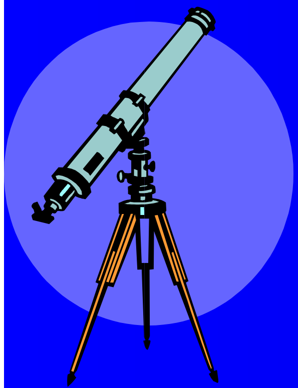
STATISTICS

CHAPTER 1

THE NATURE OF STATISTICAL DATA



CHAPTER 1

THE NATURE OF STATISTICAL DATA

ORDINAL

INTERVAL

NOMINAL

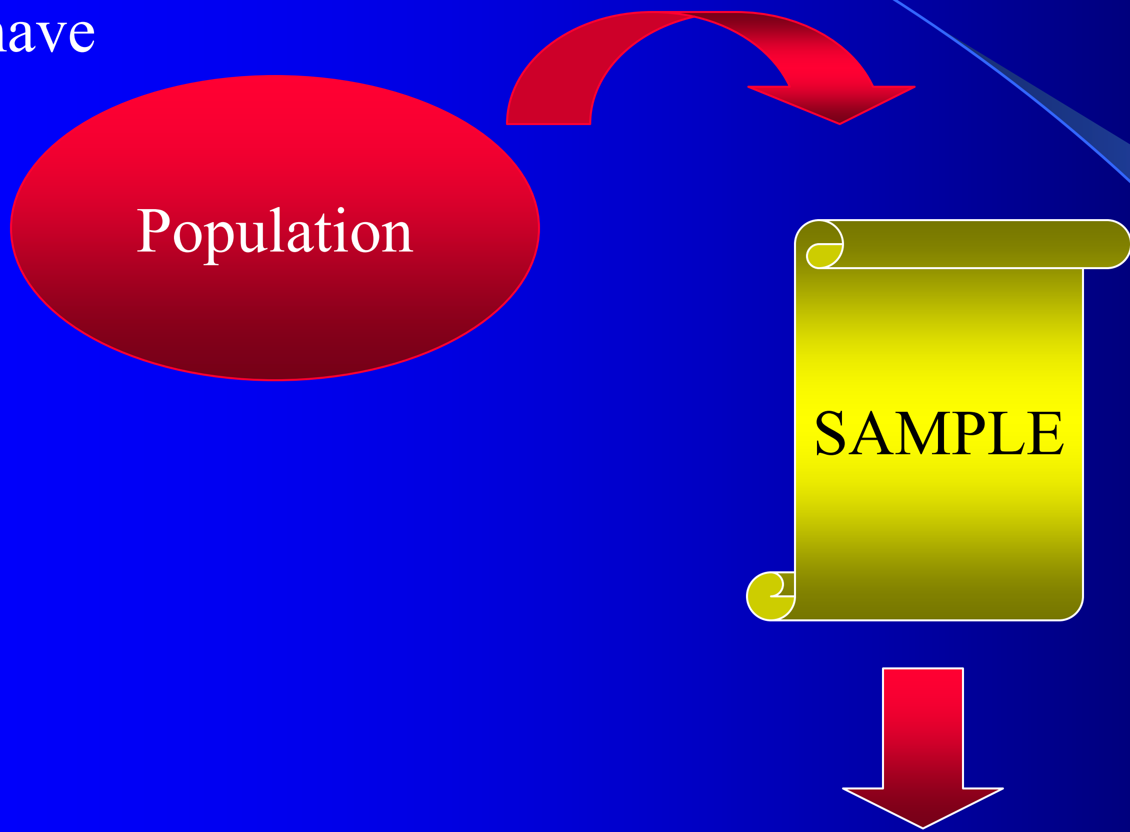
RATIO

- The distinction is important because nature of the data suggests the statistical technique we should use

CHAPTER 2

DATA COLLECTION AND SAMPLING

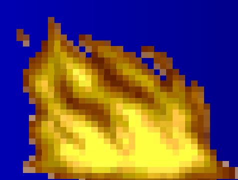
We have



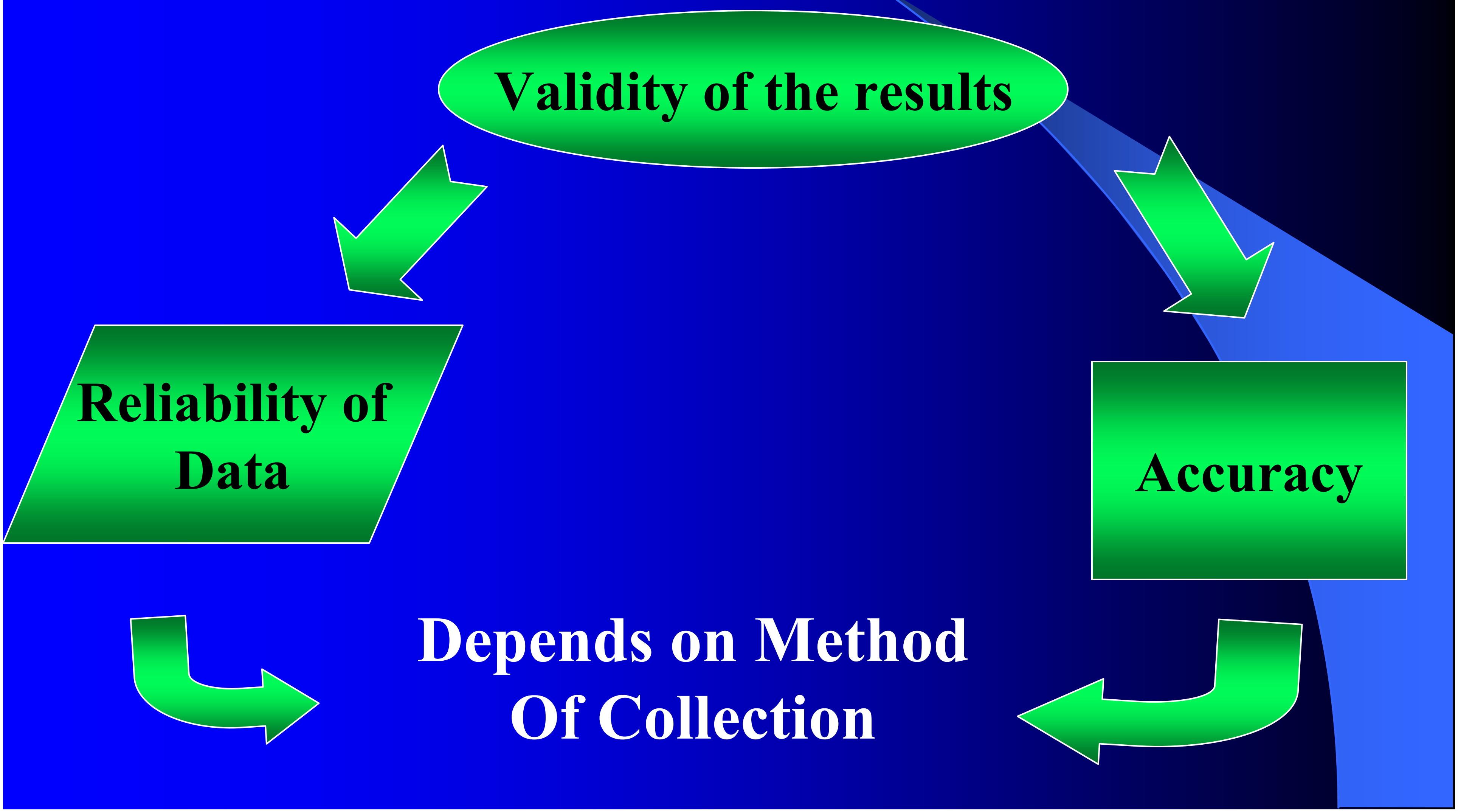
Use it to get info about population

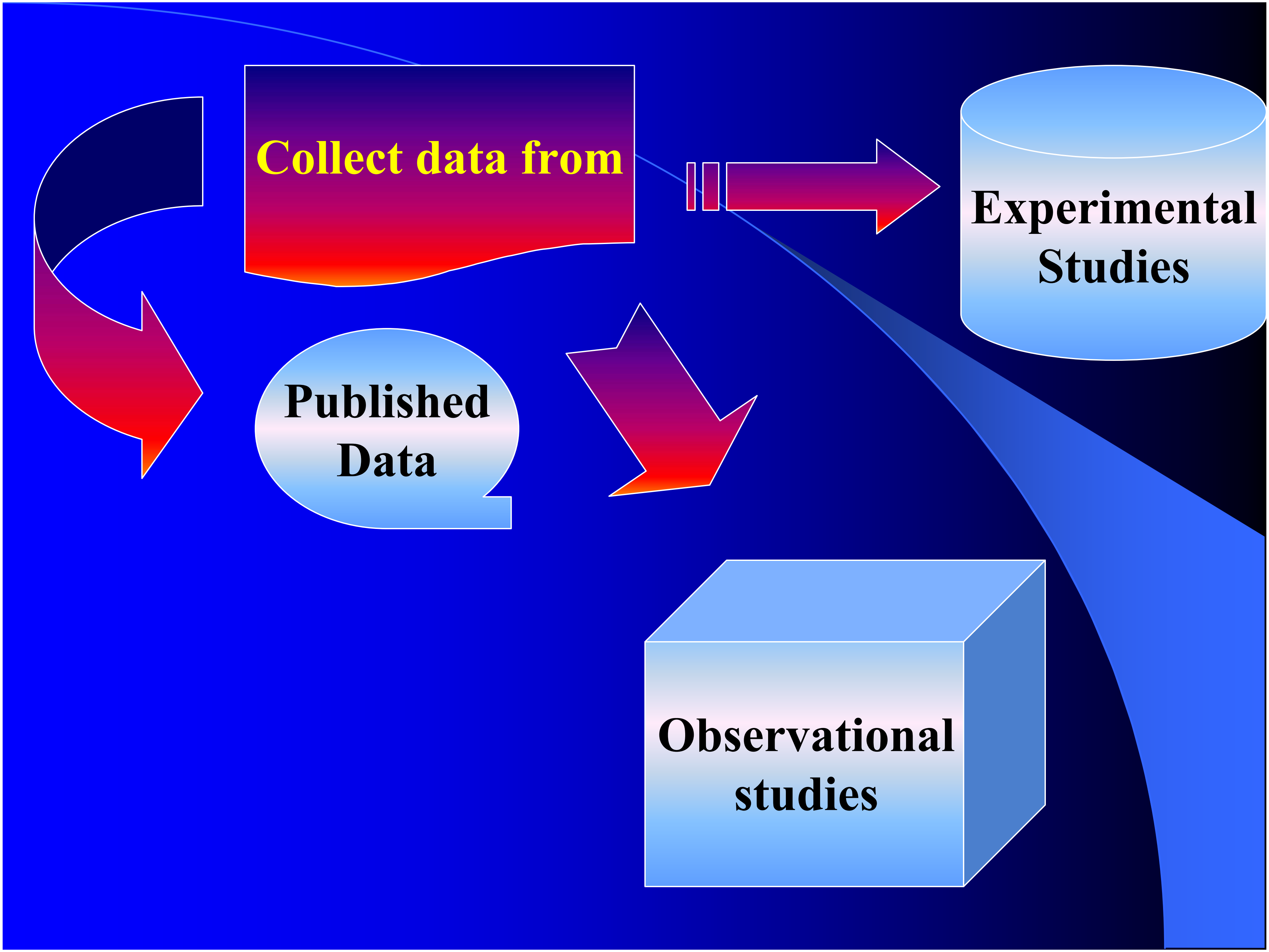
WHY?

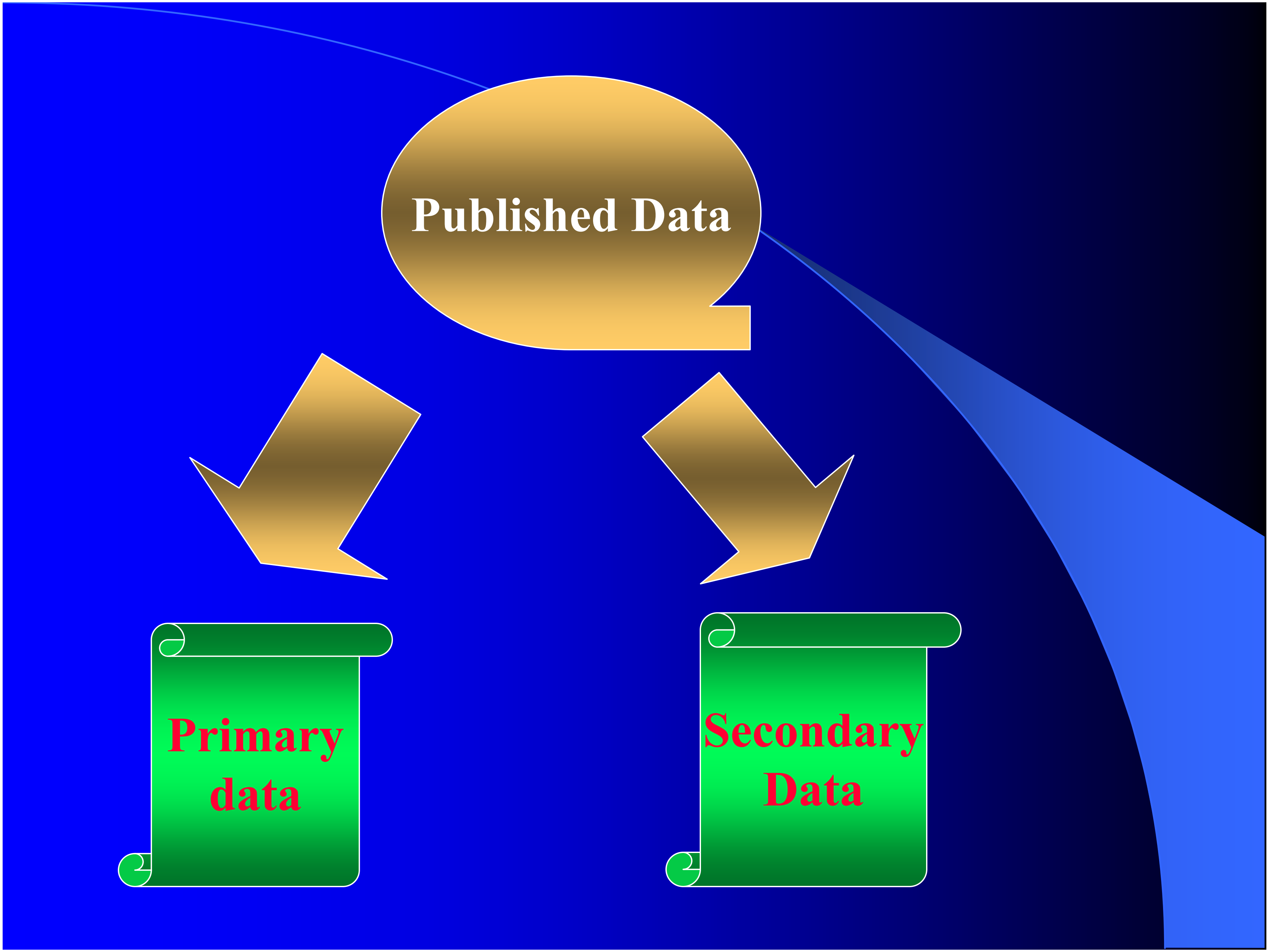
- EXPENSIVE
- IMPRACTICAL

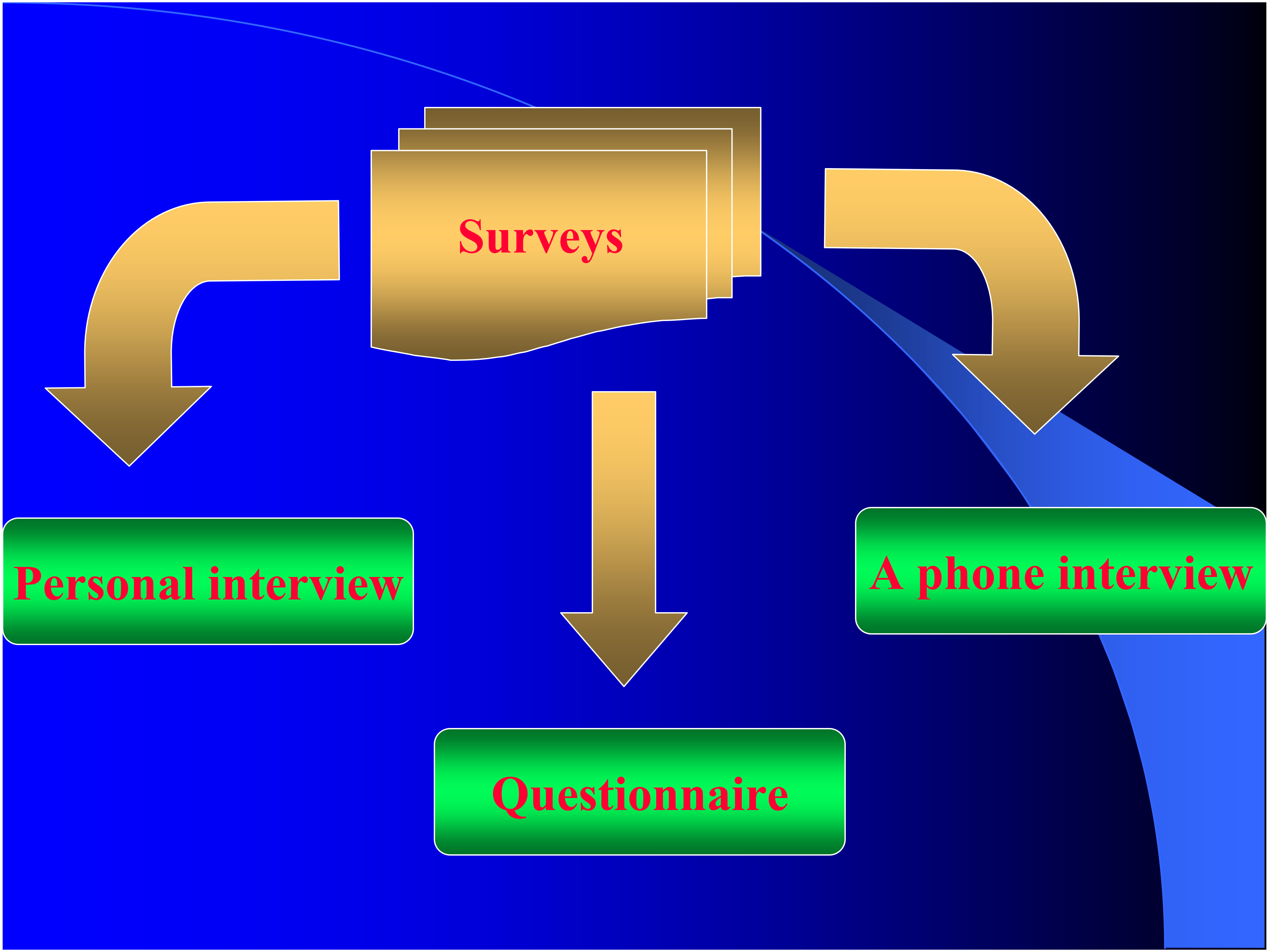


SOURCES OF DATA




Collect data from
Published Data
Experimental Studies
Observational studies


Published Data
Primary data
Secondary Data


Surveys
Personal interview
Questionnaire
A phone interview

Personal interview	A phone interview interview
•E (response) high	•E (response) low
•Cost high	•Cost low

Questionnaire

- Short
- Simple words
- Yes / No
- Avoid Leading Questions
- Pretest questionnaire

Sampling

Why? → COST

- Want to calculate population parameter
- Estimate that using a sample

Simple random sample (you can use Minitab and Excel to generate random number)

Stratified random sample

Separating population into:

1. Sex
2. Age
3. Occupation
4. Income

Cluster sampling

→ Simple groups

Sample size ↑ → Accuracy ↑

ERRORS IN SAMPLING

E.g: $\mu - \overline{\chi}$ ⟵ For sample

For population

SAMPLING ERROR:

$= \mu - \overline{\chi}$

To reduce it ⟶ Take larger sample

NON-SAMPLING ERROR

1. In data
2. Non response error
3. Selection bias

CHAPTER 3

SUMMARIZING DATA
LISTING AND GROUPING

Listing numerical data

Listing is the first task in any kind of statistical analysis

Stem-And-Leaf-Display

Example

To illustrate this technique consider the following data on the number of rooms occupied each day in a resort hotel during a recent month of June.

55	49	37	57	46	40	64	35	73	62
61	43	72	48	54	69	45	78	46	59
40	58	56	52	49	42	62	53	46	81

The smallest and largest values are 35 and 81, so that a dot diagram would allow for 47 possible values.

STEP 1

37	35									
49	46	40	43	48	45	46	40	49	42	46
55	57	54	59	58	56	52	53			
64	62	61	69	62						
73	72	78								
81										

STEP 2

3	7 5
4	9 6 0 3 8 5 6 0 9 2 6
5	5 7 4 9 8 6 2 3
6	4 2 1 9 2
7	3 2 8
8	1

And this is what we refer to as a stem-and-leaf display. In this arrangement, each row is called a stem, each number on a stem to the left of the vertical line is called a stem label, and each number on a stem to the right of the vertical line is called a leaf.

STEP 3

3	5	7									
4	0	0	2	3	5	6	6	6	8	9	9
5	2	3	4	5	6	7	8	9			
6	1	2	2	4	9						
7	2	3	8								
8	1										

FREQUENCY DISTRIBUTIONS

The frequency does not show much detail.

The construction of a frequency distribution consists essentially of three steps:

1- Choosing the classes (intervals or categories)
2- Sorting or tallying the data into these classes
3- Counting the number of items in each class

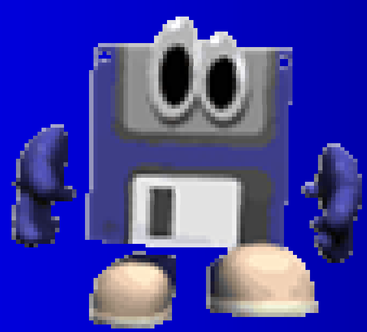
We seldom use fewer than 5 steps or more than 15 classes; the exact number we use in a given situation depends largely on how many measurements or observations there are.

We always make sure that each item (measurement or observation) goes into one and only one class.

Make these ranges multiples of numbers that are easy to work with, such as 5, 10, 100

Example

Use the following numbers to construct a frequency distribution.



81	83	94	73	78	94	73	89	112	80
94	89	35	80	74	91	89	83	80	82
91	80	83	91	89	82	118	105	64	56
76	69	78	42	76	82	82	60	73	69
91	83	67	85	60	65	69	85	65	82
53	83	62	107	60	85	69	92	40	71
82	89	76	55	98	74	89	98	69	87
74	98	94	82	82	80	71	73	74	80
60	69	78	74	64	80	83	82	65	67
94	73	33	87	73	85	78	73	74	83
83	51	67	73	87	85	98	91	73	108

<table>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></tr>
<tr><td>30-39</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td>40-49</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td>50-59</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>4</td></tr>
<tr><td>60-69</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td>19</td></tr>
<tr><td>70-79</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td>24</td></tr>
<tr><td>80-89</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td>||||</td><td></td><td>39</td></tr>
<tr><td>90-99</td><td>||||</td><td>||||</td><td>||||</td><td></td><td></td><td></td><td></td><td></td><td></td><td>15</td></tr>
<tr><td>100-109</td><td>|||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>3</td></tr>
<tr><td>110-119</td><td>||</td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>2</td></tr>
<tr><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td>Total</td><td>110</td></tr>
</table>

Frequency distribution.

30-39	2
40-49	2
50-59	4
60-69	19
70-79	24
80-89	39
90-99	15
100-109	3
110-119	2
Total	110

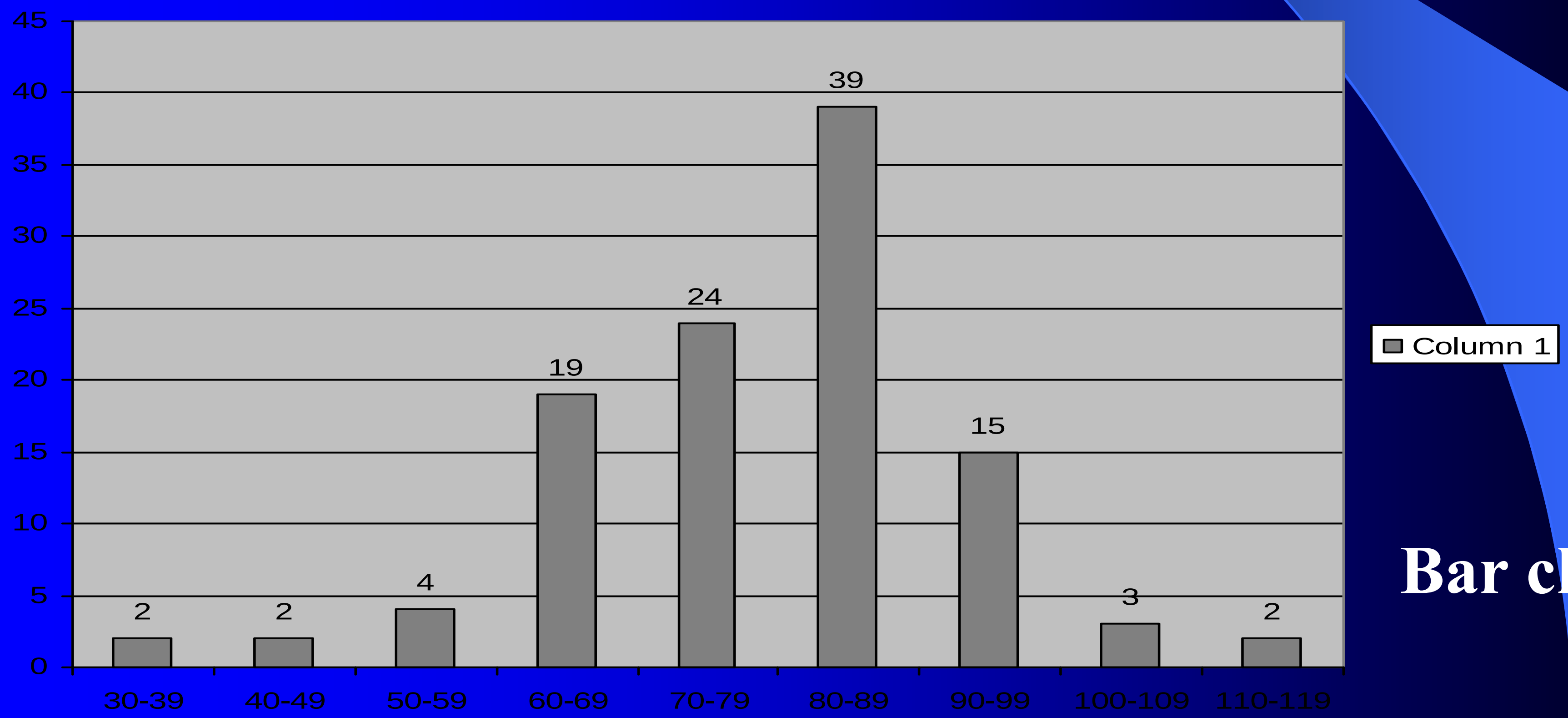
Percentage Frequency Distribution:

Classes	Frequency	Percentage
30-39	2	1.82%
40-49	2	1.82%
50-59	4	3.64%
60-69	19	17.27%
70-79	24	21.82%
80-89	39	35.45%
90-99	15	13.64%
100-109	3	2.73%
110-119	2	1.82%
Total	110	100%

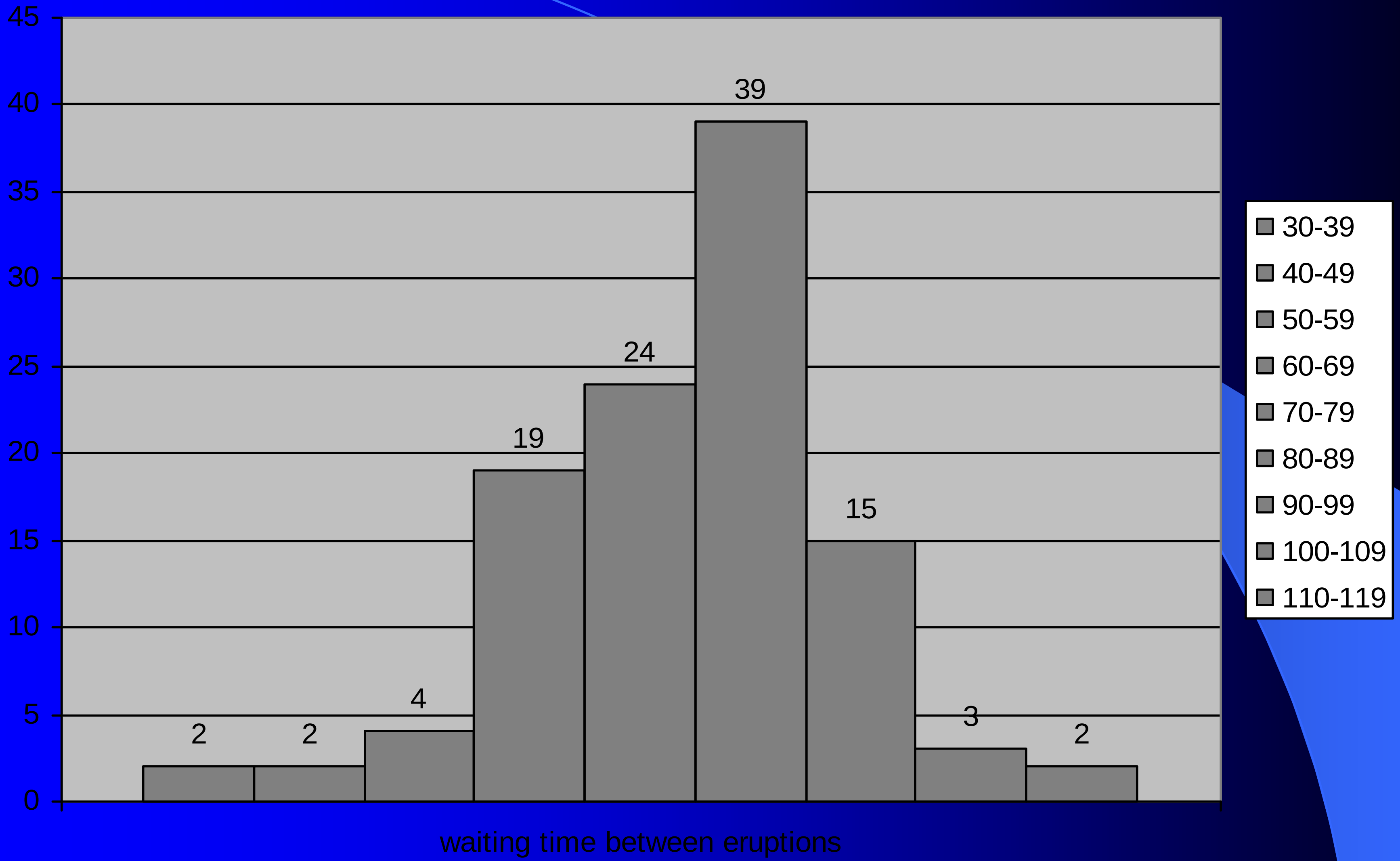
Example

Convert the distribution of the last example into a cumulative "less than" distribution.

Graphical Representation



Bar chart


45
40
35
30
25
20
15
10
5
0
2
2
4
19
24
39
15
3
2
waiting time between eruptions
30-39
40-49
50-59
60-69
70-79
80-89
90-99
100-109
110-119

Histogram

CHAPTER 4

Summarizing Data: Measures of Location

The Mean

$$\text{Sample Mean} = \frac{X_1 + X_2 + X_3 + \ldots\ldots + X_n}{n}$$

$$\bar{X} = \frac{\Sigma x}{n}$$

$$\mu = \frac{\Sigma x}{N}$$

THE MEAN:

- It always exists
- Unique
- The means of several sets of data can always be combined into the overall mean of all the data
- Means of repeated samples drawn from the same population usually do not fluctuate, or vary, widely

Overall Mean of combined data

$$\overline{\overline{X}} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + \ldots\ldots + n_k\overline{x}_k}{n_1 + n_2 + \ldots\ldots + n_k} = \frac{\Sigma n \cdot \overline{x}}{\Sigma n}$$

The Median

The median is the value of the middle item when n is odd, and the mean of the 2 middle items when n is even.

EXAMPLE 10

In five recent weeks, a town reported 36, 29, 42, 25 and 29 burglaries. Find the median number of burglaries for these Weeks.

Solution:

The data must first be arranged according to size

25 29 29 36 42

25 29 29 36 42

It can be seen that the middle one, the median, is 29

EXAMPLE 11

However where n is even as in the set of numbers below, we find that the median is mean of the two values nearest to the middle

30 32 35 37 38 40

$$\frac{35 + 37}{2} = 36$$

The Mode

The mode is defined simply as the value that occurs with the highest frequency.

The mean in the case of ungrouped data:

$$\bar{X} = \frac{\Sigma x \,.\, f}{n}$$

Where: x $\longrightarrow$ Refer to midpoint

F $\longrightarrow$ Refer to frequency

See example page 69

Classes	Frequency	x	x.f
30-39	2	34.5	69.0
40-49	2	44.5	89.0
50-59	4	54.5	218.0
60-69	19	64.5	1225.5
70-79	24	74.5	1788.0
80-89	39	84.5	3295.5
90-99	15	94.5	1417.5
100-109	3	104.5	313.5
110-119	2	114.5	229.0
Total	110		8645.0

$$\text{Then: } \bar{x} = \frac{8645.0}{110} = 78.59$$

CHAPTER 5

Summarizing data: Measures of variation

The Range

The **range** is defined as the difference between the largest and smallest values in a set of data.

The variance and standard deviation

Sample standard deviation

$$S = \left(\frac{\sum (X - \overline{X})^2}{n - 1} \right)^{½}$$

Sample variance

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

Population standard deviation

$$\sigma = \left(\frac{\Sigma(X - \overline{\mu})^2}{N} \right)^{½}$$

Computing formulae for the sample standard deviation

$$S^2 = \frac{\Sigma x^2 - \frac{(\Sigma x)^2}{n}}{n-1}$$

Coefficient of variation

$$V = \frac{S}{\bar{X}} \cdot 100\%$$

Or

$$V = \frac{\sigma}{\mu} \cdot 100\%$$

The variane in the case of ungrouped data:

$$s^2 = \Sigma x^2 . f - \frac{(\Sigma x.f)^2}{\Sigma f}$$

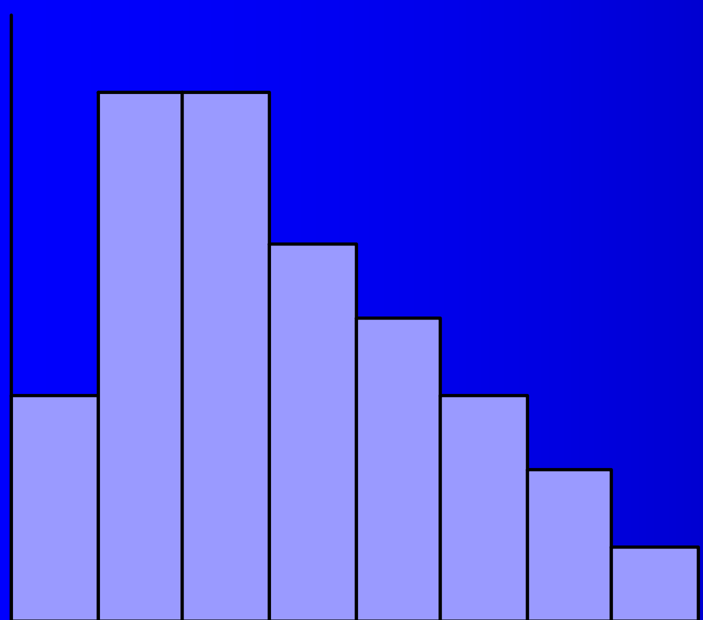
See example page 69

Classes	Frequency	x	x.f	X^2.f
30-39	2	34.5	69.0	2380.5
40-49	2	44.5	89.0	3960.5
50-59	4	54.5	218.0	11881
60-69	19	64.5	1225.5	79044.75
70-79	24	74.5	1788.0	133206
80-89	39	84.5	3295.5	278469.75
90-99	15	94.5	1417.5	133760.75
100-109	3	104.5	313.5	32760.75
110-119	2	114.5	229.0	26220.5
Total	110		8645.0	701877.5

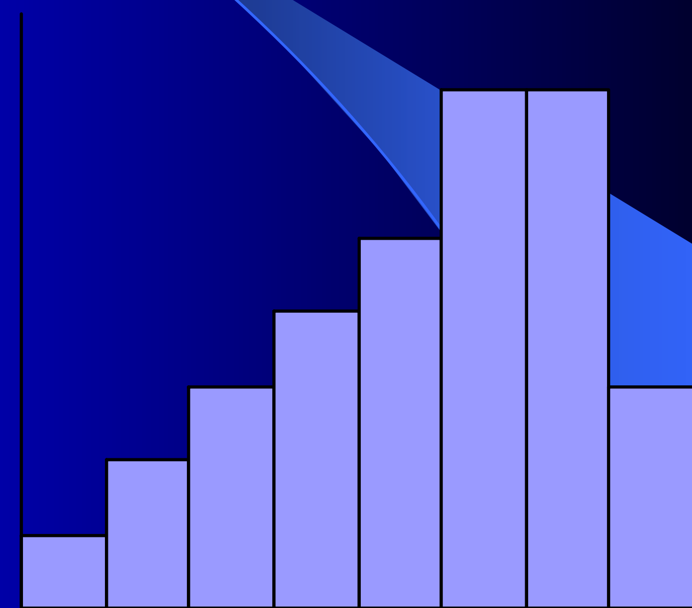
$$S^2 = 701877.5 - \frac{(8645)^2}{110} = 224593.1$$

$$S = \sqrt{224593.1} = 14.35$$

The Description of Grouped Data



Positive skewed



Negative skewed

Skewed Distributions

Pearsonian Coefficient of kewness

$$SK = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard deviation}}$$

$$-3 < SK < 3$$

Measures of Association

<u>Covariance</u>

$$\text{Population covariance} = COV(X,Y) = \frac{\sum (X_i - \mu_x)(y_i - \mu_x)}{N}$$

$$\text{Sample covariance} = COV(X,Y) = \frac{\sum (X_i - \overline{X})(y_i - \overline{y})}{n - 1}$$

Coefficient of Correlation

$$\rho = \frac{COV\,(X,Y)}{\sigma_x \, \sigma_y}$$

$$r = \frac{COV\,(X,Y)}{S_x \, S_y}$$

$$-1 < r < 1$$

Example

Let: $\overline{X} = 18.0$

$S_x = 4.02$

$\overline{y} = 217.0$

$S_y = 63.9$

$n = 15$

$$COV(X,Y) = \frac{\sum (X_i - \overline{X})(y_i - \overline{y})}{n-1} = \frac{2,859.2}{14} = 204.2$$

$$r = \frac{COV(X,Y)}{S_x S_y} = \frac{204.2}{4.02*63.9} = 0.796$$

Chapter 6

Simple Linear Regression And Correlation

Model

First-Order Linear Model

$$y = \beta_o + \beta_1 x + \in$$

where

y = dependent variable

x = independent variable

β_0 = y-intercept

β_1 = slope of the line

The slide of the line is defined as the ratio rise/run or change in y/change in x

Least Squares Method

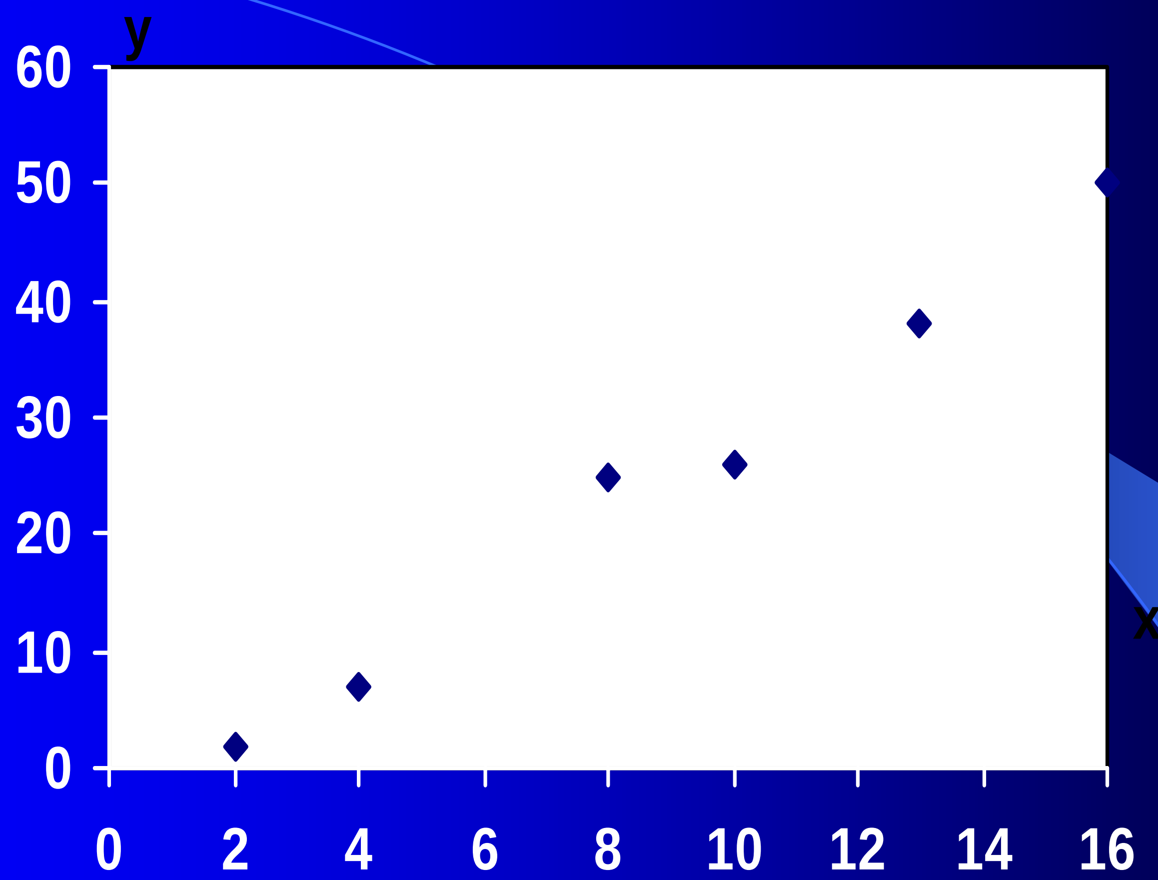
Example

Given the following six observations of variables *x* and y, determine the straight line that fits these data.

x	**2**	**4**	**8**	**10**	**13**	**16**
y	**2**	**7**	**25**	**26**	**38**	**50**

Solution:

As a first step we graph the data



we want to determine the line that minimizes

$$\sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

where y_i represents the observed value of y and represents the value of y calculated from the equation of the line. That is

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Calculation of $\hat{\beta}_1$ and $\hat{\beta}_0$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Shortcut Formulas for SS_x and SS_{xy}

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n}$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$$

Returning to our example we find

$$\sum x_i = 53$$

$$\sum y_i = 148$$

$$\sum x_i^2 = 609$$

$$\sum x_i y_i = 1{,}786$$

$$SS_x = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n} = 609 - \frac{(53)^2}{6} = 140.833$$

$$SS_{xy} = \sum x_i y_i - \frac{\sum x_i \sum y_i}{n} = 1{,}786 - \frac{53 \times 148}{6} = 478.667$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_x} = \frac{478.667}{140.833} = 3.399$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = \frac{148}{6} - (3.399 \times \frac{53}{6}) = -5.336$$

Thus, the least squares line is

$$\hat{y} = -5.356 + 3.399x$$

Using The Regression Equation

we can use it to forecast and estimate values of the dependent variable.