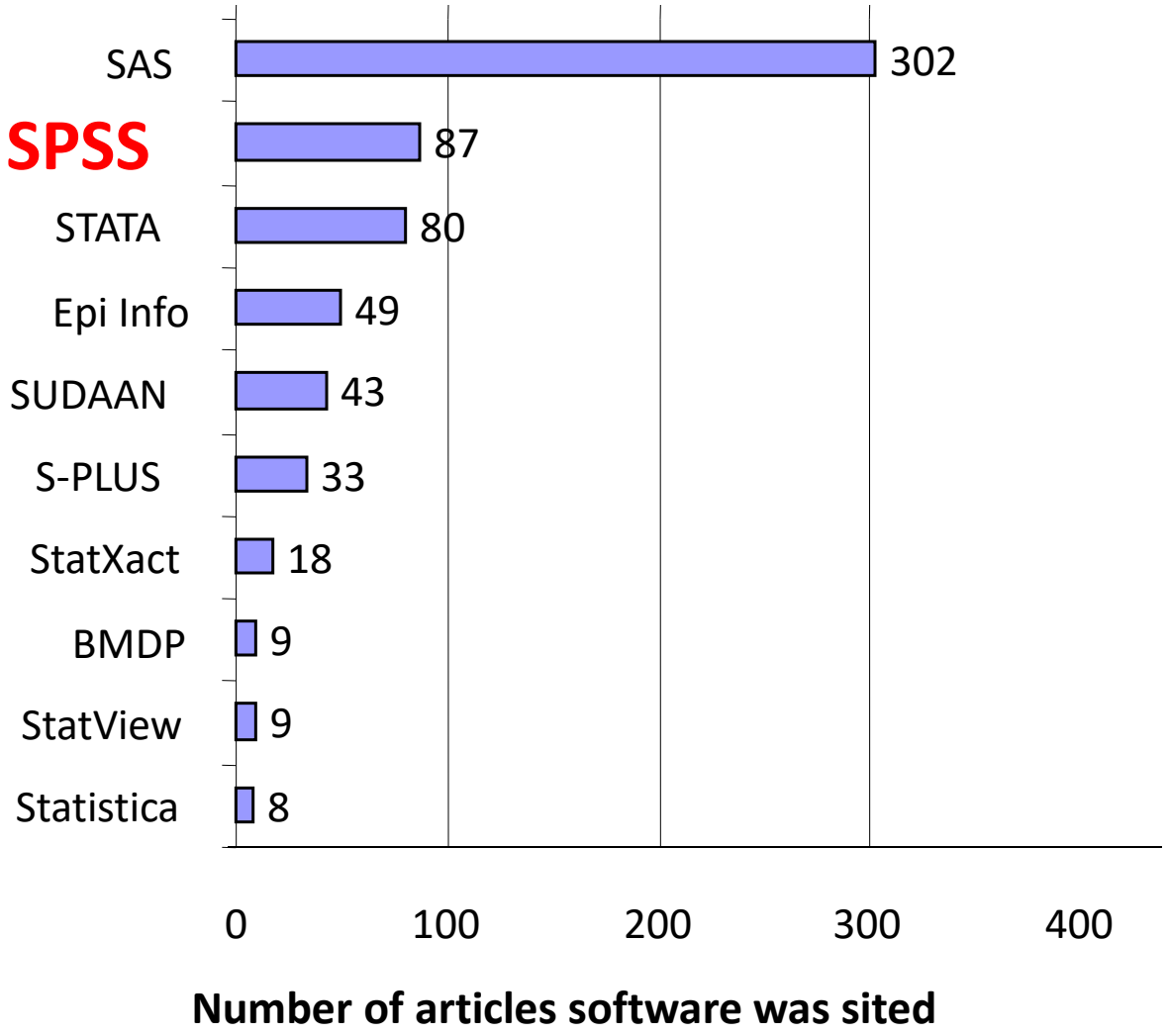
SPSS

Statistical Package for the Social Sciences

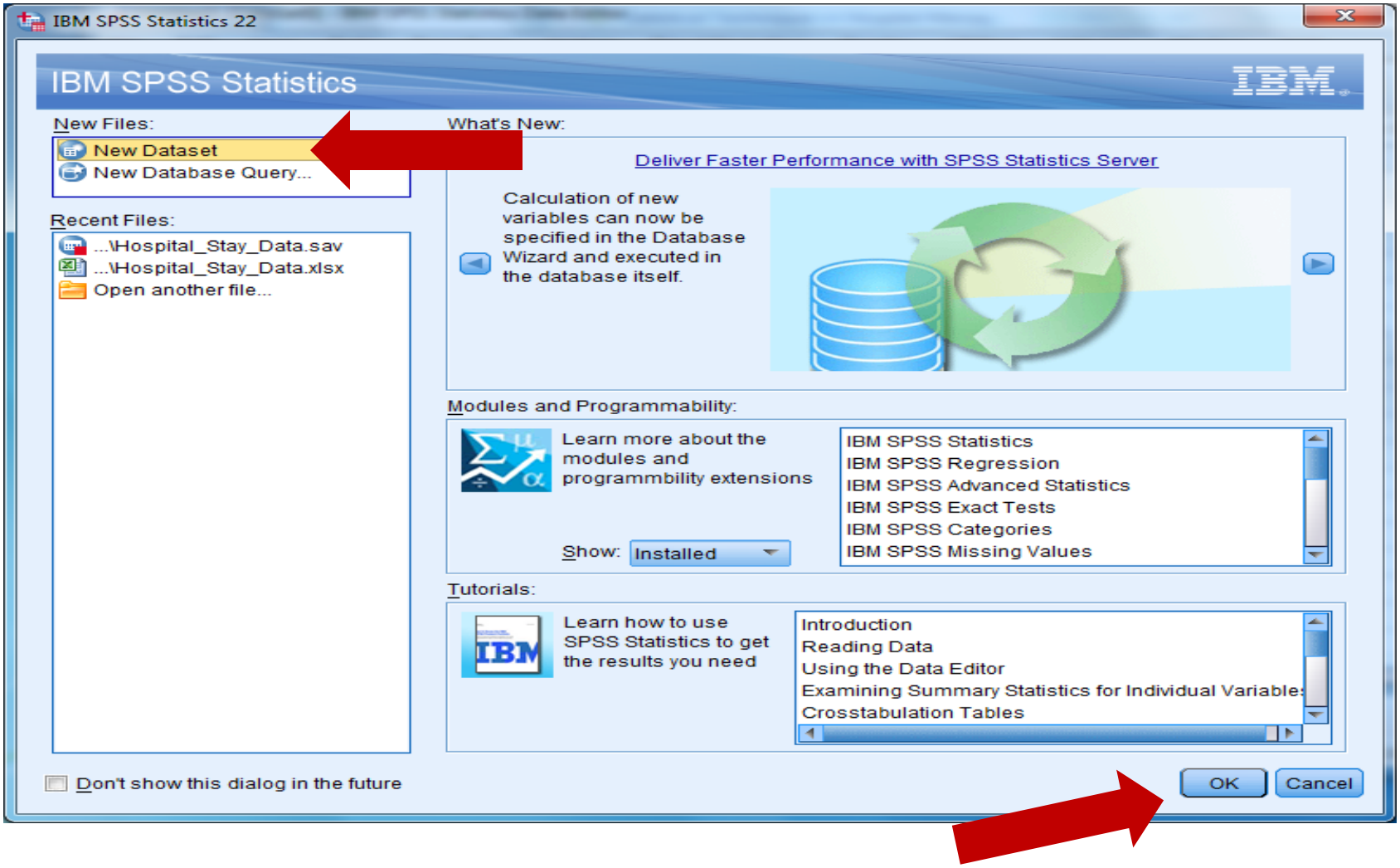
" الحزمة الاحصائية للعلوم الاجتماعية "

One of the most popular statistical packages which can perform highly complex data manipulation and analysis with simple instructions.

Statistical Software Packages Most Commonly Cited in the NEJM and JAMA between 1998 and 2002



Enter data in SPSS directly



SPSS interface

- **SPSS Windows has 3 windows:**

- **Data Editor :** Viewer or Draft Viewer which displays the output files.

Syntax Editor, which displays syntax files.

- **The Data Editor has two parts:**

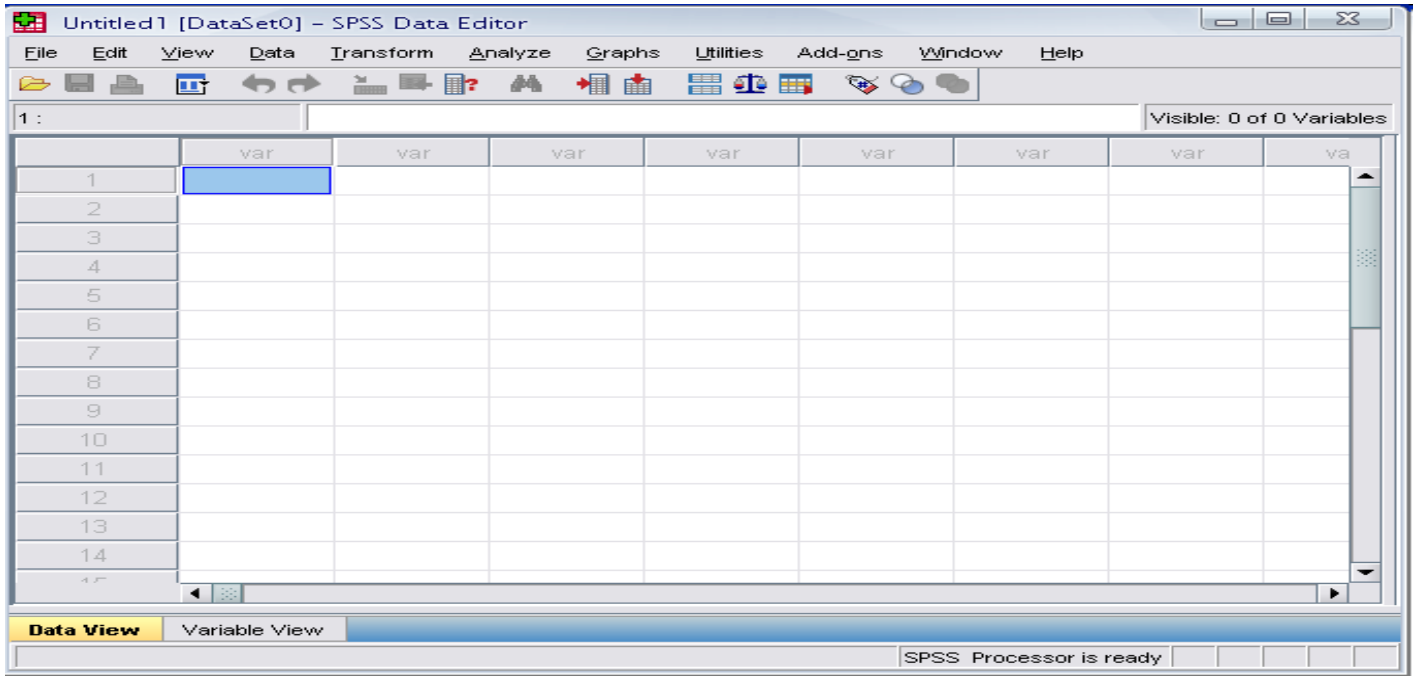
- **Data view window**, which displays data from the active file in spreadsheet format.
  - The place to enter data
  - **Columns**: variables
  - **Rows**: records

- **Variable View window**, which displays metadata or information about the data in the active file, such as variable names and labels, value labels, formats, and missing value indicators.
  - The place to enter variables
  - List of all variables
  - Characteristics of all variables

Opening SPSS

- The default window will have the data editor
- There are two sheets in the window:

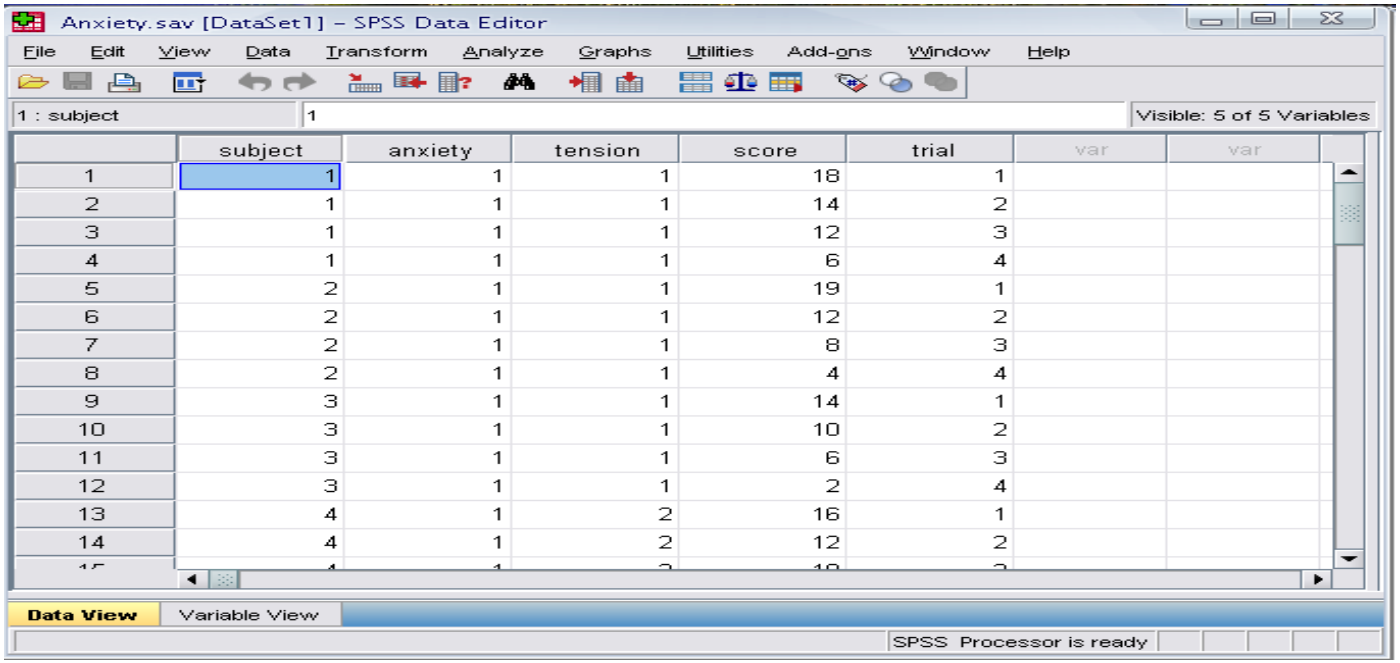
1. Data view
2. Variable view



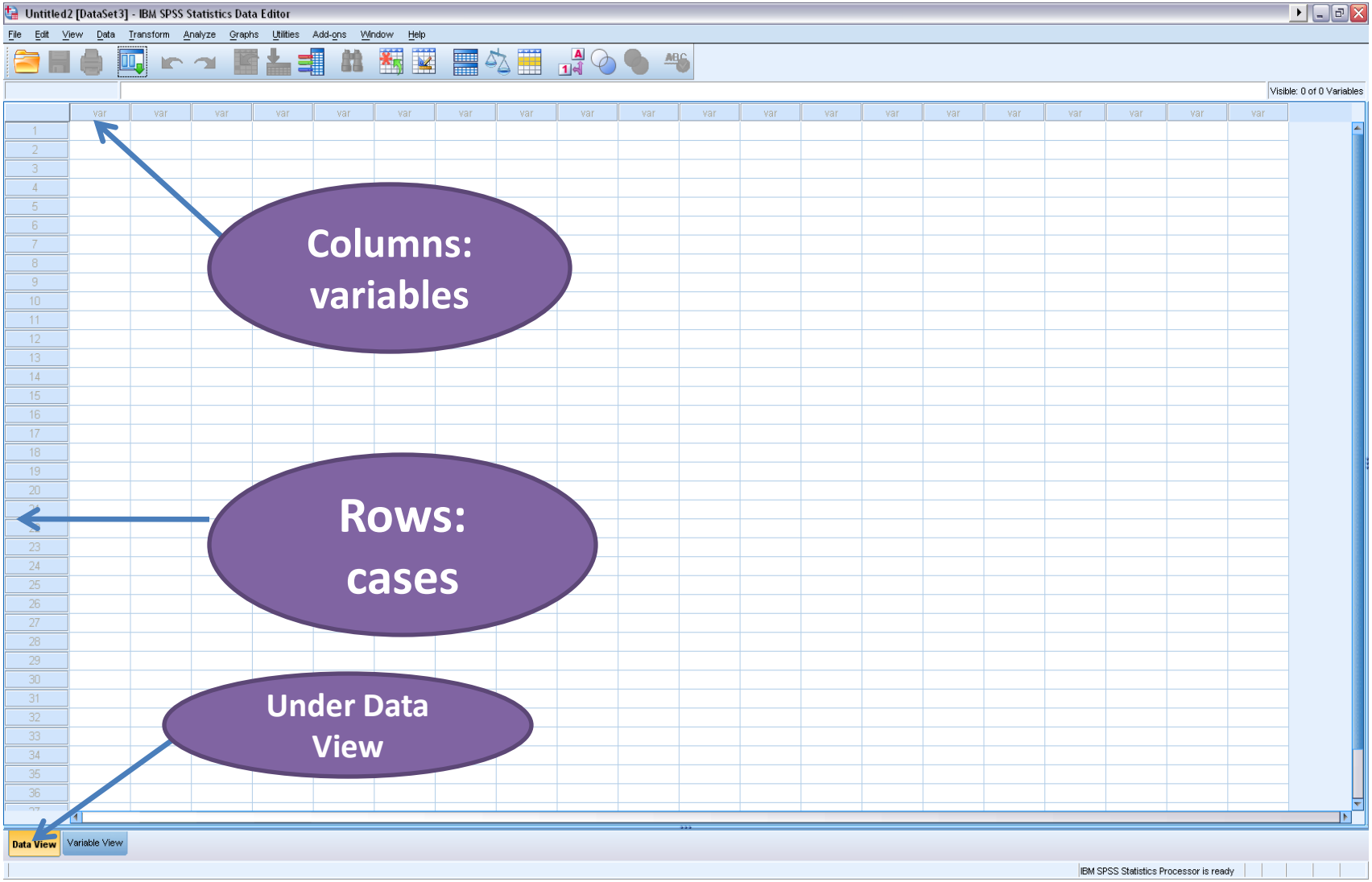
Data Editor

- Data Editor

  Spreadsheet-like system for defining, entering, editing, and displaying data. Extension of the saved file will be “sav.”



	subject	anxiety	tension	score	trial	var	var
1	1	1	1	18	1		
2	1	1	1	14	2		
3	1	1	1	12	3		
4	1	1	1	6	4		
5	2	1	1	19	1		
6	2	1	1	12	2		
7	2	1	1	8	3		
8	2	1	1	4	4		
9	3	1	1	14	1		
10	3	1	1	10	2		
11	3	1	1	6	3		
12	3	1	1	2	4		
13	4	1	2	16	1		
14	4	1	2	12	2		

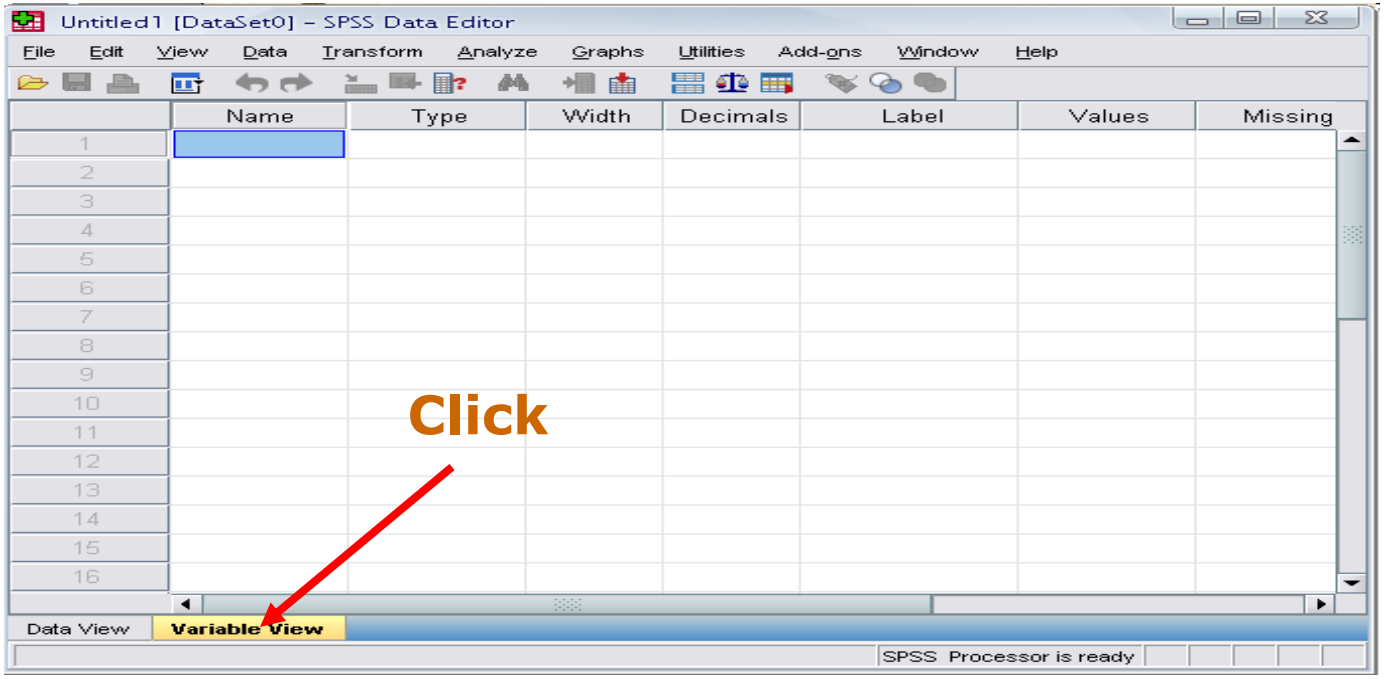

Untitled2 [DataSet3] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Visible: 0 of 0 Variables
Columns: variables
Rows: cases
Under Data View
Data View
Variable View
IBM SPSS Statistics Processor is ready

Data View window

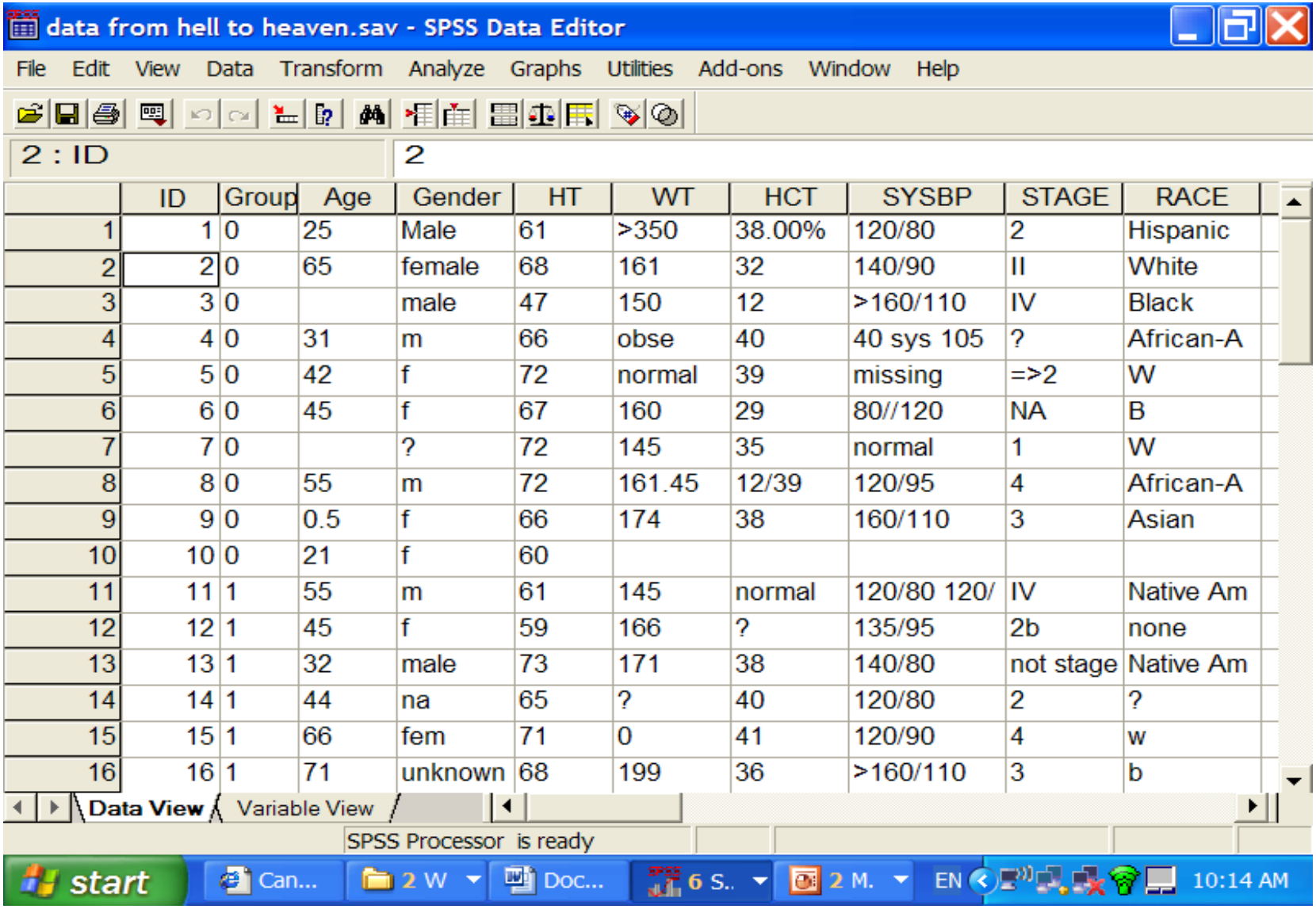
- The Data View window

  This sheet is visible when you first open the Data Editor and this sheet contains the data

- Click on the tab labeled Variable View



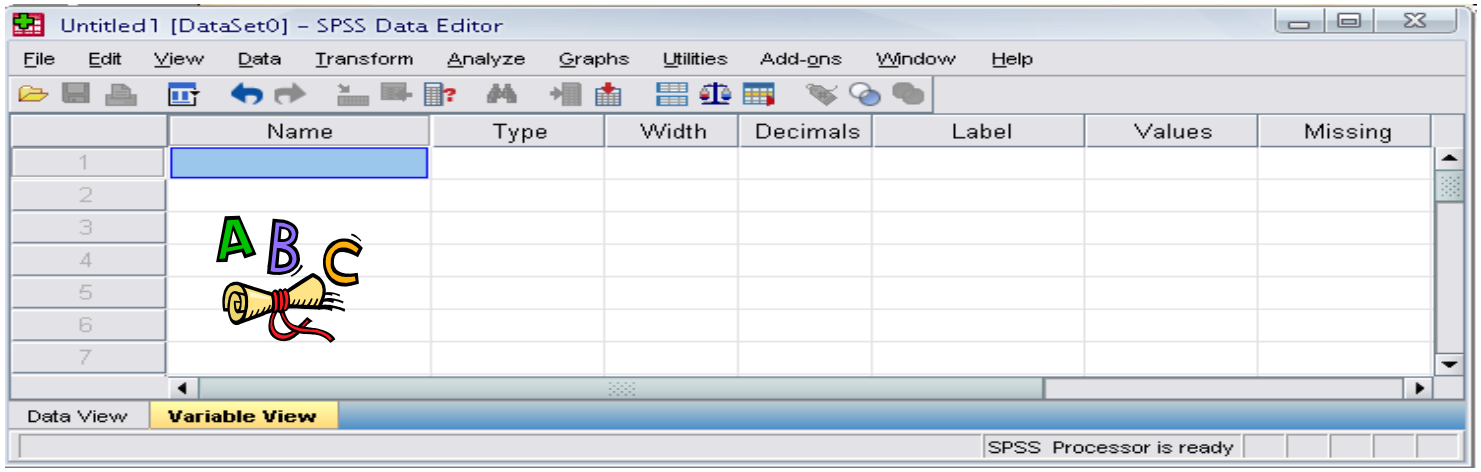
SPSS Data View



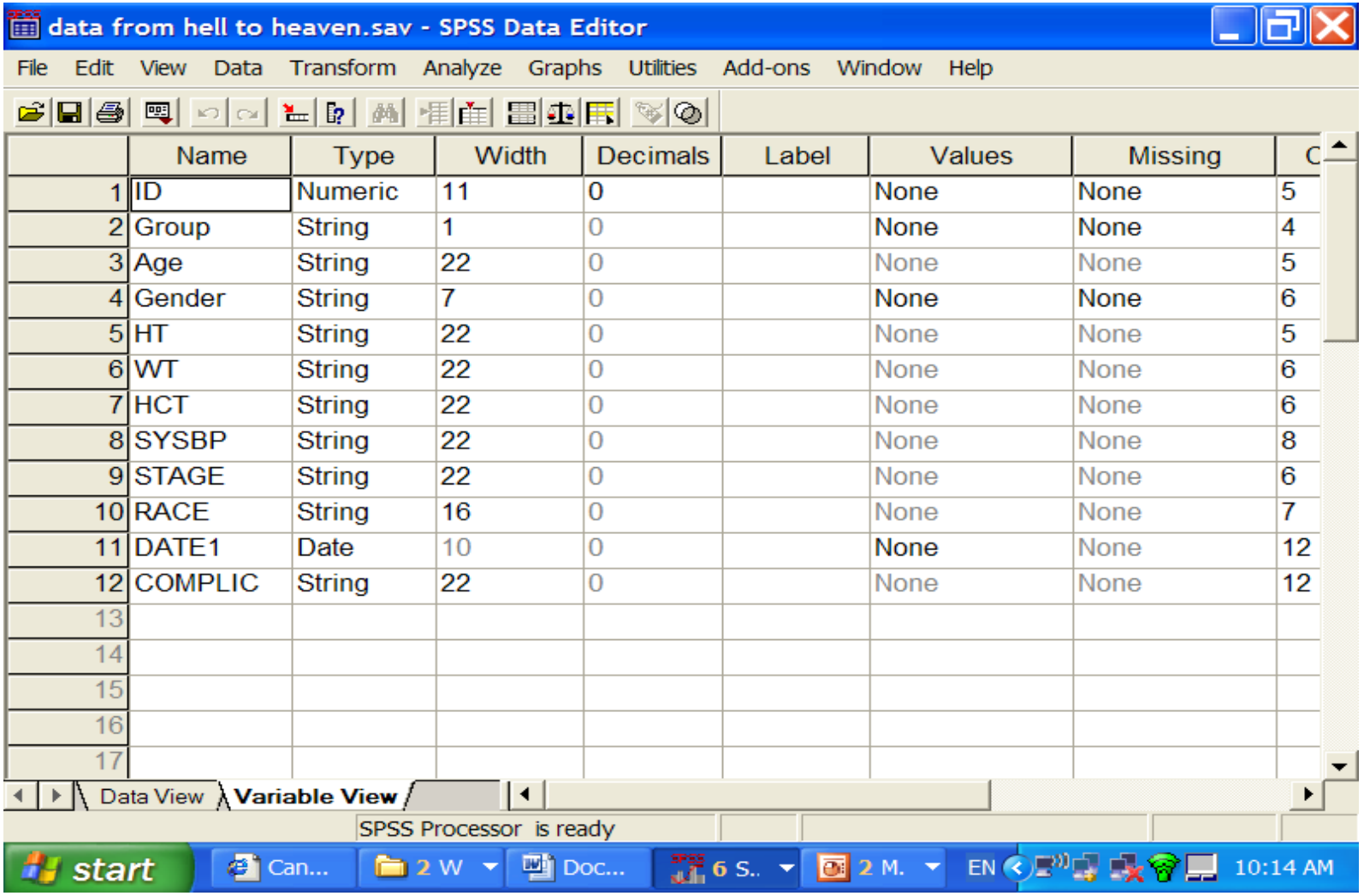
	ID	Group	Age	Gender	HT	WT	HCT	SYSBP	STAGE	RACE
1	1	0	25	Male	61	>350	38.00%	120/80	2	Hispanic
2	2	0	65	female	68	161	32	140/90	II	White
3	3	0		male	47	150	12	>160/110	IV	Black
4	4	0	31	m	66	obse	40	40 sys 105	?	African-A
5	5	0	42	f	72	normal	39	missing	=>2	W
6	6	0	45	f	67	160	29	80//120	NA	B
7	7	0		?	72	145	35	normal	1	W
8	8	0	55	m	72	161.45	12/39	120/95	4	African-A
9	9	0	0.5	f	66	174	38	160/110	3	Asian
10	10	0	21	f	60					
11	11	1	55	m	61	145	normal	120/80 120/	IV	Native Am
12	12	1	45	f	59	166	?	135/95	2b	none
13	13	1	32	male	73	171	38	140/80	not stage	Native Am
14	14	1	44	na	65	?	40	120/80	2	?
15	15	1	66	fem	71	0	41	120/90	4	w
16	16	1	71	unknown	68	199	36	>160/110	3	b

Variable View window

- This sheet contains information about the data set that is stored with the dataset
- Name
  - The first character of the variable name must be alphabetic
  - Variable names must be unique, and have to be less than 64 characters.
  - Spaces are NOT allowed.



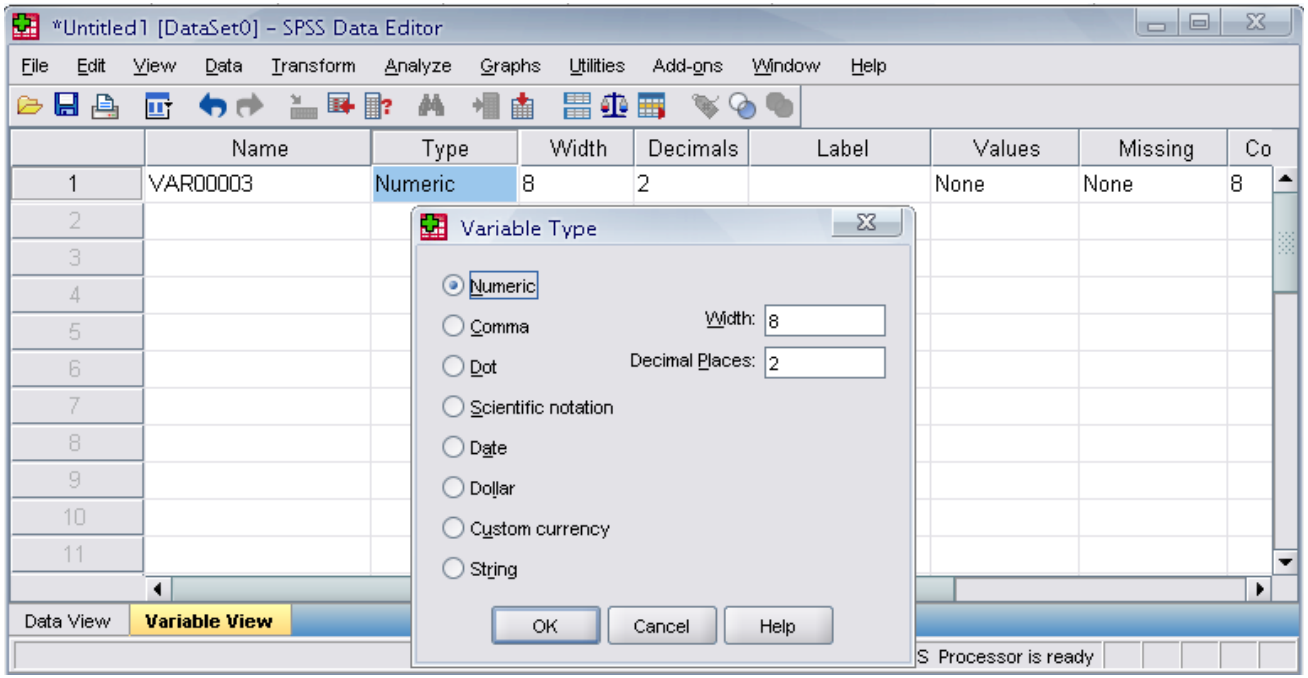
SPSS Variable View



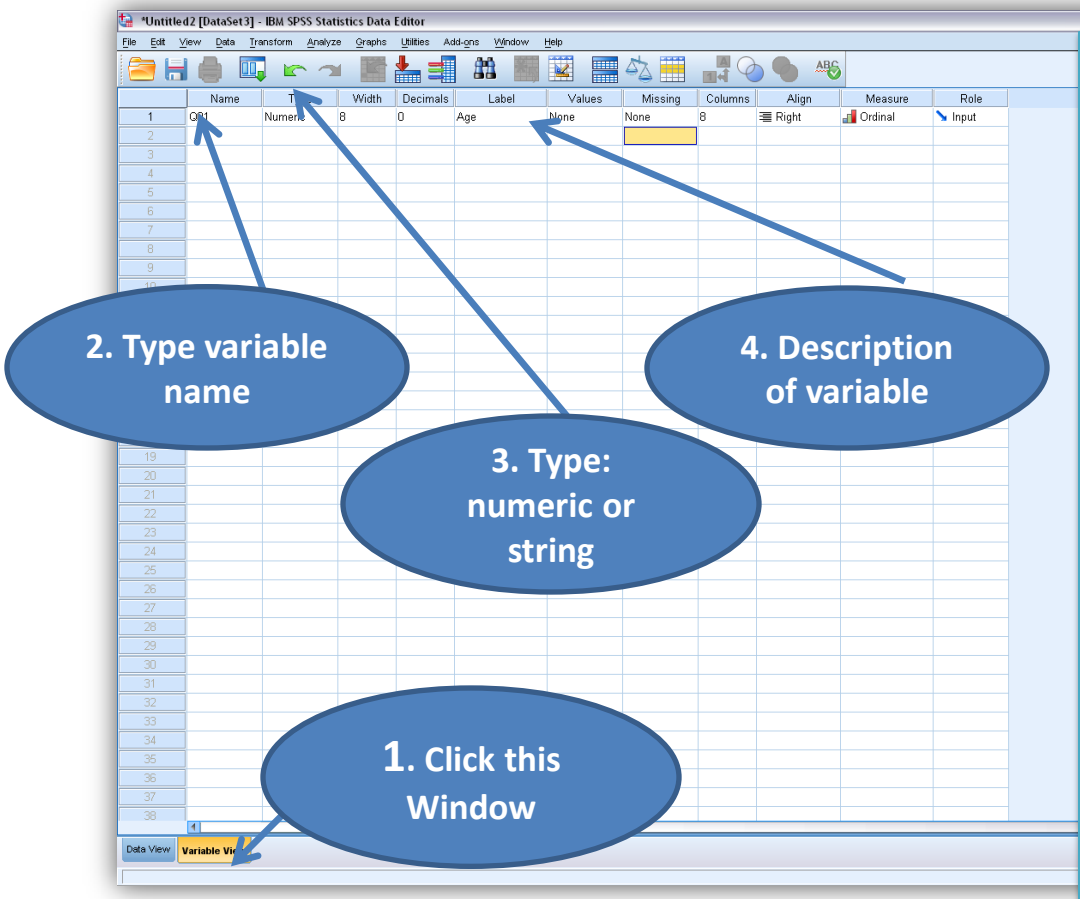
	Name	Type	Width	Decimals	Label	Values	Missing	C
1	ID	Numeric	11	0		None	None	5
2	Group	String	1	0		None	None	4
3	Age	String	22	0		None	None	5
4	Gender	String	7	0		None	None	6
5	HT	String	22	0		None	None	5
6	WT	String	22	0		None	None	6
7	HCT	String	22	0		None	None	6
8	SYSBP	String	22	0		None	None	8
9	STAGE	String	22	0		None	None	6
10	RACE	String	16	0		None	None	7
11	DATE1	Date	10	0		None	None	12
12	COMPLIC	String	22	0		None	None	12
13								
14								
15								
16								
17								

Variable View window: Type

- **Type**
  - Click on the ‘type’ box. The two basic types of variables that you will use are numeric and string. This column enables you to specify the type of variable.



Enter Variables



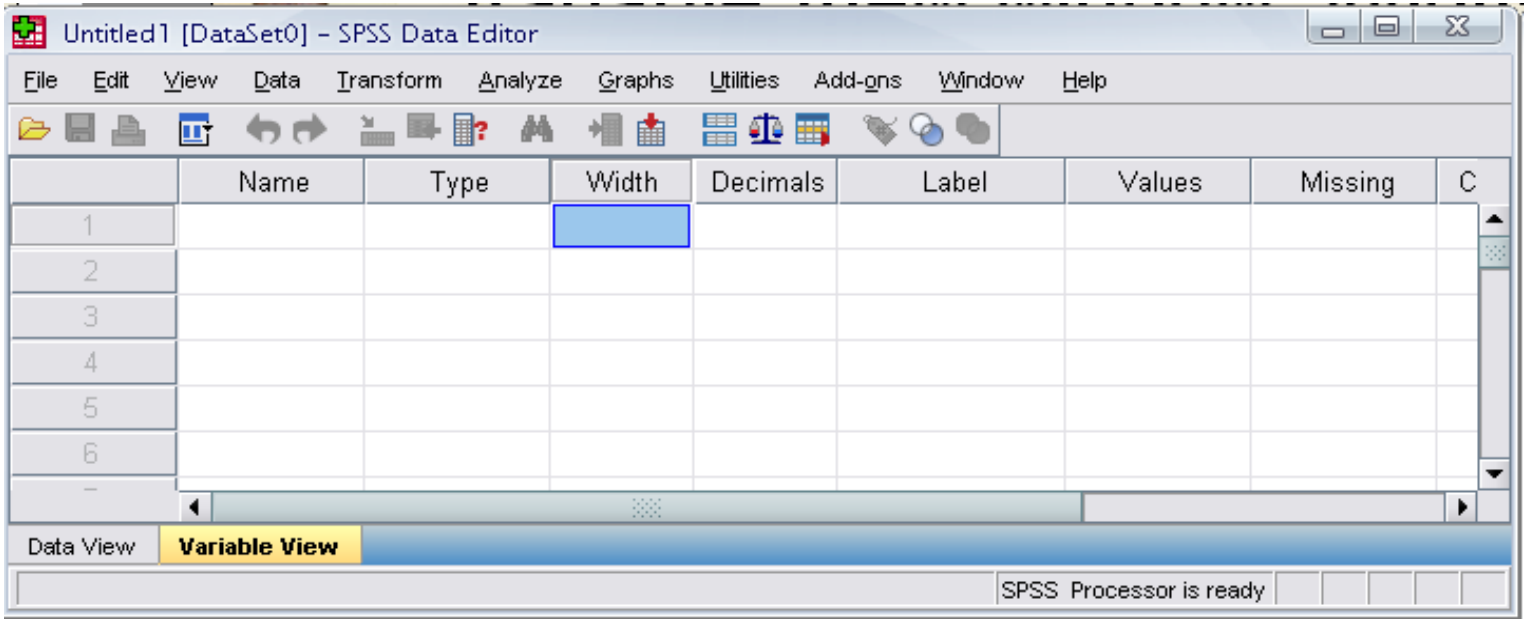
1. Click Variable View
2. Type variable name under Name column (AGE).

NOTE: Variable name can be 64 bytes long, and the first character must be a letter or one of the characters @, #, or $.

3. Type: Numeric, string, etc.
4. Label: description of variables.

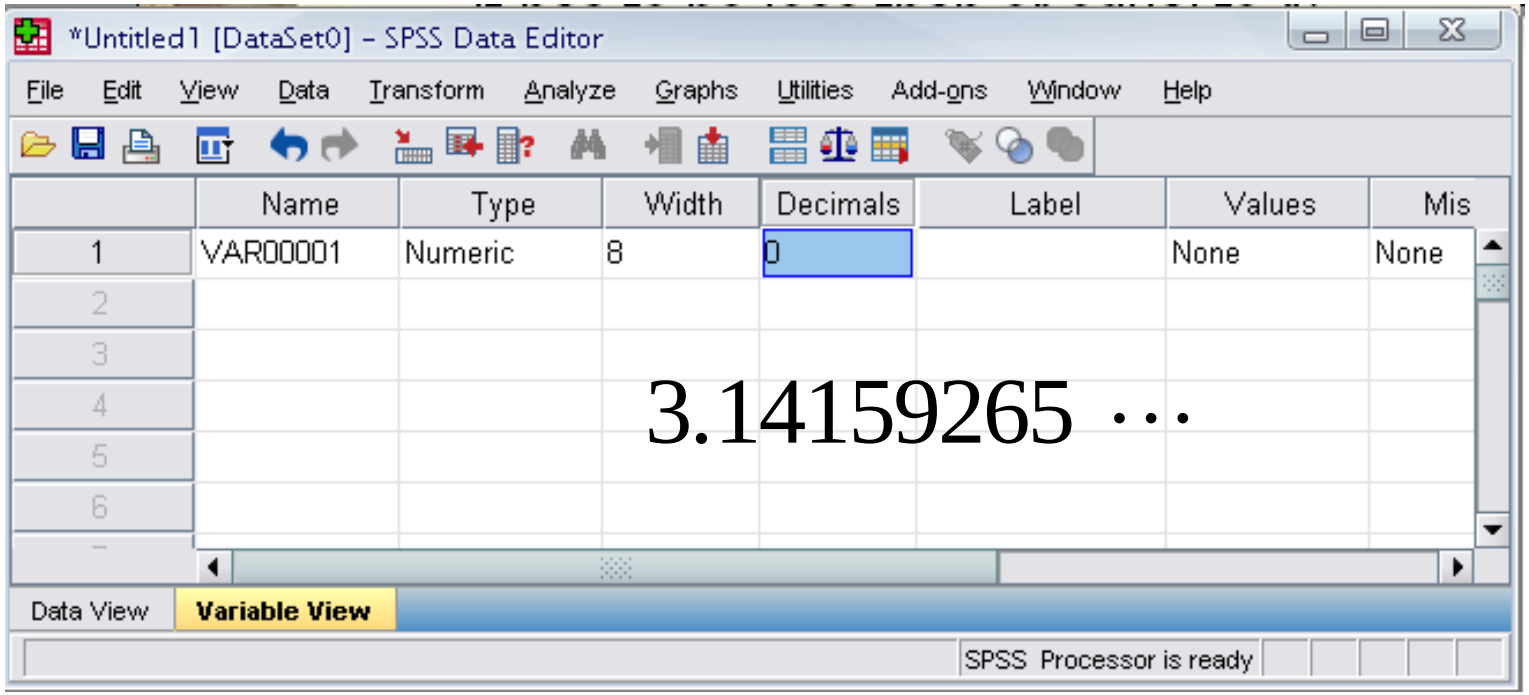
Variable View window: Width

- **Width**
  - Width allows you to determine the number of characters SPSS will allow to be entered for the variable



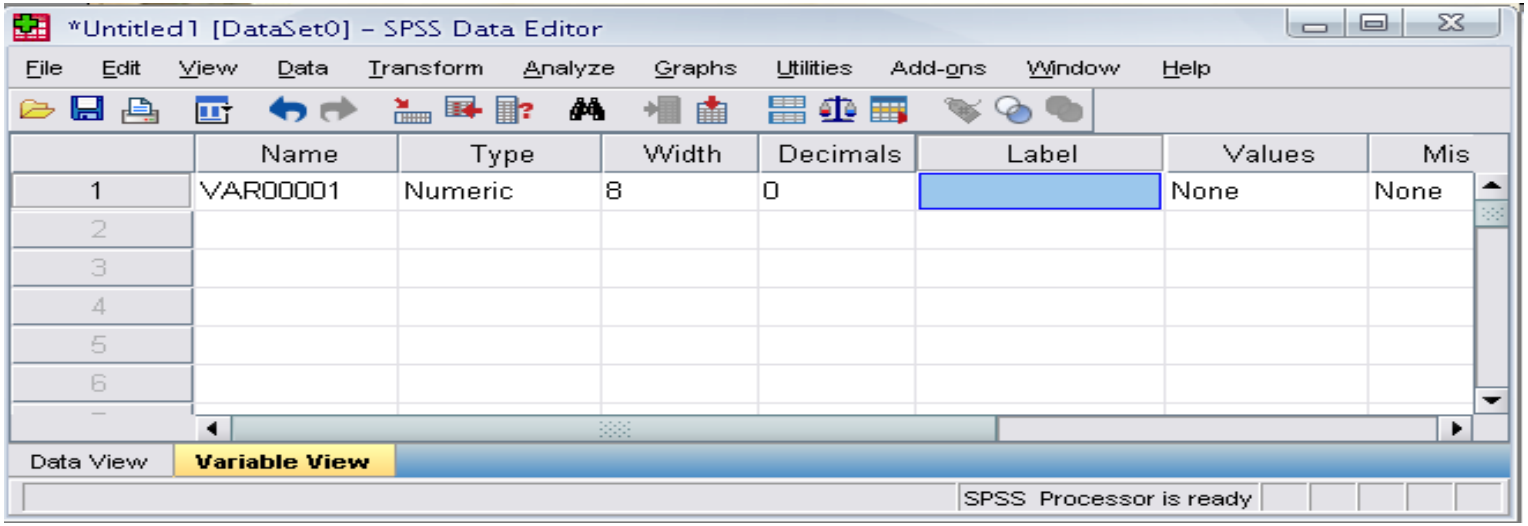
Variable View window: Decimals

- **Decimals**
  - Number of decimals
  - It has to be less than or equal to 16



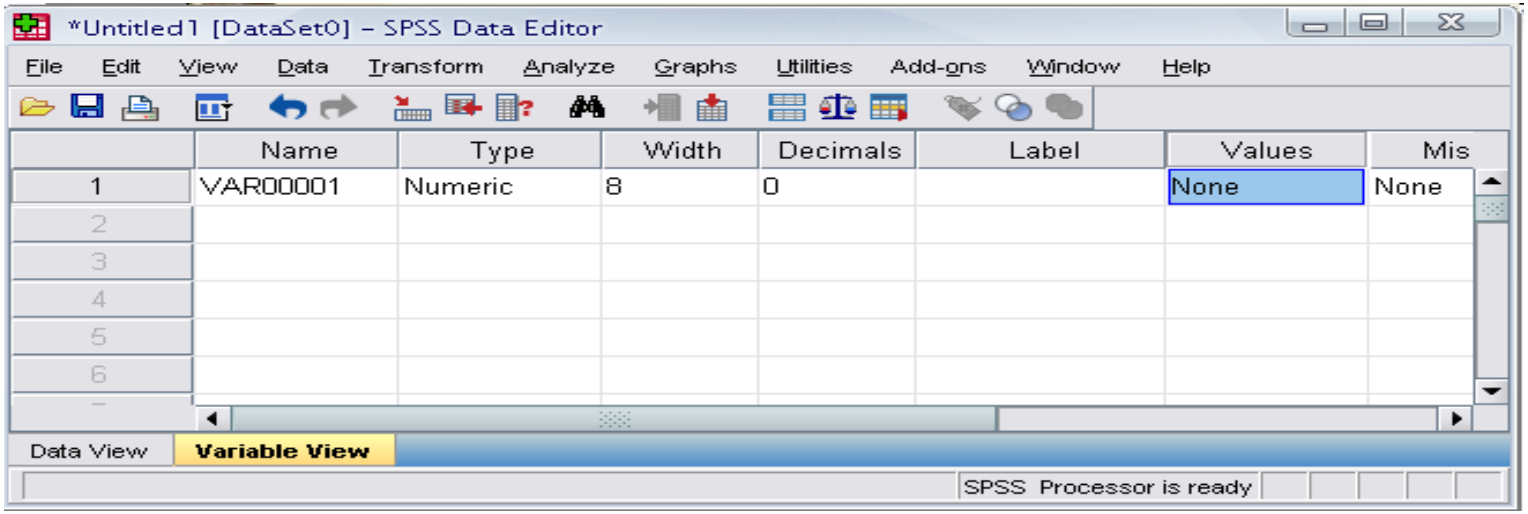
Variable View window: Label

- **Label**
  - You can specify the details of the variable
  - You can write characters with spaces up to 256 characters



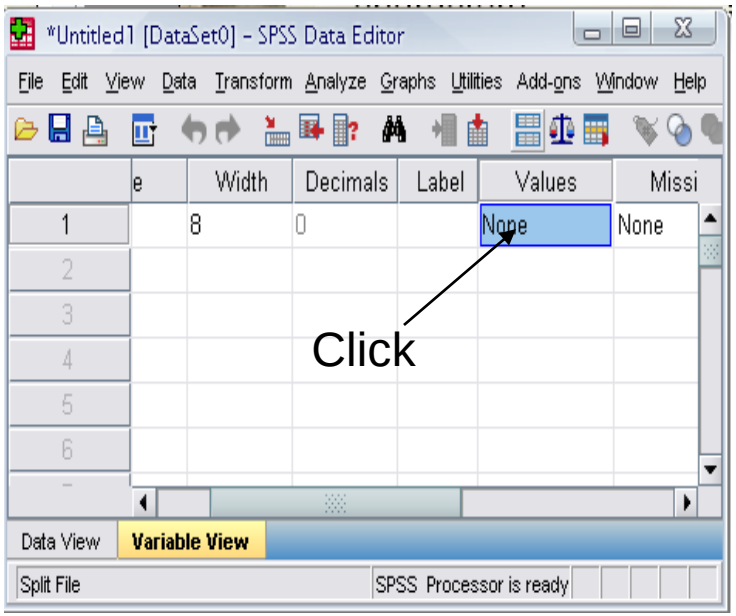
Variable View window: Values

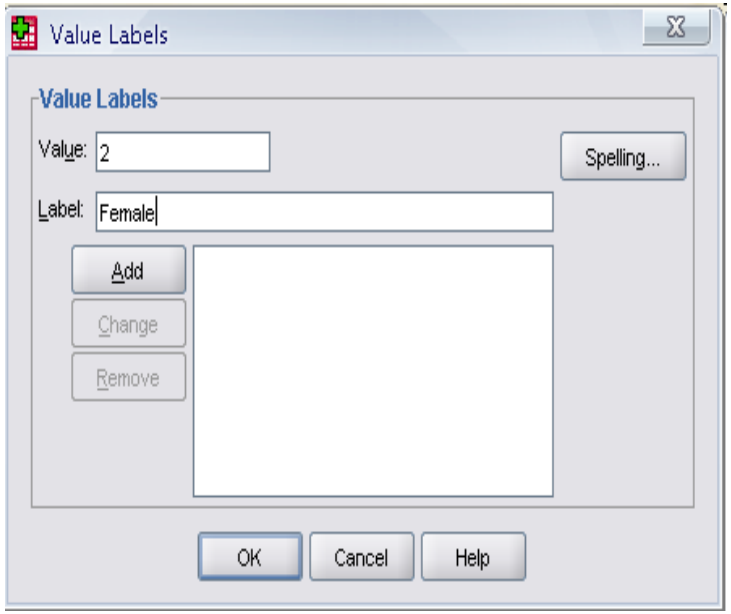
- **Values**
  - This is used and to suggest which numbers represent which categories when the variable represents a category



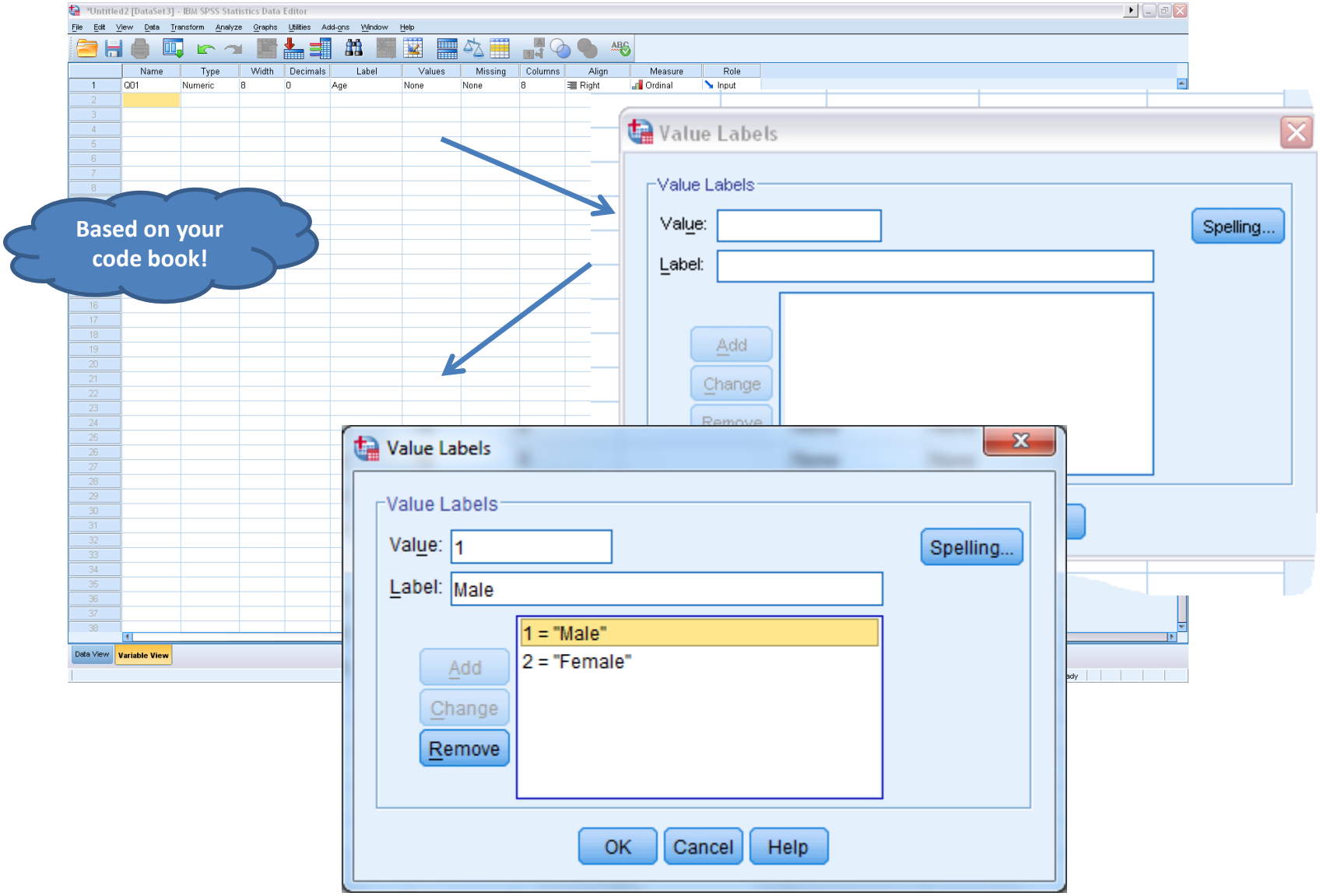
Defining the value labels

- Click the cell in the values column as shown below
- For the value, and the label, you can put up to 60 characters.
- After defining the values click add and then click OK.





Enter variables



Data Entry into SPSS

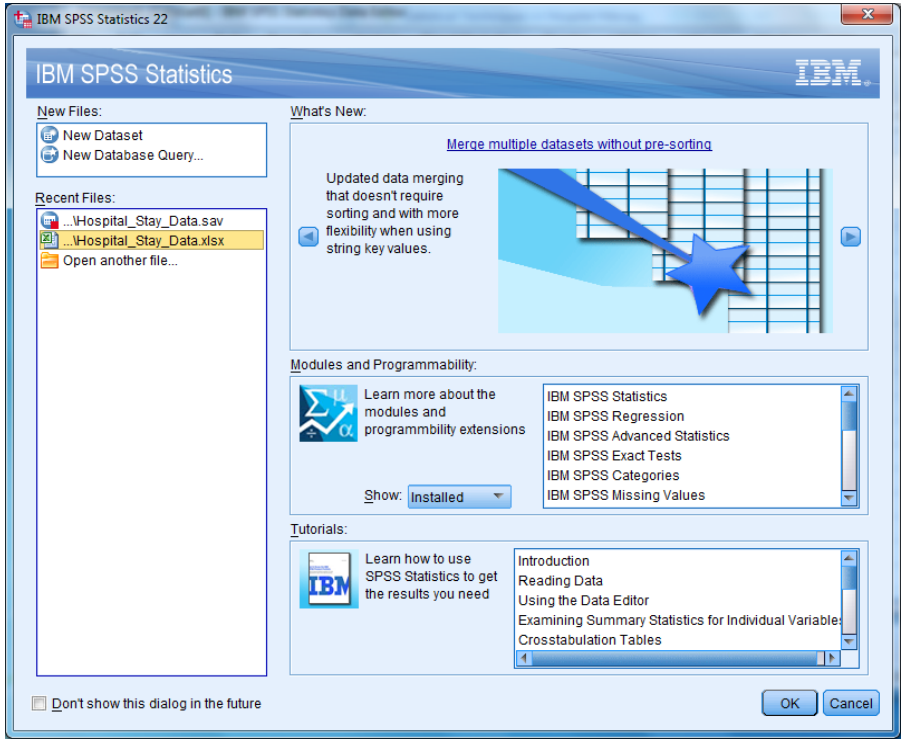
- **There are 2 ways to enter data into SPSS:**

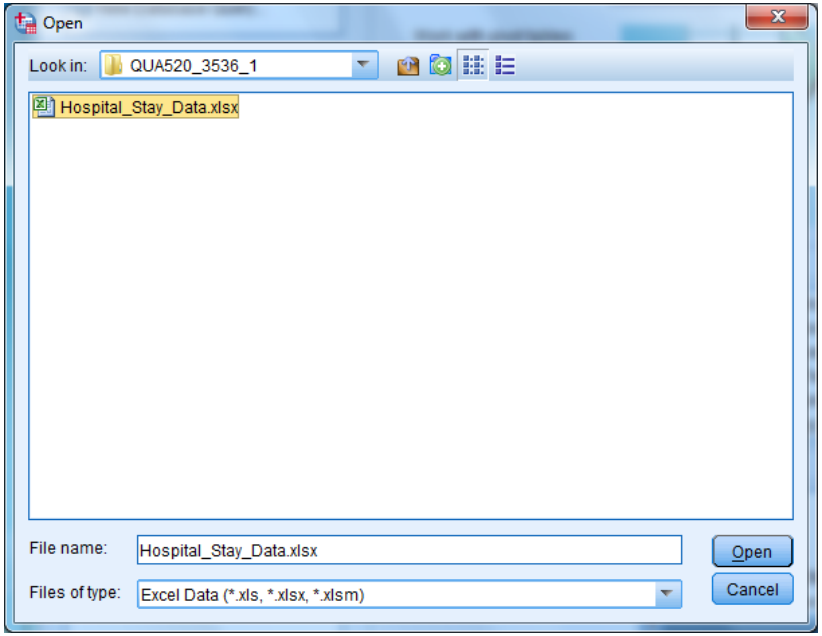
1. Directly enter in to SPSS by typing in Data View.

2. Enter into other database software such as Excel then import into SPSS.

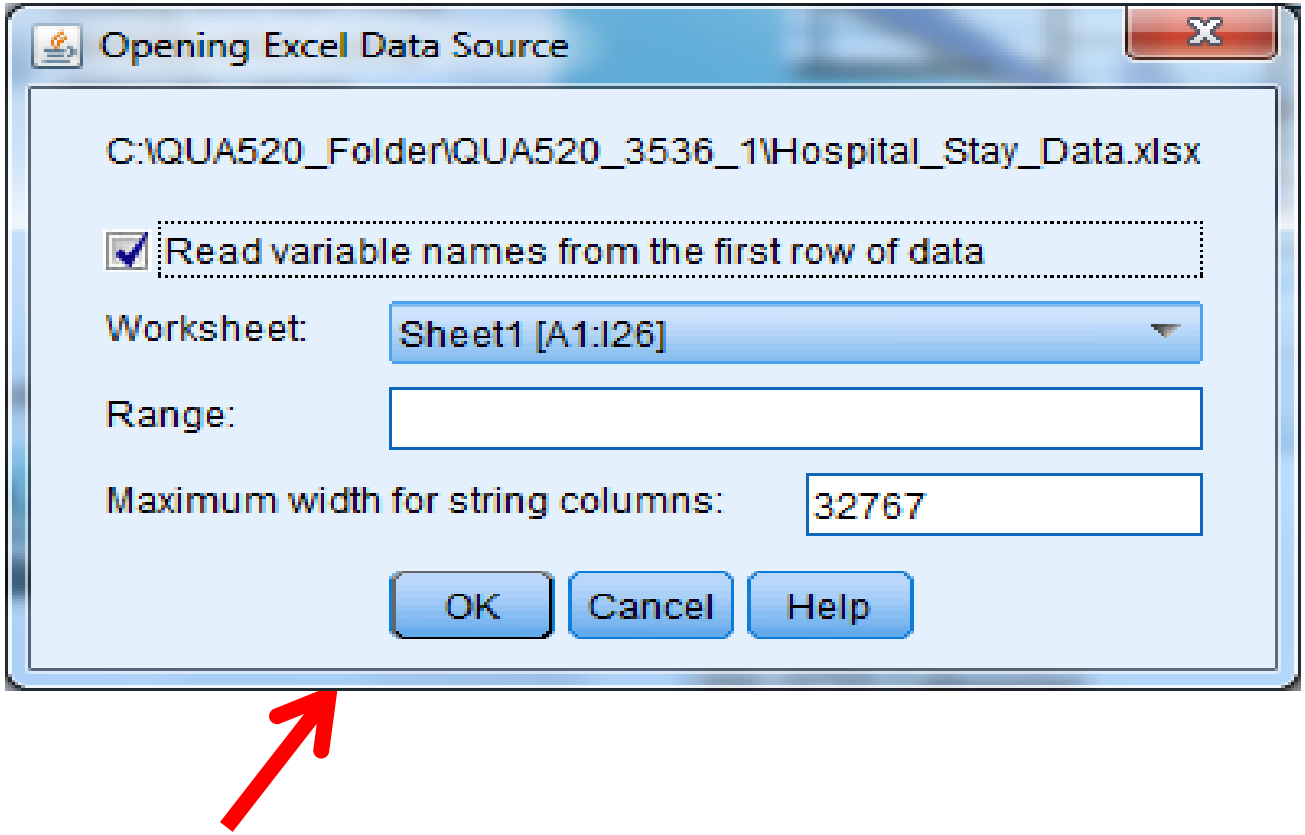
Import data from Excel

- Select File → Open → Data
- Choose **Excel** as file type
- Select the file you want to import
- Then click Open

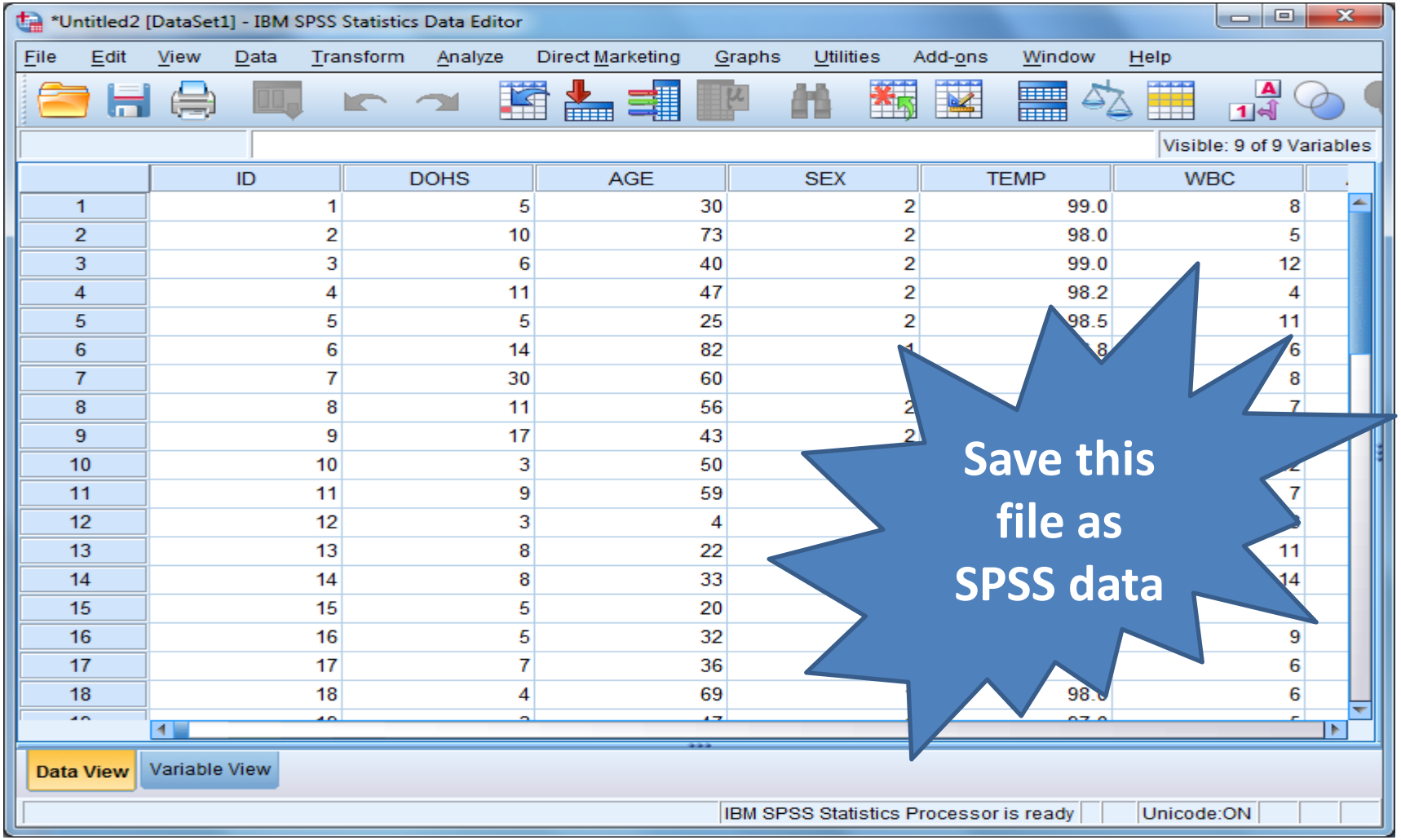




Open Excel files in SPSS



Continue …


*Untitled2 [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Visible: 9 of 9 Variables
ID DOHS AGE SEX TEMP WBC
1 1 5 30 2 99.0 8
2 2 10 73 2 98.0 5
3 3 6 40 2 99.0 12
4 4 11 47 2 98.2 4
5 5 5 25 2 98.5 11
6 6 14 82 1 8 6
7 7 30 60 8
8 8 11 56 2 7
9 9 17 43 2
10 10 3 50
11 11 9 59 7
12 12 3 4
13 13 8 22 11
14 14 8 33 14
15 15 5 20
16 16 5 32 9
17 17 7 36 6
18 18 4 69 98.0 6
Save this file as SPSS data
Data View
Variable View
IBM SPSS Statistics Processor is ready
Unicode:ON

Clean data after import data files

- Run **cases summaries** for all variables.
- Run **frequency** for qualitative variables and **Descriptive** for quantitative variables.
- Check outputs to see if you have variables with wrong values.
- Check missing values and physical surveys if you use paper surveys, and make sure they are real missing.
- Sometimes, you need to recode string variables into numeric variables

General guidelines for data entry

- Give each variable a valid name (8 characters or less with no spaces or punctuation, beginning with a letter not a numeric number). Short, easy to remember word names.

- Avoid the following variable names: *ALL, BY, EQ, GE, GT, LE, LT, NE, NOT, OR, TO, WITH, TEST*. These are used in the SPSS syntax and if they were permitted, the software would **not be able** to distinguish between a command and a variable.

- Each variable name must be unique; duplication is not allowed. Variable names are not case sensitive. The names NEWVAR, NewVar, and newvar are all considered identical.

General guidelines for data entry……..

- Encode categorical variables. **Convert letters and words to numbers**.

- **Avoid mixing symbols with data**. Convert them to numbers.

- Give each patient a unique, sequential case number (ID). Place this ID number in the first column on the left.

- **Do not** make columns **wider then 8 characters**, unless absolutely essential.

General guidelines for data entry…….

- **Each variable** should be in its own column.

Avoid this:

Animal	
Control1	
Control2	
Experiment1	
Experiment2	

Change to:

Animal	Group
1	0
2	0
3	1
4	1

❖ **Do not combine variables in one column**.

❖ It is recommended to use 0/1 for 2 groups with 0 as a reference group.

General guidelines for data entry……..

- All data for a project should be in **one spreadsheet**. Do not include graphs or summary statistics in the spreadsheet.

- Each patient should be entered **on a single line or row**. Do not copy a patient's information to another row to perform subgroup analysis.

- **Put ordinal variables into one column** if they are mutually exclusive

Avoid:

Pain		
Mild	Moderate	Severe
1	0	0
0	1	0
0	0	1

Preferred:

Pain
1
2
3

General guidelines for data entry……..

- **For yes/no questions, enter** "0" for no and "1" for yes. Do not leave blanks for no. Do not enter "?", "*", or "NA" for missing data because this indicates to the statistical program than the variable is a **string variable**.

- String variables cannot be used for any arithmetic computation.

General guidelines for data entry……..

- However when data are repeatedly collected over a patient, it's recommended to have patient-day observation on a simple line to ease data management.
- SPSS has a nice feature to convert from the longitudinal format to horizontal format. When the number of repeats are few 2 or 3, horizontal format may be preferred for simplicity.

Longitudinal data entry

Date	ID	SYSBP
1/2/2005	1	130
1/3/2005	1	120
1/4/2005	1	120
3/1/2005	2	110
3/2/2005	2	140

Horizontal data entry

ID	SYSBP1	SYSBP2	SYSBP3
1	130	120	120
2	110	140	

Ways to Display Data

Types of Data Displays

- Pictograph
- Tally Chart
- Bar Graphs
- Line Graph
- Pie Chart
- Stem and Leaf Plot
- Histograms
- Line Plot
- Box and Whisker Plot (Box Plot)
- Scatter Plot
- Map Graph (Cosmography)
- Venn Diagram

Pictograph

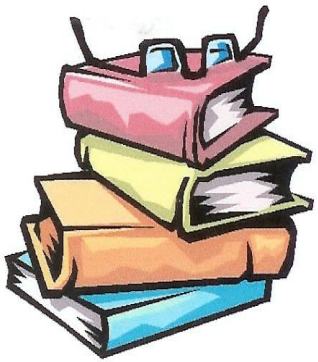
- **A pictograph** uses pictures or symbols to compare data.

A picture graph uses pictures or symbols to show data. One picture often stands for more than one vote so a key is necessary to understand the symbols.

Number of Books Read

Jeremy	📖📖📖📖📖
Kevin	📖📖📖📖
David	📖📖📖
Kelly	📖📖📖📖📖📖
Emily	📖📖

Each 📖 stands for 5 books.



FAVORITE COLOR

COLOR	Number of Students
red	● ● ●
yellow	●
blue	● ● ● ● ●
green	● ●
orange	●
purple	● ● ●

key: Each ● = 2 student votes

- **A pictograph** has **a key** that tells the value of each picture.

FAVORITE COLOR

COLOR	Number of Students
red	● ● ●
yellow	●
blue	● ● ● ● ●
green	● ●
orange	●
purple	● ● ●

KEY

key: Each ● = 2 student votes

- **A pictograph** is **similar** to the bar graph and histogram because it is also used best to compare data.

Pictographs Summary

Pictograph	Advantages	Disadvantages
▪ A pictograph uses an **icon to represent a quantity of** data values in order to decrease the size of the graph. ▪ A key must be used to explain the icon.	▪ **Easy to read.** ▪ **Visually appealing..** ▪ **Handles large data sets easily using keyed icons.**	▪ **Hard to quantify partial icons.** ▪ **Icons must be of consistent size.** ▪ **Best for only 2 – 6 categories.** ▪ **Very simplistic.**

Tally Chart

- A **tally chart** is a **table** with **tally** marks to show a valuable data set.

- A **tally chart** is one method of collecting data with **tally** marks.

- **Tally** marks are frequencies, occurrences, or total numbers being measured for a specific category in a data set.

Tally Chart

"Favorite Pets"

Pet	Tally Marks
	𝍸 𝍸
	𝍸
	𝍸 |

Favorite Pets		
Pet	Tally Marks	Number
	𝍸 𝍸	10
	𝍸	4
	𝍸 |	6

Bar Graphs

- **A bar graph** is a chart that uses bars to show **comparisons** between discrete categories, e.g. patients on the caseload by town.

- A bar chart shows data in **separate columns**.

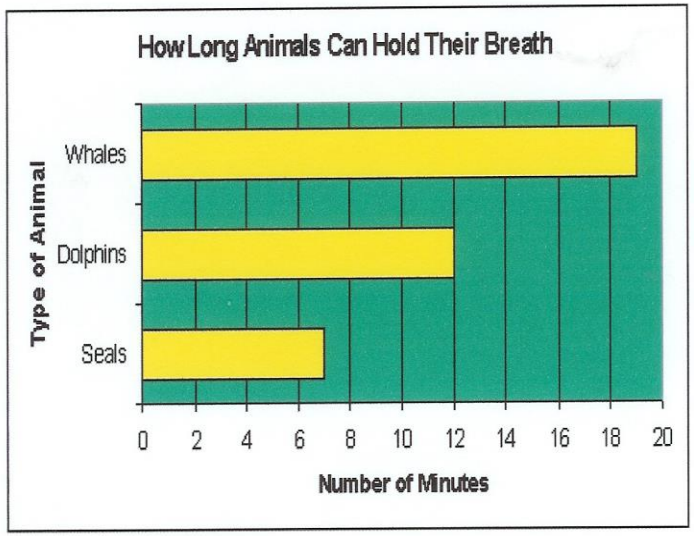
- It is more **clear** than the table and the data is **easier** to be **compered** & make some **conclusions**.

Bar Graphs

Bar graph	Advantages	Disadvantages
▪ A bar graph displays discrete data in separate columns. ▪ A double bar graph can be used to compare two data sets. ▪ Categories are considered unordered and can be rearranged alphabetically, by size, etc.	▪ Visually strong. ▪ Can easily compare two or three data sets.	▪ Graph categories can be reordered to emphasize certain effects. ▪ Use only with discrete data

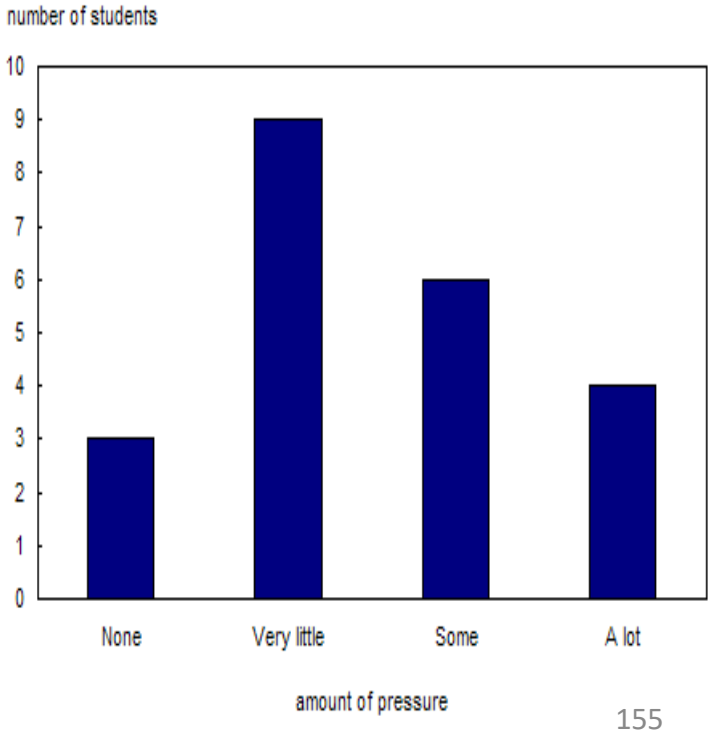
Bar Graphs Example

A bar graph uses bars to show data. The bars can be vertical (up and down), or horizontal (across). The data can be in words or numbers.



- **Horizontal Bar Graph**
  - Useful when category names are **too long** to fit at the foot of a column.

- **Vertical Bar Graph**
  - Displays data **better** than horizontal bar graphs, and is preferred when possible.

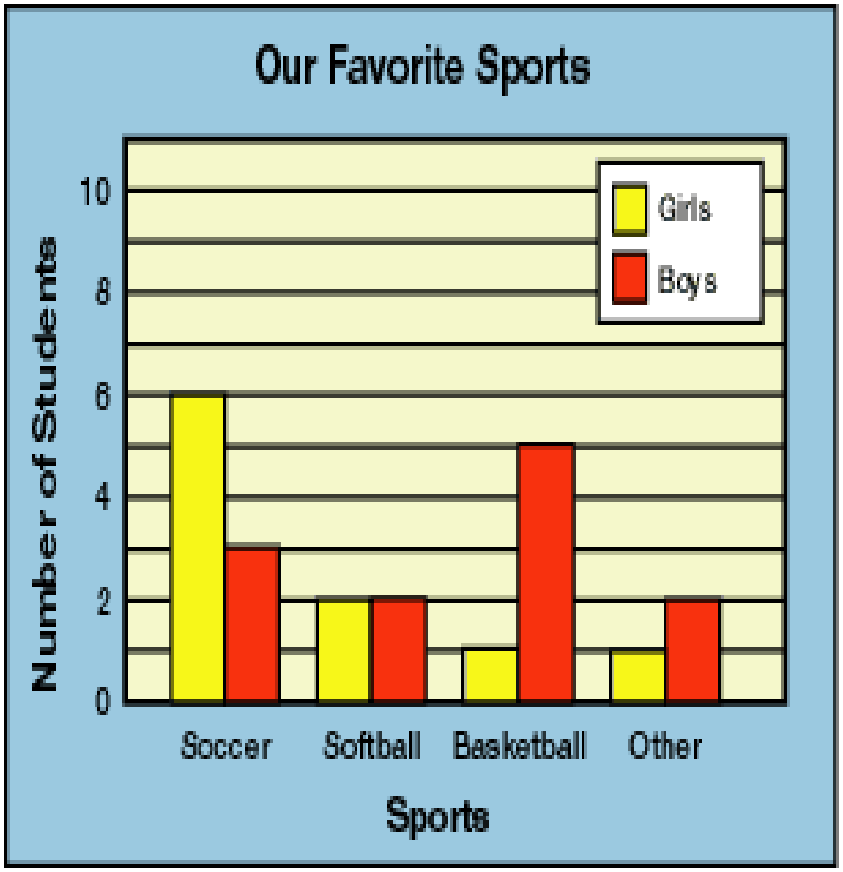


Bar Graphs

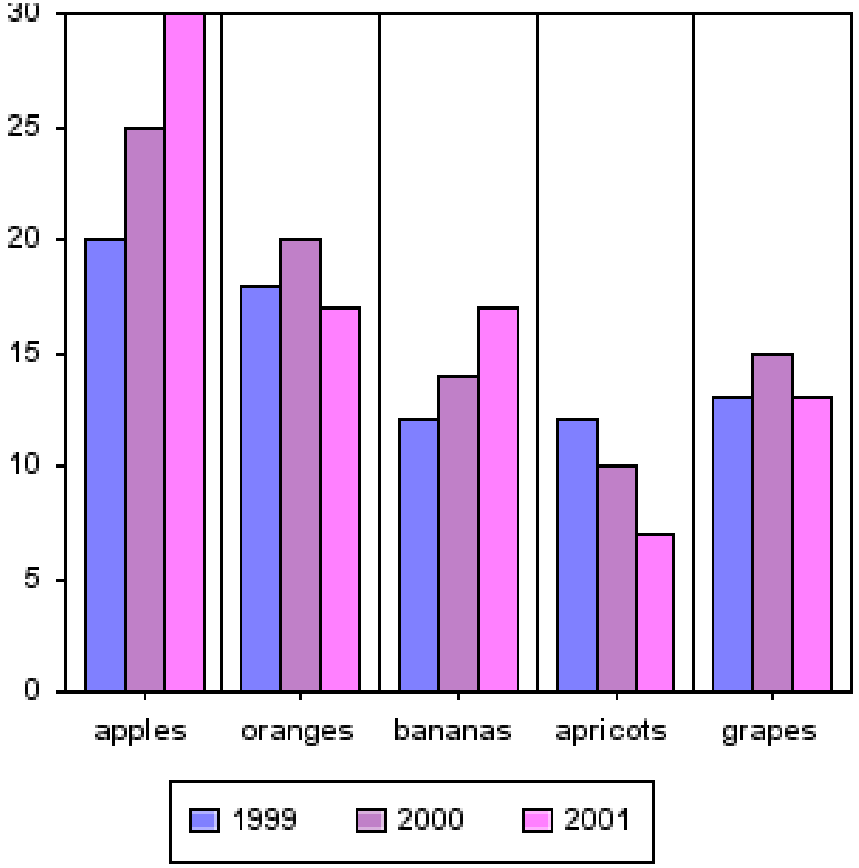
(Grades 2, 3, 4, 5)

- **Grade 2**: Single Bar Graph
- **Grade 3**: Single Bar Graph
- **Grade 4**: Double Bar Graph
- **Grade 5**: Multi-Bar Graph

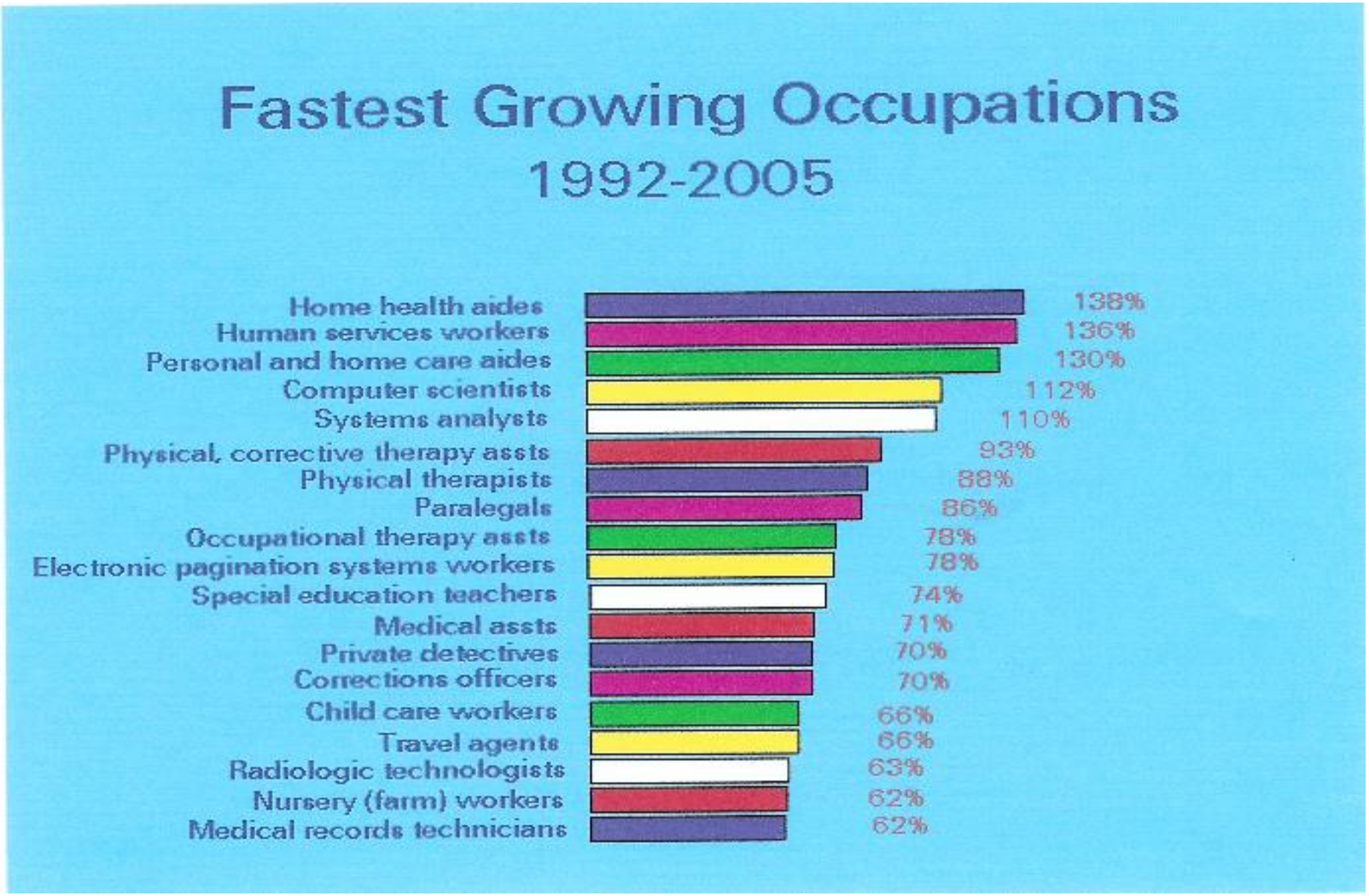
Double Bar Graph
(Grade 4)



Multi-Bar Graph
(Grade 5)

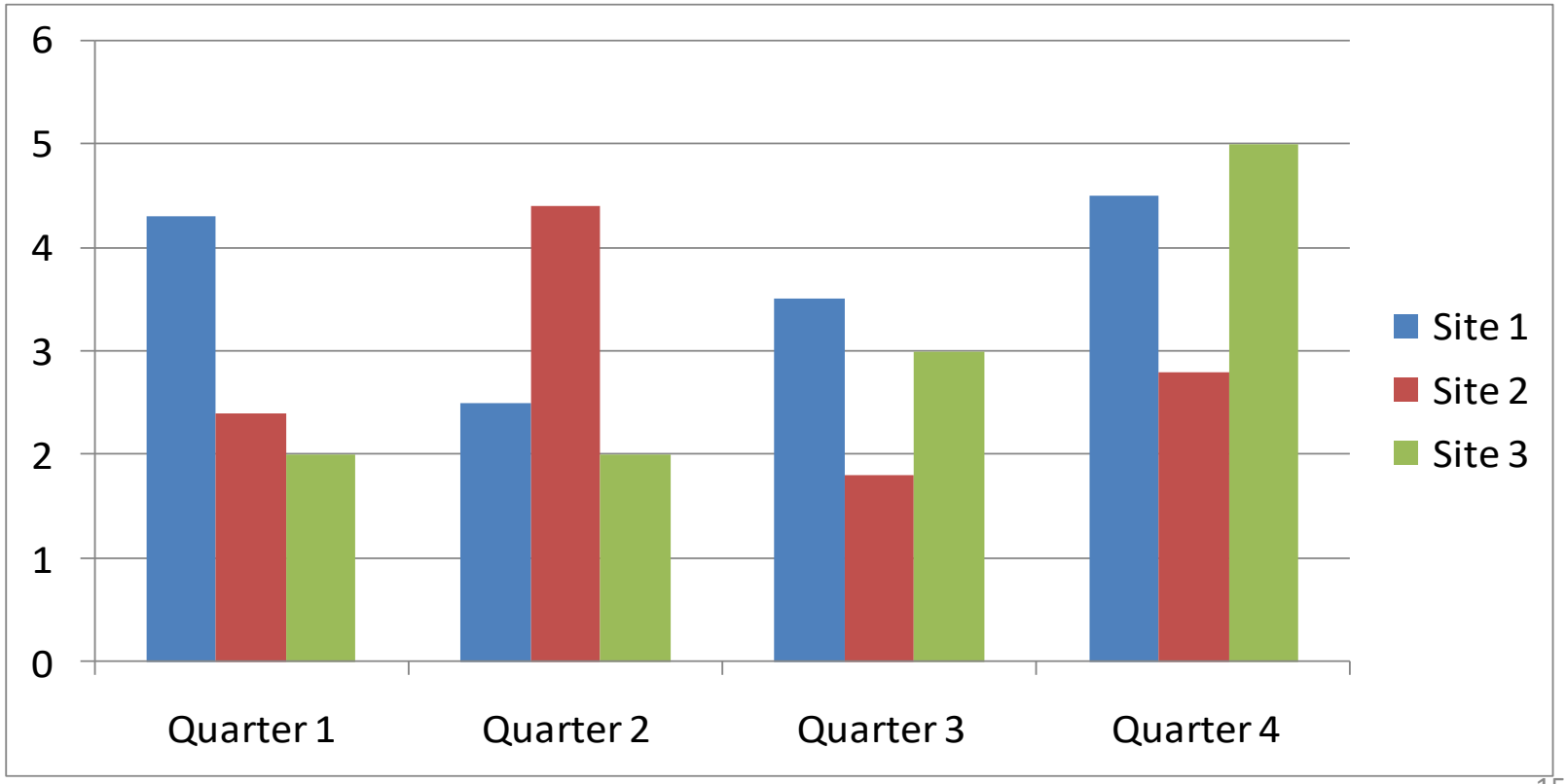


Horizontal Bar Graph Example

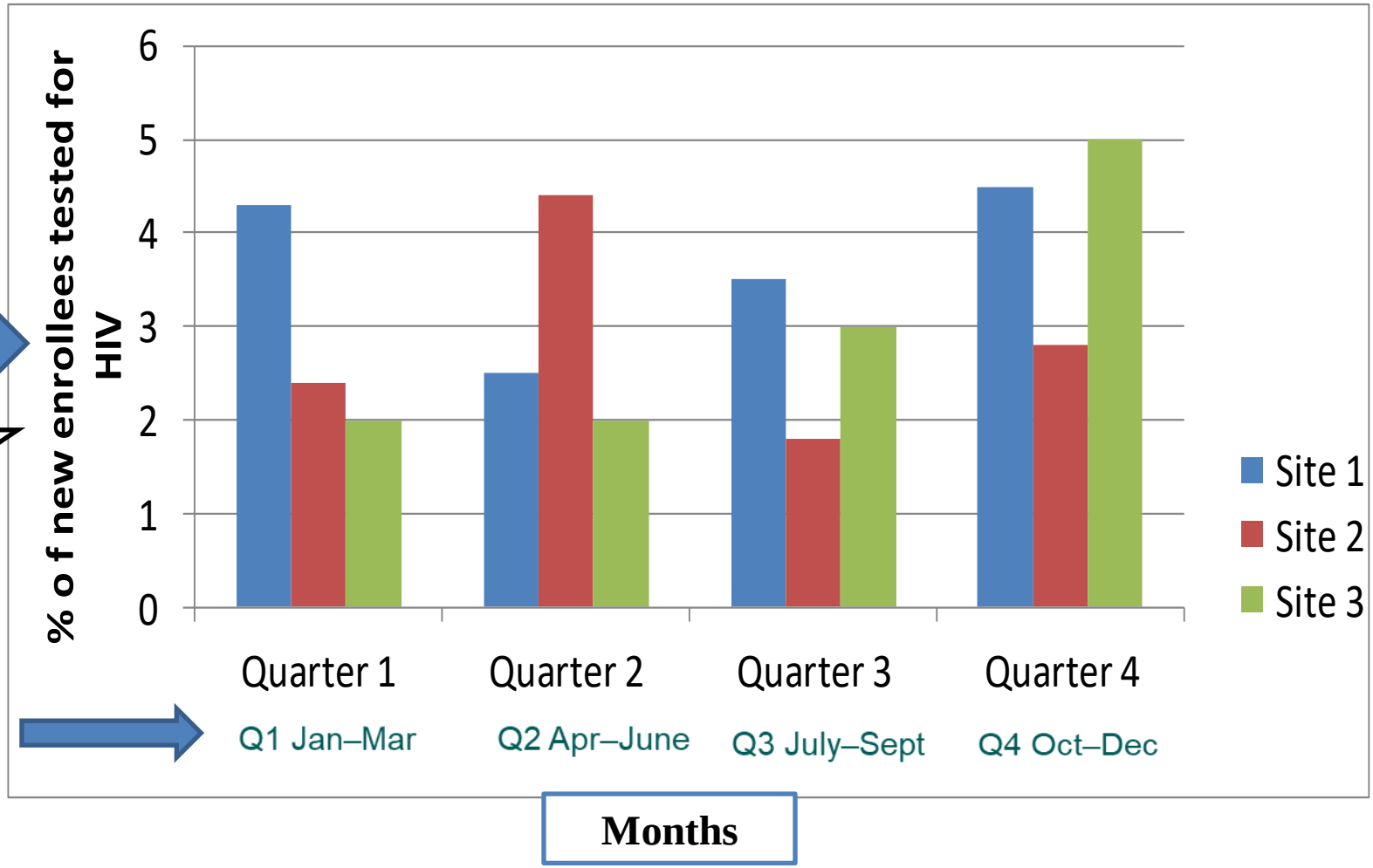


Bar chart
Comparing categories

- In this bar chart, we're comparing the categories of data, which are the different sites. You see a comparison between sites by quarters and between quarters over time.



Percentage of new enrollees tested for HIV at each site, by quarter (Population)



Label

%

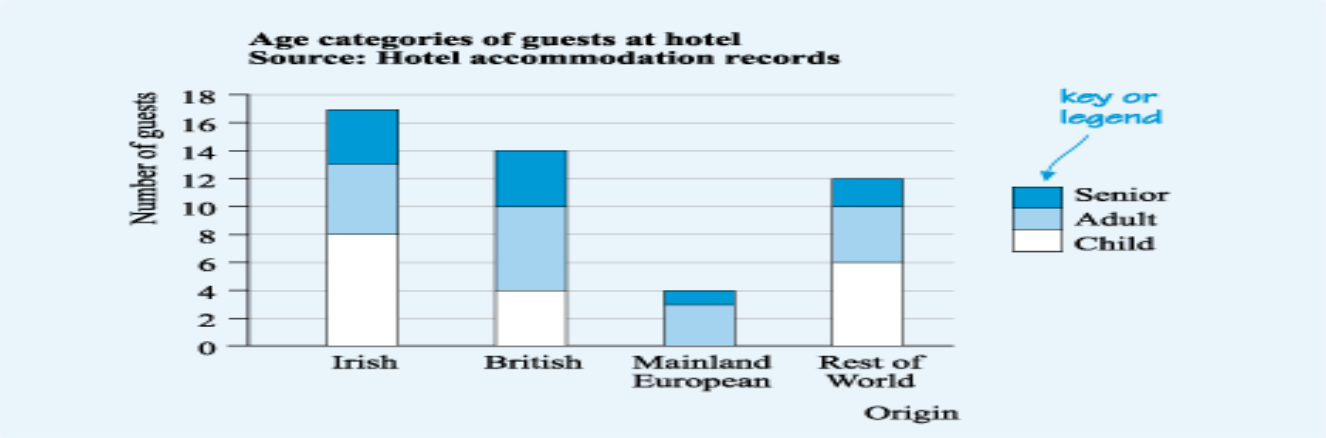
Data Source: Program records, AIDS Relief, January 2009 – December 2009.

Compound bar diagram

- A **compound bar chart** is a **graph** which combines two or more types of information in one **chart**.

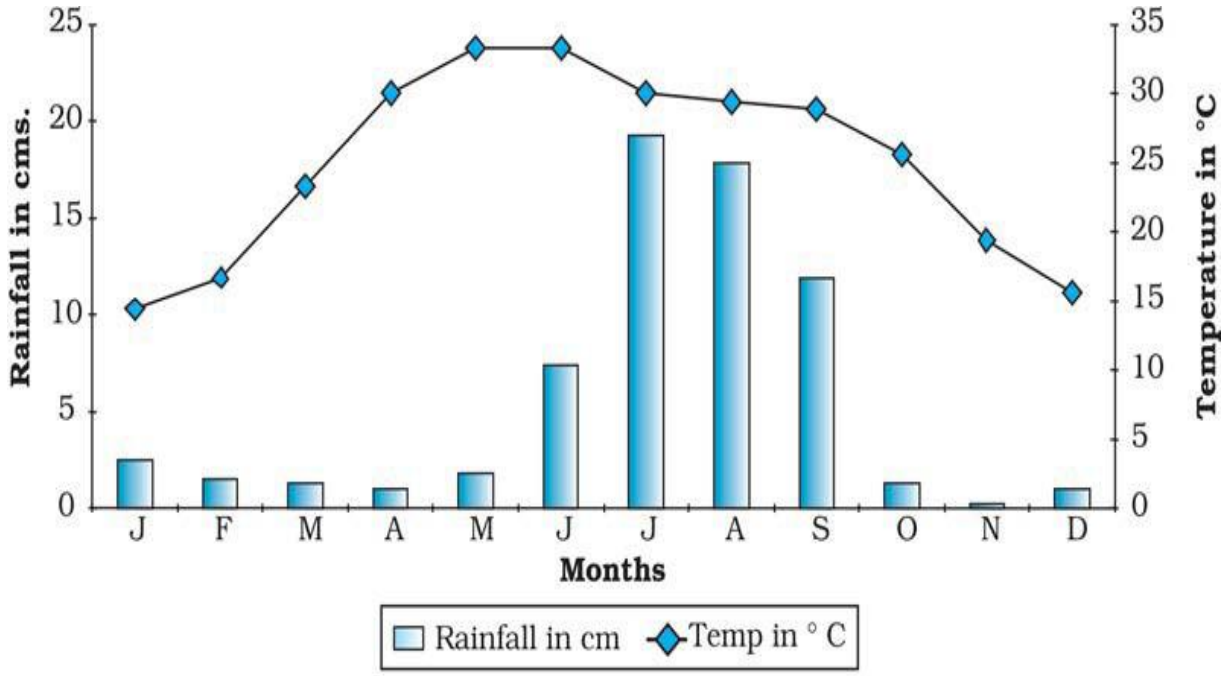
- When different components are grouped in **one set of variable** or **different variables** of one component are put together.

- It is a type of **bar chart** where columns can be split into sections to show breakdown of data.



Polybar diagram

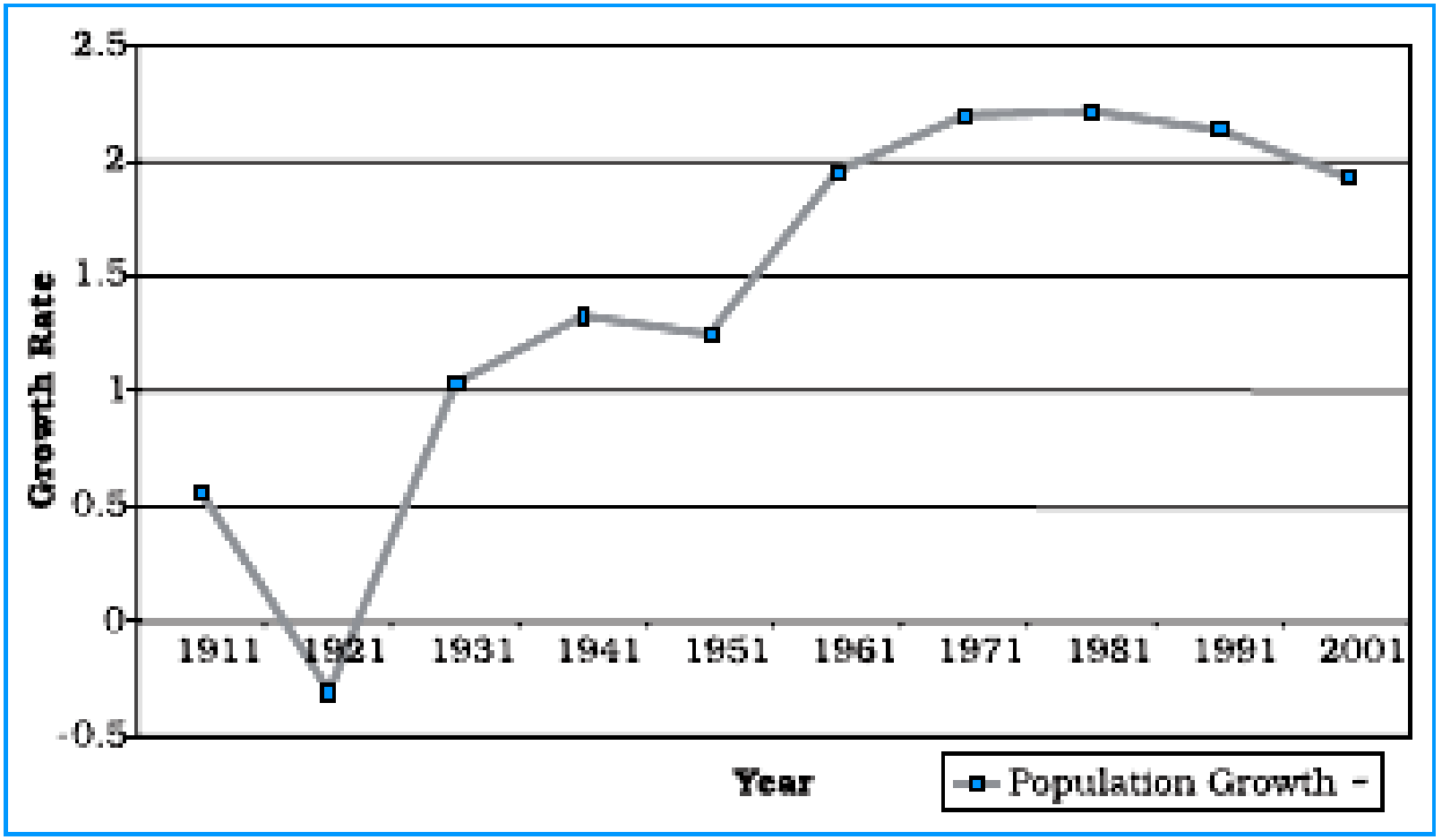
- The **line** and **bar graphs** as drawn separately may also be combined to *show the data related to some of the closely associated characteristics* such as the climatic data of mean monthly temperatures and rainfall.



Line Graph

Line graph	Advantages	Disadvantages
▪ Line graphs are used to illustrate trends over time for continuous data. ▪ A line graph plots continuous data as points and then joins them with a line. ▪ Multiple data sets can be graphed together, but a key must be used.	▪ Can compare multiple continuous data sets easily. ▪ Interim data can be inferred from graph line.	▪ Use only with continuous data.

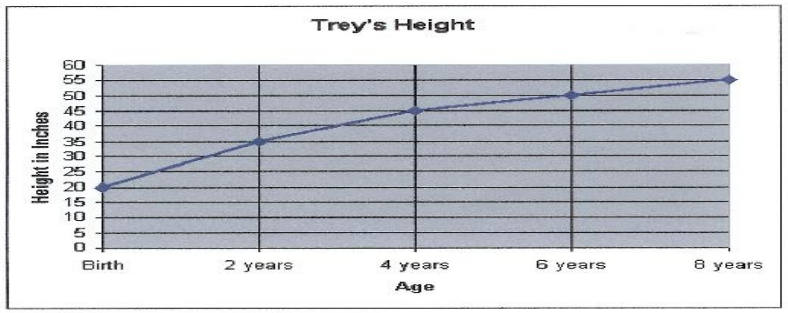
Line Graph


2.5
2
1.5
1
0.5
0
-0.5
Growth Rate
1911
1921
1931
1941
1951
1961
1971
1981
1991
2001
Year
Population Growth -

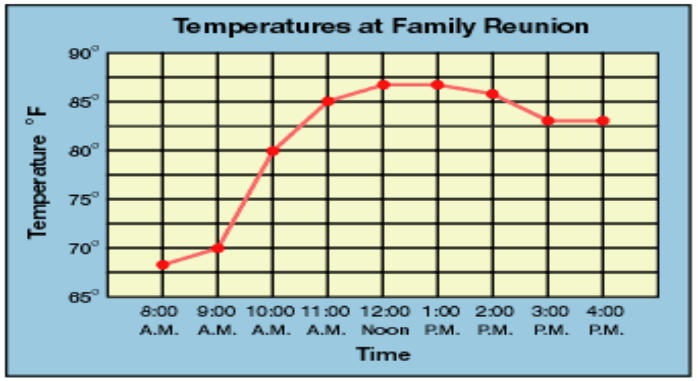
Line Graph...

A line graph shows points plotted on a graph. The points are then connected to form a line.

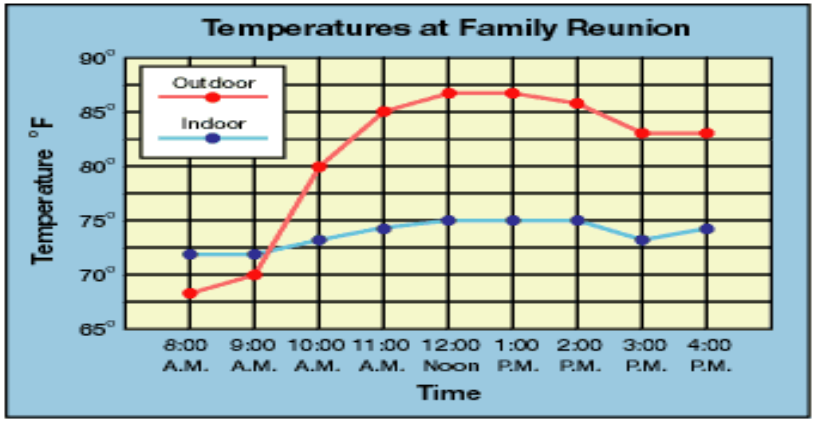
- **Single Line Graph**



- **Single Line Graph**

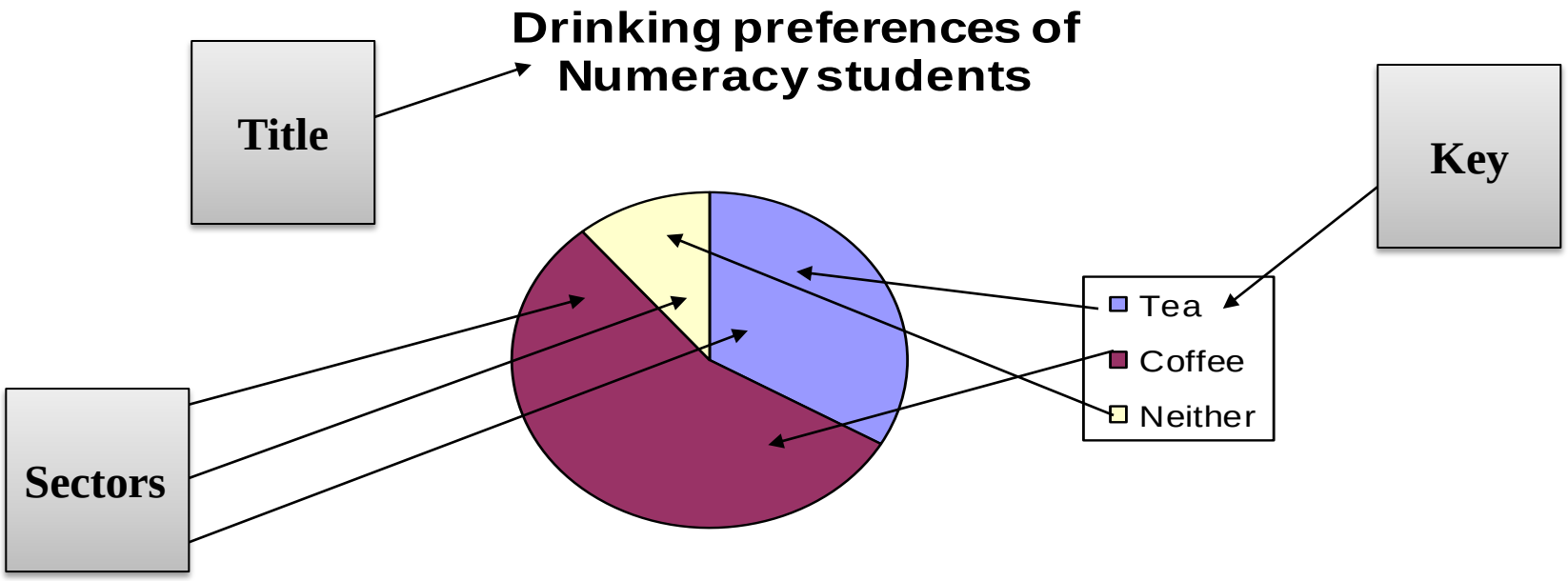


- **Double Line Graph**



Pie Chart – Circle Graph

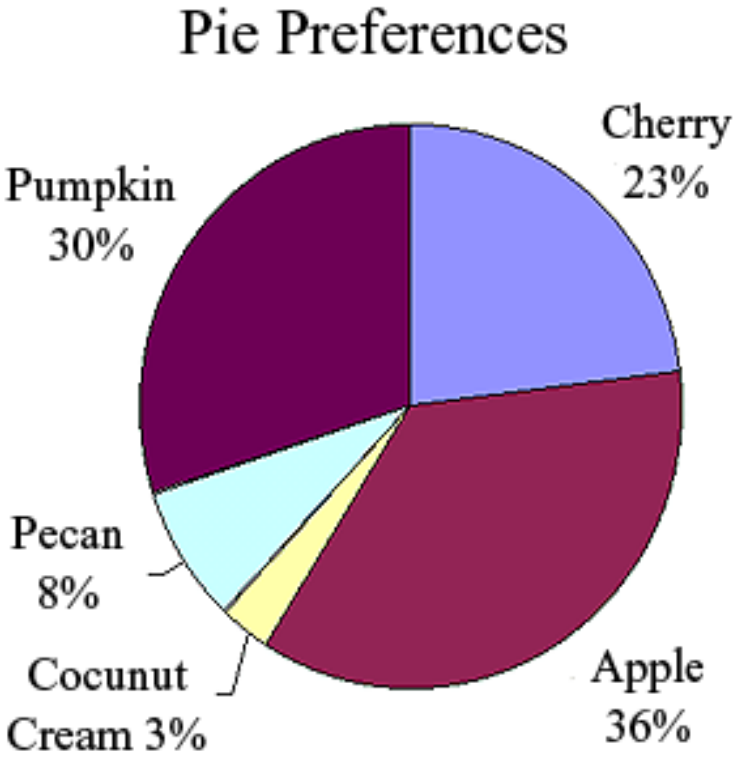
- The data is arranged in a circle and each type.
- Every pie chart has : A title , A key and Sectors.

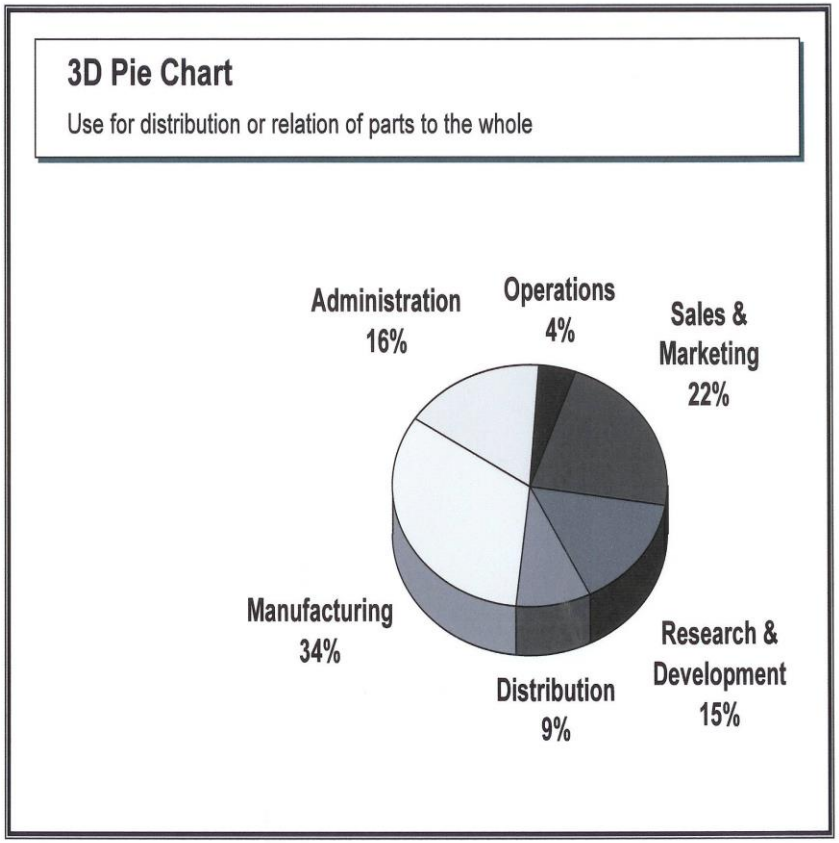


Pie Chart – Circle Graph

Pie chart	Advantages	Disadvantages
▪ A pie chart displays **data as a percentage of the whole.** ▪ Each pie section should have a label and percentage. ▪ A total data number should be included.	▪ Visually appealing. ▪ Shows percent of total for each category.	▪ No exact numerical data. ▪ Hard to compare 2 data sets. ▪ "Other" category can be a problem. ▪ Total unknown unless specified. ▪ Best for 3 – 7 categories. ▪ **Use only with discrete data.**

Pie Chart – Circle Graph Example





Pie (circle) charts - more info

- A way of summarizing a set of **categorical** data or displaying the different values of a given variable (e.g. percentage distribution).

- A circle is divided into a series of segments. Each segment represents a particular category.

- The area of each segment is the same proportion of a circle's area as the category is of the total data set.

- **Quite popular.** Circle provides a visual concept of the whole (100%).

Pie (circle) charts…

- Best used for **displaying statistical information when there are no more than six components** – otherwise, the resulting picture will be **too complex** to understand.

- Pie charts **are not useful when the values of each component are similar** because it is difficult to see the differences between slice sizes.

Stem and Leaf Plot

- **A stem-and-leaf plot** displays and organizes numerical data by separating the digits of each number into a stem and a leaf.

Example

10, 15, 22, 25, 28, 23, 29, 31, 36, 45, 48

Stem	Leafs
1	0 5
2	2 5 8 3 9
3	1 6
4	5 8

stem	leaf	
5	6	56
6	7, 7, 9	67,67,69
7	2, 4, 7, 7, 8	72 ,74 ,77 ,77 ,78
8	1, 2, 2, 3, 4, 8	81,82,82,83,84,88
9	0, 2, 3, 4	90,92,93,94

Stem and Leaf Plot

Stem and Leaf Plot	Advantages	Disadvantages
▪ Stem and leaf plots record data values in rows, and can easily be made into a histogram. ▪ Large data sets can be accommodated by splitting stems.	▪ Concise representation of data ▪ Can handle extremely large data sets	▪ Not visually appealing ▪ Does not easily indicate measures of centrality for large data sets.

Histograms

- **A histogram** is a special kind of bar graph that shows how ranges (or **intervals**) of data differ from one another. There are **no spaces** between the bars of a histogram.

Histograms

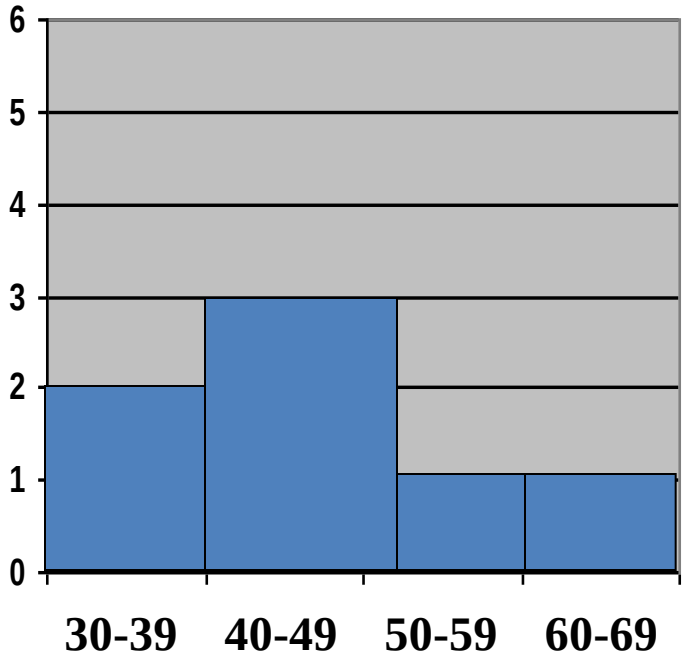
Histogram	Advantages	Disadvantages
▪ The most famous one for frequencies. ▪ A histogram is a type of bar graph that displays continuous data in ordered columns. ▪ Categories are of continuous measure such as time, inches, temperature, etc.	▪ Visually strong ▪ Can compare to normal curve ▪ Usually vertical axis is a frequency count of items falling into each category.	▪ Cannot read exact values because data is grouped into categories. ▪ Use only with continuous data.

Histogram Example

Minutes Spent on Homework

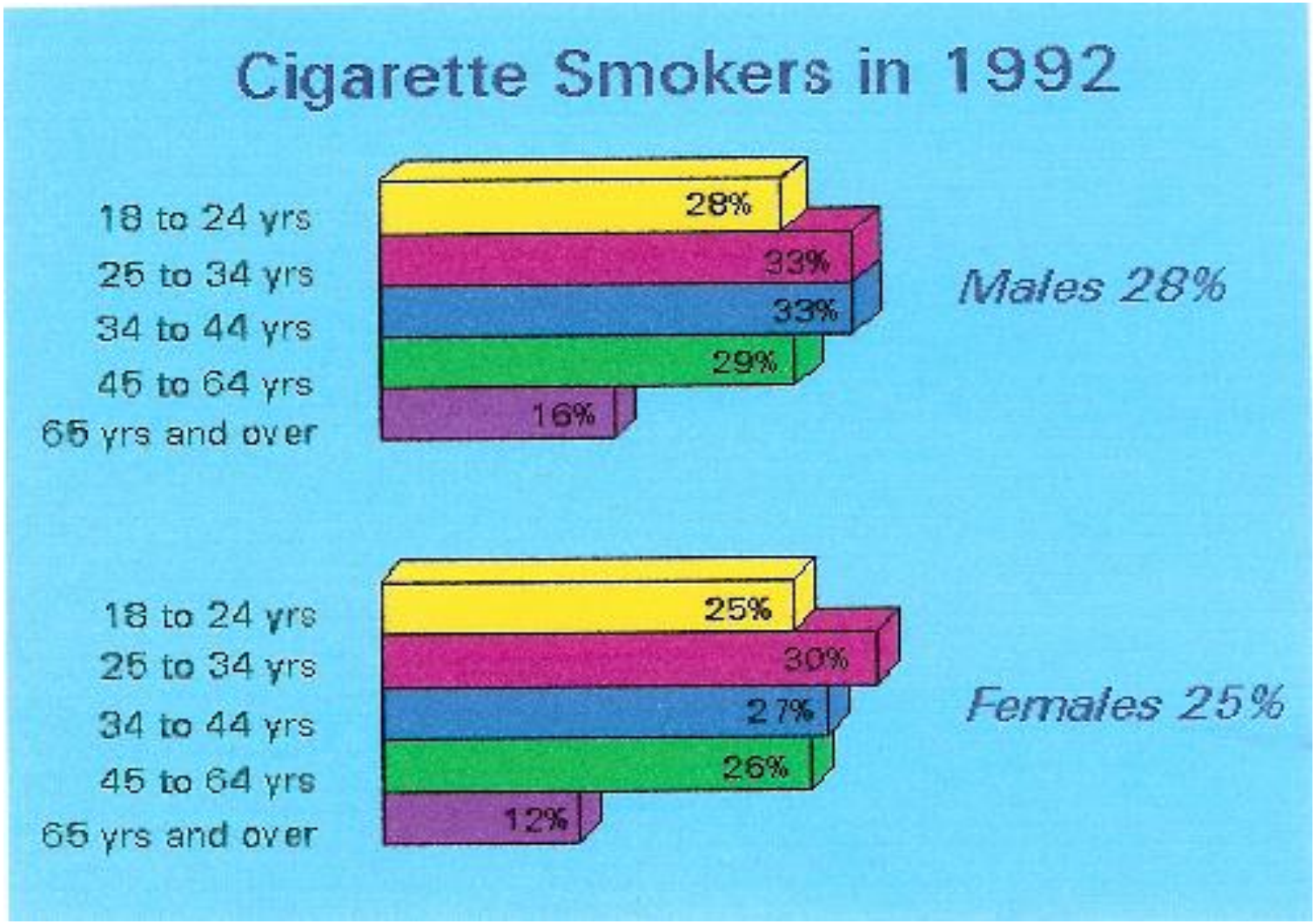
# of Minutes	Tally	Frequency
30-39	||	2
40-49	|||	3
50-59	|	1
60-69	|	1

Histogram

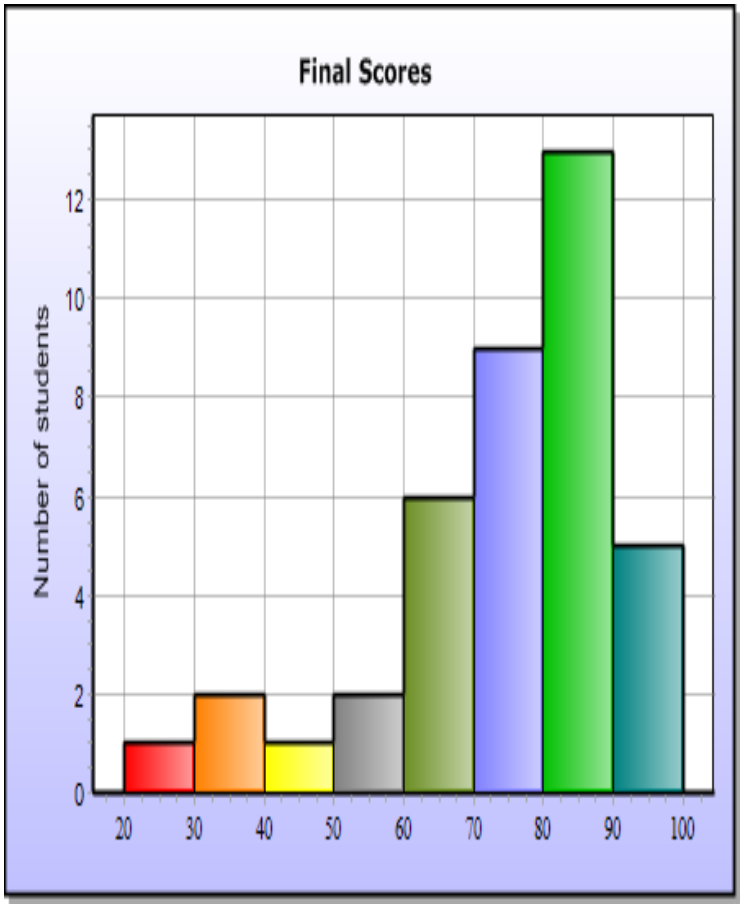


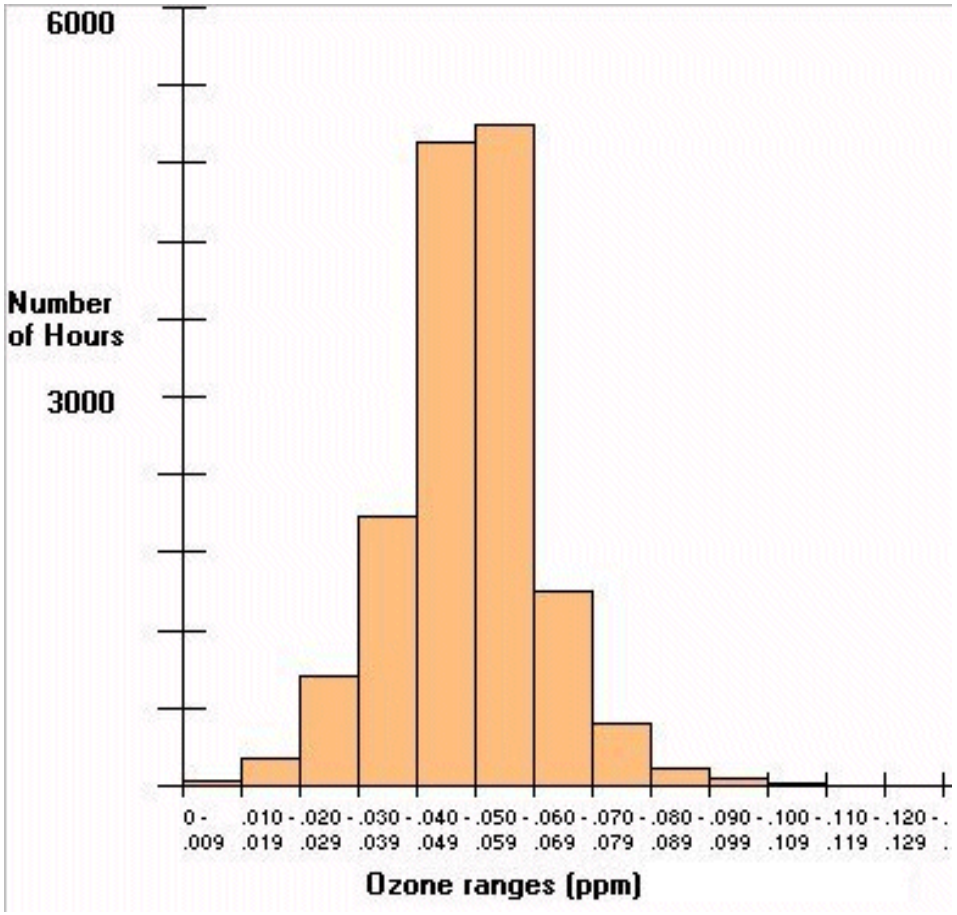
Histogram Example

(A type of bar graph)



Histogram..


Final Scores
Number of students
0
2
4
6
8
10
12
20
30
40
50
60
70
80
90
100

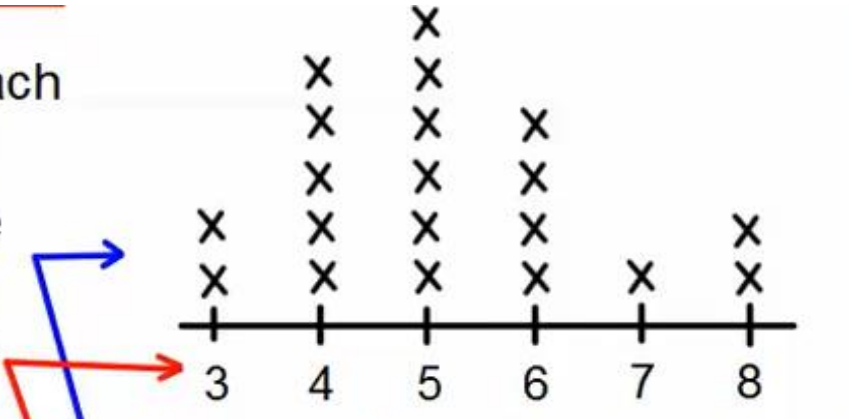

6000
Number of Hours
3000
0 - .009
.010 - .019
.020 - .029
.030 - .039
.040 - .049
.050 - .059
.060 - .069
.070 - .079
.080 - .089
.090 - .099
.100 - .109
.110 - .119
.120 - .129
Ozone ranges (ppm)

Line Plot

A **line plot** uses marks to record each piece of data above a number line.

Some students took a survey to see how many pencils each had in their desk. They recorded the data in the line plot.

How many students have 6 pencils in their desk?



Find 6 on the number line.

The 6 stands for 6 pencils .

There are 4 X's above the 6.

Each X stands for 1 student.

The numbers show the number of pencils in a desk.

So, 4 students have 6 pencils in their desk.

Line Plot

Line plot	Advantages	Disadvantages
▪ A line plot can be used as an **initial record of discrete data values**. ▪ The range determines a number line which is then plotted with X's (or something similar) for each data value.	▪ **Quick analysis of data.** ▪ **Shows range, minimum & maximum, gaps & clusters, and outliers easily.** ▪ **Exact values retained.**	▪ **Not as visually appealing.** ▪ **Best for under 50 data values.** ▪ **Needs small range of data.**

Example: Line Plot

						X					
					X	X		X			
			X		X	X		X			
		X	X		X	**X**	X	X	X		
X		X	X	X	X	X	X	X	X		X
12	13	14	15	16	17	18	19	20	21	22	23

- **Notice:** the cluster at 17 & 18 as well as the gap at 13 and 22. The **mode** is 18, the **median** is the second X from the bottom for number 18, and the **mean** is 17.68 or 18.

*12,14,14,15,15,15,16,17,17,17,17,18,18,18,18,18,19,19,20,20,20,20,21,21,23

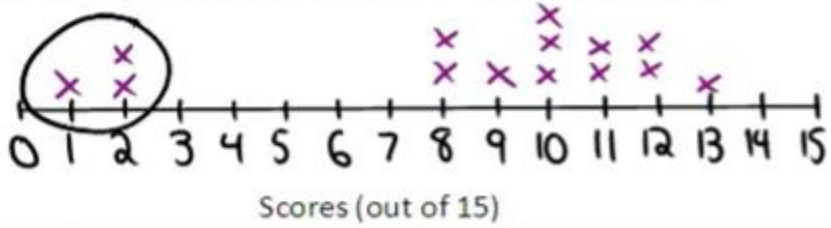
Exercise: Line Plots

(Mean, Median, Mode, Range, outliers)

Mr. Martens' Foundations 11 class wrote a quiz out of 15. These are the scores:

9, 10, 11, 11, 13, 8, 10, 12, 12, 8, 2, 10, 1, 2

a) Represent the scores on a line plot



b) Determine the mean of the quiz scores. (Answer: 8.5)

9 + 10 + ... + 2 = 119 ÷ 14 = 8.5

c) Determine the median of the quiz scores. (Answer: 10)

1, 2, 2, 8, 8, 9, 10, 10, 10, 11, 11, 12, 12, 13

average of 10 and 10 is 10

d) Determine the mode of the quiz scores. (Answer: 10)

10 is most!

e) Determine the dispersion (range) of the quiz scores. (Answer: 12)

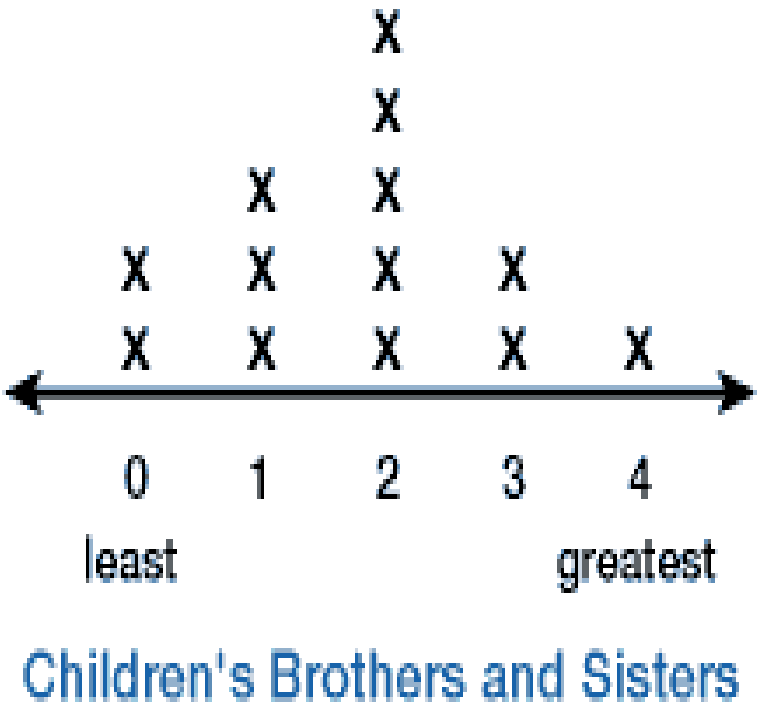
13 − 1 = 12

f) What are the outliers of the quiz scores? (Answer: 1 &

1 + 2

Line plot made from a Tally Chart

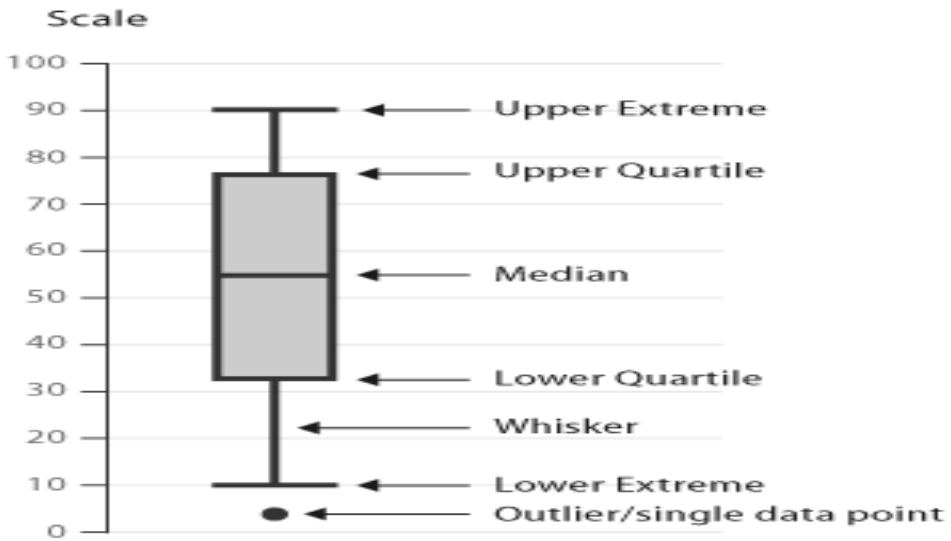
Brothers and Sisters	Number of Children
0	//
1	///
2	////
3	//
4	/



Box and Whisker Plot

Sometimes called a boxplot

- The ends of the **box** are the upper and lower quartiles, so the **box** spans the interquartile range.
- the median is marked by a vertical line inside the **box**.
- the **whiskers** are the two lines outside the **box** that extend to the highest and lowest observations.



Box and Whisker Plot..

- Named so because there is box in the middle & whiskers in the sides.

- Depends on finding the quartiles.

❑ **The interpretation of the plot:**

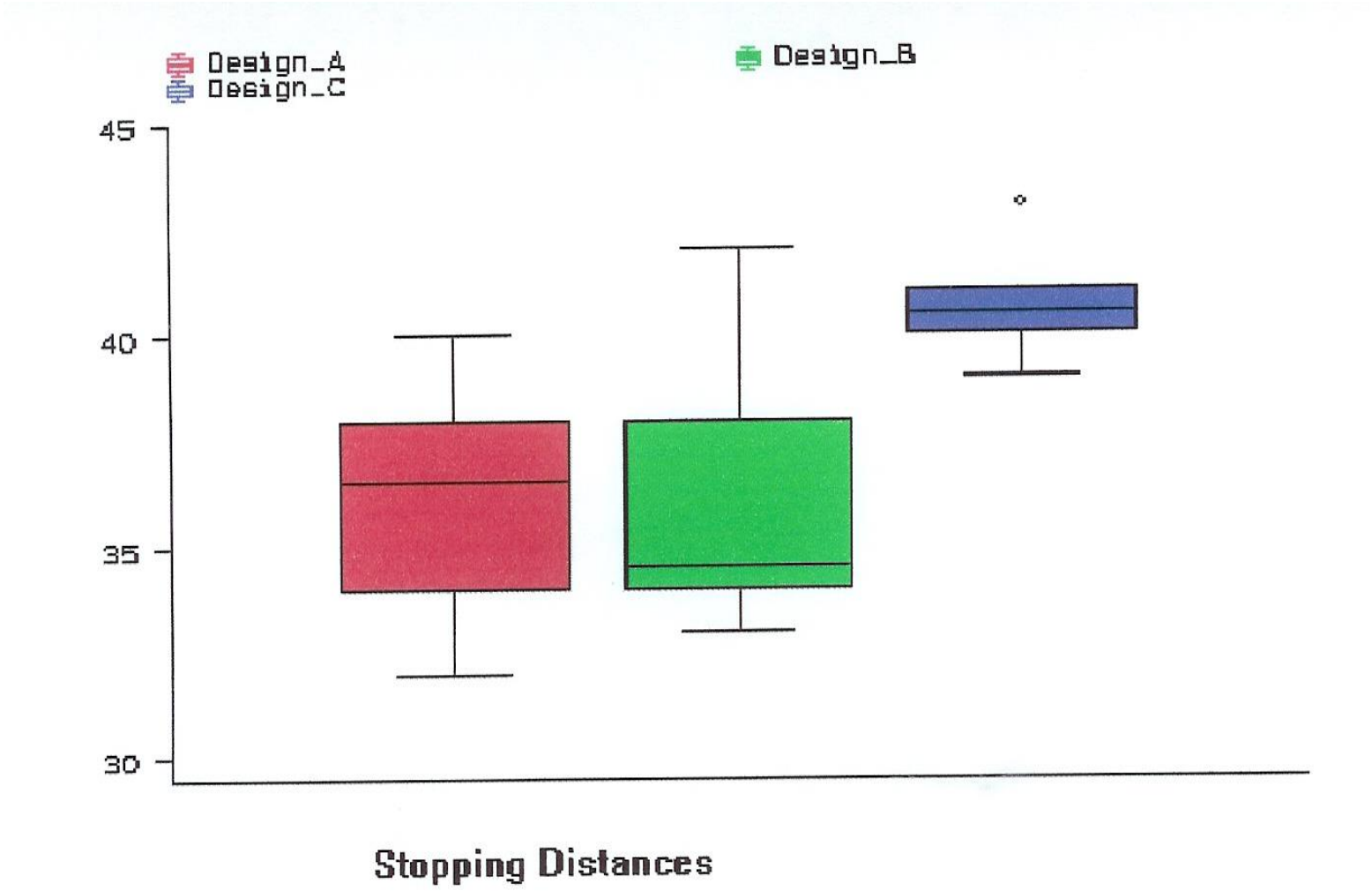
➢ The **line at the center** represent the median of the data.

➢ The **top and bottom** of the box represent the 3^{rd} and 1^{st} quartiles of the data, respectively.

Box and Whisker Plot

Box plot	Advantages	Disadvantages
▪ A box plot is a concise graph **showing the five point summary.**	▪ **Shows 5-point summary and outliers.**	▪ **Not as visually appealing as other graphs**
▪ Multiple box plots can be drawn side by side to compare more than one data set.	▪ **Easily compares two or more data sets.** ▪ **Handles extremely large data sets easily.**	▪ **Exact values are not retained.**

Box & Whisker Graph Example



Histograms V.s Box Plot

- **Histograms:**

➢ Usually provide more information about the distribution.

- **Box plots:**

➢ Allow you to immediately see the median and quartiles and they may also reveal skewness/extreme values more easily.

➢ Can easily compare two data sets using side-by-side box plots "e.g.: comparing the heights of Jordanian students & international students.

Scatter Plot

Scatter plot	Advantages	Disadvantages
▪ A scatter plot displays the **relationship between two factors of the experiment**. ▪ A trend line is used to determine **positive**, **negative** or **no correlation**.	▪ Shows a **trend in the data relationship.** ▪ Plots **several box plot** side-by-side & this is useful if we want to compare several different groups. ▪ Retains exact data values and sample size. ▪ Shows **minimum /maximum and outliers.**	▪ Hard to visualize results in large data sets. ▪ Flat trend line gives inconclusive results. ▪ Data on both axes should be **continuous.**

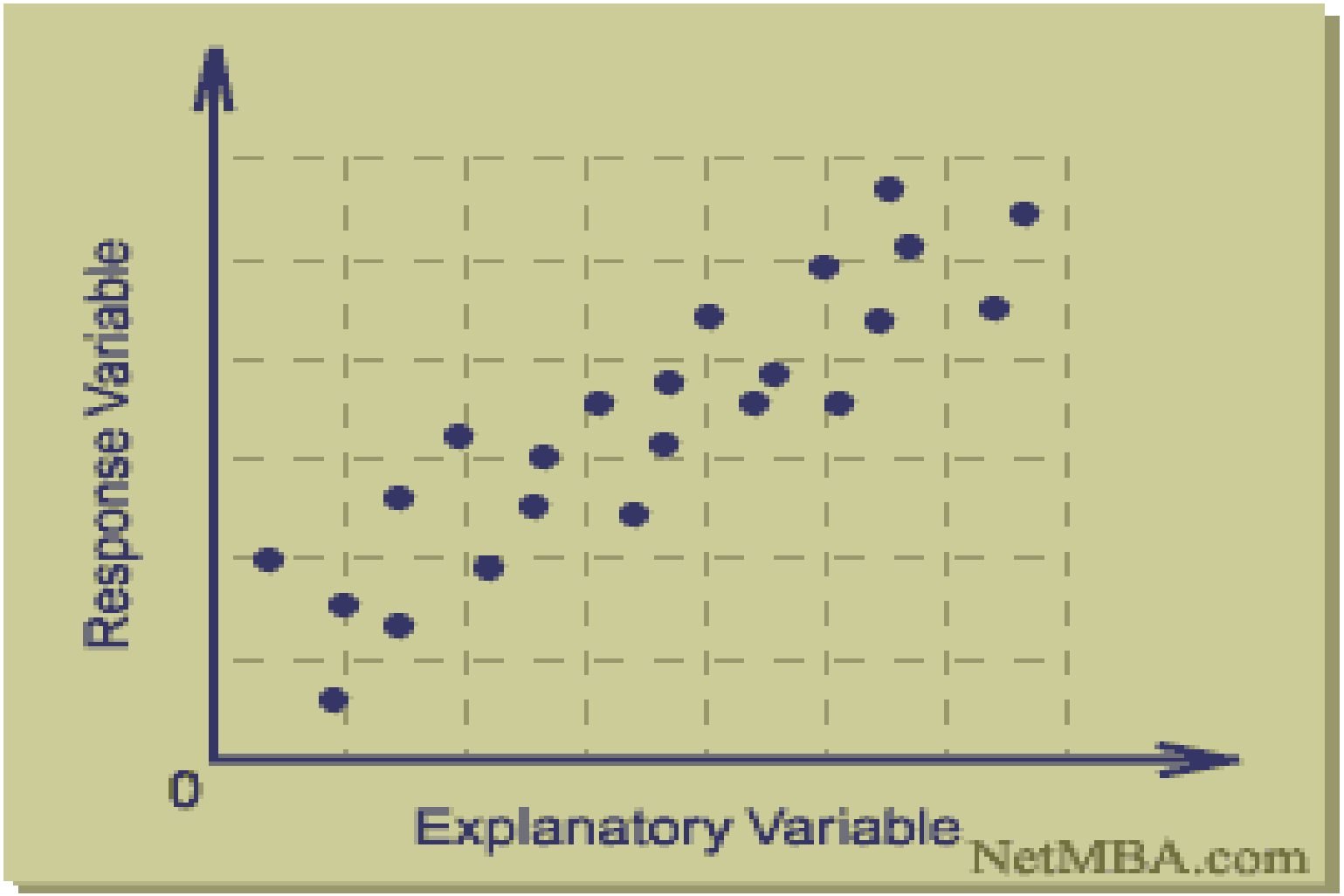
❖ **When plotting two quantitative variables there will be a crossing point "relation" in between.**

Describing/Interpreting scatterplots

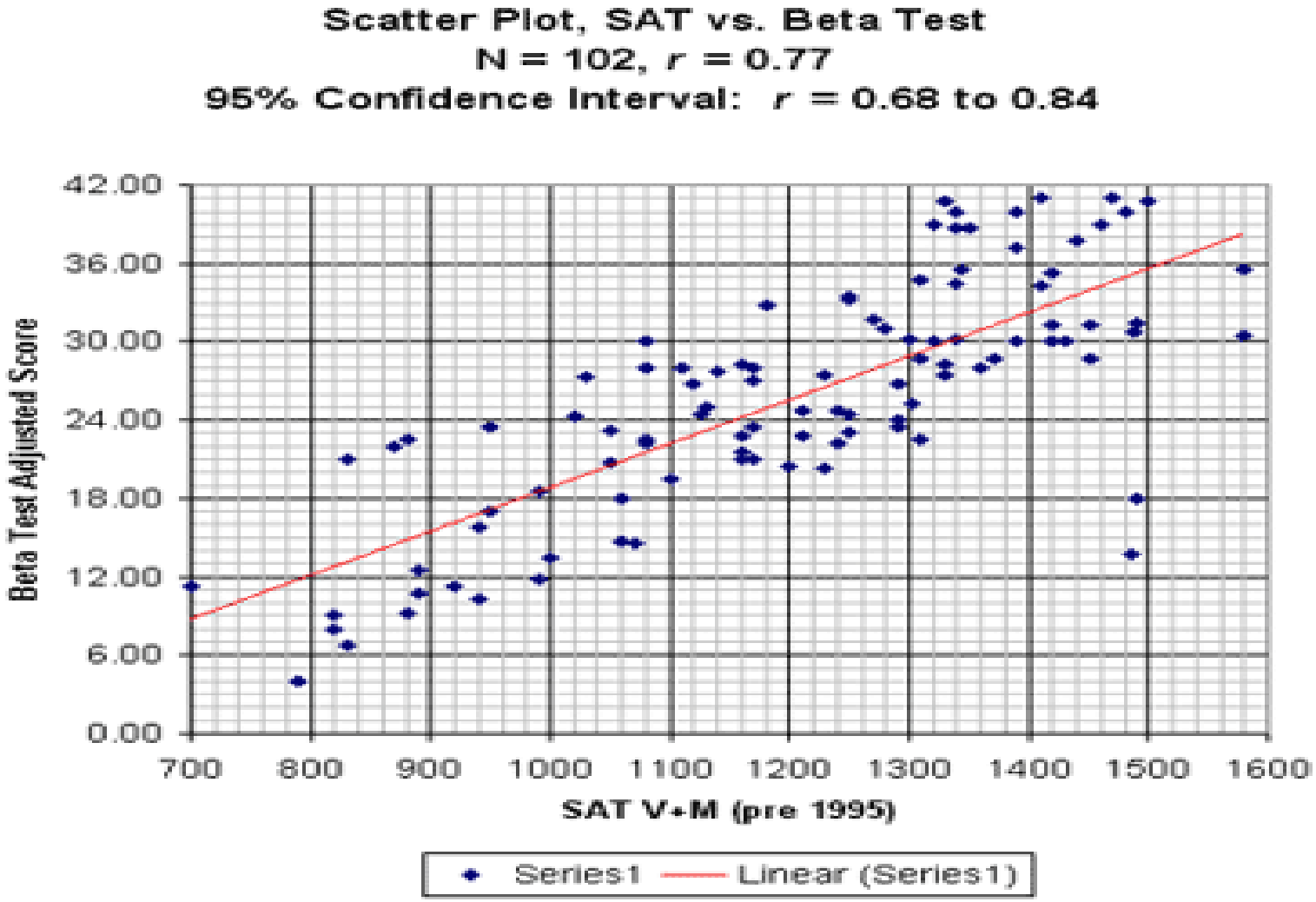
- When describing a scatterplot, we describe the relationship by examining the **form, direction,** and **strength** of the association. We look for an overall pattern …

- **Form**: linear (a straight line), curved, clusters, no pattern.
- **Direction**: positive, negative, no direction.
- **Strength**: how closely the points fit the "form".

Scatter Plot

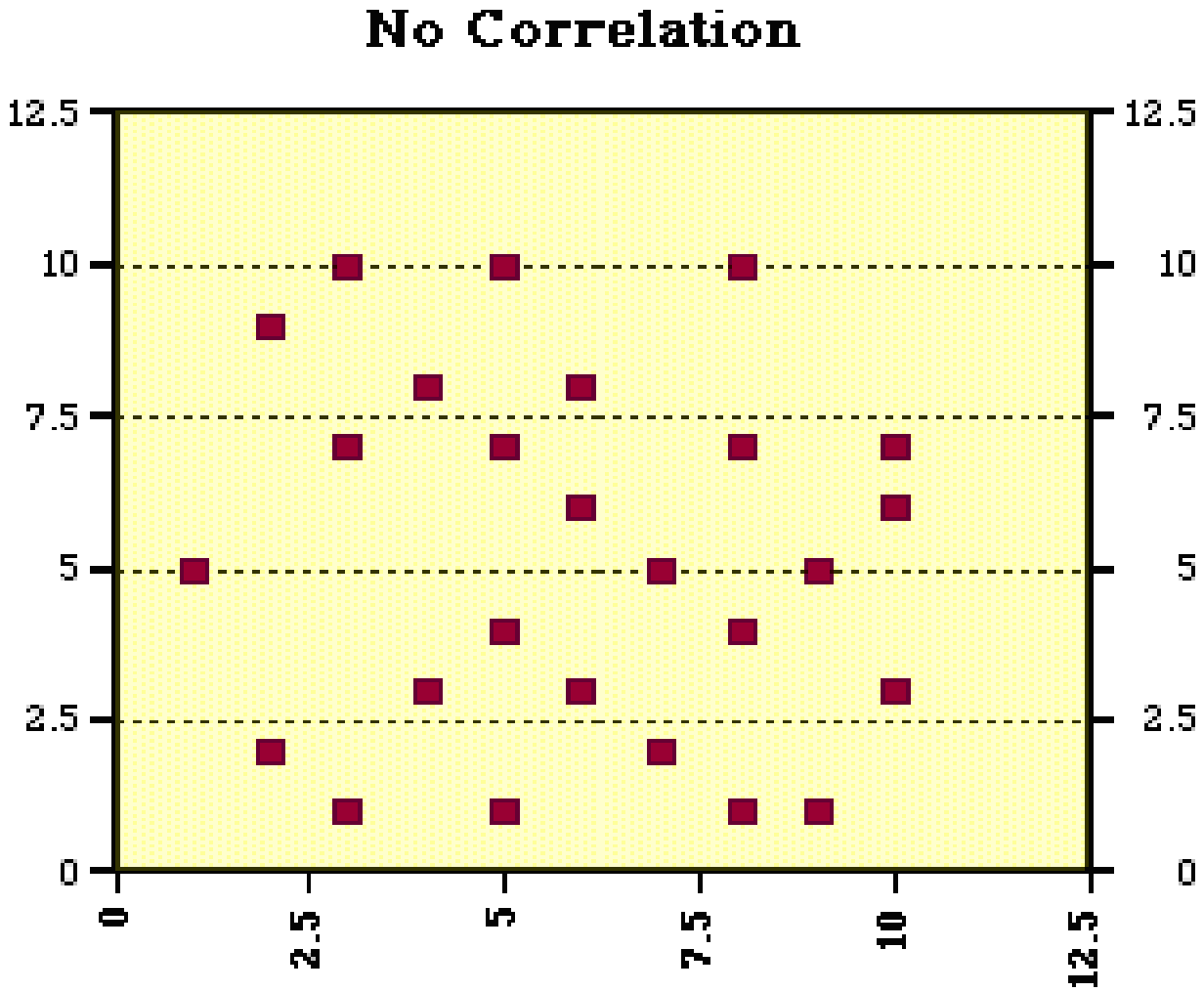


Scatter Plot Example



No Correlation

If there is absolutely **no correlation** present, the value given is 0.



Direction of a linear association

Positive or Negative

- A ***linear*** relationship is given a directional description of Positive or Negative.

- **Positive association:** High values of one variable tend to occur together with high values of the other variable.

- **Negative association:** High values of one variable tend to occur together with low values of the other variable.

Positive correlation

- **Example:** If you look at the age of a child and the child's height, you will find that as the child gets older, the child gets taller. Because both are going up, it is **positive correlation**.

Age	1	2	3	4	5	6	7	8
Height “	25	31	34	36	40	41	47	55

Negative correlation

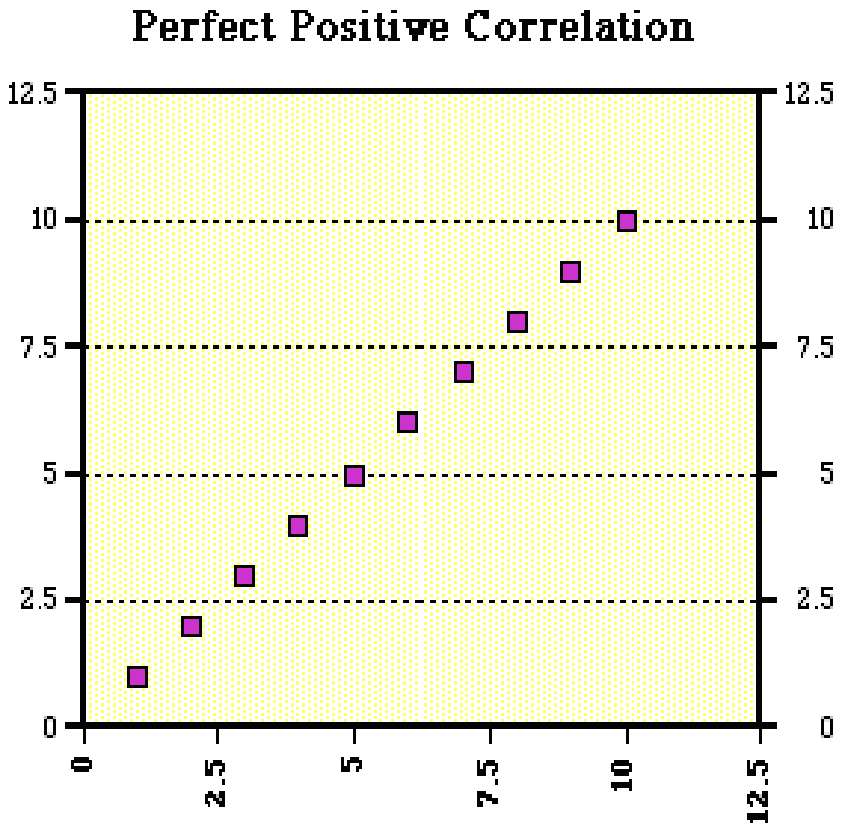
- **Example:** If you look at the age of your family's car and its value, you will find as the car gets older, the car is worth less. This is **negative correlation**.

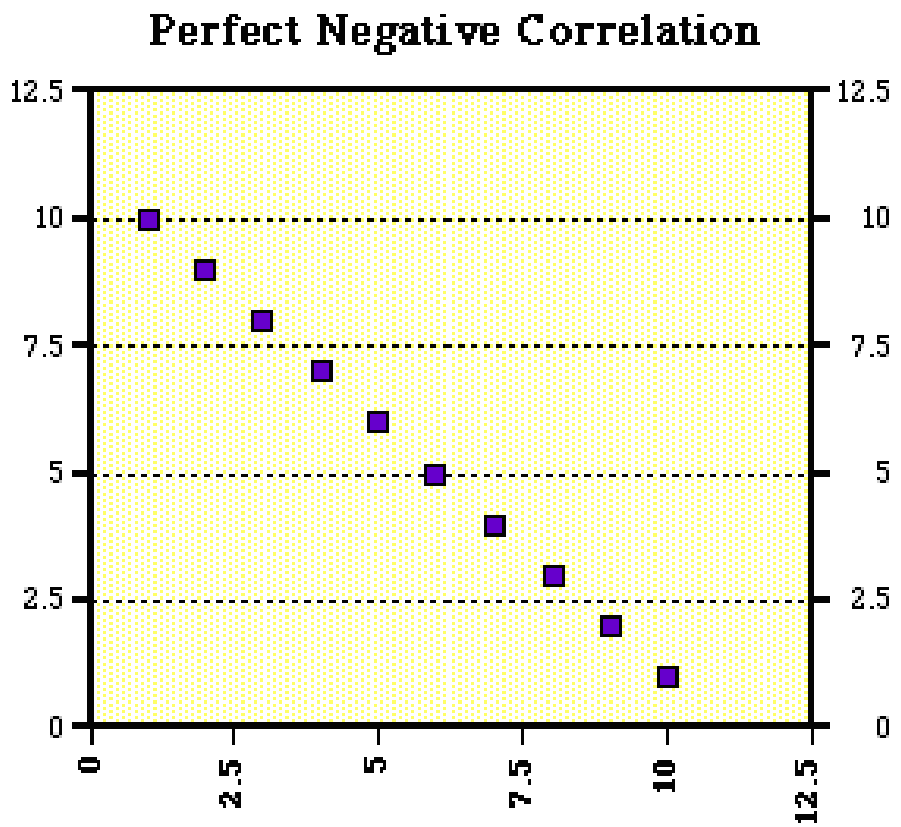
Age of car	1	2	3	4	5
Value	$30,000	$27,000	$23,500	$18,700	$15,350

Perfect linear correlation:

A **perfect positive** correlation is given the value of 1.

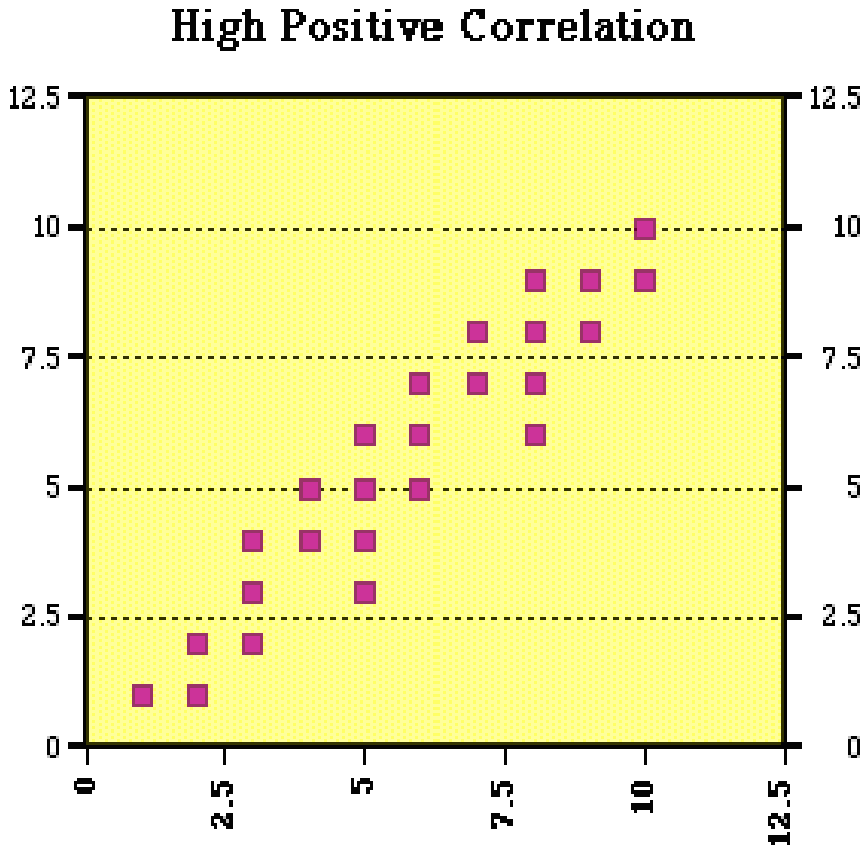
A **perfect negative** correlation is given the value of -1.

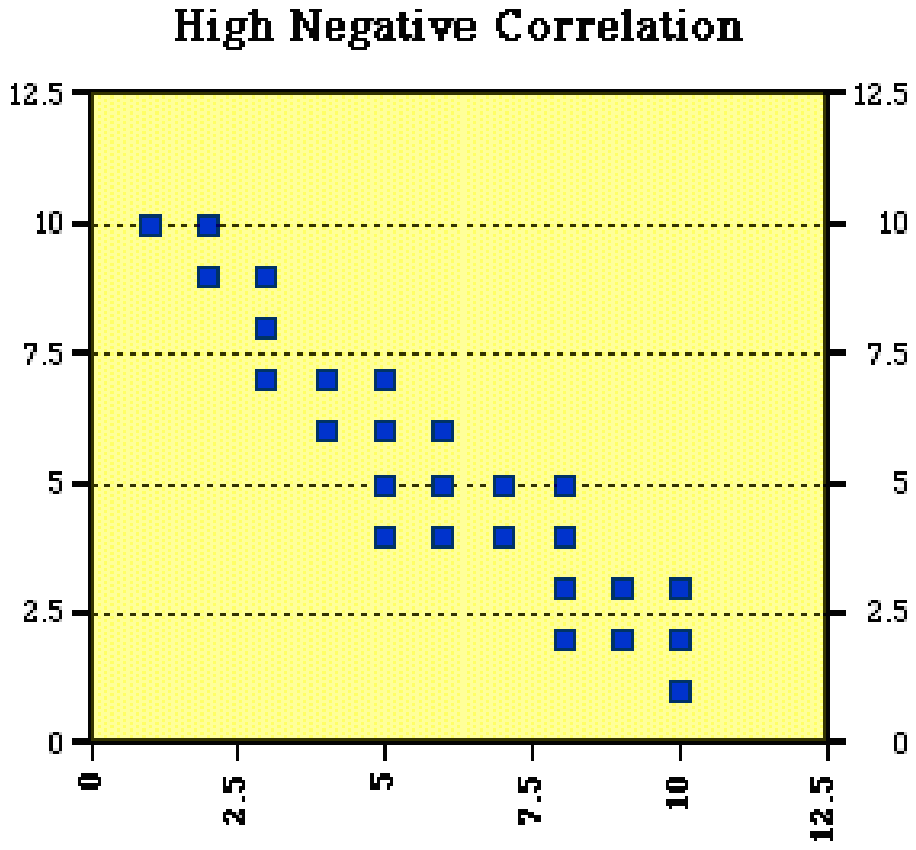




Strong linear correlation:

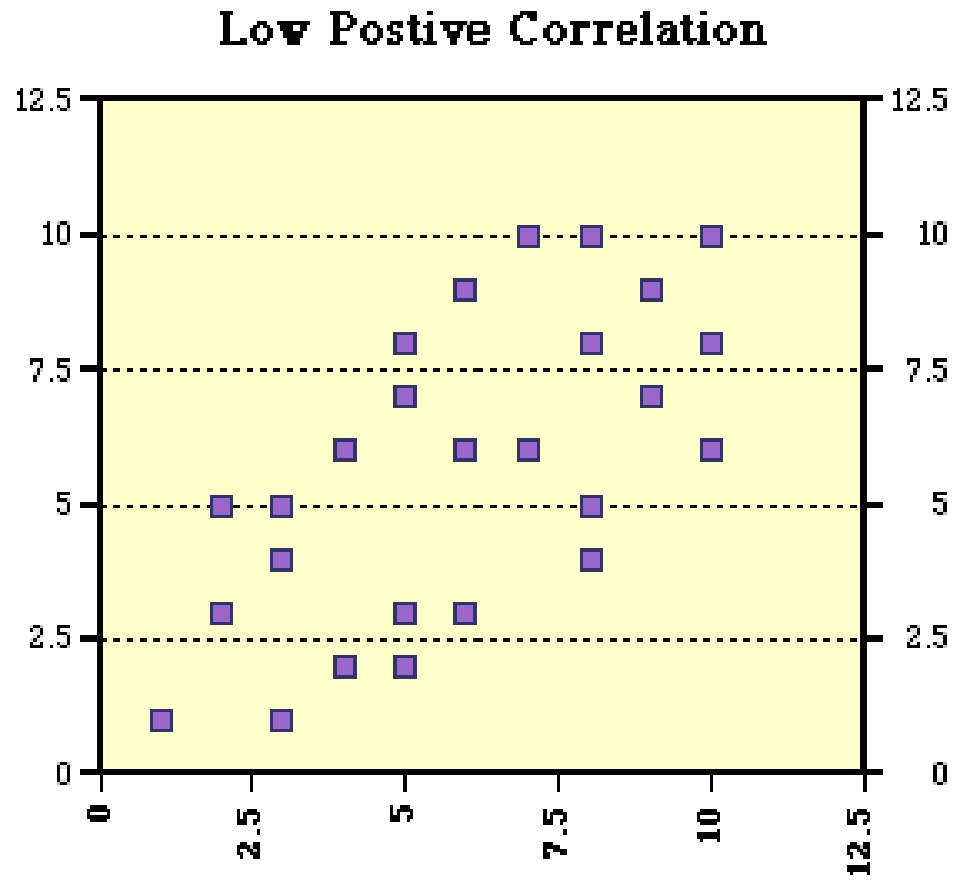
The closer the number is to 1 or -1, the **stronger** the correlation, or the stronger the relationship between the variables.

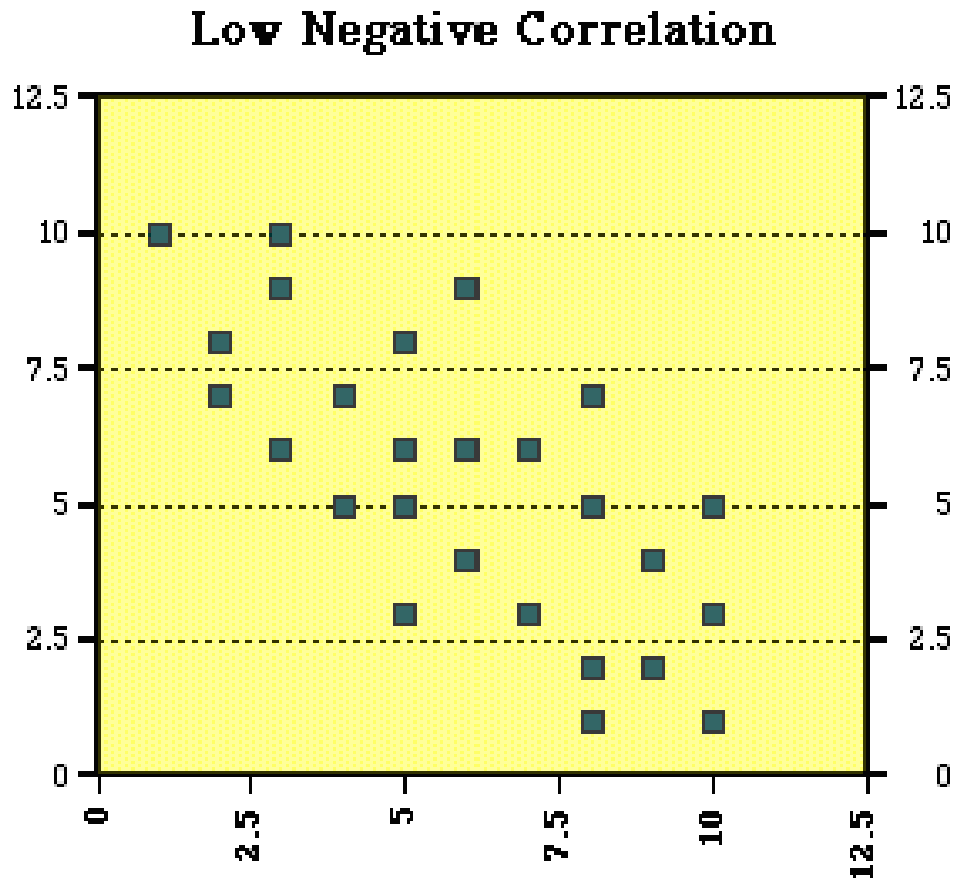




Weak linear correlation:

The closer the number is to 0, the **weaker** the correlation.

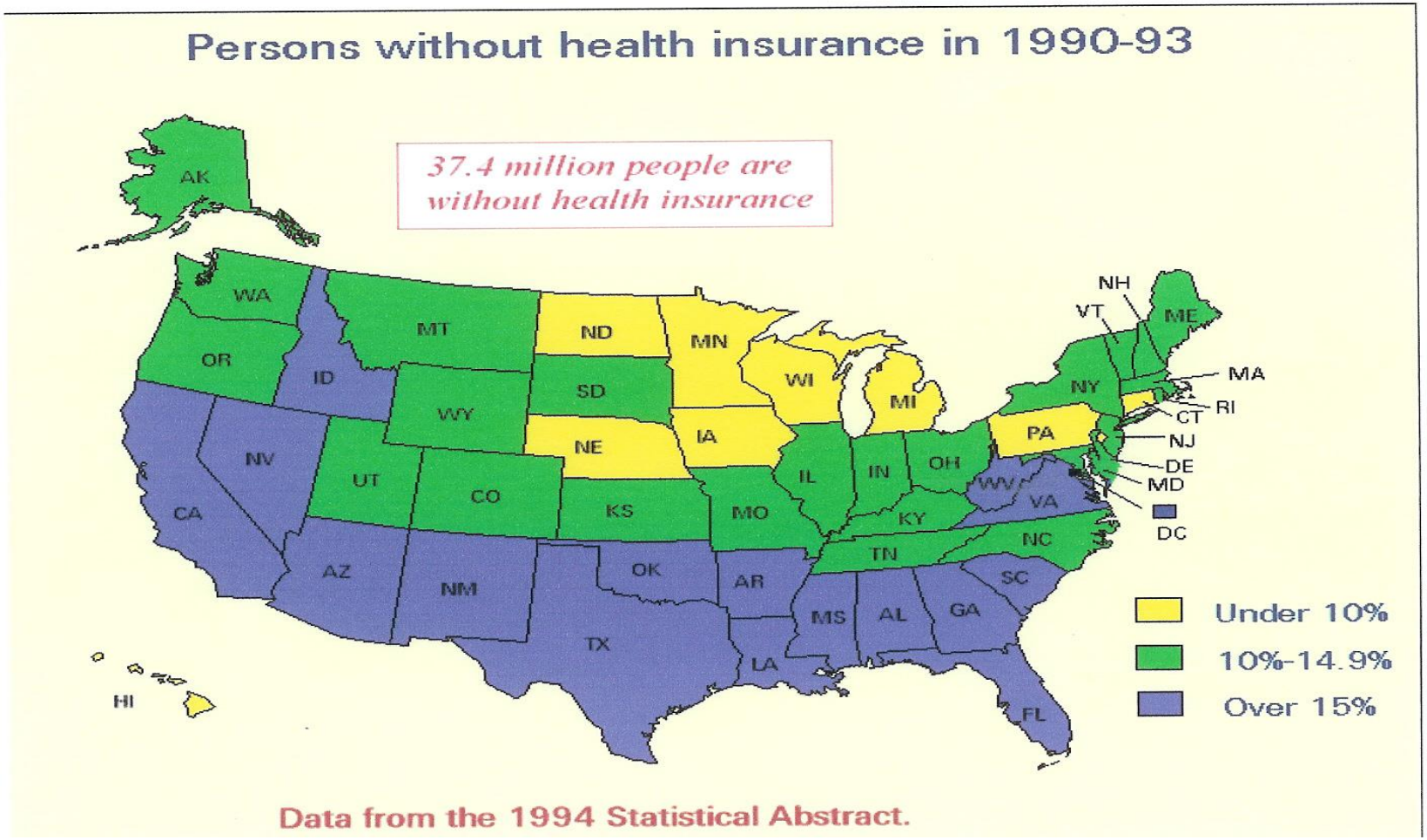




Map Graph
Cosmography

Map chart	Advantages	Disadvantages
▪ A map chart displays data by shading sections of a map, and must include a key. ▪ A total data number should be included.	▪ Good visual appeal ▪ Overall trends show well.	▪ Needs limited categories ▪ No exact numerical values ▪ Color key can skew visual interpretation.

Map Chart
Cosmography …

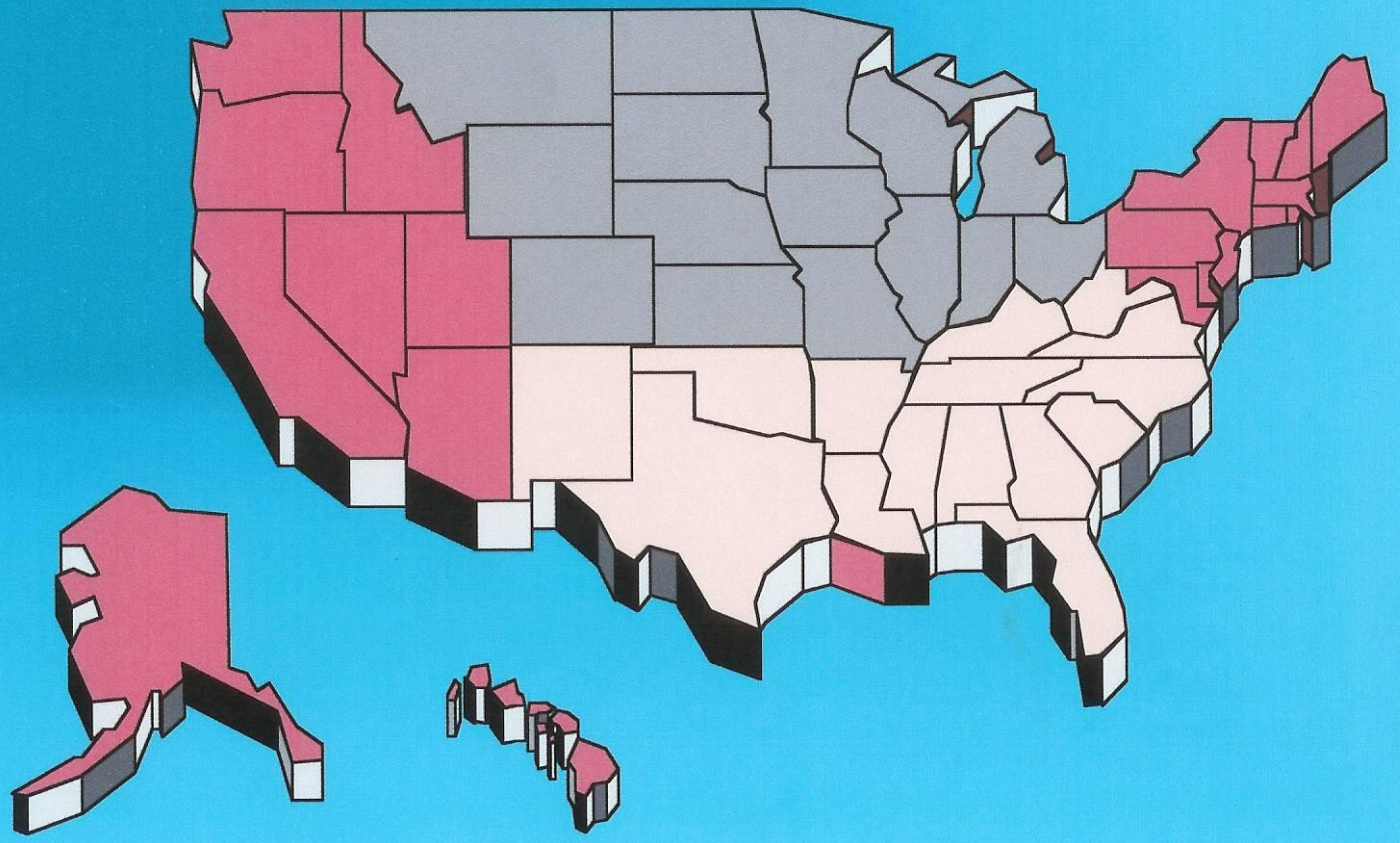


Map Graph …

- Parts of whole so similar to a pie graph.
- **Less numerical** and more graphic.

Cosmograph

Depicts parts to the whole, but less numerical than pie graphs

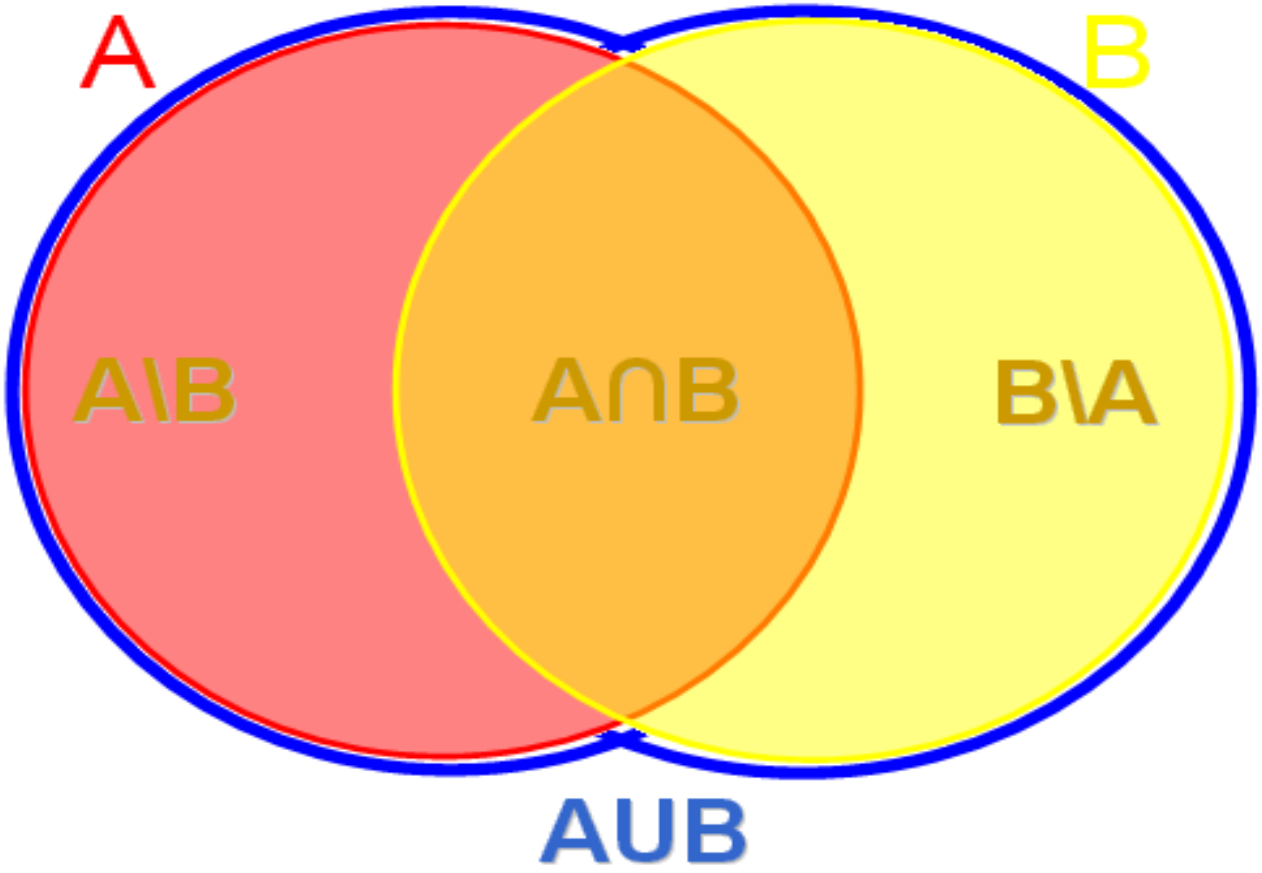


Venn Diagram

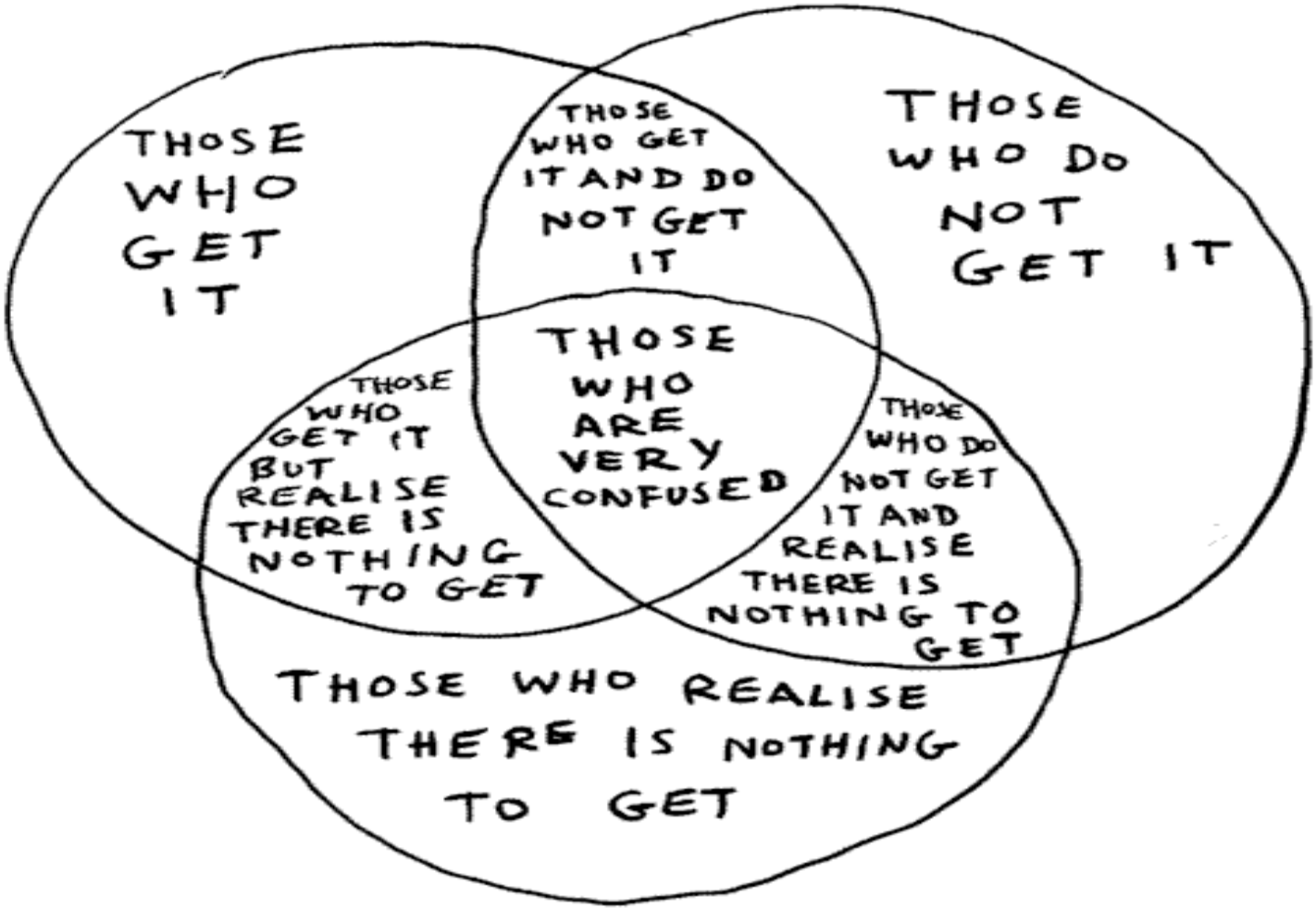
- Venn diagrams enable students to organize information visually so they are able to see the **relationships** between two or three sets of items.

- They can then identify similarities and differences.

Venn Diagram..



Venn Diagram…



Frequency table

- **A frequency table** shows how often something occurs.
- The frequency may be shown by tally marks or the number.
- Data is displayed numerically.

Frequency Table

Data Intervals	Frequency	Cumulative Frequency	Relative Frequency (%)	Cumulative Relative Frequency (%)
10-19	5	5		
20-29	18	23		
30-39	10	33		
40-49	13	46		
50-59	4	50		
60-69	4	54		
70-79	2	56		
Total				