Data Exploration

Agenda

- **Introduction to Statistics**
- Type Of Statistics
- Organizing Numerical and Categorical Data
- Data Collection : Sampling Technique
- Sampling error and non sampling error

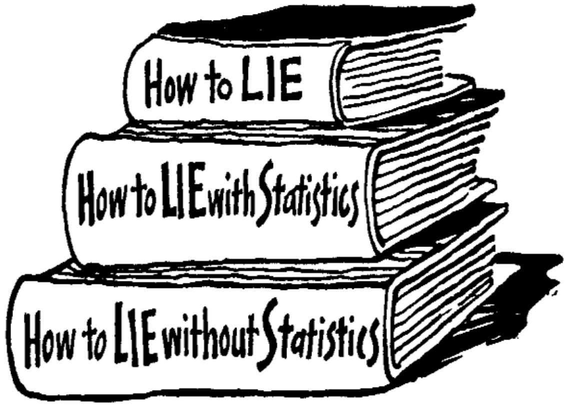
Introduction to Statistics

“There are three kinds of lies: lies, damned lies, and statistics“

(B.Disraeli)

What is Statistics?

- “…a set of procedures and rules…for reducing large masses of data to manageable proportions and for allowing us to draw conclusions from those data”



Introduction to Statistics

Why learn statistics?

- Data are everywhere
- Statistical techniques are used to make many decisions that affect our lives
- No matter what your career, you will make professional decisions that involve data. An understanding of statistical methods will help you make these decisions efectively

Applications of statistical concepts in the world

- Finance
  - correlation and regression, index numbers, time series analysis
- Marketing
  - hypothesis testing, chi-square tests, nonparametric statistics
- Personnel
  - hypothesis testing, chi-square tests, nonparametric tests
- Operating management
  - hypothesis testing, estimation, analysis of variance, time series analysis

Introduction to Statistics

Statistics definition

- Statistics is the science of conducting studies to collect, organize, summarize, analyze and draw conclusions from data.
- Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to assist in making more effective decisions
- Statistical analysis – used to manipulate summarize, and investigate data, so that useful decision-making information results.

Key Definitions

- A **population** (universe) is the collection of things under consideration
- A **sample** is a portion of the population selected for analysis
- A **parameter** is a summary measure computed to describe a characteristic of the population
- A **statistic** is a summary measure computed to describe a characteristic of the sample

Agenda

- Introduction to Statistics
- **Type Of Statistics**
- Organizing Numerical and Categorical Data
- Data Collection : Sampling Technique
- Sampling error and non sampling error

Type of Statistics

- **Starts with data**
  - Nominal, Ordinal, Interval, and Ratio
- **Descriptive statistics**
  - Exploring, visualizing, and summarizing data without fitting the data to any models than Collecting, presenting, and describing data
- **Inferential statistics**
  - Identification of a suitable model than Testing either predictions or hypotheses of the model and Drawing conclusions and/or making decisions concerning a population based only on sample data

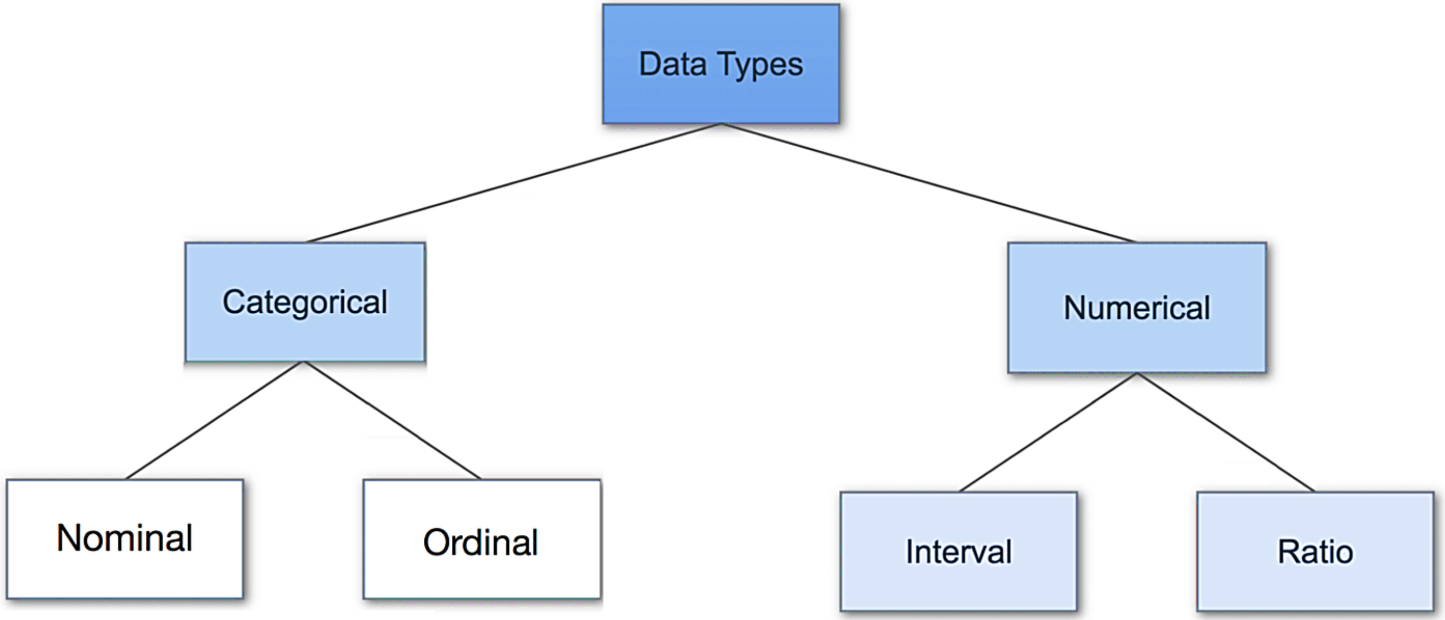
Databases are highly structured for storage but do not automatically reveal patterns and insights.

We explore databases in a five-step process:

- Understand the data
- Organize and subset the database
- Examine individual variables and their distributions
- Calculate summary measures for individual variables
- Examine relationships among variables

Type of Statistics

Understand the data



Type of Statistics

Understand the Data

- **Nominal data,** which simply names the category of record.
  - **Example: A GENDER field, with only two variables (male and female) and The DESCRIPTION field in previous slides, with numerous variables (e.g., ADVIL, TYLENOL X/STRGTH LIQ).**

- **Ordinal data,** also identifies category of record but with a natural order to the values.
  - **Example: High, Medium and Low, than Numerical rankings, where 5 = most preferred, 1 = least preferred.**

- **Interval data,** which conveys a sense of the difference between values.
  - **Example: The Fahrenheit scale.**

- **Ratio data,** based on a scale with a meaningful zero point.
  - **Example: Monetary units, ages.**

Agenda

- Introduction to Statistics
- Type Of Statistics
- **Organizing Numerical and Categorical Data**
- Data Collection : Sampling Technique
- Sampling error and non sampling error

Data Exploration:
Organizing Numerical and Categorical Data

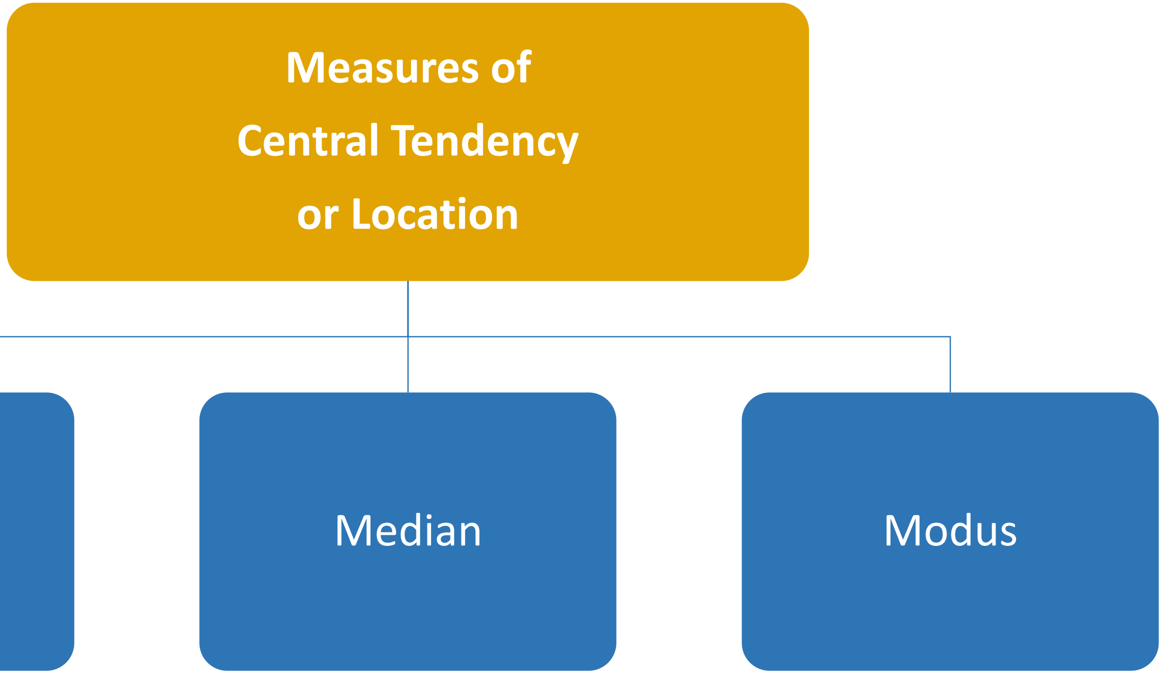
- Numerical data is information that is something that is measurable. It is always collected in number form, although there are other types of data that can appear in number form. An example of numerical data would be the number of people that attended the movie theater over the course of a month

- Examples of categorical variables are race, sex, age group, and educational level. While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.

Data Exploration:
Organizing Numerical and Categorical Data

- Numerical data is information that is something that is measurable. It is always collected in number form, although there are other types of data that can appear in number form. An example of numerical data would be the number of people that attended the movie theater over the course of a month
- Examples of categorical variables are race, sex, age group, and educational level. While the latter two variables may also be considered in a numerical manner by using exact values for age and highest grade completed, it is often more informative to categorize such variables into a relatively small number of groups.

Measures of
Central Tendency or Location



Measures of

Central Tendency

or Location

Mean

Median

Modus

Mean

Average, the sum of the observed values divided by the number of observations.

Population Mean

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

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x}{n}$$

Median

Middle value of data when sorted in order of magnitude, **50th percentile**

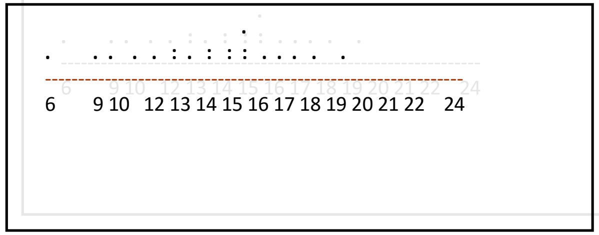
Sales	Sorted Sales
9	6
6	9
12	10
10	12
13	13
15	14
16	14
14	15
14	16 ← Median
16	16
17	16
16	17
24	17
21	18
22	18
18	19
19	20
18	21
20	22
17	24

(20+1)50/100=10.5

16 + (.5)(0) = 16

Mode

Most frequently- occurring value



Mode = 16

Mean

Average, the sum of the observed values divided by the number of observations.

Population Mean

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

Sample Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x}{n}$$

Median

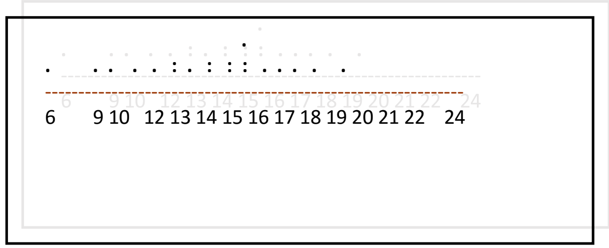
Middle value of data when sorted in order of magnitude, **50th percentile**

Sales	Sorted Sales
9	6
6	9
12	10
10	12
13	13
15	14
16	14
14	15
14	16 ← Median
16	16
17	16
16	17
24	17
21	18
22	18
18	19
19	20
18	21
20	22
17	24

(20+1)50/100=10.5
16 + (.5)(0) = 16

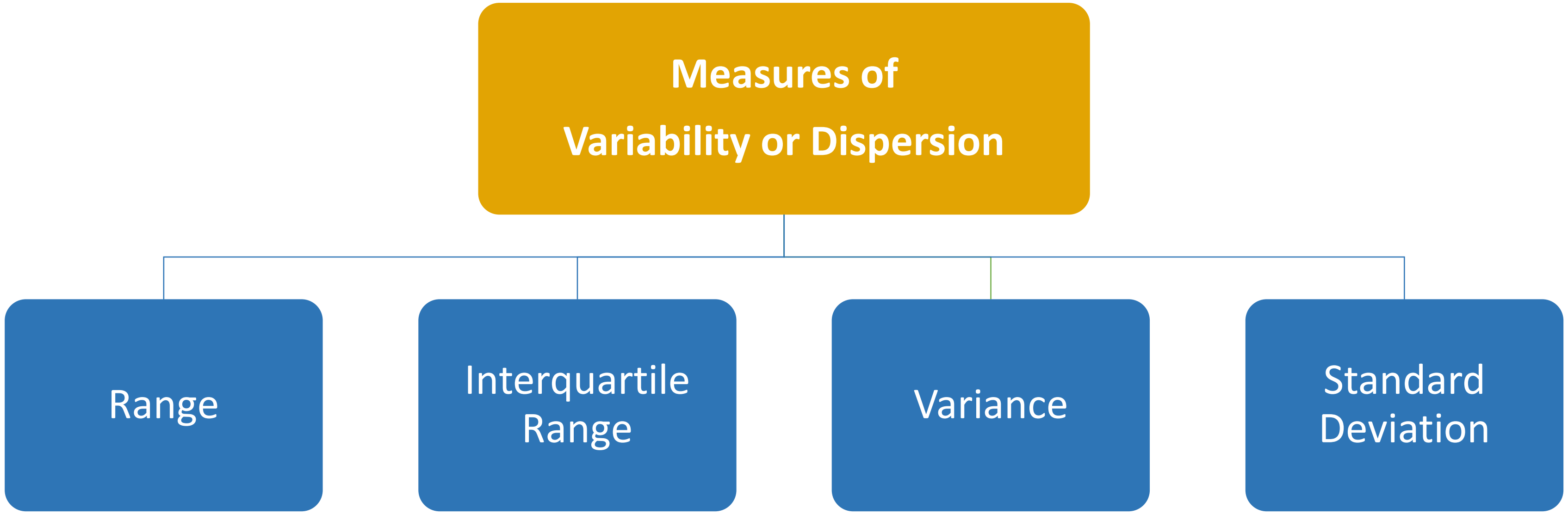
Mode

Most frequently- occurring value



Mode = 16

Measures of
Variability or Dispersion



Measures of
Variability or Dispersion

Range

Interquartile Range

Variance

Standard Deviation

Range : Difference between maximum and minimum values

Interquartile Range : Difference between third and first quartile **(Q_3 - Q_1)**

Sales	Sorted Sales	Rank	
9	6	1	← **Minimum**
6	9	2	
12	10	3	
10	12	4	
13	13	5	← **First Quartile**
15	14	6	
16	14	7	
14	15	8	
14	16	9	
16	16	10	
17	16	11	
16	17	12	
24	17	13	
21	18	14	
22	18	15	← **Third Quartile**
18	19	16	
19	20	17	
18	21	18	
20	22	19	
17	24	20	← **Maximum**

Range Maximum - Minimum =
24 - 6 =
18

Q_1 = 13 + (.25)(1) = 13.25

Q_3 = 18+ (.75)(1) = 18.75

Interquartile Range Q3 - Q1 =
18.75 - 13.25 = 5.5

Variance : Mean* squared deviation from the mean

Standard Deviation : Square root of the variance

* Definitions of population variance and sample variance differ slightly.

Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x-\mu)^2}{N}$$

$$= \frac{\sum_{i=1}^{N} x^2 - \frac{\left(\sum_{i=1}^{N} x\right)^2}{N}}{N}$$

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

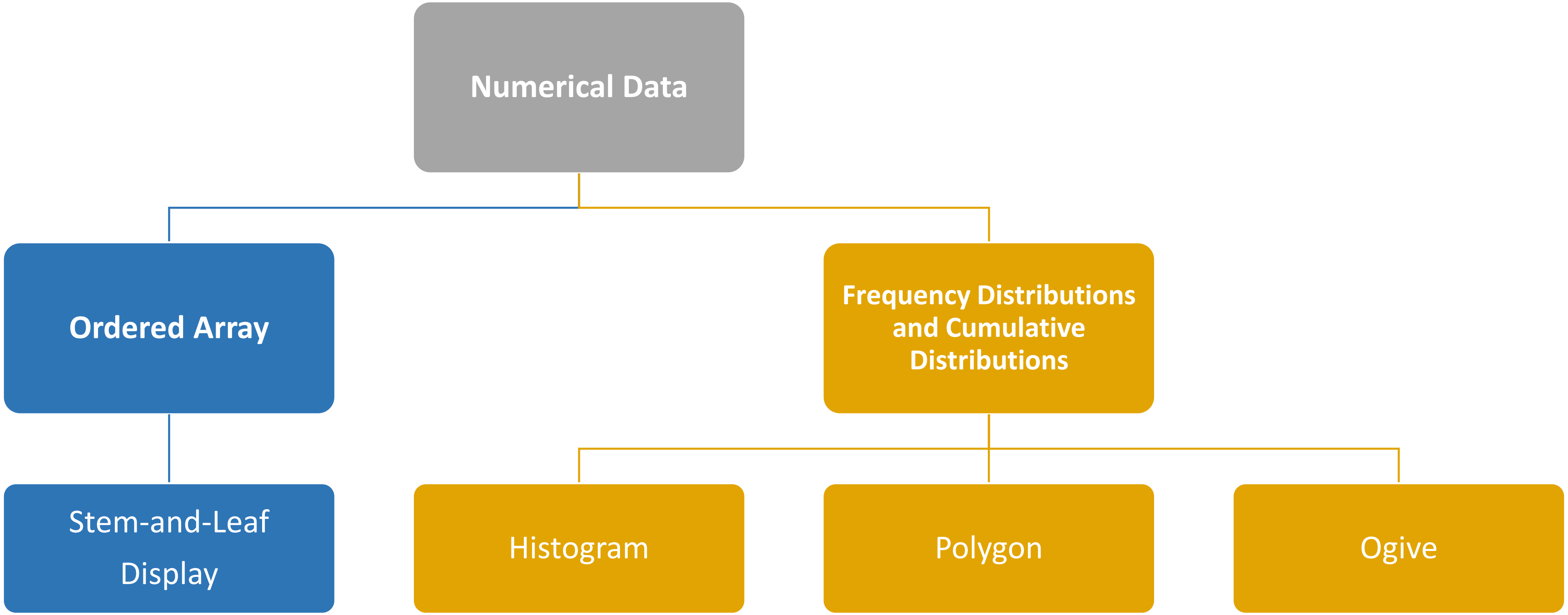
Sample Variance

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

$$= \frac{\sum_{i=1}^{n} x^2 - \frac{\left(\sum_{i=1}^{n} x\right)^2}{n}}{(n-1)}$$

$$s = \sqrt{s^2}$$

Visualizing
Numerical Data



Visualizing Numerical Data

Frequency Distribution Example

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Relative Freq	Percentage
10 but less than 20	3	0.15	15
20 but less than 30	6	0.30	30
30 but less than 40	5	0.25	25
40 but less than 50	4	0.20	20
50 but less than 60	2	0.10	10
Total	20	1.00	100

Visualizing Numerical Data

Cumulative Frequency

Data in ordered array:

12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

Class	Frequency	Percentage	Cumulative Frequency	Cumulative Percentage
10 but less than 20	3	15	3	15
20 but less than 30	6	30	9	45
30 but less than 40	5	25	14	70
40 but less than 50	4	20	18	90
50 but less than 60	2	10	20	100
Total	20	100		

Visualizing Numerical Data

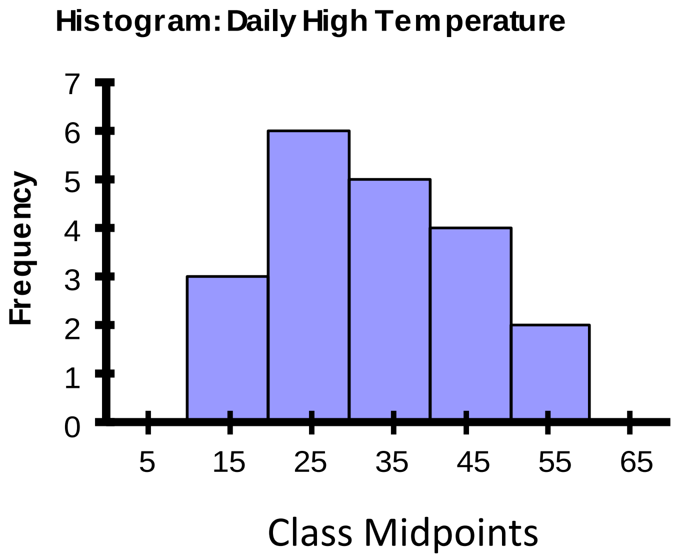
Histogram

Histogram is a chart made of bars of different heights.

- **Widths and locations of bars correspond to widths and locations of data groupings**
- **Heights of bars correspond to frequencies or relative frequencies of data groupings**

Class	Class Midpoint	Frequency
10 but less than 20	15	3
20 but less than 30	25	6
30 but less than 40	35	5
40 but less than 50	45	4
50 but less than 60	55	2

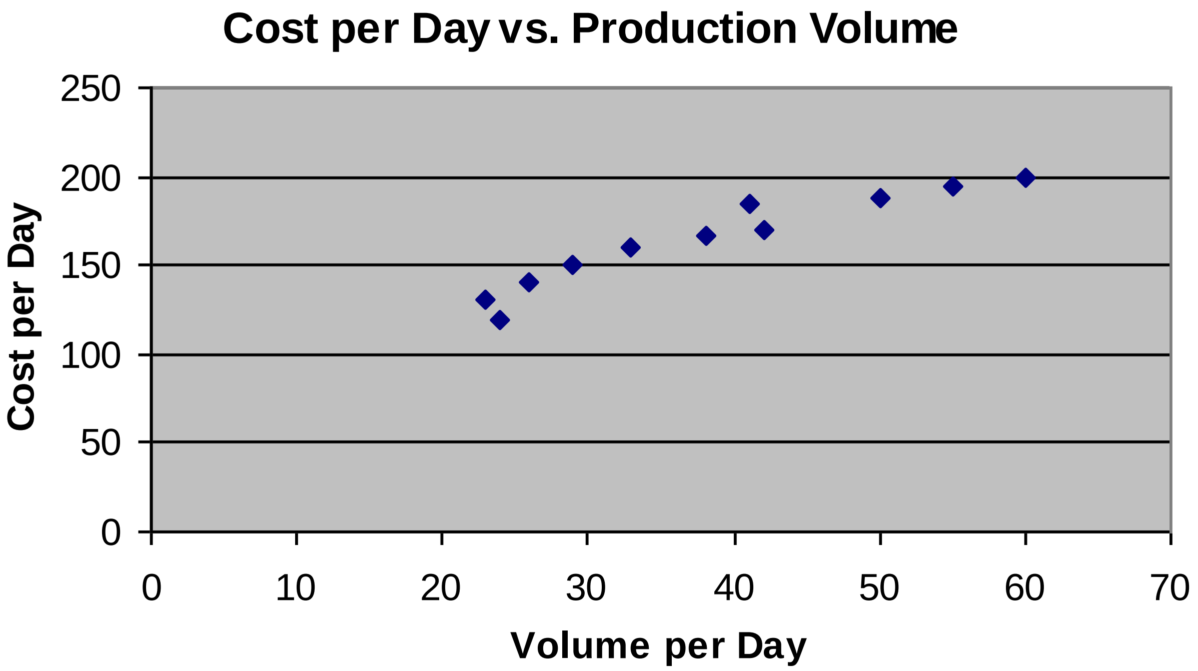
(No gaps between bars)



Visualizing Numerical Data

Scatter Diagram

Volume per day	Cost per day
23	131
24	120
26	140
29	151
33	160
38	167
41	185
42	170
50	188
55	195
60	200



Skewness and Kurtosis

Skewness

Measure of asymmetry of a frequency distribution

- **Skewed to left**
- **Symmetric or unskewed**
- **Skewed to right**

Kurtosis

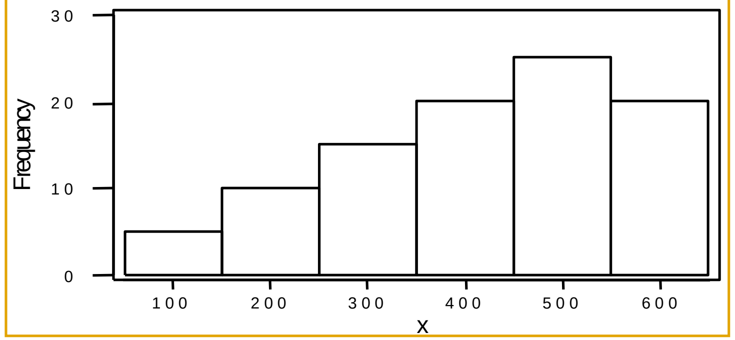
Measure of flatness or peakedness of a frequency distribution

- **Platykurtic (relatively flat)**
- **Mesokurtic (normal)**
- **Leptokurtic (relatively peaked)**

Skewness

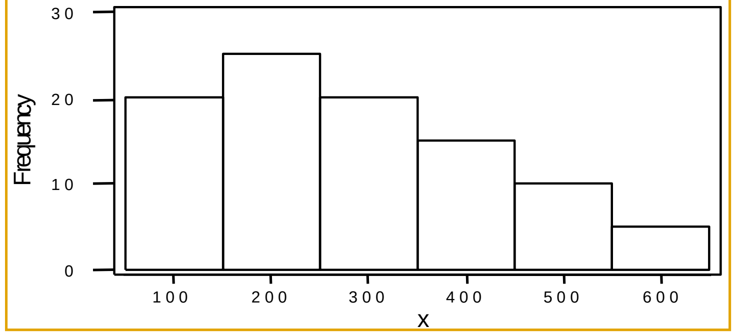
Skewed to left

Mean < median < mode



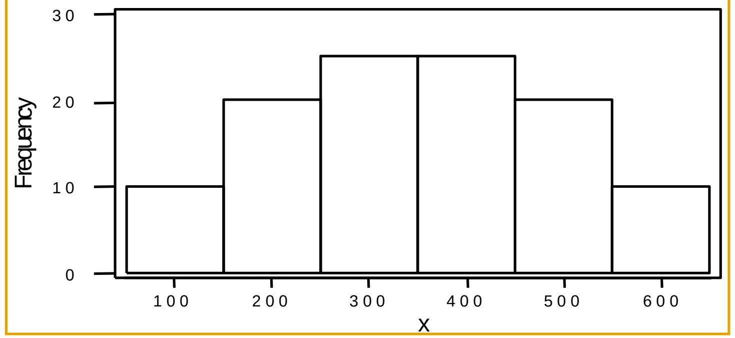
Skewed to right

Mode > median > mean



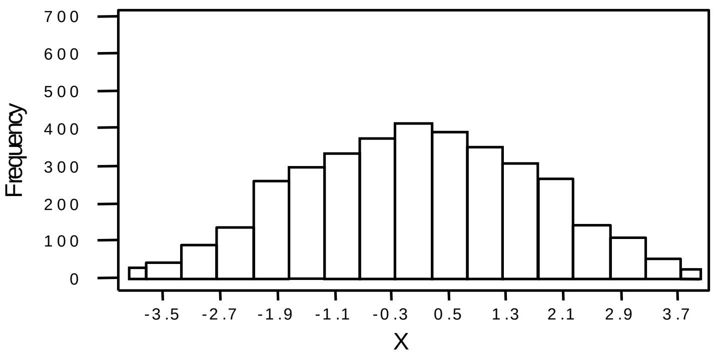
Symmetric

Mean = median = mode

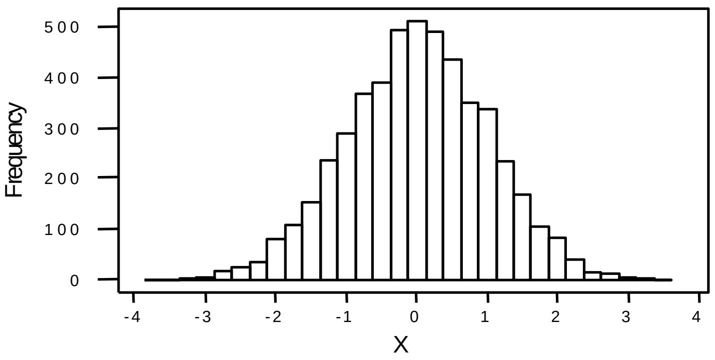


Kurtosis

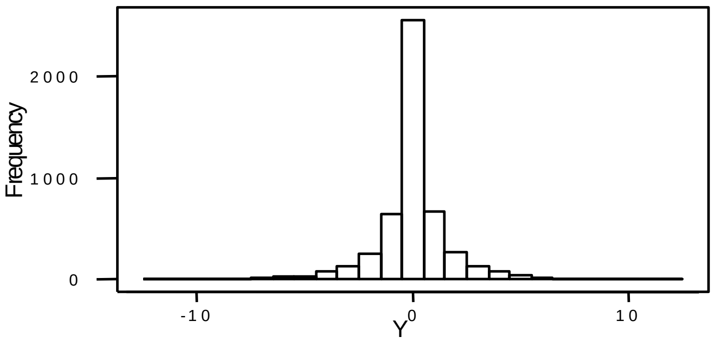
Platykurtic - flat distribution



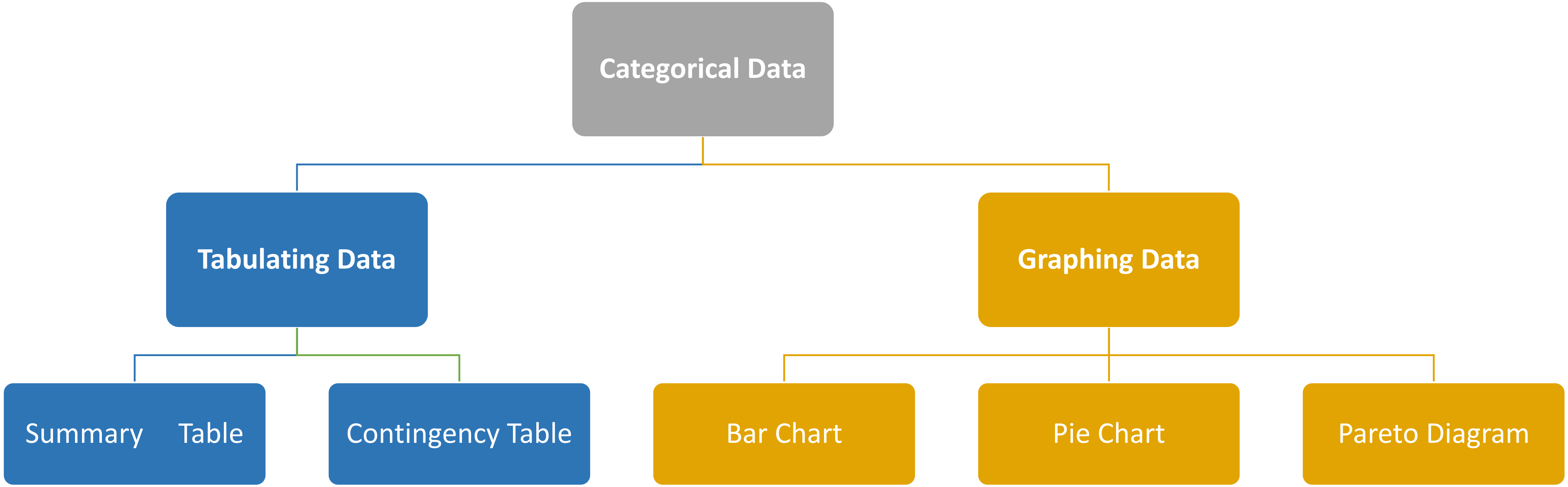
Mesokurtic - not too flat and not too peaked



Leptokurtic - peaked distribution



Organizing
Categorical Data



Organizing Categorical Data:
Summary Table

- The Summary Table is a visualization that summarizes statistical information about data in table form. The Summary Table automatically updates the values displayed to reflect the current selection.
- All visualizations can be set up to show data limited by one or more markings in other visualizations only (details visualizations). Summary tables can also be limited by one or more filtering. Another alternative is to set up a summary table without any filtering at all.

Summarize data by category

Example: Current Investment Portfolio

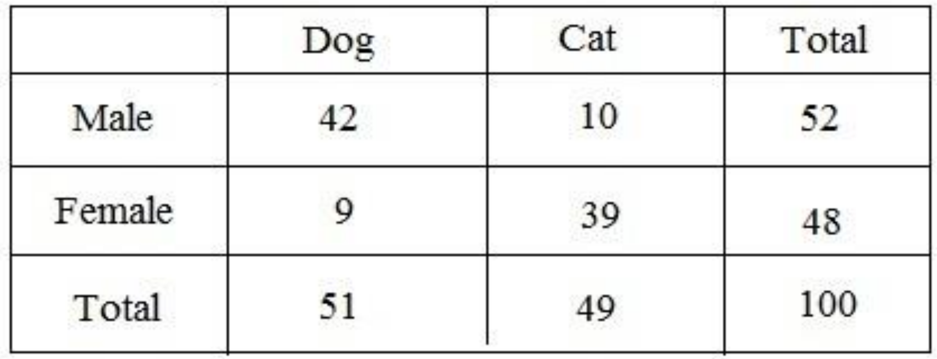
Investment Type	Amount (in thousands $)	Percentage (%)
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	110.0	100

(Variables are Categorical)

Organizing Categorical Data:
The Contingency Table

- In statistics, a contingency table is a type of table in a matrix format that displays the frequency distribution of the variables. They are heavily used in survey research, business intelligence, engineering and scientific research.
- A contingency table, sometimes called a two-way frequency table, is a tabular mechanism with at least two rows and two columns used in statistics to present categorical data in terms of frequency counts. More precisely, an contingency table shows the observed frequency of two variables, the observed frequencies of which are arranged into rows and columns. The intersection of a row and a column of a contingency table is called a cell.

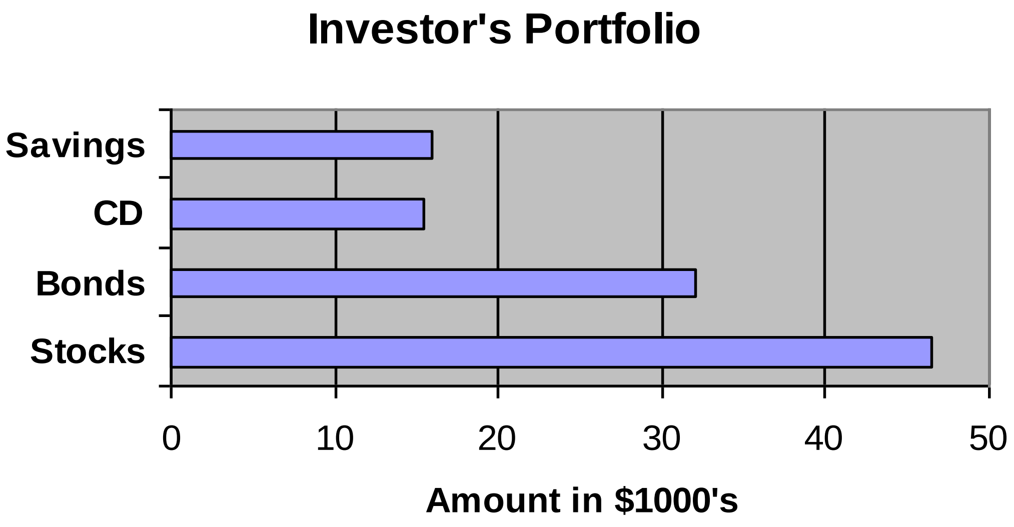
- Useful in situations involving multiple population proportions
- Used to classify sample observations according to two or more characteristics
- Also called a cross-classification table.



	Dog	Cat	Total
Male	42	10	52
Female	9	39	48
Total	51	49	100

Organizing Categorical Data:

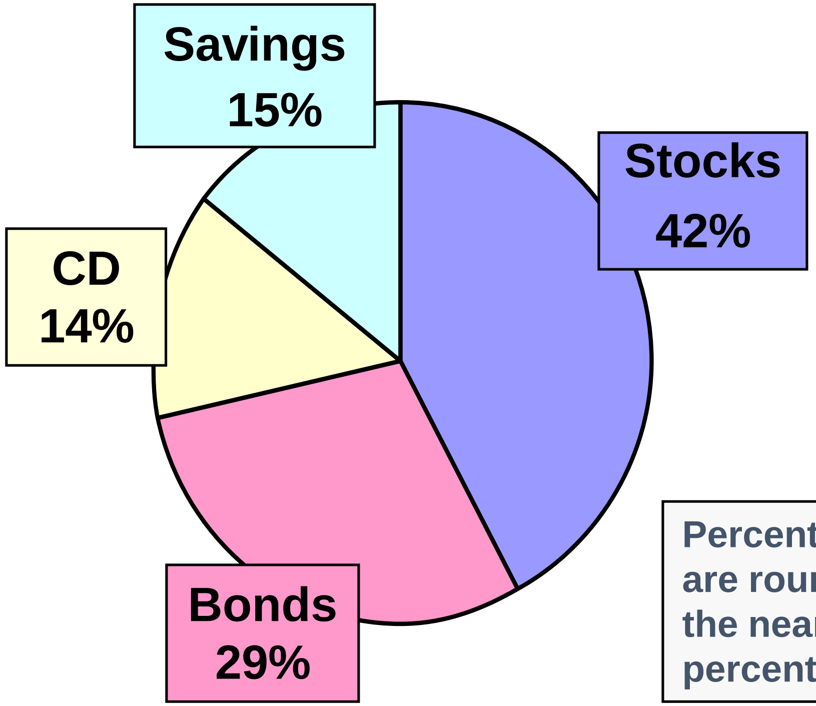
Bar Chart Example



Current Investment Portfolio

Investment Type	Amount (in thousands $)	Percentage (%)
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	110.0	100

Organizing Categorical Data:
Pie Chart Example



Percentages are rounded to the nearest percent

Current Investment Portfolio

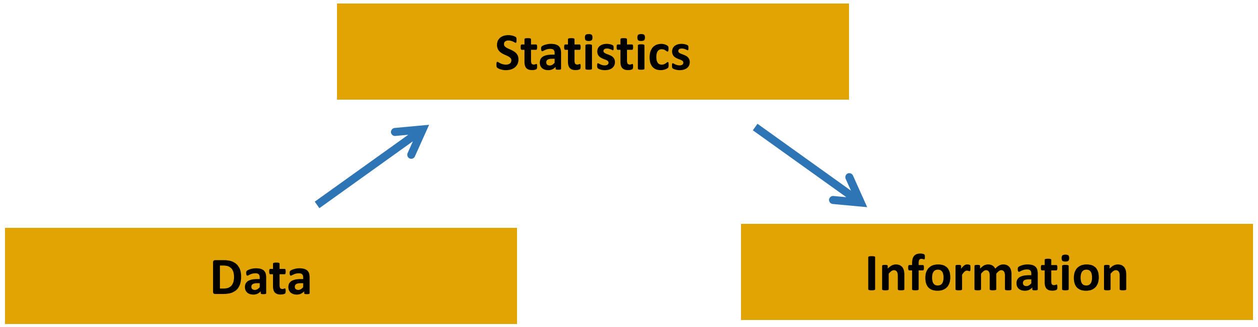
Investment Type	Amount (in thousands $)	Percentage (%)
Stocks	46.5	42.27
Bonds	32.0	29.09
CD	15.5	14.09
Savings	16.0	14.55
Total	110.0	100

Agenda

- Introduction to Statistics
- Type Of Statistics
- Organizing Numerical and Categorical Data
- **Data Collection : Sampling Technique**
- Sampling error and non sampling error

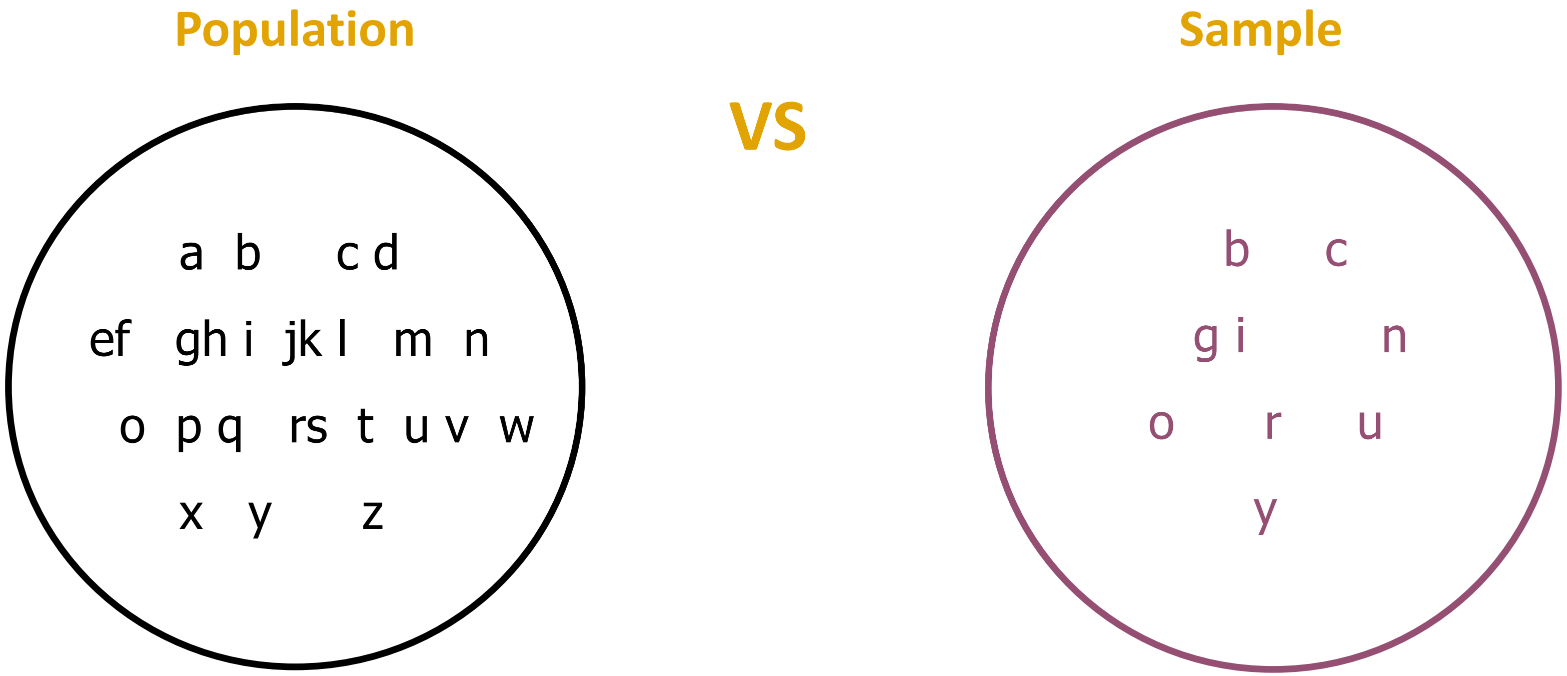
Data Collection and Sampling

Statistics is a tool for converting *data* into *information*



- But where then does data come from? How is it gathered? How do we ensure its accurate? Is the data reliable? Is it representative of the population from which it was drawn? This chapter explores some of these issues.
- There are many methods used to collect or obtain data for statistical analysis. Three of the most popular methods are:
- Direct Observation, Experiments and Surveys.

Data Collection and Sampling



Population

a b c d

ef gh i jk l m n

o p q rs t u v w

x y z

VS

Sample

b c

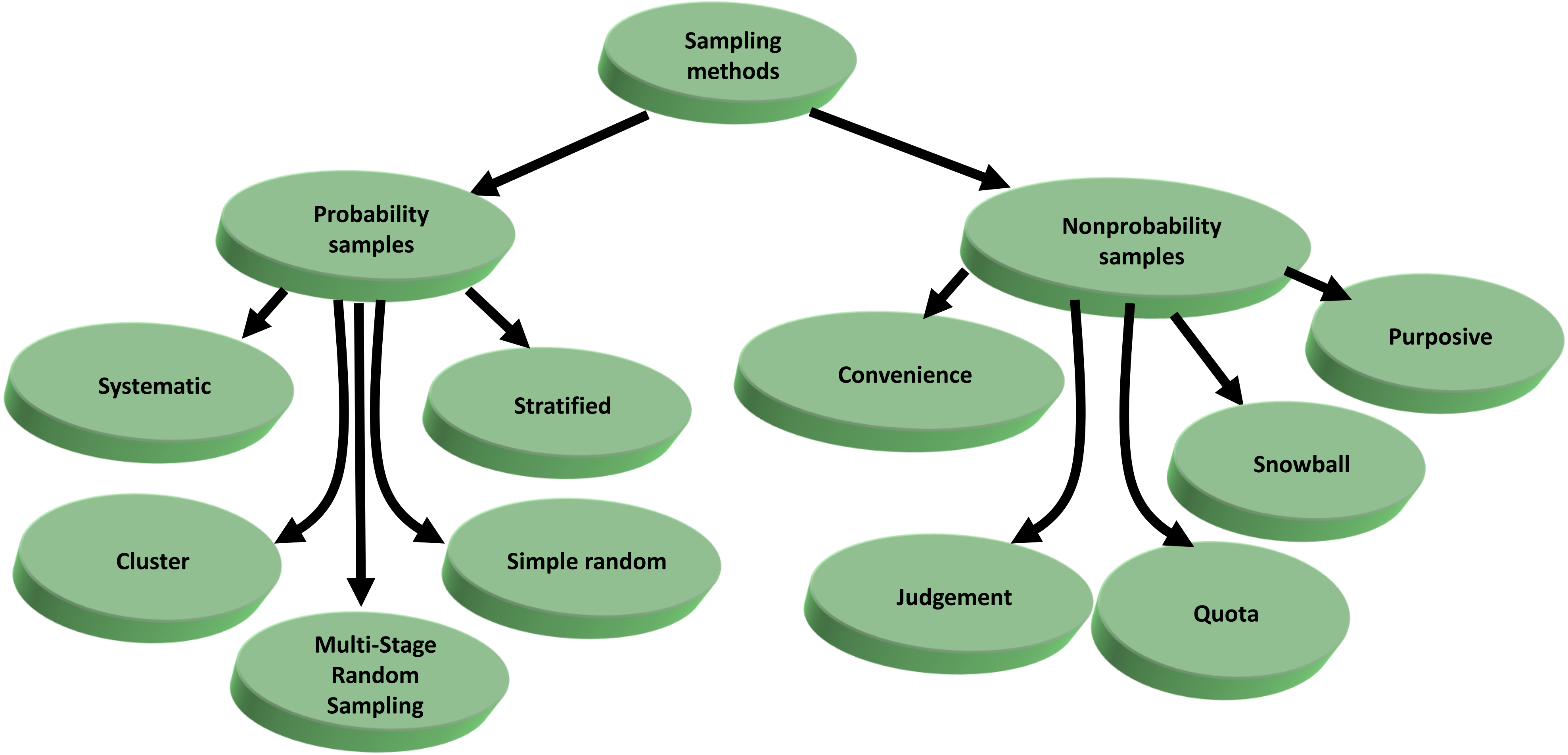
g i n

o r u

y

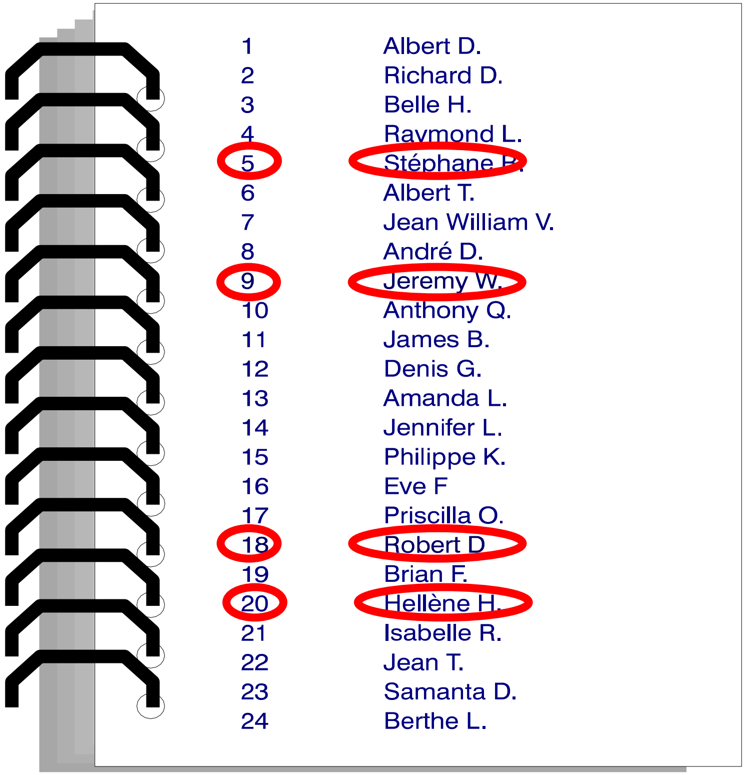
Data Collection and Sampling

Classification of Sampling Technique



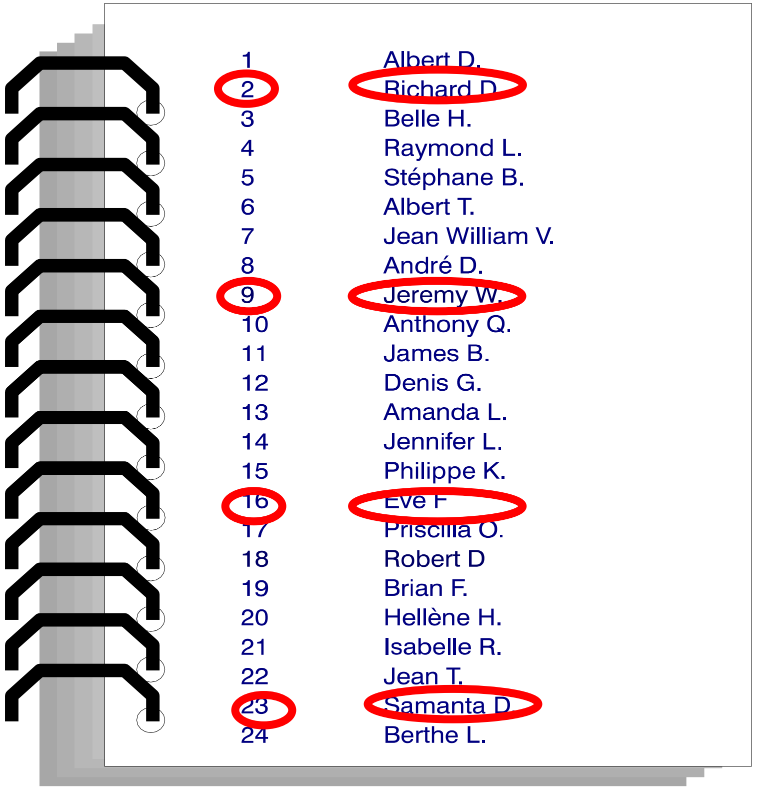
Probability
Samples

Simple Random Sampling



1	Albert D.
2	Richard D.
3	Belle H.
4	Raymond L.
[✓ 5]	[✓ Stéphane B.]
6	Albert T.
7	Jean William V.
8	André D.
[✓ 9]	[✓ Jeremy W.]
10	Anthony Q.
11	James B.
12	Denis G.
13	Amanda L.
14	Jennifer L.
15	Philippe K.
16	Eve F
17	Priscilla O.
[✓ 18]	[✓ Robert D]
19	Brian F.
[✓ 20]	[✓ Hellène H.]
21	Isabelle R.
22	Jean T.
23	Samanta D.
24	Berthe L.

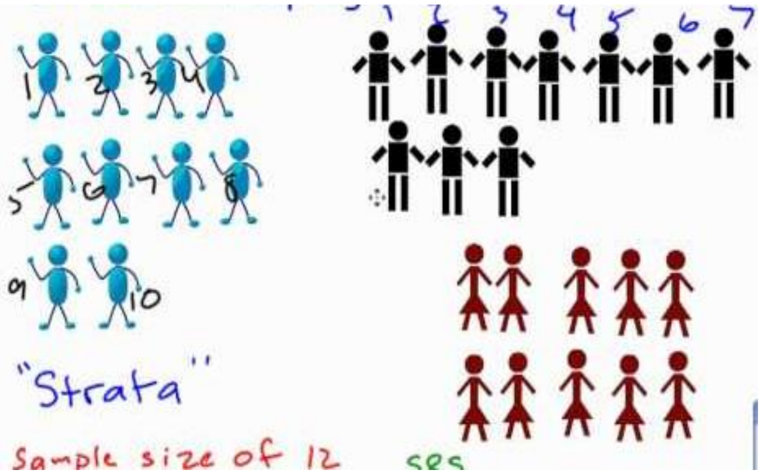
Systematic Sampling



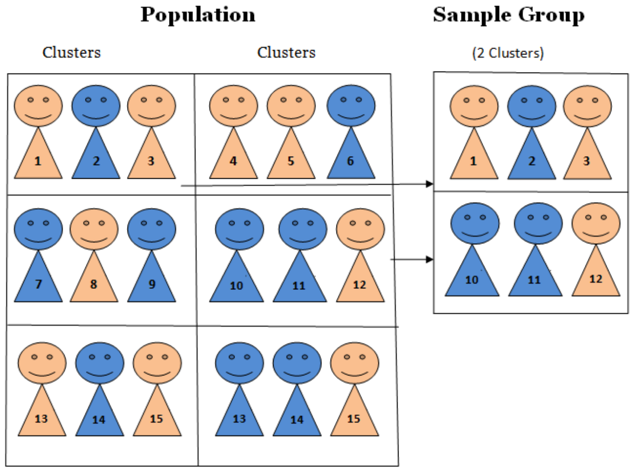
1	Albert D.
[✓ 2]	[✓ Richard D.]
3	Belle H.
4	Raymond L.
5	Stéphane B.
6	Albert T.
7	Jean William V.
8	André D.
[✓ 9]	[✓ Jeremy W.]
10	Anthony Q.
11	James B.
12	Denis G.
13	Amanda L.
14	Jennifer L.
15	Philippe K.
[✓ 16]	[✓ Eve F]
17	Priscilla O.
18	Robert D
19	Brian F.
20	Hellène H.
21	Isabelle R.
22	Jean T.
[✓ 23]	[✓ Samanta D.]
24	Berthe L.

Probability
Samples

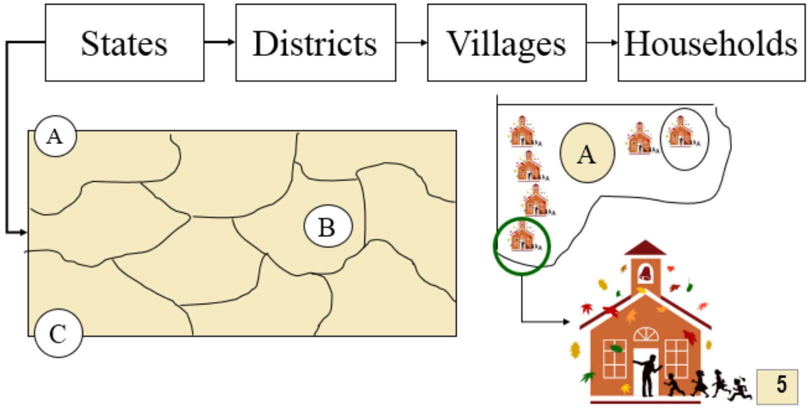
Stratified Sampling



Cluster Sampling

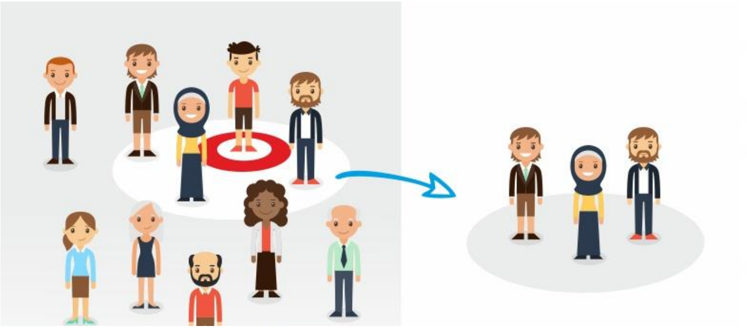


Multistage Sampling

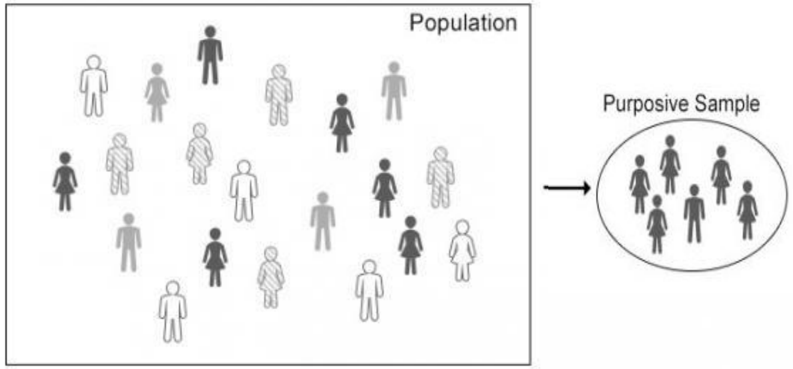

States
Districts
Villages
Households

Non-Probability Samples

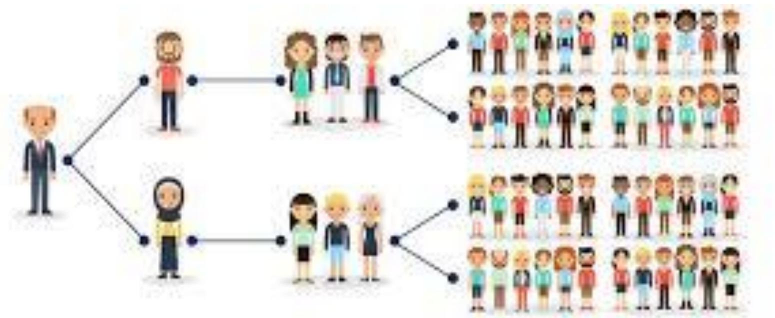
Convenience Sampling



Purposive Sampling

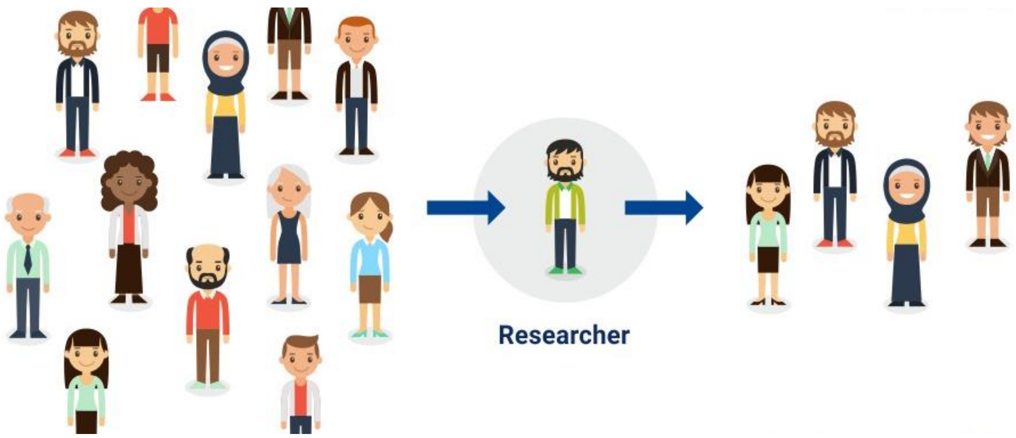


Snowball Sampling

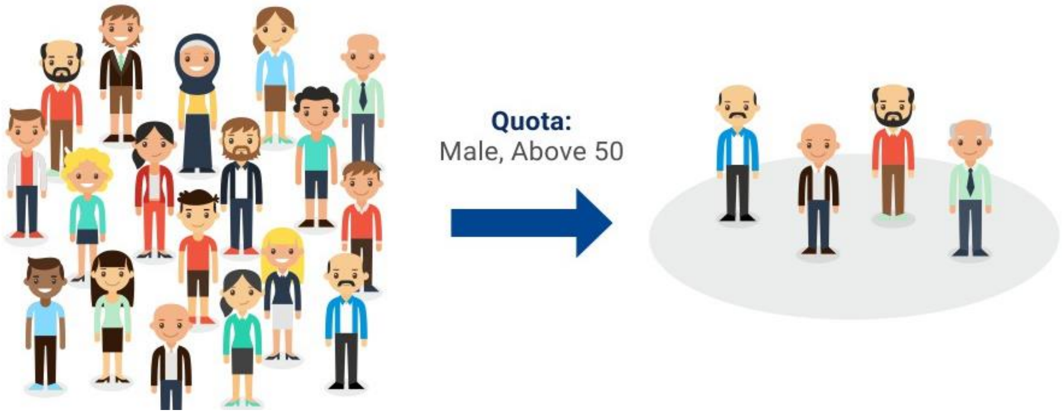


Non-Probability Samples

Judgemental Sampling



Quota Sampling



Agenda

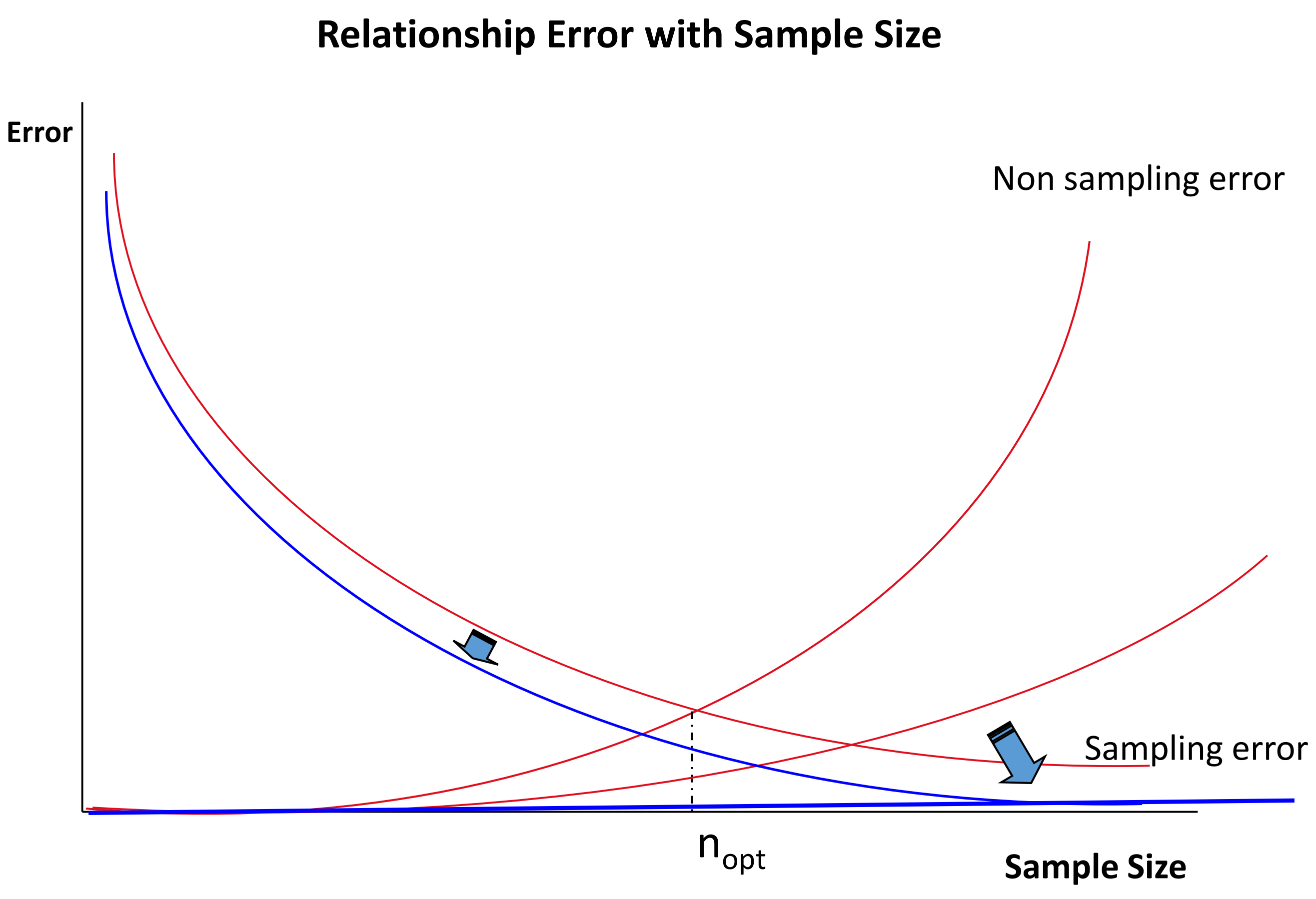
- Introduction to Statistics
- Type Of Statistics
- Organizing Numerical and Categorical Data
- Data Collection : Sampling Technique
- **Sampling error and non sampling error**

Sampling Error and Non-Sampling Error

- Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.
  - Noted : Increasing the sample size will reduce this type of error.

- Non-sampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly.
  - Three types of non-sampling errors:
    - Errors in data acquisition,
    - Nonresponse errors
    - Selection bias.
  - Note: increasing the sample size will not reduce this type of error.

Sampling Error and Non-Sampling Error


Relationship Error with Sample Size
Error
Non sampling error
Sampling error
n_{opt}
Sample Size