Biostatistics

Kabtamu Tolosie

Email: kabtamuto@gmail.com

Arsi University, College of Health Science,

Department of Public Health

Asella, Ethiopia

August, 2025

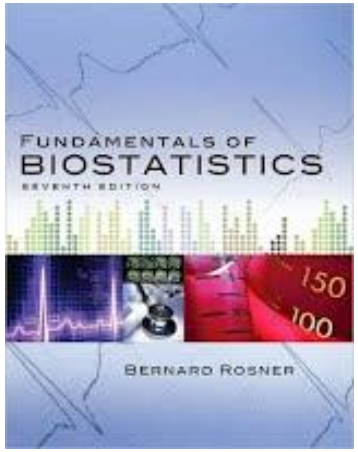
Everything You Wanted to Know about Biostatistics but Were Afraid to Ask

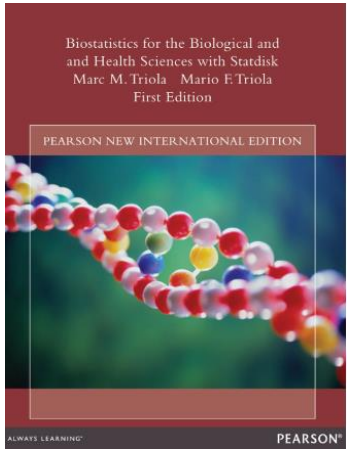
Course Outline

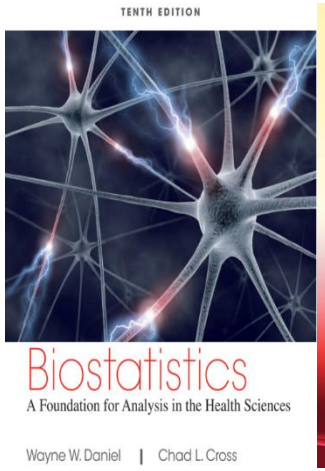
- Introduction to Biostatistics
- Descriptive Biostatistics
- **Demography and vital statistics**
- Probability
- Discrete Probability Distributions
- Continuous Probability Distributions
- Sampling Theory
- Estimation
- Hypothesis Testing

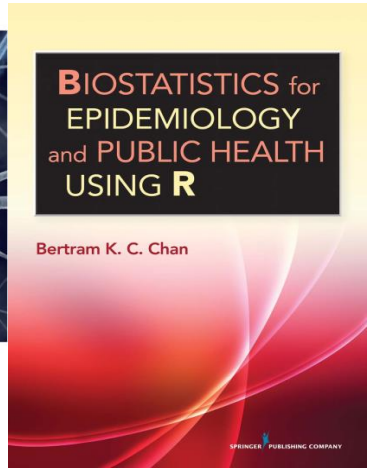
Reference Textbooks

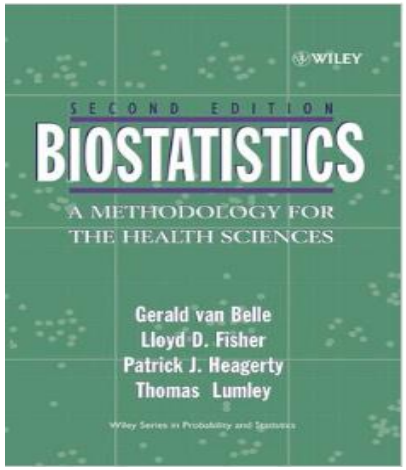
1. Rosner, B., 2015. *Fundamentals of biostatistics*. Cengage learning.
2. Triola, M.M. and Triola, M.F., 2013. *Biostatistics for the Biological and Health Sciences with Statdisk: Pearson New International Edition*. Pearson Higher Ed.
3. Daniel, W.W. and Cross, C.L., 2018. *Biostatistics: a foundation for analysis in the health sciences*. Wiley.
4. Chan, B.K., 2015. *Biostatistics for epidemiology and public health using R*. Springer Publishing Company.
5. Van Belle, G., Fisher, L.D., Heagerty, P.J. and Lumley, T., 2004. *Biostatistics: a methodology for the health sciences*. John Wiley & Sons.











Lesson 1: Introduction to Biostatistics

Lecture Outline

Introduction

Some basic concepts

Measurement and measurement scales

Introduction

STATISTICS - HISTORICAL PERSPECTIVES

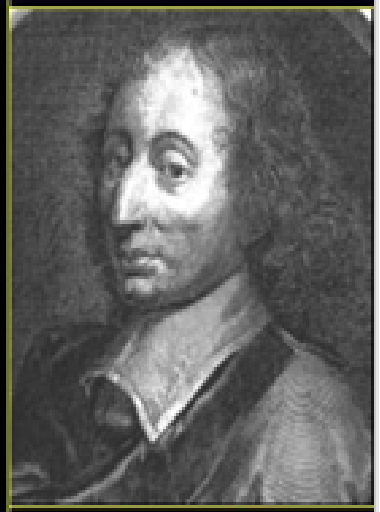
- Statistics seems to be derived from Latin word '*Status*' or Italian word '*Statista*' or German word '*Statistik*' or French word '*Statistique*' which all meaning '*political state*'.

- In ancient time the king used to collect information about total population, land, wealth, soldiers of the country and thus statistics served as an index of a country's overall condition. In olden days, statistics was regarded as 'the science of kings'.

Introduction

- In mid 17th century, the theoretical development in modern statistics came with the introduction of 'Theory of Probability' and 'Theory of Games and Chances'.

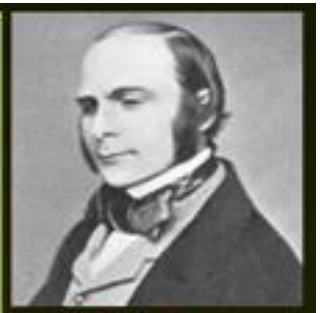
- Gambling, in the form of games of chance, led to this theory of probability being originated by the French mathematician Pascal (1623-1662).



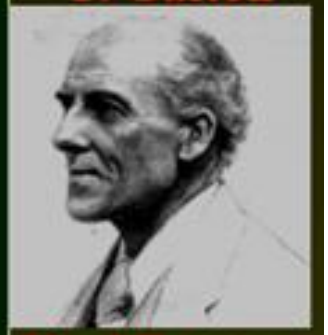
Pascal

Introduction

- Francis Galton (1822-1921) introduced the concept of regression line.
- Galton and his friend Karl Pearson later introduced correlation analysis and chai-square test which play an important role in modern theory of statistics.
- W.S. Gosset (1876-1937), student of Karl Pearson, introduced 'Student *t*-test' is the basic tool of statistical analysis.



F. Galton



Karl Pearson

W.S. Gosset

Introduction

- Sir R.A. Fisher (1890-1962), known as the father of statistics, introduced a number of statistical procedures such as Analysis of Variance (ANOVA) and design of experiments and so on.

R.A. Fisher

Definition of Statistics

Croxton and Cowden have given a very simple and concise definition of statistics.

Statistics is the science of

- collection
- organization
- presentation
- analysis and
- interpretation of data

Definition of Biostatistics

- BIOSTATISTICS is derived from Greek word Bios (Life) & Metron (Measure).
- Thus biostatistics is the term used when tools of statistics are applied to the data that derived from biological and medical science.

Biostatistics is the science of

- Collection
- Organization
- Presentation
- analysis and
- interpretation of data that is derived from biological sciences such as medicine.

Biostatistics

Biostatistics can be defined as the application of the mathematical tools used in statistics to the fields of biological sciences and medicine.

or

Biostatistics is the study of statistics as applied to biological areas. Biological laboratory experiments, medical research (including clinical research), and health services research all use statistical methods.

or

Cont'd

- The tools of statistics are employed in many fields—business, education, psychology, agriculture, and economics, to mention only a few. When the data analyzed are derived from the biological sciences and medicine, we use the term ***biostatistics*** to distinguish this particular application of statistical tools and concepts.

Uses of Biostatistics

- Hospital utility statistics, Resource allocation, Vaccination uptake, Magnitudes of a disease/condition, Assessing risk factors, Disease frequency, Making diagnosis and choosing an appropriate treatment (implicit/probability).

Reasons to know about biostatistics

- Medicine is becoming increasingly quantitative.
- The planning, conduct and interpretation of much of medical research are becoming increasingly reliant on the statistical methodology.
- Statistics pervades the medical literature.

CLINICAL MEDICINE

- Documentation of medical history of diseases.
- Planning and conduct of clinical studies.
- Evaluating the merits of different procedures.
- In providing methods for definition of “normal” and “abnormal”.

Role of Biostatistics in patient care

- In increasing awareness regarding diagnostic, therapeutic and prognostic uncertainties and providing rules of probability to delineate those uncertainties
- In providing methods to integrate chances with value judgments that could be most beneficial to patient
- In providing methods such as sensitivity-specificity and predictivities that help choose valid tests for patient assessment
- In providing tools such as scoring system and expert system that can help reduce epistemic uncertainties

PREVENTIVE MEDICINE

- To provide the magnitude of any health problem in the community.
- To find out the basic factors underlying the ill-health.
- To evaluate the health programs which was introduced in the community (success/failure).
- To introduce and promote health legislation.

Role of Biostatics in Health Planning and Evaluation

- In carrying out a valid and reliable health situation analysis, including in proper summarization and interpretation of data.

- In proper evaluation of the achievements and failures of a health programme

Role of Biostatistics in Medical Research

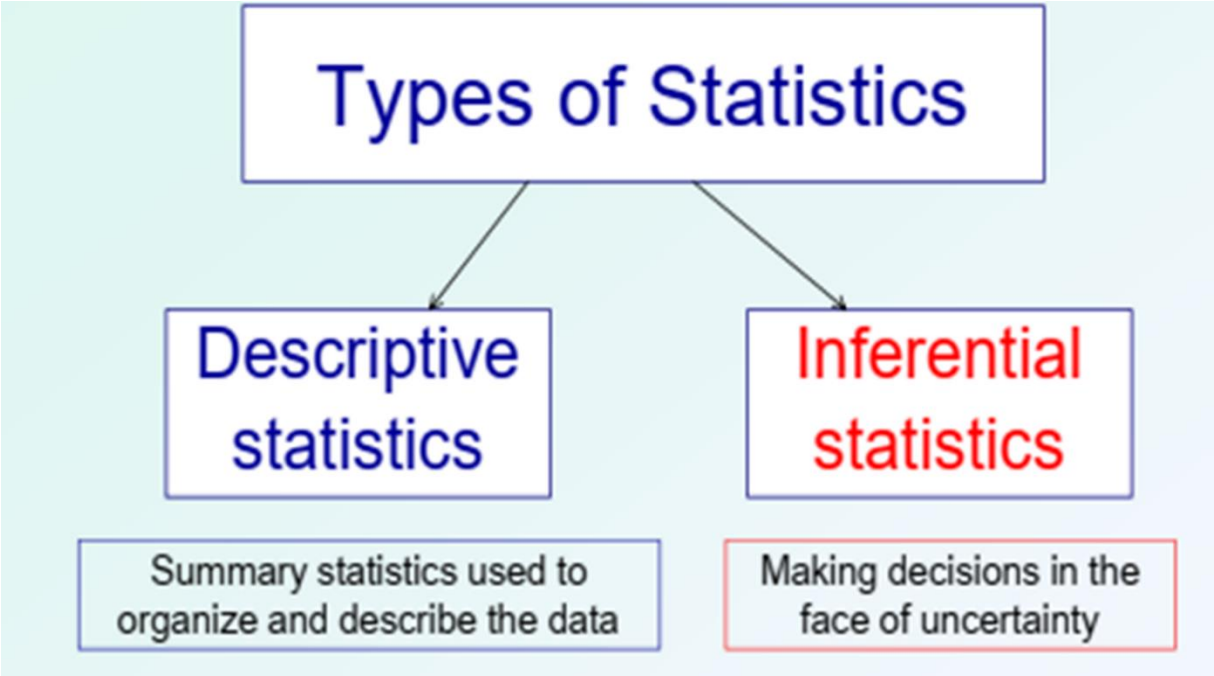
- In developing a research design that can minimize the impact of uncertainties
- In assessing reliability and validity of tools and instruments to collect the information
- In proper analysis of data

Concerns of Biostatistics

Biostatistics is a field of study concerned with;

1) The collection, organization, summarization, and analysis of data. (**Descriptive Statistics**)

2) The drawing of inferences (decision) about a body of data when only a part of the data is observed. (**Inferential Statistics**)



Cont'd

- Descriptive statistics
- Which summary statistics to use to organize and describe the data?
  - Proportion, mean, median, SD, percentiles
- Descriptive statistics do not generalize beyond the available data
- Inferential statistics
- Generalize from the sample.
  - Hypothesis testing, confidence intervals
    - t-test, Fisher's Exact, ANOVA, survival analysis
    - Bayesian approaches
- Making decisions in the face of uncertainty

General steps in a research process

What does Biostatistics cover?	Planning
	Design
	Data collection
	Data Processing
	Data Presentation
	Data Analysis
	Interpretation
	Publication

WHAT ARE THE ISSUES?

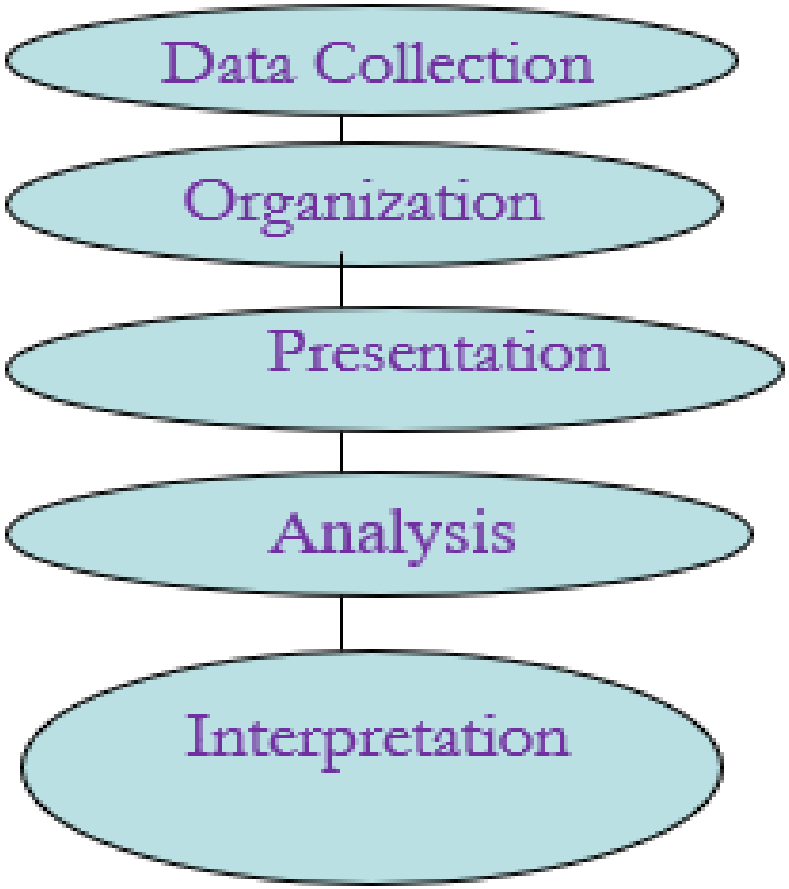
- Appropriately conducting and interpreting biostatistical applications require attention to a number of important issues.

1. Clearly defining the objective or research question
2. Choosing an appropriate study design (i.e., the way in which data are collected)
3. Selecting a representative sample, and ensuring that the sample is of sufficient size
4. Carefully collecting and analyzing the data

Cont'd

5. Producing appropriate summary measures or statistics
6. Generating appropriate measures of effect or association
7. Quantifying uncertainty
8. Appropriately accounting for relationships among characteristics
9. Limiting inferences to the appropriate population

Stages in Statistical Investigation



Some Basic Concepts

Population

- The average person thinks of a population as a collection of entities, usually people.

- A population or collection of entities may, however, consist of animals, machines, places, or cells.

- A **population** of values as the largest collection of values of a random variable for which we have an interest at a particular time/situation/study.
- Populations may be finite or infinite.

Cont'd

If a population of values consists of a fixed number of these values, the population is said to be **finite**.

If, on the other hand, a population consists of an endless succession of values, the population is an **infinite** one.

Example:

we are interested in the weights of all the children enrolled in a certain county elementary school system, our population consists of all these weights.

Target population: A collection of items that have something in common for which we wish to draw conclusions at a particular time.

Study Population: The specific population from which data are collected.

Sample

A ***sample*** may be defined simply as a part of a population.

Example:

- Suppose our population consists of the weights of all the elementary school children enrolled in a certain county school system. If we collect for analysis the weights of only a fraction of these children, we have only a part of our population of weights, that is, we have a *sample*.

Cont'd

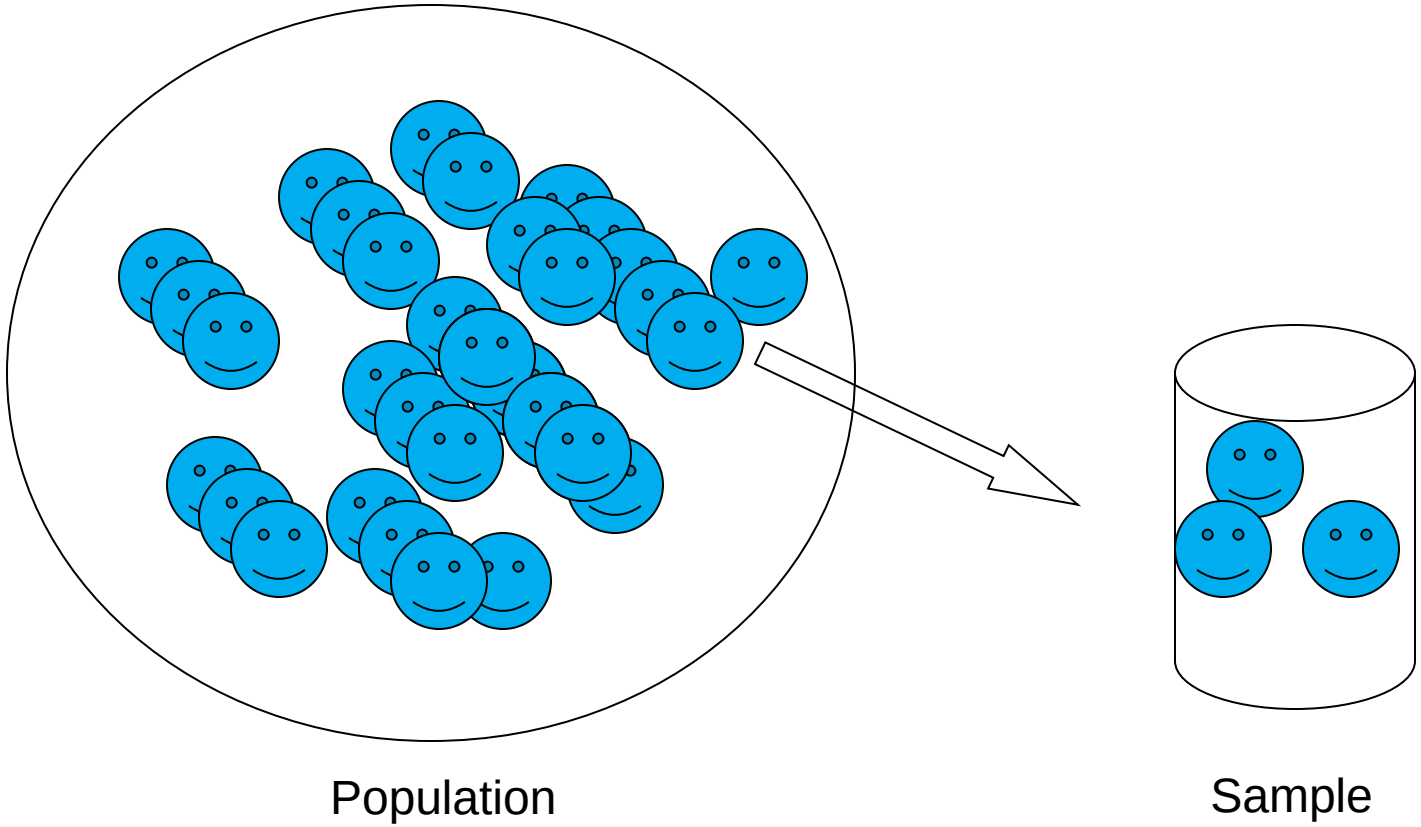
Generalizability: is a two-stage procedure: we want to generalize conclusions from the sample to the study population and then from the study population to the target population.

Example: In a study of the prevalence of Kat chewing among secondary students in Ethiopia a random sample of Secondary students in Asella were taken.

- **Target Population: All secondary students in Ethiopia**
- **Study population: All secondary students in Asella**
- **Sample: secondary students in Asella**

- **Sampling:** the process of taking sample from a population

Sample vs. Population



Cont'd

For each of the following situations, answer questions a through d:

(a) What is the population?

(b) What is the sample in the study?

(c) What is the variable of interest?

(d) What is the type of the variable?

Situation A: A study of 300 households in a small southern town revealed that if she has school-age child present.

Cont'd

(a) Population: All households in a small southern town.

(b) Sample: 300 households in a small southern town

(c) Variable: Does households had school age child present.

(d) Variable is qualitative nominal.

Cont'd

Situation B: A study of 250 patients admitted to a hospital during the past year revealed that, on the average, the patients lived 15 miles from the hospital.

(a) Population: All patients admitted to a hospital during the past year.

(b) Sample: 250 patients admitted to a hospital during the past year.

(c) Variable: Distance the hospital live away from the hospital

(d) **Variable is** Quantitative continuous.

Parameter and Statistic

Parameter: A descriptive measure computed from the data of a population. (Quantity calculated from population).

Statistic: A descriptive measure computed from the data of a sample. (Quantity calculated from the sample).

Data

Data are numbers which can be obtained from *measurements* or by *counting*.

Example (measurements):

A nurse weighs a patient or takes a patient's temperature, a measurement, consisting of a number such as 150 pounds or 100 degrees or Fahrenheit, is obtained.

Example (counting):

A hospital administrator counts the number of patients—perhaps 20—discharged from the hospital on a given day.

Each of the above three numbers is a *datum*, and the three taken together are *data*.

Sources of Data

Data are usually available from one or more of the following sources:

1. **Routinely Kept Records**
2. **Surveys**
3. **Experiments**
4. **Reports**

Sources of Data (continue)

1. **Routinely Kept Records**

- It is difficult to imagine any type of organization that does not keep records of day-to-day transactions of its activities.

Example:

- Hospital medical records, for example, Hospital medical records contain immense amounts of information on patients, while hospital accounting records contain a wealth of data on the facility's business activities.

Cont'd

2. **Surveys**

- If the data needed to answer a question are not available from routinely kept records, the logical source may be a survey

Example:

- The administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic. If admission forms do not contain a question on mode of transportation, we may conduct a survey among patients to obtain this information.

Sources of Data (continue)

3. **Experiments**

- Frequently the data needed to answer a question are available only as the result of an experiment. A

Example:

- A nurse may wish to know which of several strategies is best for maximizing patient compliance (medicine). The nurse might conduct an experiment in which the different strategies of motivating compliance are tried with different patients. Subsequent evaluation of the responses to the different strategies might enable the nurse to decide which is most effective.

Cont'd

4. **Reports**

Example:

- The data needed may already exist in the form of published reports, commercially available data banks, or the research literature.
- In other words, we may find that someone else has already asked the same question, and the answer obtained may be applicable to our present situation.

Types of Data

There are **Two** types of data (or numbers): **constants** and **variables**.

1. **Constants**

- A ***constant***, as its name suggests, is something that does not vary or change (or that may not be susceptible to variation or change).
- A constant has only one attribute or value.
- Any variable can be made into a constant by reducing its expression to only one of its values.

Cont'd

A constant has no use in statistics. That is, anything that is or remains constant cannot be subjected to statistical analysis.

Example:

we could keep the temperature of a room constant at 35° C during an experiment. In this case, temperature stops being a variable.

Types of Data (continue)

2. **Variables**

- If, as we observe a characteristic, we find that it takes on different values in different persons, places, or things, we label the characteristic a ***variable***.

- A ***variable*** is a quantity that may vary from object to object.

Example:

- Blood pressure, heart rate, the heights of adult males, the weights of preschool children, and the ages of patients seen in a dental clinic.

Types of Variable

There are **Two** types of a variable.

1) **Quantitative Variable**
2) **Qualitative Variable**

1. **Quantitative Variable**

- A ***quantitative variable*** is one that can be measured in the usual sense.
- A ***quantitative variable*** has values that are intrinsically numerical.
- Measurements made on quantitative variables convey information regarding amount/ quantity.

Cont'd

Example: We can obtain measurements on;

- The heights of adult males
- The weights of pre-school children
- The ages of patients seen in a clinic
- The number of children in a family

2. **Qualitative Variable**

- A ***qualitative variable*** has values that are intrinsically non-numerical (categorical).
- Measurements made on qualitative variables convey information regarding attribute

Types of Variable (continue)

Example: Qualitative (or categorical) variables and their values are:

- Color of a person's hair (black, gray, red, . . ., brown)
- Gender of child (male, female)
- Cause of death of newborn (congenital malformation, asphyxia, . . .)
- Province of residence of a Ethiopia (…..)

Random Variable

- Whenever we determine the height, weight, or age of an individual, the result is frequently referred to as a value of the respective variable.
- When the values obtained arise as a result of chance factors, so that they cannot be exactly predicted in advance, the variable is called a random variable.
- Observations or measurements are used to obtain the value of a random variable.

Example: Adult Height.

- When a child is born, we cannot predict exactly his or her height at maturity.
- Attained adult height is the result of numerous genetic and environmental factors.

Types of Random Variable

A ***Random variable*** may be **discrete** or **continuous**.

1. **Discrete Random Variable**

- A discrete random variable is characterized by gaps or interruptions in the values that it can assume.

- These gaps or interruptions indicate the absence of values between particular values that the variable can assume.

Example:

- The number of daily admissions to a general hospital is a discrete random variable since the number of admissions each day must be represented by a whole number, such as 0, 1, 2, or 3. The number of admissions on a given day cannot be a number such as 1.5, 2.997, or 3.333.

Types of Random Variable (continue)

2. Continuous Random Variable

- A continuous random variable does not possess the gaps or interruptions characteristic of a discrete random variable.
- A continuous random variable can assume any value within a specified relevant interval of values assumed by the variable.

Example:

- The various measurements that can be made on individuals such as height, weight, and skull circumference. No matter how close together the observed heights of two people, for example, we can, theoretically, find another person whose height falls somewhere in between.

Measurement and Measurement Scales

Measurement may be defined as the assignment of numbers to objects or events according to a set of rules.

Measuring scales are different according to the degree of precision involved.

Measurement Scales

There are **Four** *Measurement Scales* result from the fact that measurement may be carried out under different sets of rules.

1. **The Nominal Scale**
2. **The Ordinal Scale**
3. **The Interval Scale**
4. **The Ratio Scale**

Measurement Scales (continue)

1. The Nominal Scale

- The lowest scale is the ***nominal scale***.
  - Categorical
  - Qualitative
  - No sense of order
  - Can be stored with text, color or numeric code but code do not assign an order.
- As the name implies it consists of “naming” observations or classifying them into various mutually exclusive and collectively exhaustive categories.

Cont'd

- The practice of using numbers to distinguish among the various medical diagnoses constitutes measurement on a nominal scale.

Examples: include such dichotomies (contrasts or oppositions) as male–female, well–sick, under 65 years of age–65 and over, child–adult, and married–not married.

Measurement Scales (continue)

2. The Ordinal Scale

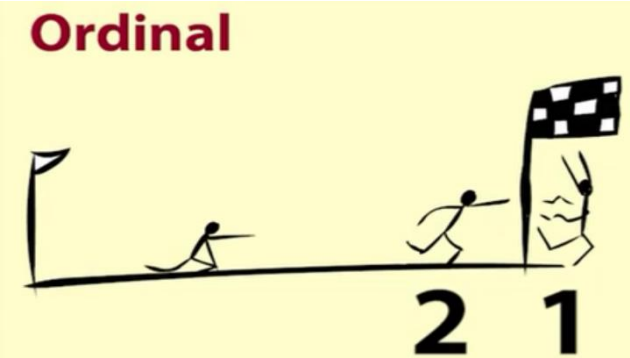
Whenever observations are not only different from category to category but can be ranked according to some criterion, they are said to be measured on an ***ordinal scale***.

- May have meaningful order but Interval between the values may not be equal.

Cont'd

Example:

- Recovery of a patient may be characterized as unimproved, improved, and much improved.
- Individuals may be classified according to socioeconomic status as low, medium, or high.
- The intelligence of children may be above average, average, or below average.



Measurement Scales (continue)

3. The Interval Scale

- The interval scale is a more sophisticated scale than the nominal or ordinal scale.
- Things can be measured rather classified or ordered.
- In the ***interval scale***, it is possible to order measurements, also the distance between any two measurements is known.
- The interval scale unlike the nominal and ordinal scales is a truly quantitative scale.

Cont'd

- We know, that the difference between a measurement of 20 and a measurement of 30 is equal to the difference between measurements of 30 and 40.
- The ability to do this implies the use of a unit distance and a zero point, both of which are arbitrary.
- The selected zero point is not necessarily a true zero in that it does not have to indicate a total absence of the quantity being measured.
- **Example: Temperature Scale** The unit of measurement is the degree, and the point of comparison is the arbitrarily chosen "zero degrees," which does not indicate a lack of heat.

Measurement Scales (continue)

4. The Ratio Scale

- The highest level of measurement is the *ratio scale*.
- Things can be measured rather classified or ordered.
- The ***ratio scale*** is characterized by the fact that equality of ratios as well as equality of intervals may be determined.
- Fundamental to the ratio scale is a true zero point.
- The measurement of such familiar individualities as height, weight, and length makes use of the ratio scale.

Lesson 2: Descriptive Biostatistics

Methods of summarizing and displaying data

Lecture Outline

- Descriptive Measures
- Measures of Central Tendency
  - The Mean
  - The Median
  - The Mode
  - Data Distribution (symmetric and skewed distribution)
- Measures of Dispersion
  - The Range
  - The Variance
  - The Standard Deviation
  - The Coefficient of Variance
  - The Percentiles
  - The Interquartile Range
  - Outliers
  - Kurtosis
- Grouped Data: The Frequency Distribution
- Graphic Methods

Descriptive Biostatistics

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

- Measurements that have not been organized, summarized, or otherwise manipulated are called ***raw data***.

Descriptive Measures

- The ability to summarize the data by means of a single number called a ***descriptive measure***.
- Descriptive measures may be computed from the **data of a sample** or the **data of a population**.
- A descriptive measure computed from the data of a sample is called ***a statistic***.
- A descriptive measure computed from the data of a population is called ***a parameter***.

Frequencies and frequency distribution tables

- **Frequency distribution**: is a table showing a listing of all observed values of the variable being studied and how many times each value is observed.

- The number of times that something occurs is known as its *frequency*

- The notation fx is used to denote the f*requency* or number of times the value x occurs

- The *relative frequency* is just the frequency divided by the sample size n.

Categorical or Qualitative Frequency Distributions

- What is a categorical frequency distribution?

A categorical frequency distribution represents data that can be placed in specific categories, such as gender, blood group, & hair color, etc.

Example: The blood types of 25 blood donors are given below. Summarize the data using a frequency distribution.

AB	B	A	O	B
O	B	O	A	O
B	O	B	B	B
A	O	AB	AB	O
A	B	AB	O	A

Categorical Frequency Distribution for the Blood Types

Class (Blood Type)	Frequency, f
A	5
B	8
O	8
AB	4
Total	n = 25

Note: The classes for the distribution are the blood types.

Quantitative Frequency Distributions -- Ungrouped

▶ What is an ungrouped frequency distribution?

An ungrouped frequency distribution simply lists the data values with the corresponding frequency counts with which each value occurs.

Example: The at-rest pulse rate for 16 athletes at a meet were 57, 57, 56, 57, 58, 56, 54, 64, 53, 54, 54, 55, 57, 55, 60, and 58. Summarize the information with an ungrouped frequency distribution.

Quantitative Frequency Distributions – Ungrouped --

Example Continued

Class (pulse Rate)	Frequency, f
53	1
54	3
55	2
56	2
57	4
58	2
60	1
64	1
Total	n = 16

Note: The (ungrouped) classes are the observed values themselves.

Relative Frequency Distribution

Class (pulse Rate)	Frequency, f	Relative Frequency
53	1	0.0625
54	3	0.1875
55	2	0.1250
56	2	0.1250
57	4	0.2500
58	2	0.1250
60	1	0.0625
64	1	0.0625
Total	n = 16	1.0000

Note: The relative frequency for a class is obtained by computing f/n.

Cumulative Frequency and Cumulative Relative Frequency

Class (pulse Rate)	Frequency, f	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
53	1	0.0625	1	0.0625
54	3	0.1875	4	0.2500
55	2	0.1250	6	0.3750
56	2	0.1250	8	0.5000
57	4	0.2500	12	0.7500
58	2	0.1250	14	0.8750
60	1	0.0625	15	0.9375
64	1	0.0625	16	1.0000
Total	n = 16	1.0000		

Note: Table with relative and cumulative relative frequencies.

Grouped data frequency distribution

- In order to present data using grouped frequency distribution, it is not as simple as that of ungrouped. In this case we need to compute some values. These values are given below:
  - Number of class: The number of categories the table will have
  - The approximate class width and number of class can be computed using Sturge's rule as:

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

$$w = \frac{L\arg est\ obs\ value - Smallest\ obs}{k}$$

  - K= number of class
  - Where n=sample size. Then the width of each class, W, can be computed as:

Grouped data frequency distribution cont'd….

- Class limit: The range for each class/ The smallest and largest values that can go into any class; they can be either lower- or upper-class limits.
  - Lower class limit: Smallest observation of the dataset
  - Upper class limit: Smallest observation plus width of the class minus one.

Cont'd

- We always make sure that each item (measurement or observation) goes into one and only one class, i.e. classes should be mutually exclusive (that the classes do not overlap, namely, that successive classes have no values in common).

- To this end we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into possible gaps between successive classes,

- **Note that**: the Sturge's rule should not be regarded as final, but should
- be considered as a guide only. The number of classes specified by the rule should be increased or decreased for convenient or clear presentation.

Cont'd

- **Class Boundaries/True Limits:** are those limits, which are determined mathematically to make an **interval of a continuous variable** in both directions, and no gap exists between classes.
  - ✓ Lower class boundary
  - ✓ Upper class boundary
- Class mark/ Mid-point (Xc) of an interval is the value of the interval which lies mid-way between the lower true limit (LTL) and the upper true limit (UTL) of a class.

  It is calculated as: The average of lower- and upper-class limit.

$$X_c = \frac{\text{Upper Class Limit} + \text{Lower Class Limit}}{2}$$

Cont'd

NB: The constructed frequency distribution expected to be (for continuous variables):

- Class intervals should be continuous, non overlapping, mutually exclusive and exhaustive.

- Class intervals generally should be of the same width
- open indeed class intervals should be avoided. These are classes like less then 10, greater than 65, and so on.

Cont'd

Example : the following data is about current age of women and it was collected from 240 women.

25	27	30	20	20	35	42	24	45	42	27	46	24	27	18	15	18	29	23	29	34	23	16
28	25	30	17	48	26	48	34	41	32	22	22	39	22	22	23	49	30	24	20	24	27	16
30	19	29	38	16	18	21	46	45	34	32	30	17	16	49	32	46	25	30	20	26	37	29
21	21	38	39	23	44	34	25	20	44	38	22	27	28	25	45	20	40	35	25	20	29	40
23	34	19	19	29	34	20	28	20	30	17	23	23	24	44	21	38	22	40	38	25	39	38
28	42	22	22	20	22	20	22	29	26	38	20	21	32	26	22	24	23	33	45	29	30	48
22	38	25	24	41	23	26	37	23	22	33	25	23	28	37	21	36	21	25	37	19	15	25
35	37	35	45	42	48	27	22	42	27	42	48	26	29	35	19	22	17	20	20	25	28	26
45	31	24	31	25	30	24	35	16	33	20	25	17	29	19	23	27	20	28	23	33	35	31
30	41	20	28	29	49	31	24	23	21	22	23	41	22	38	25	31	27	40	25	41	24	33
32	30	47	21	25	44	24	30	43	46													

Cont'd

- As we can see the dataset, it is too hard to get quick information as it is. Hence we need to make the data so that quick information can be grasped from it.

- If we tried to use the value of age at first marriage as a class and count the number of occurrence to get the frequency, the table will too long and hence the objective of summarization does not meet.

- Thus, we need to get range for each class with the number of subjects falling on each class will be taken as the corresponding frequency. To do that we need to get the width of each class, lower and upper class limit and the maximum number of classes.

Cont'd

The constant k can be computed using Sturge's rule as:

$k = 1 + 3.322\log(240) = 8.9$

Therefore, the width W of each class can be computed as:

$w = (\text{largest} - \text{smallest}) / k = (49-15)/9 = 4$

Thus the width of each class can be 4 and the lower-class limit for the first class will be the minimum observation from the dataset.

Cont'd

- In our data, the minimum observation is 15, and hence this value is the lower-class limit for the first class denoted by LC1. While the upper-class limit, UC1,of the first class can be computed using the formula as:

$$UC1 = LC1 + W - 1$$

- The lower-class limit for the second class will be obtained by adding 1 on the upper-class limit of the first class, that is LC2=UC1+1 implying that the upper-class limit for the second class is:

$$UC2 = LC2 + W - 1 \quad \text{and so on.}$$

- The lower-class boundary for each class can be obtained by subtracting 0.5 from lower class limit of the corresponding class, while the upper-class boundary for each class can be obtained by adding 0.5 on the upper-class limit of the corresponding class.

Example

Thus, the grouped frequency distribution for the data can be constructed as:

Class Limit	Class Boundary	Class Mark	Frequency	RF	CF
15-18	14.5-18.5	16.5	15	6.25	15
19-22	18.5-22.5	20.5	49	20.41	64
23-26	22.5-26.5	24.5	51	21.25	115
27-30	26.5-30.5	28.5	40	16.67	155
31-34	30.5-34.5	32.5	21	8.75	176
35-38	34.5-38.5	36.5	22	9.17	198
39-42	38.5-42.5	40.5	18	7.50	216
43-46	42.5-46.5	44.5	15	6.25	231
47-50	46.5-50.5	48.5	9	3.75	240

Frequency Distribution (Example-1)

- TABLE: Ordered Array of Ages of Subjects

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

- TABLE: Frequency Distribution of Ages of 189 Subjects Shown in the above tables.

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

Frequency Distribution (Example-2)

Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110
133	115	127	135	89	121	112	135	115	64

Table: Grouped frequency distribution of birthweight (oz) from 100 consecutive deliveries.

Group interval	Frequency
$29.5 \le x < 69.5$	5
$69.5 \le x < 89.5$	10
$89.5 \le x < 99.5$	11
$99.5 \le x < 109.5$	19
$109.5 \le x < 119.5$	17
$119.5 \le x < 129.5$	20
$129.5 \le x < 139.5$	12
$139.5 \le x < 169.5$	6
	100

Charts and tables used to present qualitative data

1. **Pie charts**

2. Bar charts (**simple and clustered bar charts**)

3. Relative frequency (percentage) table

- These **two charts** are used for presentation of qualitative data.

Pie charts

- Pie charts are typically used to present the **relative frequency** of qualitative data.
- In most cases the data are **nominal**, but **ordinal** data can also be displayed in a pie chart

Cont'd

- **The complete circle represents the total** number of measurements.
- **Partition into slices -** one for each category.
- **The size of a slice is proportional to the** relative frequency of that category.
- **Determine the angle of each slice by** multiplying **the relative frequency** by 360 degree. (Recall a circle spans 360)

Cont'd

Steps to create a pie-chart

- Construct a frequency table
- Calculate relative frequency % (percentage)
- Change the percentages into degrees, where: degree = Percentage X 360o.
- Draw a circle and divide it accordingly

For single variable:

- For example, in a class of 40 students, 15 are boys and 25 are girls. (See the pie chart)

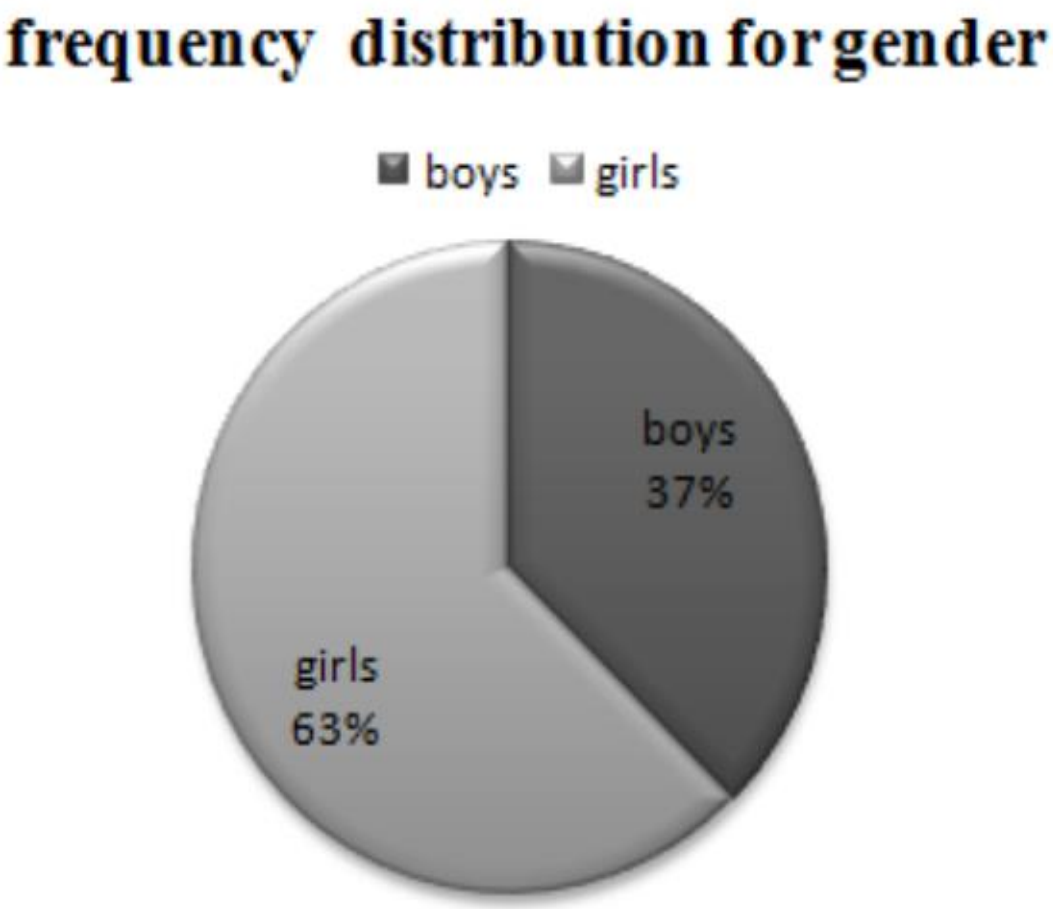
Cont'd

- ***Frequency:*** number of times that something occurs.
- **Relative frequency** = frequency divide by sum of all frequencies

$$\textit{Relative frequency} = \frac{\textit{Frequency}}{\textit{Sum of all frequencies}}$$

Gender	Frequency	Relative frequency %
Boys	15	$\frac{15}{40} * 100 = 37.5\%$
Girls	25	$\frac{25}{40} * 100 = 62.5\%$
	40	

Cont'd


frequency distribution for gender
boys
girls
boys
37%
girls
63%

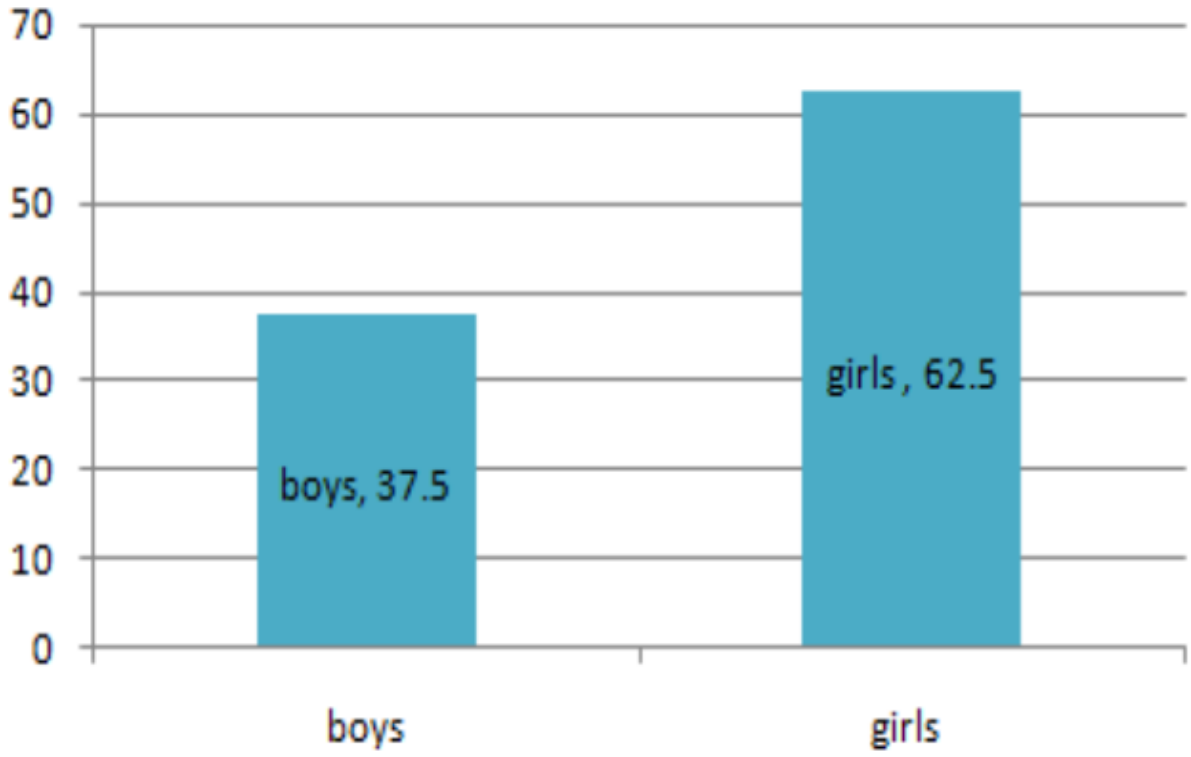
Bar Chart (Bar Graph)

- **Place categories on the horizontal axis.**
- **Place frequency (or relative frequency)** on the vertical axis.
- **Construct vertical bars of equal width,** one for each category.
- Its **height** is proportional to the frequency (or **relative frequency**) of the category.

Cont'd

Simple bar chart

frequency distribution for gender (Bar chart)



Two variables (cross tabulation)

Cross tabulation or cross tabs are often used in presenting the counts of two qualitative variables.

Suppose the variables of interest are :

- Gender and
- wearing spectacles.

The are presented in this table.

	Wearing spectacles		Total
	yes	No	
Boy	5	10	15
Girls	10	15	25
Total	15	25	40

Cont'd

Two variables (qualitative)
We cross tabulation

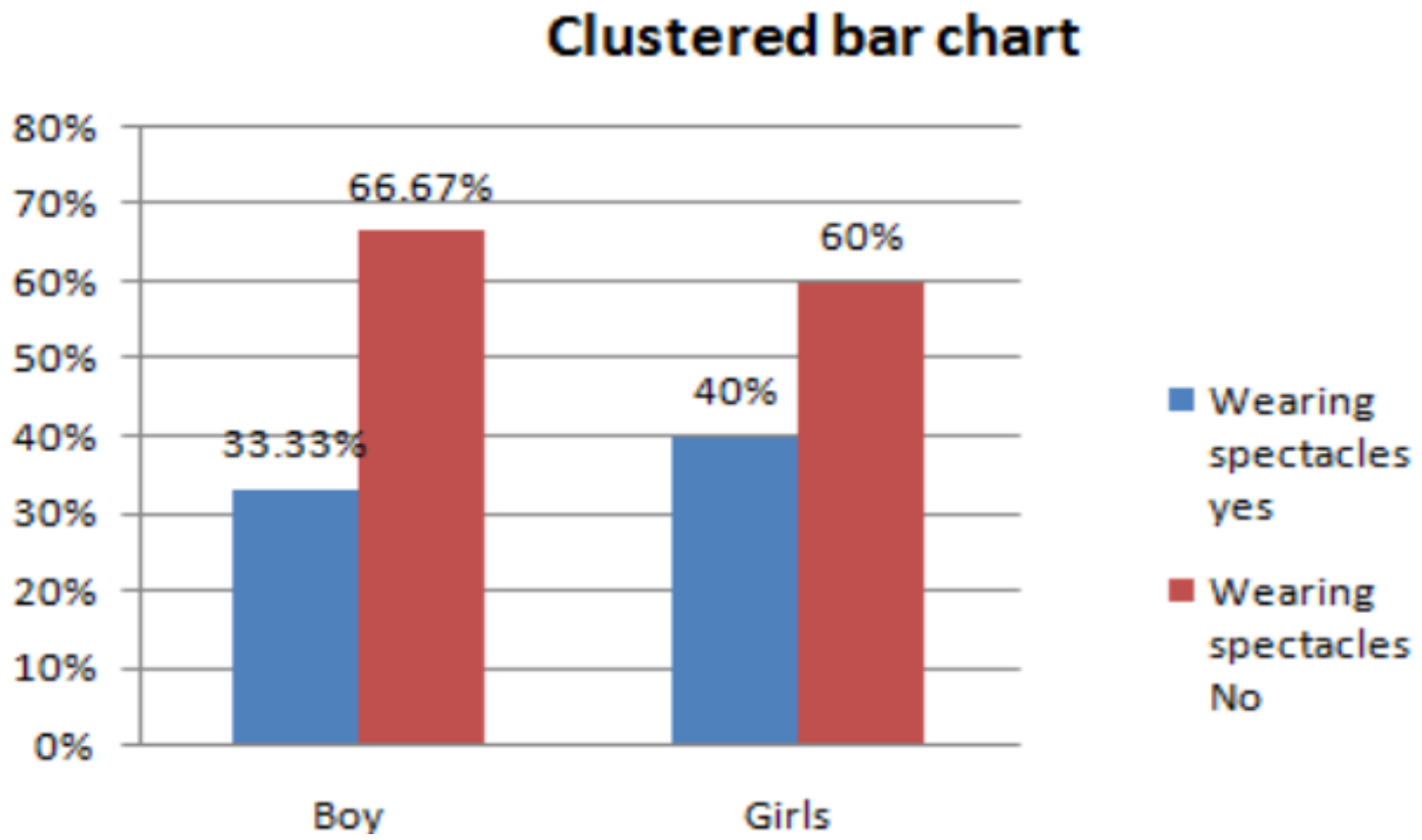
	Wearing spectacles		Total
	yes	No	
Boy	5	10	15
Girls	10	15	25
Total	15	25	40

Cont'd

	Wearing spectacles		Total
	yes	No	
Boy	33.33%	66.67%	100%
Girls	40%	60%	100%
Total	37.50%	62.50%	100%

Cont'd

Table showing the percentage of Gender and wearing spectacles.



Cont'd

Crosstabs and clustered bar chart

Expressed in percentage. 33.33% of the boys and 40% of the girls wear spectacles

Calculate the percentages

Smoking	Lung cancer		Total
	YES	NO	
YES	70	100	
NO	3	70	

Graphic Methods for displaying quantitative data

- In this section, certain commonly used graphic methods for displaying data are presented.

- The purpose of using graphic displays is to give a quick overall impression of data, which is sometimes difficult to obtain with numeric measures.

- The commonly used graphic methods are;

i. **Histogram**

ii. **Frequency Polygon**

iii. **Stem-and-Leaf Displays**

iv. **Box-and-Whisker Plots**

Histogram

- The ***Histogram*** is a special type of bar graph.
- In a histogram, the values of the variable under consideration are represented by the horizontal axis, while the vertical axis represent the frequency of occurrence.
- Above each class interval on the horizontal axis a rectangular bar, or cell, is erected.
- The height of the bar indicates the frequency.
- In a histogram, the bars must touch to indicate that there are no data in the data set that are missing from the histogram.

Histogram

TABLE: Ordered Array of Ages of Subjects

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

TABLE: Frequency Distribution of Ages of 189 Subjects Shown in the above tables.

True Class Limits	Frequency
29.5–39.5	11
39.5–49.5	46
49.5–59.5	70
59.5–69.5	45
69.5–79.5	16
79.5–89.5	1
Total	189

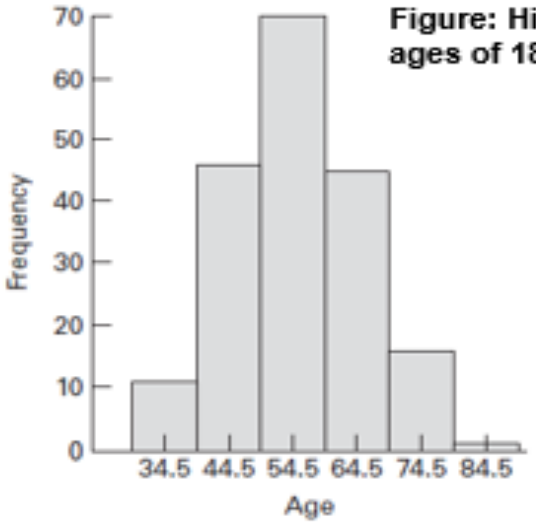


Figure: Histogram of ages of 189 subjects

Frequency Polygon

- The ***frequency polygon*** is a special kind of line graph.
- To draw a frequency polygon, we first place a dot above the midpoint of each class interval represented on the horizontal axis of a graph.
- The height of a given dot above the horizontal axis corresponds to the frequency of the relevant class interval.
- Connecting the dots by straight lines produces the frequency polygon.

Cont'd

- Both ends of the frequency polygon are attached on the x-axis.
- To accomplish this, we must utilize a class interval before the first one, and a class interval after the last one, each containing no data points.
- This allows for the total area to be enclosed.
- The total area under the frequency polygon is equal to the area under the histogram.

Frequency Polygon

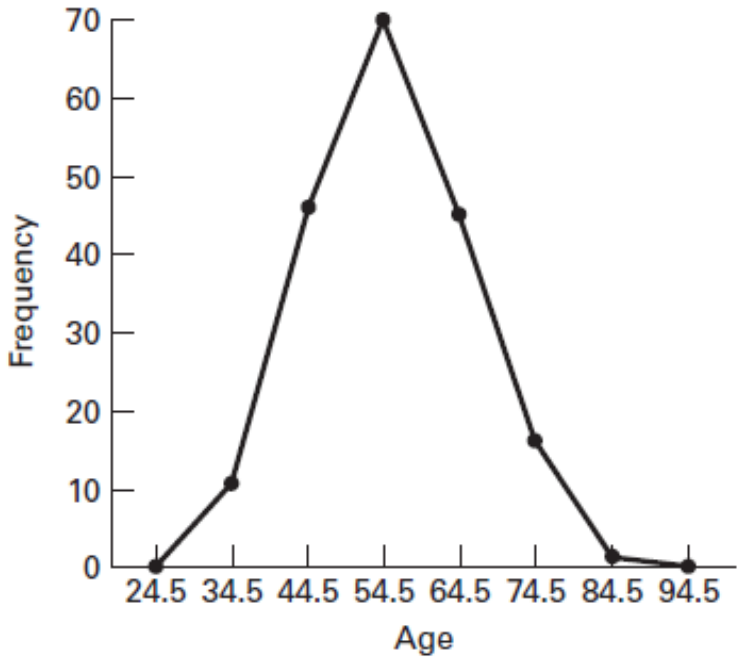


Figure: Frequency polygon for the ages of 189 subjects.

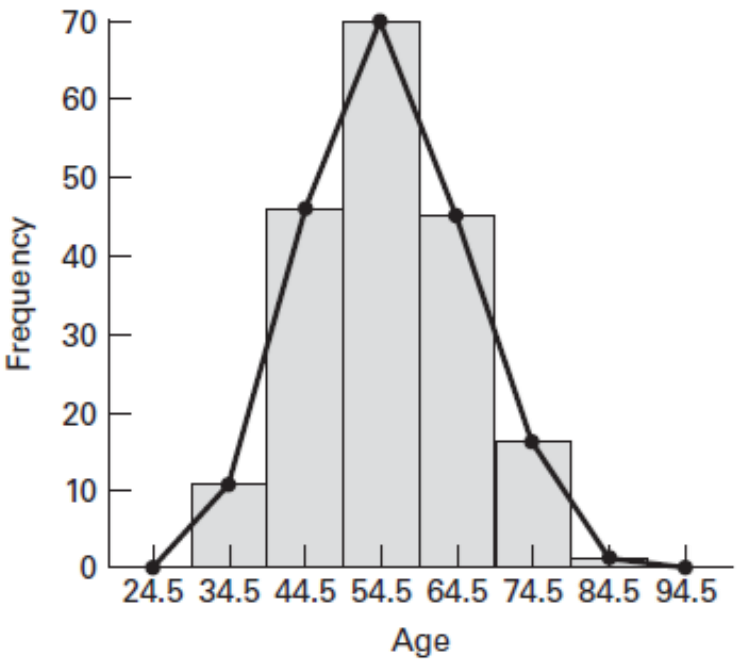


Figure: Histogram and Frequency polygon for the ages of 189 subjects.

Stem-and-Leaf Displays

- The ***stem-and-leaf display*** is useful for representing quantitative data sets.

- A quick way to obtain an informative visual representation of the data set is to construct a *stem-and-leaf display.*

- A stem-and-leaf display bears a strong resemblance to a histogram.

- A properly constructed stem-and-leaf display, like a histogram, provides information regarding the range of the data set, shows the location of the highest concentration of measurements, and reveals the presence or absence of symmetry.
- An advantage of the stem-and-leaf display over the histogram is the fact that it preserves the information contained in the individual measurements. Such information is lost when measurements are assigned to the class intervals of a histogram.

Cont'd

- Another advantage of stem-and-leaf displays is the fact that they can be constructed during the tallying process, so the intermediate step of preparing an ordered array is eliminated.
- To construct a stem-and-leaf display we partition each measurement into two parts.
- The first part is called the ***stem***, and the second part is called the ***leaf***.
- The stem consists of one or more of the initial digits of the measurement.
- The leaf is composed of one or more of the remaining digits.
- Thus the stem of the number 483 is 48, and the leaf is 3.

Stem-and-Leaf Displays

- All partitioned numbers are shown together in a single display; the stems form an ordered column with the smallest stem at the top and the largest at the bottom.
- The rows of the display contain the leaves, ordered and listed to the right of their respective stems.
- When leaves consist of more than one digit, all digits after the first may be deleted.
- Decimals when present in the original data are omitted in the stem-and-leaf display.
- The stems are separated from their leaves by a vertical line.
- A stem-and-leaf display is also an ordered array of the data.

Stem-and-Leaf Displays (Example-1)

Since the measurements are all two-digit numbers, we will have one-digit stems and one-digit leaves.

For Example: The measurement 30 has a stem of 3 and a leaf of 0.

TABLE: Ordered Array of Ages of 189 Subjects

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

```
Stem   Leaf
-------------------------------------------------------------------------------
   3 | 04577888899
   4 | 002233333344444445556666667777778888888999999
   5 | 00000000111122222233333333333333333344444444444555666666777777788999999
   6 | 000011111111111222222233444444556666667888999
   7 | 0111111123567888
   8 | 2
```

(stem unit = 10, leaf unit = 1).

(Example-2)

- In some stem-and-leaf plots the leaf can consist of more than one digit.
- In this case, the leaf would consist of the rightmost two digits.
- The stem the leftmost two digits.
- The pairs of digits to the right of the vertical bar would be underlined to distinguish between two different leaves.
- The stem-and-leaf display for the data in the table is shown in the figure.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

20	69
21	
22	
23	
24	
25	81
26	
27	59
28	41 38 34
29	
30	31
31	01
32	65 60 45 00 48
33	23 14
34	84
35	41
36	49 09
37	
38	
39	
40	
41	46

(Example-3)

- The point 5|8 represents 58, 11|8 represents 118, and so forth.
- Notice how this plot gives an overall feel for the distribution without losing the individual values.
- Also, the cumulative frequency count from either the lowest or the highest value is given in the first column.
- For the 11 stem, the absolute count is given in parentheses (17) instead of the cumulative total because the highest or lowest value would exceed 50% (50).

Stem-and-leaf of birthwgt N = 100
Leaf Unit = 1.0

1	3	2
1	4	
2	5	8
5	6	478
5	7	
15	8	3556788999
26	9	12344568889
45	10	0123444445567888899
(17)	11	00122235555556889
38	12	01112222344445567788
18	13	222334557888
6	14	0146
2	15	5
1	16	1

Sample of birthweights (oz) from 100 consecutive deliveries at a Boston hospital

58	118	92	108	132	32	140	138	96	161
120	86	115	118	95	83	112	128	127	124
123	134	94	67	124	155	105	100	112	141
104	132	98	146	132	93	85	94	116	113
121	68	107	122	126	88	89	108	115	85
111	121	124	104	125	102	122	137	110	101
91	122	138	99	115	104	98	89	119	109
104	115	138	105	144	87	88	103	108	109
128	106	125	108	98	133	104	122	124	110

Box-and-Whisker Plots

- The construction of a ***Box-and-Whisker plot*** (or ***boxplot***) makes use of the quartiles of a data set and may be accomplished by following these five steps:

1. Represent the variable of interest on the horizontal axis.

2. Draw a box in the space above the horizontal axis in such a way that the left end of the box aligns with the first quartile Q_1 and the right end of the box aligns with the third quartile Q_3.

3. Divide the box into two parts by a vertical line that aligns with the median.

4. Draw a horizontal line called a *whisker* from the left end of the box to a point that aligns with the smallest measurement in the data set.

Cont'd

1. Draw another horizontal line, or whisker, from the right end of the box to a point that aligns with the largest measurement in the data set.

- Examination of a box-and-whisker plot for a set of data reveals information regarding the amount of spread, location of concentration, and symmetry of the data.

(Example-1)

Evans et al. examined the effect of velocity on ground reaction forces (GRF) in dogs with lameness (a condition in which the animal fails to travel in a regular and sound manner on all four feet) from a torn cranial cruciate ligament (disease). The dogs were walked and trotted (run) over a force platform, and the GRF was recorded during a certain phase of their performance. The Table given below contains 20 measurements of force where each value shown is the mean of five force measurements per dog when trotting.

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

(Example-1)

- The smallest and largest measurements are **14.6** and **44**, respectively.

- The first quartile is the Q1 = (20 + 1)/4 = 5.25th measurement.
- The 25th measurement is equal to 27.2 + (0.25)(27.4 – 27.2) = **27.25**.

- The second quartile or median is the Q2 = (20 + 1)/2 = 10.5th measurement.
- The 10.5th measurement is equal to 30.7 + (0.5)(31.5 – 30.7) = **31.1**.

- The third quartile is the Q3 = 3(20 + 1)/4 = 15.75th measurement.
- The 15.75th measurement is equal to 33.3 + (0.75)(33.6 – 33.3) = **33.525**.

Cont'd

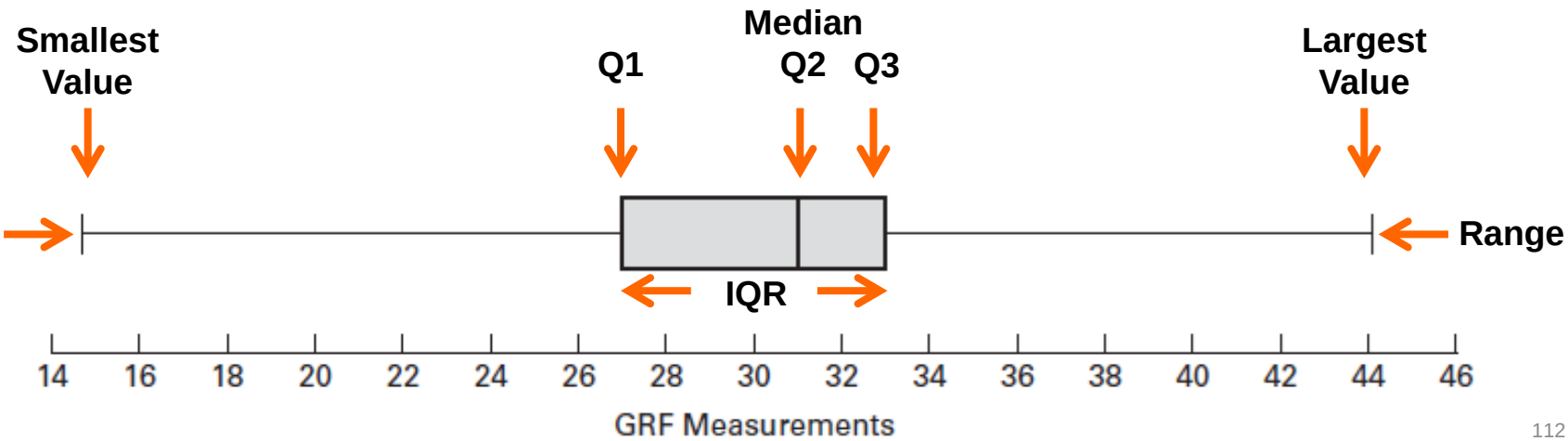
- The interquartile range is IQR = 33.525 – 27.25 = **6.275**.
- The range is R = 44 – 14.6 = **29.4**.
- The IQR 100(6.275/29.4) = **21 percent of the range**.

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

(Example-1)

- The resulting box-and-whisker plot is shown in the figure.
- Examination of figure reveals that 50 percent of the measurements are between about 27 and 33, the approximate values of the first and third quartiles, respectively.
- The vertical bar inside the box shows that the median is about **31**.



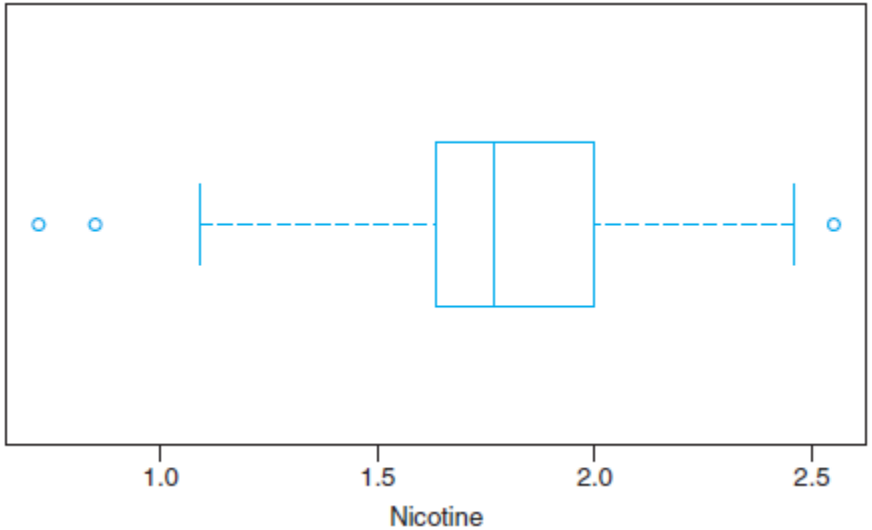
(Example-2)

Nicotine content was measured in a random sample of 40 cigarettes. The data are displayed in Table.

Nicotine Data

1.09	1.92	2.31	1.79	2.28	1.74	1.47	1.97
0.85	1.24	1.58	2.03	1.70	2.17	2.55	2.11
1.86	1.90	1.68	1.51	1.64	0.72	1.69	1.85
1.82	1.79	2.46	1.88	2.08	1.67	1.37	1.93
1.40	1.64	2.09	1.75	1.63	2.37	1.75	1.69

Box-and-whisker plot



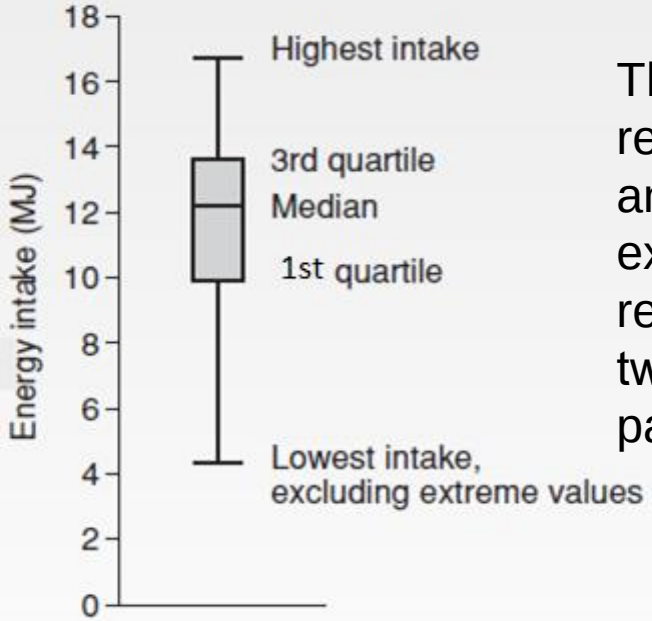
(Example-2)

Stem-and-leaf plot for the nicotine data.

```
The decimal point is 1 digit(s) to the left of the |
 7 | 2
 8 | 5
 9 |
10 | 9
11 |
12 | 4
13 | 7
14 | 07
15 | 18
16 | 3447899
17 | 045599
18 | 2568
19 | 0237
20 | 389
21 | 17
22 | 8
23 | 17
24 | 6
25 | 5
```

Example

A dietician measured the energy intake over 24 hours of 50 patients on a variety of wards. One ward had two patients that were “nil by mouth”. The median was 12.2 megajoules, IQR 9.9 to 13.6. The lowest intake was 0, the highest was 16.7. This distribution is represented by the box and whisker plot



The ends of the whiskers represent the maximum and minimum values, excluding extreme results like those of the two “nil by mouth” patients.

Types of Descriptive Measures

Several ***types of descriptive measures*** can be computed from a set of data. However, the **Two** important types are;

1. **Measures of Central Tendency**
2. **Measures of Dispersion**

SUMMARIZING DATA: Measures of location

A measure of location or central tendency or average is a single value used to represent a set of data.

Objective of average:

1. **To get single value that represent the entire data.**
2. **To facilitate comparison between groups of data of similar nature.**

Important measures of central tendency are:

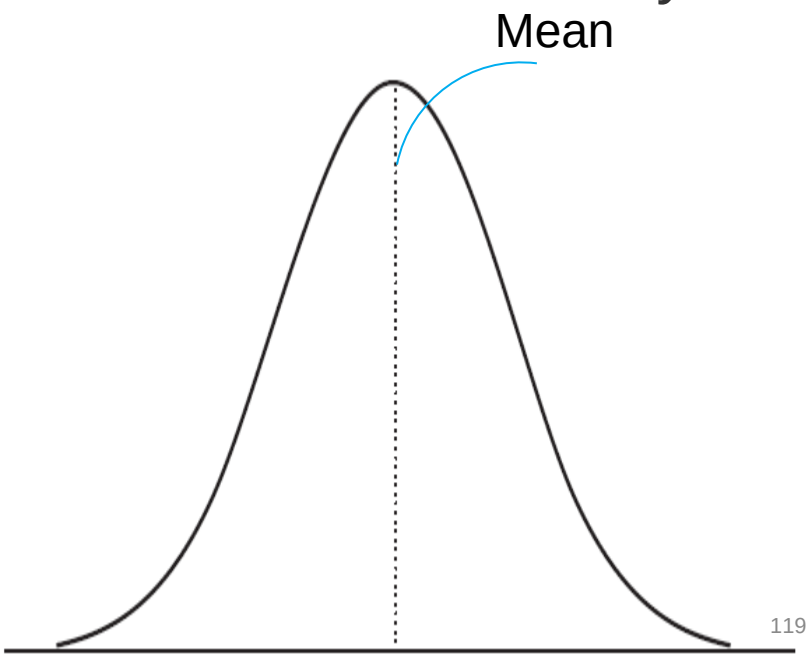
1. **Mean**
2. Median
3. Mode

Characteristics of a good MCT

1. It should be based on all the observations
2. It should not be affected by the extreme values
3. It should be as close to the maximum number of values as possible
4. It should have a definite value
5. It should not be subjected to complicated and tedious calculations
6. It should be capable of further algebraic treatment
7. It should be stable with regard to sampling fluctuation

MEAN

- It is important to have an understanding of how it is calculated.
- It is used when the spread of the data is fairly similar on each side of the mid point.
- for example, when the data are “normally distributed”.
- The “normal distribution” is symmetrical, bell-shaped distribution of data as shown.
- The mean is the sum of all the values, divided by the number of values.



Example

Five women in a study on lipid-lowering agents are aged 52, 55, 56, 58 and 59 years.

Add these ages together:

$$52 + 55 + 56 + 58 + 59 = 280$$

Now divide by the number of women:

$$\frac{280}{5} = 56$$

So the mean age is 56 years.

The Mean

The **three** types of mean are;

1. **Arithmetic Mean**
2. **Geometric Mean**
3. **Harmonic Mean**

▶ The most familiar measure of central tendency is the arithmetic mean.

▶ Since geometric and harmonic means are not the scope of this course, ***the arithmetic mean*** simply referred as ***the mean***.

The Arithmetic Mean

- The ***arithmetic mean*** is colloquially called the ***average***.
- The arithmetic mean (or mean or sample mean) is usually denoted by $\bar{x}$.
- The arithmetic mean is the sum of all the observations divided by the number of observations.
- The arithmetic mean is written in statistical terms as;

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

The Arithmetic Mean

The Sample Mean

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

The Finite Population Mean

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

The Arithmetic Mean

Example: What is the arithmetic mean for the sample of birth-weights in the table.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

$$\bar{x} = (3265 + 3260 + \cdots + 2834)/20 = 3166.9 \text{ g}$$

Grouped Data

- Here the assumption is all values in a particular class interval are located at the mid point of the interval
- And the mean is calculated as follows;

$$\bar{x} = \frac{\sum_{i=1}^{k} m_i f_i}{\sum_{i=1}^{k} f_i}$$

where,

k = the number of class intervals

m_i = the mid-point of the ith class interval

f_i = the frequency of the ith class interval

Cont'd

- ***Example: Compute the mean age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.***

Mean=5810.5/169=34.48years*(What does it imply?)*

Class interval	Mid-point (m_i)	Frequency (f_i)	$m_i f_i$
10-19	14.5	4	58.0
20-29	24.5	66	1617.0
30-39	34.5	47	1621.5
40-49	44.5	36	1602.0
50-59	54.5	12	654.0
60-69	64.5	4	258.0
Total	__	169	5810.5

The Arithmetic Mean

Limitations:

- The arithmetic mean is, in general, a very natural measure of location.
- One of its main limitations, however, is that it is oversensitive to extreme values.
- In this instance, it may not be representative of the location of the great majority of sample points.

The Arithmetic Mean

Example-1: The Arithmetic Mean Limitation

- If the first infant in the Table happened to be a premature infant weighing 500 g rather than 3265 g, then the arithmetic mean of the sample would fall to 3028.7 g.

- In this instance, 7 of the birth-weights would be lower than the arithmetic mean, and 13 would be higher than the arithmetic mean.

- It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean.

- In these types of samples, the arithmetic mean is a poor measure of central location because it does not reflect the center of the sample.

- Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

The Arithmetic Mean

- **Example-2: The Arithmetic Mean Limitation**

- Suppose the five physicians who practice in an area are surveyed to determine their charges for a certain procedure.

- Assume that they report these charges: $75, $75, $80, $80, and $280.

- The mean charge for the five physicians is found to be $118, a value that is not very representative of the set of data as a whole.

- The single atypical value had the effect of inflating the mean.

Properties of the Mean

1. **Uniqueness**

For a given set of data there is one and only one mean.

2. **Simplicity**

The mean is easy to calculate.

3. **Affected by Extreme Values**

The mean is influenced by each value equally. Therefore, extreme values can distort the mean.

Mean

Merits:

- It is the most popular average easy to understand and easy to calculate.
- It takes all the observation into account.
- The mean is used in computing other statistics (such as the variance, standard deviation etc)

Limitation:

- **Mean is affected by extremely high or low values.**
- **It is not a good measure of average in extremely asymmetric distribution of observations.**

The Median

- An alternative measure of location is ***the median*** or, more precisely, ***the sample median***.

- The median of a finite set of values is that value which divides the set into two equal parts.

- The median is defined differently when n is **even** and **odd**.

- Samples with an ***odd sample size*** have a unique central point, when all values have been arranged in order of magnitude.

- **Example:** For samples of size 7, the fourth largest point is the central point in the sense that 3 points are smaller than it and 3 points are larger.

Cont'd

- Samples with an ***even sample size*** have no unique central point, and the middle two values must be averaged, when all values have been arranged in the order of their magnitudes.

- **Example:** For samples of size 8 the fourth and fifth largest points would be averaged to obtain the median, because neither is the central point.

The Median

Suppose there are *n* observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

The sample median is;

(1) The $\left(\frac{n+1}{2}\right)$th largest observation if n is odd

(2) The average of the $\left(\frac{n}{2}\right)$th and $\left(\frac{n}{2}+1\right)$th largest observations if n is even

The Median

- Example-1: Compute the sample median for the sample in the given table.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

- First, arrange the sample in ascending order:
- 2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Cont'd

Because *n* is even,

Sample median = average of the 10th and 11th largest observations

Sample median = (3245 + 3248)/2 = **3246.5 g**

The Median

- **Example-2:** The data set in the table consists of white-blood counts taken on admission of all patients entering a small hospital in Allentown, Pennsylvania, on a given day. Compute the median white-blood count.

Sample of admission white-blood counts (× 1000) for all patients entering a hospital in Allentown, PA, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Cont'd

First, order the sample as follows: 3, 5, 7, 8, 8, 9, 10, 12, 35.

Because *n* is odd, the sample median is given by the fifth largest point, which equals **8 or 8000** on the original scale.

Strength:

- The main strength of ***the sample median*** is that it is ***insensitive*** to very large or very small values.

- In particular, if the second patient in ***Table of Example-2*** had a white count of 65,000 rather than 35,000, the sample median would remain unchanged, because the fifth largest value is still 8000.

The Median

- Conversely, the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample.

Weakness:

- The main weakness of the sample median is that it is determined mainly by the middle points in a sample and is less sensitive to the actual numeric values of the remaining data points.

The Median

Example-3: A simple random sample of 10 subjects from the population of subjects are shown in the table. find the median age of the subjects.

Table: Sample of 10 Ages Drawn from the Ages of a population

Random Number	Sample Subject Number	Age
137	1	43
114	2	66
155	3	61
183	4	64
185	5	65
028	6	38
085	7	59
181	8	57
018	9	57
164	10	50

Cont'd

- Arraying the 10 ages in order of magnitude from smallest to largest gives 38, 43, 50, 57, 57, 59, 61, 64, 65, 66.
- Since we have an even number of ages, there is no middle value. The two middle values, however, are 57 and 59.
- The sample median is (57 + 59)/2 = 58.

Grouped Data

- we assume that the values within a class-interval are evenly distributed through the interval.
- The first step is to locate the class in which the median is located, using the following procedure.
- Find n/2 and see a class with a minimum cumulative frequency which contains n/2.
- Then, use the following formula;

Cont'd

$$\tilde{x} = L_m + \left(\frac{\frac{n}{2} - F_c}{f_m} \right) W$$

where,

L_m = lower class boundary of the median class

F_c = cumulative frequency of the class preceding the median class

f_m = frequency of the median class

W= width of the median class

n = total number of observations

Cont'd

- ***Example: Compute the median age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.***

n/2=169/2=84.5

Class interval	Mid-point (m_i)	Frequency (f_i)	Cum. freq
10-19	14.5	4	4
20-29	24.5	66	70
30-39	34.5	47	117
40-49	44.5	36	153
50-59	54.5	12	165
60-69	64.5	4	169
Total		169	

Cont'd

- n/2 = 84.5 = in the 3rd class (the median class is the 3rd class)
- Lower class boundary = 29.5,
- Frequency of the median class = 47
- $(n/2 - F_c) = 84.5-70 = 14.5$

- **Median = 29.5 + (14.5/47)10 = 32.58 ≈ 33**
- ***What does it imply?***

Properties of the Median

Properties of the median include the following:

1. **Uniqueness**

As is true with the mean, there is only one median for a given set of data.

2. **Simplicity**

The median is easy to calculate.

3. **Less Affected by Extreme Value**

It is not as drastically affected by extreme values as is the mean.

Median

Merits:

- Median is easy to understand and easy to calculate.
- It is not affected by extremely high or low values.

Limitation:

- It is not based on all the observations. It is a position average and thus it is not determined by each and every observation.
- It is less reliable average than mean when number of observation is small.

The Mode

- The ***mode*** is the most frequently occurring value among all the observations in a sample.
- The mode is another widely used measure of location.
- If all the values are different then there is no mode.
- Some distributions have more than one mode.
- In fact, one useful method of classifying distributions is by the number of modes present.
- A distribution with one mode is called ***unimodal***; two modes, ***bimodal***; three modes, ***trimodal***; and so forth.

The Mode

Example-1: Compute the mode of the distribution in the table.

Sample of admission white-blood counts ($\times$ 1000) for all patients entering a hospital in Allentown, PA, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

The mode is **8×1000 = 8000** because it occurs more frequently than any other white blood count.

The Mode

- Example-2: Find the modal age of the subjects whose ages are given in the table.

Table: Ordered Array of Ages of 189 Subjects Who Participated in a Study on Smoking Cessation

30	34	35	37	37	38	38	38	38	39	39	40	40	42	42
43	43	43	43	43	43	44	44	44	44	44	44	44	45	45
45	46	46	46	46	46	46	47	47	47	47	47	47	48	48
48	48	48	48	48	49	49	49	49	49	49	49	50	50	50
50	50	50	50	50	51	51	51	51	52	52	52	52	52	52
53	53	53	53	53	53	53	53	53	53	53	53	53	53	53
53	53	54	54	54	54	54	54	54	54	54	54	54	55	55
55	56	56	56	56	56	56	57	57	57	57	57	57	57	58
58	59	59	59	59	59	59	60	60	60	60	61	61	61	61
61	61	61	61	61	61	61	62	62	62	62	62	62	62	63
63	64	64	64	64	64	64	65	65	66	66	66	66	66	66
67	68	68	68	69	69	69	70	71	71	71	71	71	71	71
72	73	75	76	77	78	78	78	82						

A count of the ages in the table reveals that the age 53 occurs most frequently (17 times).

The mode for this population of ages is 53.

The Mode

Example-3: Compute the mode of the distribution in the table.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

There is no mode of the distribution in the table, because all the values occur exactly once.

The Mode

- A distribution will be skewed to the right, or positively skewed, if its mean is greater than its mode.

- A distribution will be skewed to the left, or negatively skewed, if its mean is less than its mode.

Grouped data

- To find the mode of grouped data, we usually refer to the modal class, where the modal class is the class with the highest frequency.

- Once the modal class is/are identified it can be calculated using the following formula.

Cont'd

$$\hat{X} = L_m + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right) * W$$

Where,

W= width of the modal class

L_m = lower class boundary of the modal class

f_1 = frequency of the class preceding the modal class

f_2 = frequency of the modal class

f_3 = frequency of the class succeeding the modal class

$$\Delta_1 = f_2 - f_1$$

$$\Delta_2 = f_2 - f_3$$

Cont'd

- ***Example: Compute the most frequent age at diagnosis of 169 cervical cancer patients; use the grouped frequency table below.***

Class interval	Mid-point (m_i)	Frequency (f_i)	Cum. freq
10-19	14.5	4	4
20-29	24.5	66	70
30-39	34.5	47	117
40-49	44.5	36	153
50-59	54.5	12	165
60-69	64.5	4	169
Total		169	

Cont'd

- the 2nd class is the modal class ($f_2 = 66$)
- $f_1 = 4$ and $f_3 = 47$
- Lower class boundary(L_m) = 19.5,
- Class width (W)= 10
- = 84.5-70 = 14.5
- $\Delta_1 = f_2 - f_1 = 62$
- $\Delta_2 = f_2 - f_3 = 19$
- **M0de = 19.5 + (62/81)*10 = 27.15 ≈ 27**
- ***What does it imply?***

Mode

Merits:

- Mode is easy to understand and easy to calculate.
- Like median, mode is not at all affected by extremely high or low values.
- When there is a large frequency in a distribution, mode happens to be meaningful as an average.

Limitation:

- It is not based on all the observations.
- It is less reliable average than mean when number of observation is small.

Data Distributions

- ***Data distributions*** may be classified on the basis of whether they are ***symmetric*** or ***asymmetric***.

- If a distribution is ***symmetric***, the left half of its graph will be a mirror image of its right half.

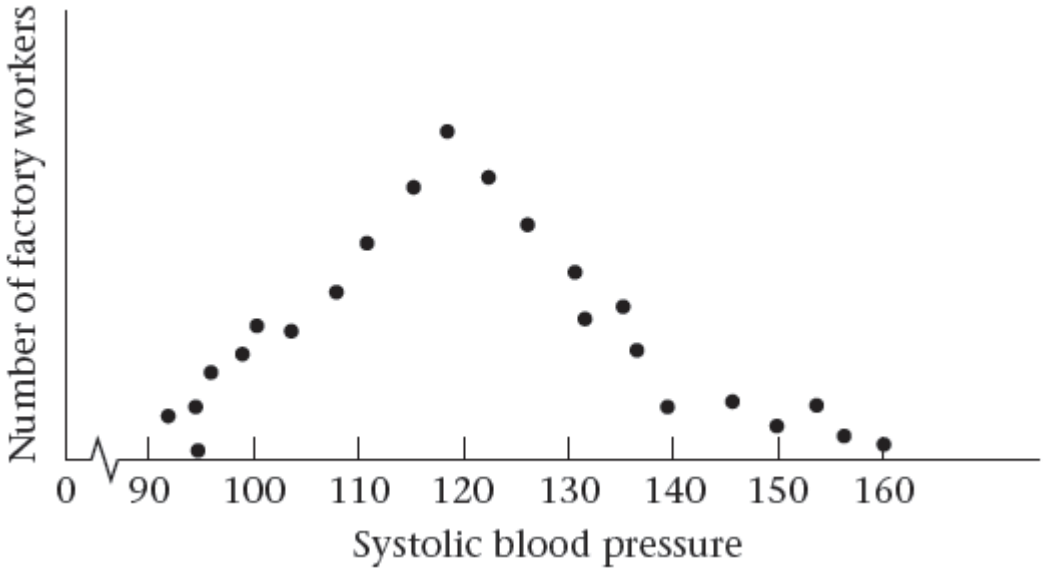
- When the left half and right half of the graph of a distribution are not mirror images of each other, the distribution is ***asymmetric***.

- If the graph of a distribution is asymmetric, then the distribution is said to be ***skewed***.

Symmetric Distribution

In ***symmetric distribution***, the relative position of the points on each side of the sample median is the same.

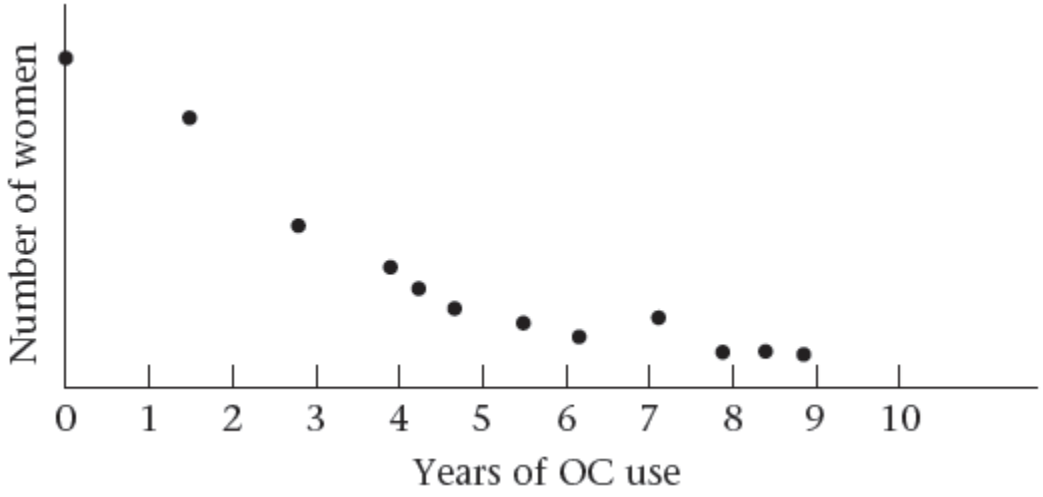
Example: A distribution that is expected to be roughly symmetric is the distribution of systolic blood-pressure measurements taken on all 30- to 39-year-old factory workers in a given workplace.



Positively Skewed Distribution

- If a distribution is not symmetric because its graph extends further to the right than to the left, that is, if it has a long tail to the right, then the distribution is ***skewed to the right*** or is ***positively skewed***.
- In ***positively skewed distribution***, the points above the median tend to be farther from the median in absolute value than points below the median.

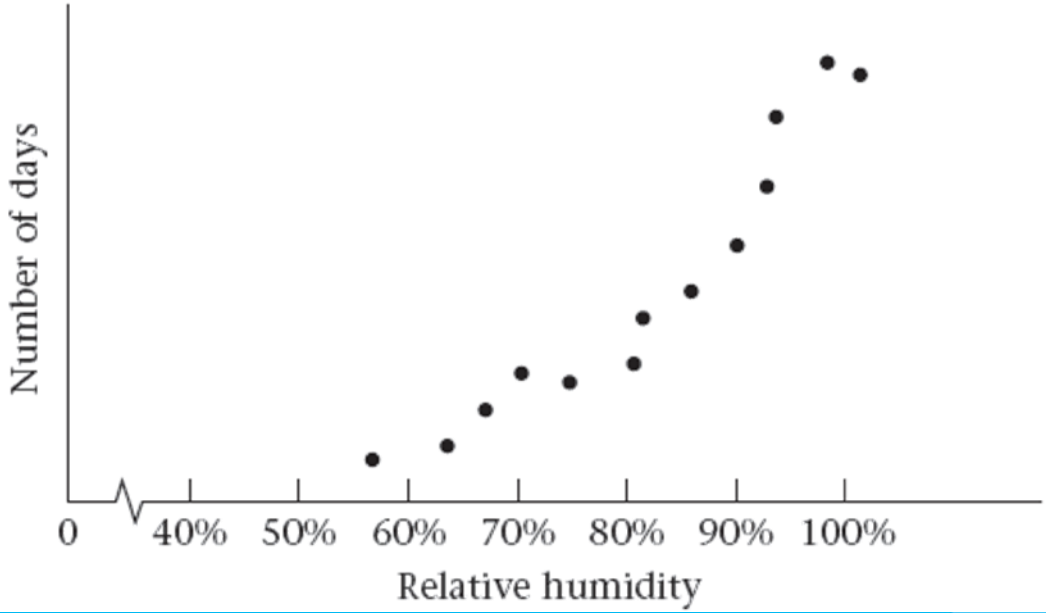
Example: The number of years of antihypertensive drugs use among a group of women ages 20 to 29 years.



Negatively Skewed Distribution

- If a distribution is not symmetric because its graph extends further to the left than to the right, that is, if it has a long tail to the left, then the distribution is ***skewed to the left*** or is ***negatively skewed***.
- In ***negatively skewed distribution***, the points below the median tend to be farther from the median in absolute value than points above the median.

Example: Relative humidities observed in a humid climate at the same time of day over a number of days. In this case, most humidities are at or close to 100%, with a few very low humidities on dry days.



Skewness

Skewness can be expressed as follows:

$$Skewness = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^{3/2}} = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-1)\sqrt{n-1}\,s^3}$$

Where *s* is the standard deviation of a sample.

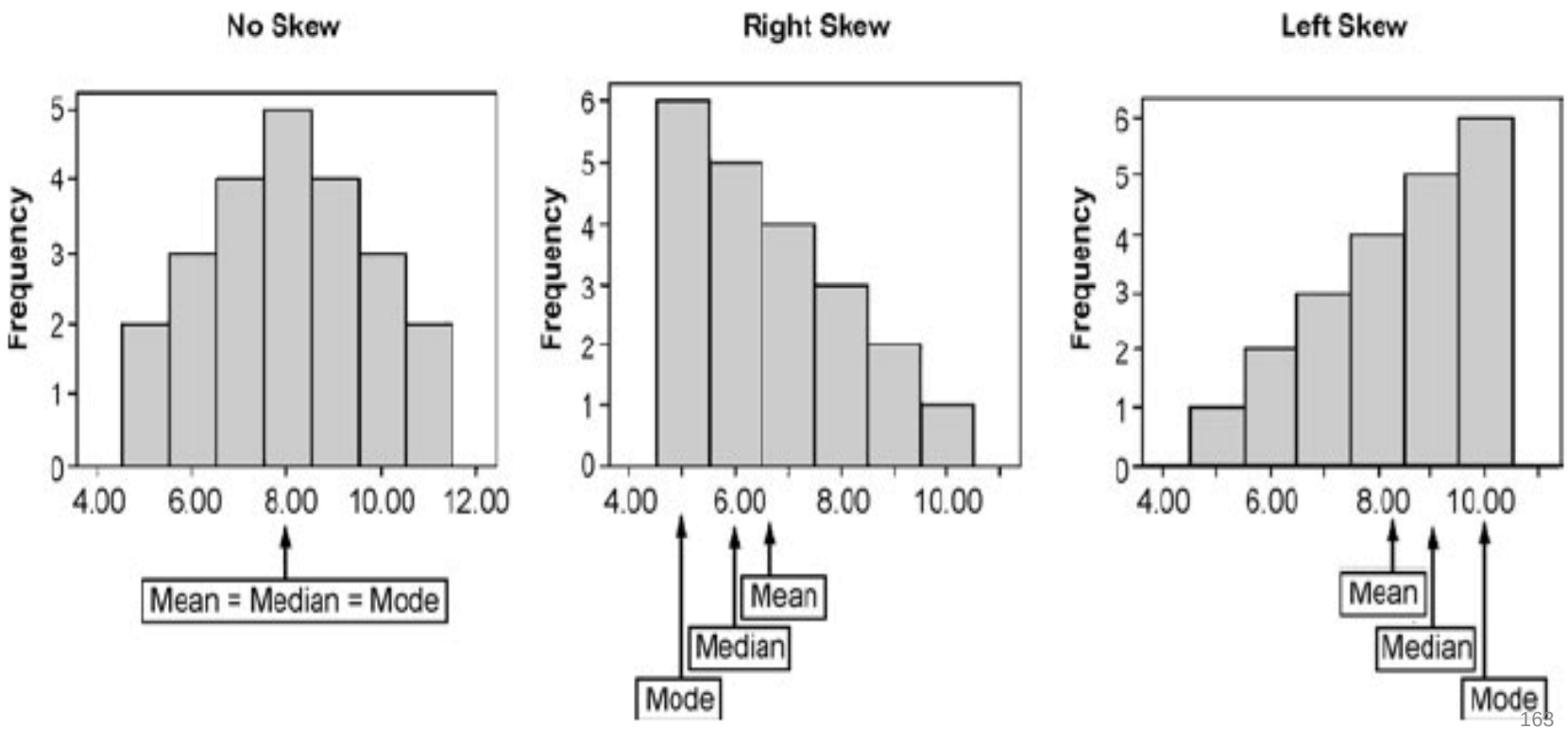
- A value of skewness > 0 indicates positive skewness.
- A value of skewness < 0 indicates negative skewness.

Relationship between the Arithmetic Mean and the Median

- In many samples, the relationship between the arithmetic mean and the sample median can be used to assess the symmetry of a distribution.
- For symmetric distributions, the arithmetic mean is approximately the same as the median.
- For positively skewed distributions, the arithmetic mean tends to be larger than the median.
- For negatively skewed distributions, the arithmetic mean tends to be smaller than the median.

Histograms Illustrating Skewness

Consider the three distributions shown in the figure . Given that the histograms represent *frequency counts*, the data can be easily re-created and entered into a statistical package.



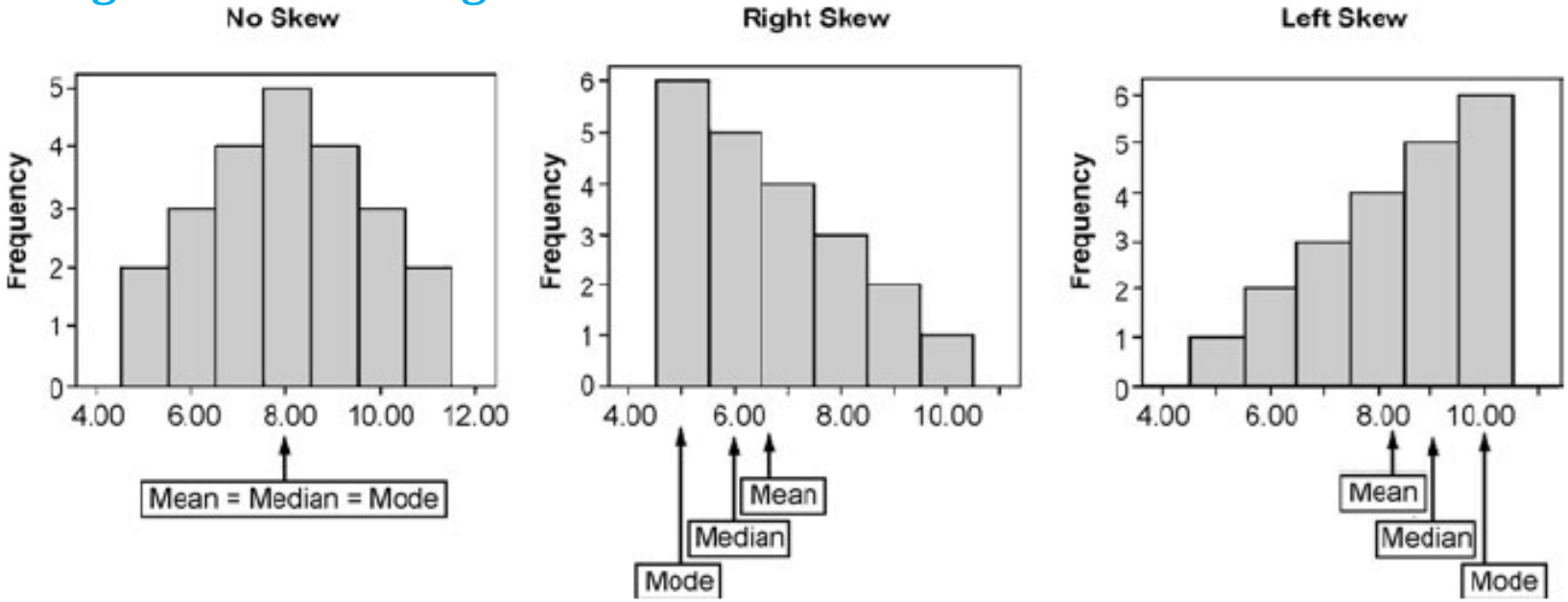
Cont'd

Example: observation of the "*No Skew*" distribution would yield the following data:

5, 5, 6, 6, 6, 7, 7, 7, 7, 8, 8, 8, 8, 8, 9, 9, 9, 9, 10, 10, 10, 11, 11.

Values can be obtained from the skewed distributions in a similar fashion.

Histograms Illustrating Skewness



	No Skew	Right Skew	Left Skew
Mean	8.0000	6.6667	8.3333
Median	8.0000	6.0000	9.0000
Mode	8.00	5.00	10.00

The descriptive statistics for these three distributions are given in the following table.

Measures of Spread or Dispersion

- Other terms used synonymously with dispersion include *variation, spread,* and *scatter*.
- The *dispersion* of a set of observations refers to the variety that they exhibit.
- A ***measure of dispersion*** conveys information regarding the amount of variability present in a set of data.
- If all the values are the same, then there is no dispersion.

Cont'd

- If the values are not all the same, then the dispersion is present in the data.
- The amount of dispersion may be small when the values, though different, are close together.
- The following Figure shows the frequency polygons for two populations that have equal means but different amounts of variability.

Measures of Spread or Dispersion

- Population B, which is more variable than population A, is more spread out.
- If the values are widely scattered, the dispersion is greater.

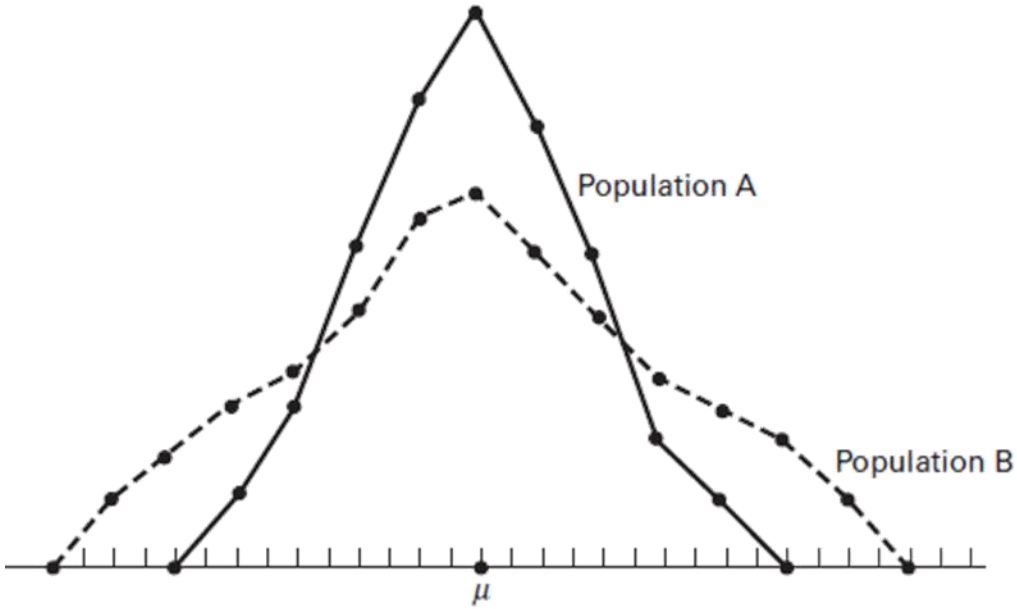


Figure: Two frequency distributions with equal means but different amounts of dispersion.

Measures of Spread or Dispersion

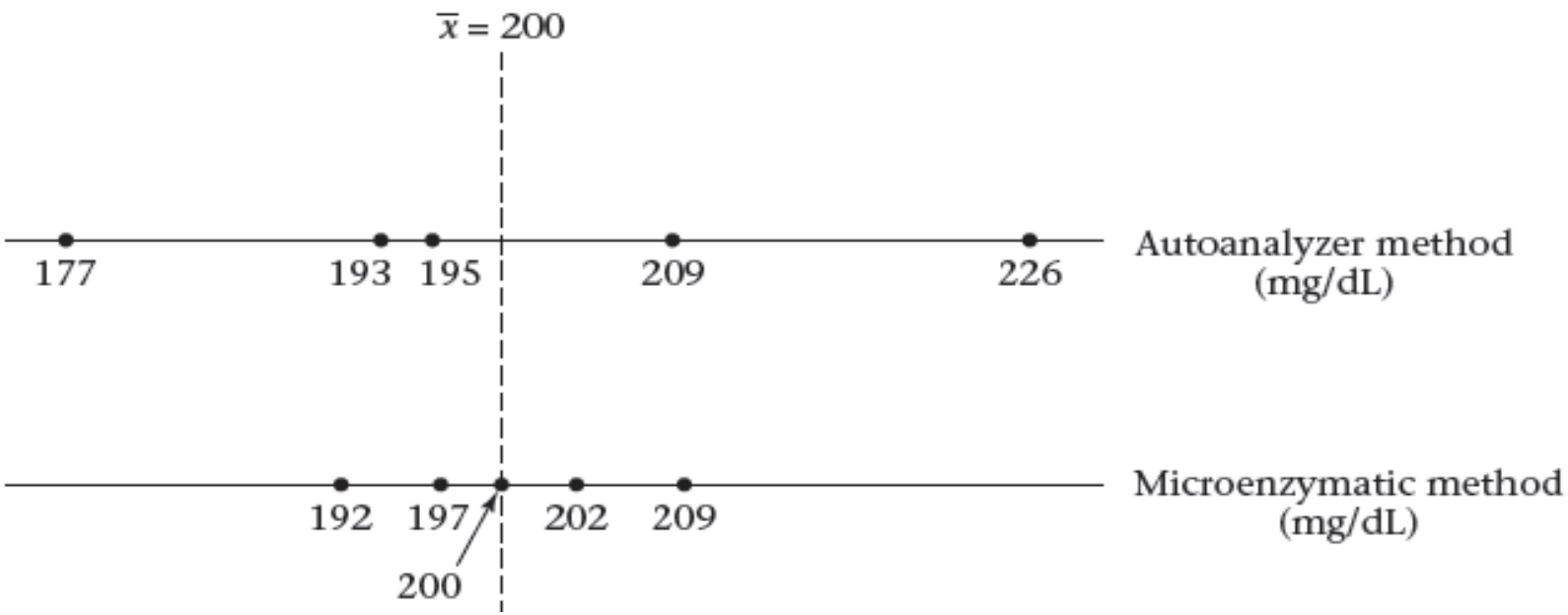
- The following figure represents two samples of cholesterol measurements, each on the same person, but using different measurement techniques.

- The arithmetic means for both samples are same, i.e., 200 mg/dL.

Measures of Spread or Dispersion

- Visually, however, the two samples appear radically different.
- This difference lies in the greater variability, or spread, of the Autoanalyzer method relative to the Microenzymatic method.

Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



The Range

- Several different measures can be used to describe the variability of a sample.
- Perhaps the simplest measure is the range.

"The ***range*** is the difference between the largest and smallest observations/values in a sample".

- If we denote the range by $\boldsymbol{R}$, the largest value by $\boldsymbol{x_L}$, and the smallest value by $\boldsymbol{x_S}$, then we can compute the range as follows:

$$\boldsymbol{R} = \boldsymbol{x_L} - \boldsymbol{x_S}$$

The Range

Example-1: Find the range in the sample of birthweights given in the table.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

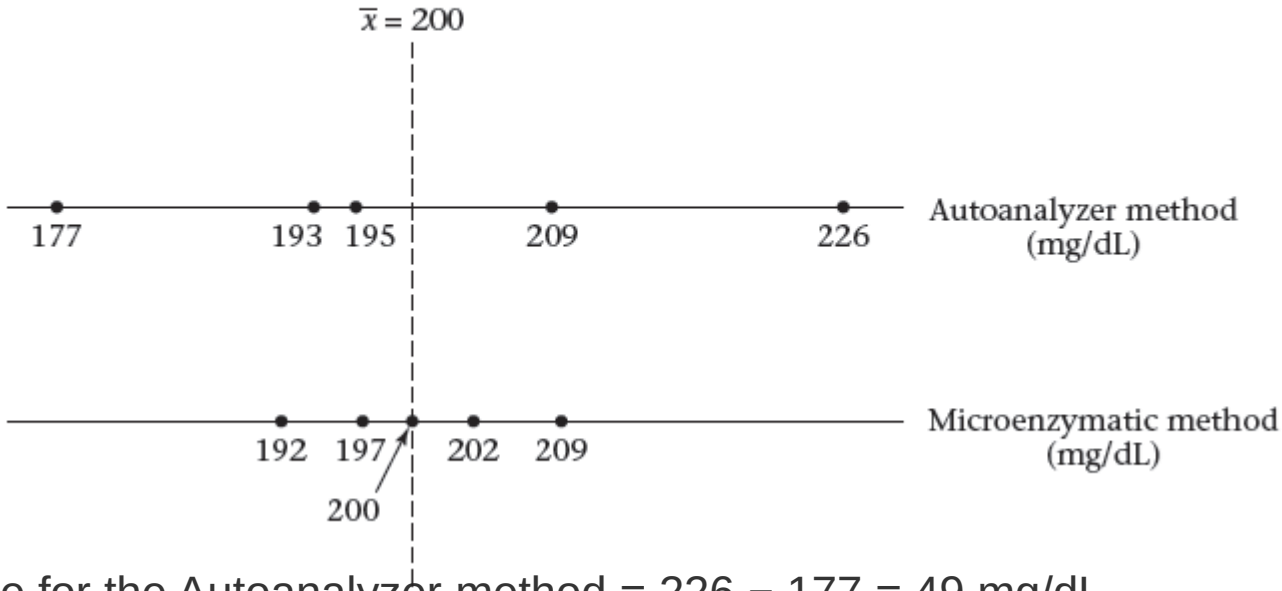
Solution: R = ? $x_L = 4146$, $x_S = 2069$

$R = x_L - x_S$, $R = 4146 - 2069$, ***R* = 2077 g**

The Range

Example-2: Compute the ranges for the Autoanalyzer- and Microenzymatic-method data in the figure and compare the variability of the two methods.

Two samples of cholesterol measurements on a given person using the Autoanalyzer and Microenzymatic measurement methods



Solution:

- The range for the Autoanalyzer method = 226 − 177 = 49 mg/dL.
- The range for the Microenzymatic method = 209 - 192 = 17 mg/dL.
- The Autoanalyzer method clearly seems more variable.

The Range

- Example-3: Compute the range of the ages of the sample subjects in the table.
- Table: Sample of 10 Ages Drawn from the Ages population

Random Number	Sample Subject Number	Age
137	1	43
114	2	66
155	3	61
183	4	64
185	5	65
028	6	38
085	7	59
181	8	57
018	9	57
164	10	50

Solution:

- The youngest subject in the sample is 38 years old and the oldest is 66 years old.
- The range to be R = 66 - 38 = 28 years.

The Range

- The usefulness of the range is limited.
- The fact that it takes into account only two values causes it to be a ***poor measure of dispersion***.
- The main advantage in using the range is the ***simplicity*** of its computation.
- Since the range, expressed as a single measure, imparts minimal information about a data set and therefore, is of ***limited use***.

Cont'd

- It is often preferable to express the range as a number pair, $[x_S, x_L]$, in which x_S and x_L are the smallest and largest values in the data set, respectively.
- One disadvantage of the range is very sensitive to extreme observations.
- Another disadvantage of the range is that it depends on the sample size (n). that is, the larger n is, the larger the range tends to be.

Range

Merits:

- **It is simple to compute and understand.**
- **It gives a rough but quick answer**

Limitation:

1. It is not a satisfactory measure as it is based only on two extreme values, ignoring the distribution of all other observations within the extremes. These extreme values vary from study to study, depending upon the size and nature of sample and type of study.

Variance & Standard deviation

- Karl Pearson introduced the concept of Standard Deviation in 1893.
- The standard deviation is a statistic that tells us how tightly all the values are clustered around the mean in a set of data.

The mean of the squares of the deviations of every observation from their mean is a measure of spread and is called the variance. The standard deviation is the square root of the variance.

$$Variance = \frac{\sum (x - \bar{x})^2}{n-1}$$

$$S.D.(\sigma) = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

It is computed as the root of average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3 the mean is 2 and the standard deviation is:

$$\frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = .667$$

$$SD = \sqrt{0.667} = 0.44$$

The Variance (s^2)

The ***variance*** is the measure of dispersion relative to the scatter of the values about their mean.

The variance or sample variance, is defined as follows:

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

where;

s^2 = variance

x = sample value

$\bar{x}$ = sample mean or arithmetic mean

n = sample size

The Variance (s^2)

Example: A simple random sample of 10 subjects from the population of subjects represented in the table. Compute the variance of the ages of the subjects from the sample.

Table: Sample of 10 Ages Drawn from the Ages of a population

Random Number	Sample Subject Number	Age
137	1	43
114	2	66
155	3	61
183	4	64
185	5	65
028	6	38
085	7	59
181	8	57
018	9	57
164	10	50

Solution:

$$s^2 = \frac{(43-56)^2 + (66-56)^2 + \cdots + (50-56)^2}{9}$$

$$s^2 = \frac{810}{9} = 90$$

The Variance of a Finite Population (σ^2)

The ***variance*** of a finite population of *N* values is defined as follows:

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

where;

σ^2 = finite population variance

x = sample value

μ = population mean or arithmetic mean

N = population size

The Standard Deviation (s)

- The variance represents squared units and, therefore, is not an appropriate measure of dispersion in terms of the original units.

- The ***standard deviation*** is the square root of the variance.

- In general, the standard deviation of a sample is given by;

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

where;

s = standard deviation

s^2 = variance

x = sample value

$\bar{x}$ = sample mean or arithmetic mean

n = sample size

The Standard Deviation (s)

- The standard deviation often abbreviated as SD or sd.
- The standard deviation measures the dispersion or spread about the mean.
- The bigger value of s shows that the more variability present in the data.
- The standard deviation can equal to zero if there is no spread.
- The units of standard deviation are the same as the units of the data.

The Standard Deviation of a Finite Population (σ)

The ***standard deviation*** of a finite population of *N* values is defined as follows:

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

where;

σ = standard deviation of a finite population

σ^2 = finite population variance

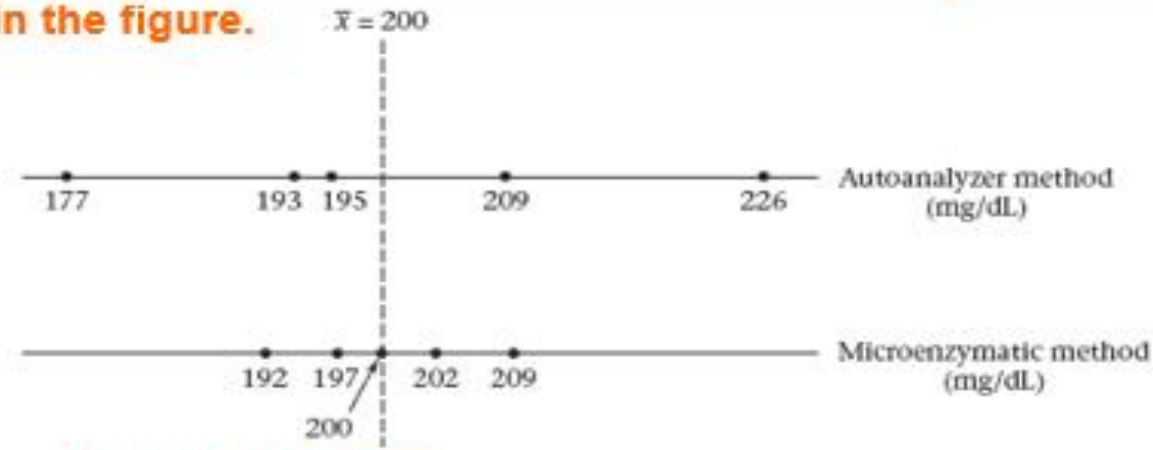
x = sample value

μ = population mean or arithmetic mean

N = population size

The Standard Deviation (s)

Example-1: Compute the variance and standard deviation for the Autoanalyzer and Microenzymatic method data in the figure.



Solution:

Autoanalyzer Method

$$s^2 = \left[(177-200)^2 + (193-200)^2 + (195-200)^2 + (209-200)^2 + (226-200)^2\right]/4$$

$$= (529+49+25+81+676)/4 = 1360/4 = 340$$

$$s = \sqrt{340} = 18.4$$

Microenzymatic Method

$$s^2 = \left[(192-200)^2 + (197-200)^2 + (200-200)^2 + (202-200)^2 + (209-200)^2\right]/4$$

$$= (64+9+0+4+81)/4 = 158/4 = 39.5$$

$$s = \sqrt{39.5} = 6.3$$

Thus, the Autoanalyzer method has a standard deviation roughly three times as large as that of the Microenzymatic method.

The Standard Deviation (s)

Example-2: Compute the variance and standard deviation of the birthweight data in the table in both grams and ounces.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

The Variance and Standard Deviation in Grams

$$s^2 = \frac{(3265-3166.9)^2+\cdots+(2834-3166.9)^2}{19}$$

$$= 3{,}768{,}147.8/19 = 198{,}323.6 \text{ g}^2$$

$$s = 445.3 \text{ g}$$

The Standard Deviation (s)

- Example-2: continue.The Variance and Standard Deviation in Ounces

$$1 \text{ oz} = 28.35 \text{ g} \quad \text{or} \quad y_i = \frac{1}{28.35} x_i$$

$$\text{Thus} \quad s^2(\text{oz}) = \frac{1}{28.35^2} s^2(\text{g}) = 246.8 \text{ oz}^2$$

$$s(\text{oz}) = \frac{1}{28.35} s(\text{g}) = 15.7 \text{ oz}$$

- Thus, if the sample points change in scale by a factor of c, the variance changes by a factor of c2 and the standard deviation changes by a factor of c.
- This relationship is the main reason why the standard deviation is more often used than the variance as a measure of spread.
- The standard deviation and the arithmetic mean are in the same units, whereas the variance and the arithmetic mean are not.

The Coefficient of Variation (CV)

- The standard deviation is useful as a measure of variation within a given set of data.

- When one desires to compare the dispersion in two sets of data, however, comparing the two standard deviations may lead to misleading results.

- It may be that the two variables involved are measured in different units.
- **For example:** we may wish to know, for a certain population, whether serum cholesterol levels, measured in milligrams per 100 ml, are more variable than measured in pounds.

Cont'd

- Even the same unit of measurement is used, the two means may be quite different.
- **For example:** If we compare the standard deviation of weights of first-grade children with the standard deviation of weights of high school freshmen, we may find that the latter standard deviation is numerically larger than the former, because the weights themselves are larger, not because the dispersion is greater.

- Also, It is useful to relate the arithmetic mean and the standard deviation to each other.

- **For example:** a standard deviation of 10 means something different conceptually if the arithmetic mean is 10 than if it is 1000.

The Coefficient of Variation (CV)

- What is needed in situations like these is a measure of relative variation rather than absolute variation.

- Such a measure is found in the *coefficient of variation,* which expresses the standard deviation as a percentage of the mean.

- The ***coefficient of variation (CV)*** is defined by;

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

- This measure remains the same regardless of what units are used because if the units change by a factor *c*, then both the mean and standard deviation change by the factor *c;* the *CV*, which is the ratio between them, remains unchanged.

- The units of standard deviation s and mean $\bar{x}$ are same, therefore the ***C.V*** is unit-less.

- Hence, the CV is a measure that is independent of the unit or scale of measurement.

The Coefficient of Variation (CV)

Example-1: Suppose two samples of human males yield the results shown in the table. Find that which is more variable, the weights of the 25-year-olds or the weights of the 11-year-olds.

	Sample 1	Sample 2
Age	25 years	11 years
Mean weight	145 pounds	80 pounds
Standard deviation	10 pounds	10 pounds

CV for the 25-year-olds

$$\text{C.V.} = \frac{10}{145}(100) = 6.9\%$$

CV for the 11-year-olds

$$\text{C.V.} = \frac{10}{80}(100) = 12.5\%$$

- A comparison of the standard deviations show that the two samples possess equal variability.
- However, It is clear from this example that variation is much higher in the sample of 11-year-olds than in the sample of 25-year-olds.

The Coefficient of Variation (CV)

Example-2: Compute the coefficient of variation for the data in the table when the birthweights are expressed in either grams or ounces.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

When the Data is Expressed in Grams

$$CV = 100\% \times (s/\bar{x}) = 100\% \times (445.3\,\text{g}/3166.9\,\text{g}) = 14.1\%$$

When the Data is Expressed in Ounces

$$CV = 100\% \times (15.7\ \text{oz}/111.71\ \text{oz}) = 14.1\%$$

Percentiles

- Percentiles are values that divide a set of observations into 100 equal parts, so there are total 99 percentiles.

- Percentiles are used for location of data on the horizontal axis.

- Percentiles have the advantage over the range of being ***less sensitive to outliers*** and of ***not being greatly affected by the sample size (n)***.

- A *Percentile* is define as follows,

"Given a set of *n* observations, x_1, x_2, x_3, … , x_n, the *p*th percentile *P* is the value of *X* such that *p* percent or less of the observations are less than *P* and (100 - *p*) percent or less of the observations are greater than *P*".

Percentiles

- Subscripts on P serve to distinguish one percentile from another.
- **For example:** The 10th percentile is designated P_{10}, the 70th is designated $P_{70}$and so on.
- 70 Percentile means that 70% values lie below the value at P_{70} while 30% of the values lie above the value at P_{70}.
- The 50th percentile is the median and is designated P_{50}.
- The median divides the lower 50% values and the higher 50% values in a data set.

Percentiles

Frequently used percentiles are;

- ***Quartiles*** divide the data set into four equal parts.
- **Example:** 25th, 50th, and 75th percentiles.

- ***Quintiles*** divide the data set into five equal parts.
- **Example:** 20th, 40th, 60th, and 80th percentiles.

- ***Deciles*** divide the data set into 10 equal parts.
- **Example:**10th, 20th, . . . ,90th percentiles.

Quartiles

- ***Quartiles*** can divide the data set into four equal parts.

- The 25th percentile is often referred to as the ***first quartile*** or ***lower quartile*** and denoted as Q_1. It contains one-quarter of the data.

- The 50th percentile (the median) is referred to as the ***second or middle quartile*** and written as Q_2. It marks the point with half of the data.

- The 75th percentile is referred to as the ***third quartile*** or ***upper quartile*** and denoted as Q_3. It contains three-quarters of the data.

Cont'd

- The quartiles for a set of data are calculated using the following formulas;

$$Q_1 = \frac{n+1}{4} \text{ th} \quad \text{ordered observation}$$

$$Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2} \text{ th} \quad \text{ordered observation}$$

$$Q_3 = \frac{3(n+1)}{4} \text{ th} \quad \text{ordered observation}$$

Cont'd

- If (n+1)/4 is an integer, then 25th percentile will be (n+1)/4 th smallest observation.

- If (n+1)/4 is not an integer, then 25th percentile is a value between two observations. For example, n=22, thus (n+1)/4 will be 5.75. The 25th percentile will be three fourth of the way between the 5th and sixth smallest observation. To find it we sum 5th smallest observation with 0.75 of the difference of 5th and 6th smallest observations.

Percentiles

Example-1: Compute the 10th and 90th percentiles for the birthweight data in the table.

Sample of birthweights (g) of live-born infants born at a private hospital in San Diego, California, during a 1-week period

i	x_i	i	x_i	i	x_i	i	x_i
1	3265	6	3323	11	2581	16	2759
2	3260	7	3649	12	2841	17	3248
3	3245	8	3200	13	3609	18	3314
4	3484	9	3031	14	2838	19	3101
5	4146	10	2069	15	3541	20	2834

Solution: The first step is to arrange data from the smallest value to the largest value.

- The $np/100 = 20 \times 0.1 = 2$, and $np/100 = 20 \times 0.9 = 18$ are integers.

Cont'd

- Therefore , the 10th and 90th percentiles are defined by;
- 10th percentile: average of the second and third largest values = (2581 + 2759)/2 = ***2670 g***
- 90th percentile: average of the 18th and 19th largest values = (3609 + 3649)/2 = ***3629 g***
- We would estimate that 80% of birthweights will fall between 2670 g and 3629 g, which gives an overall impression of the spread of the distribution.

Percentiles

Example-2: Compute the 20th percentile for the white-blood-count data in the table.

Sample of admission white-blood counts ($\times$ 1000) for all patients entering a hospital in Allentown, PA, on a given day

i	x_i	i	x_i
1	7	6	3
2	35	7	10
3	5	8	12
4	9	9	8
5	8		

Solution:

The first step is to arrange data from the smallest value to the largest value.

- The $np/100 = 9 \times 0.2 = 1.8$ is not an integer.
- Therefore, the 20th percentile is defined by the (1 + 1)th largest value.
- Hence, the 20th percentile is the second largest value = **5000**.

Percentiles

- The pth percentile can be computed as;

(1) The $(k + 1)$th largest sample point if $np/100$ is not an integer (where k is the largest integer less than $np/100$)

(2) The average of the $(np/100)$th and $(np/100 + 1)$th largest observations if $np/100$ is an integer.

Interquartile Range (IQR)

- The range provides a simple measure of the variability present in a set of data.
- A disadvantage of the range is the fact that it is computed from only two values, the largest and the smallest.
- A similar measure that reflects the variability among the middle 50 percent of the observations in a data set is the ***interquartile range***.
- The ***interquartile range (IQR)*** is the difference between the third and first quartiles.

$$\text{IQR} = Q_3 - Q_1$$

Interquartile Range

- A large IQR indicates a large amount of variability among the middle 50 percent of the relevant observations.

- A small IQR indicates a small amount of variability among the relevant observations.

- It is more informative to compare the interquartile range with the range for the entire data set.
- A comparison may be made by forming the ratio of the IQR to the range (R) and multiplying by 100. i.e., 100(IQR/R) tells us what percent the IQR is of the overall range.

Outliers or Outlying Values

An outlier or outlying value is a value x such that either

1) x > upper quartile (Q3) + 1.5 × (upper quartile (Q3) − lower quartile (Q1)) or

2) x < lower quartile (Q1) − 1.5 × (upper quartile (Q3) − lower quartile (Q1))

Outliers are unusually large and unusually small values of x in a data set.

Outliers or Outlying Values

Example: For the data shown in Table, determine how large or how small a value would have to be in order to be considered an outlier

GRF Measurements When Trotting of 20 Dogs with a Lame Ligament

14.6	24.3	24.9	27.0	27.2	27.4	28.2	28.8	29.9	30.7
31.5	31.6	32.3	32.8	33.3	33.6	34.3	36.9	38.3	44.0

Solution

Firstly, we need to compute the values of Q_1, Q_3, and IQR.

- Q1 = 27.25
- Q3 = 33.525
- IRQ = Q3 – Q1 = 33.525 – 27.525 = 6.275

The upper outlier can be computed as;

$$x > Q_3 + 1.5(Q_3 - Q_1)$$

$$x > 33.525 + 1.5(6.275) = \mathbf{42.9375}$$

Continue on next slide…

Outliers or Outlying Values

The lower outlier can be computed as;

$$x < Q_1 - 1.5(Q_3 - Q_1)$$

$$x < 27.25 - 1.5(6.275) = \mathbf{17.8375}$$

- Hence for the data given in Table, an observed value smaller than **17.8375** or larger than **42.9375** would be considered an outlier.

Kurtosis

- Just as we may describe a distribution in terms of skewness, we may describe a distribution in terms of kurtosis.

- ***Kurtosis*** **is a measure of the degree to which a distribution is "peaked" or flat in comparison to a normal distribution whose graph is characterized by a bell-shaped appearance.**

- A normal, or bell-shaped distribution, is said to be ***mesokurtic***.

- A distribution may possess an excessive proportion of observations in its tails, so that its graph exhibits a flattened appearance. Such a distribution is said to be ***platykurtic***.

- A distribution may possess a smaller proportion of observations in its tails, so that its graph exhibits a more peaked appearance. Such a distribution is said to be ***leptokurtic***.

Kurtosis

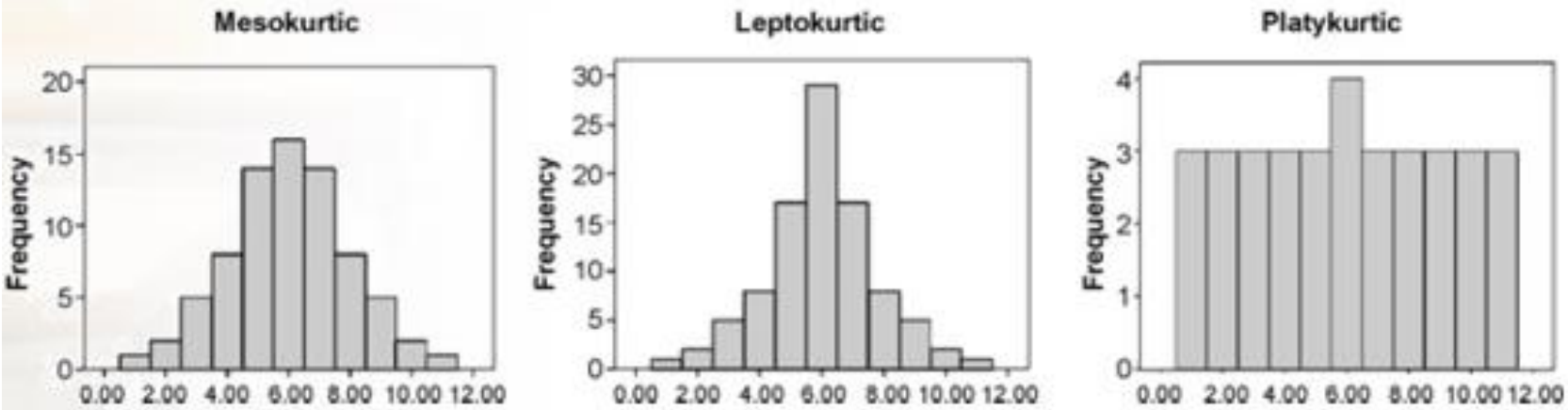
Kurtosis can be expressed as;

$$Kurtosis = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2} - 3 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{(n-1)^2 s^4} - 3$$

- A perfectly mesokurtic distribution has a kurtosis measure of 3 based on the equation.
- Most computer algorithms reduce the measure by 3, as is done in Equation, so that the kurtosis measure of a mesokurtic distribution will be equal to 0.
- A leptokurtic distribution will have a kurtosis measure > 3.
- A platykurtic distribution will have a kurtosis measure < 3.

Kurtosis

Graphs of distributions representing the three types of kurtosis are shown in the figure.



The descriptive statistics of these three distributions are shown in the following table.

	Mesokurtic	Leptokurtic	Platykurtic
Mean	6.0000	6.0000	6.0000
Median	6.0000	6.0000	6.0000
Mode	6.00	6.00	6.00
Skewness	.000	.608	−1.158

Introduction to Demography

LEARNING OBJECTIVES

At the end of this chapter, the students will be able to:

- 1. Define and understand the concepts of demographic statistics
- 2. Identify the different methods of data collection for demographic studies
- 3. Understand different ratios used to describe demographic data
- 4. Understand fertility and mortality measures
- 5. Understand methods of population projection and computation of doubling time
- 6. Understand and compute the different indices relating to hospitals

Introduction

- Demography is a science that studies human population with respect to size, distribution, composition, social mobility and its variation with respect to all the above features and the causes of such variation and the effect of all these on health, social, ethical, and economic conditions.

- **Size:** is the number of persons in the population at a given time.
  - Example: The size of Ethiopian population in 2024 is about 110 million.
  - **Distribution**: is the arrangement of the population in the territory of the nation in geographical, residential area, climatic zone, etc.
  - Example: Distribution of Ethiopian population by Zone

Cont'd

- **Composition (Structure):** is the distribution of a population into its various groupings mainly by age and sex.

- Example: The age and sex distribution of the Ethiopian population refers to the number (%) of the population falling in each age group (at each age) by sex.

- **Change**: refers to the increase or decline of the total population or its components. The components of change are birth, death, and migration.

Cont'd

- **Demographic Statistics** is the application of statistics to the study of human populations. It analyzes essential demographic variables, including:
  - Fertility
  - Mortality
  - Migration
- The **source of population variations** is studied with respect to these variables.
- To understand these variables, we must also study their relationship with socio-economic factors, such as:
  - ✓ Place of Birth
  - ✓ Religion
  - ✓ Occupation
  - ✓ Education
  - ✓ Marital Status

Sources of Demographic Data

Demographic information is acquired through two primary methods:

1. **Complete Enumerations (Census)**
   - Involves collecting data from **every inhabitant** in a specified area.
   - Provides **comprehensive** and **accurate** population data.
2. **Sample Surveys**
   - Involves **interviewing a sample** of the population to represent the whole.
   - Increasingly used due to its **cost-effectiveness** and **efficiency**.

Cont'd

Additionally, demographic data is collected through the **recording of vital events** over time, such as:

- Births
- Deaths
- Marriages
- Divorces

This continuous recording of vital events helps complete the data collected from censuses and sample surveys.

THE CENSUS

- **Definition:** A census refers to a nation-wide count of the population, obtained by directly canvassing each person or household.
- **Two Main Schemes for Enumeration:**
  - **De Jure:** Enumeration is done based on the **usual or legal place of residence.**
    - Excludes temporary residents and visitors.
    - Includes permanent residents temporarily away.
  - **De Facto:** Enumeration is based on the **actual place of residence** on the day of the census.
    - Includes temporary residents and visitors.
    - Excludes permanent residents who are away.

Cont'd

- Differences in Practice:
  - **De Facto**: Used by Britain (counts people wherever they are on the census day).
  - **De Jure**: Used by the United States (counts people by their permanent residence).

Advantages and Disadvantages of De Jure and De Facto Schemes

De Jure Scheme (Usual or Legal Place of Residence)

- Advantages:
  - Provides an accurate picture of the **permanent population**.
  - Relatively unaffected by seasonal or temporary movements.
- Disadvantages:
  - Some individuals may be **omitted** or counted **twice**.
  - Difficult to determine a person's **usual** or **legal residence** in areas with high mobility or no fixed address.
  - **Incomplete or incorrect** data for people temporarily away from home.

Cont'd

De Facto Scheme (Actual Place of Residence on Census Day)

- Advantages:
  - **Less chance of double-counting**.
  - **Reduces the chance** of omitting people from the count.
- Disadvantages:
  - Population figures may be **inflated or deflated** due to temporary residents (e.g., tourists, traveling salesmen).
  - In high migration areas, **vital statistics** can be distorted.

Conclusion:

- **Combination of both schemes** is often used depending on national needs and specific enumeration plans.

Information Collected in a Census

- Key Data Points:
  - Sex
  - Age
  - Marital Status
  - Educational Characteristics
  - Economic Characteristics
  - Place of Birth
  - Language
  - Fertility & Mortality
  - Citizenship (Nationality)
  - Living Conditions (e.g., house-ownership, type of housing)
  - Religion

Essential Features of a Census

1. **Individual Enumeration:** Each person's characteristics are recorded separately.
2. **Defined Territory:** Refers to people living within a clearly defined geographic area.
3. **Simultaneity:** Data is collected at a well-defined point in time.
4. **Regular Intervals:** Typically conducted every 10 years.
5. **Confidentiality:** Personal data collected is only used for statistical purposes.

Census Operation

Three Stages of Census Operation:

1. **Pre-enumeration:**
   - Planning and preparatory work for the census.
2. **Enumeration:**
   - Fieldwork – collection of the data from individuals and households.
3. **Post-enumeration:**
   - Editing, coding, compilation, tabulation, analysis, and publication of results.

Uses of a Census

1. **Provides a Complete and Valid Picture:**
   - A comprehensive overview of the population's composition and characteristics.
2. **Serves as a Sampling Frame:**
   - Basis for selecting samples in surveys and research.
3. **Vital Statistics:**
   - Data on fertility, mortality, and other key demographic factors.
4. **Planning and Welfare:**
   - Used for planning social welfare schemes (e.g., schools, hospitals, housing, pensions).
   - Helps determine **food requirements** and public services needs.

Common Errors in Census Data

1. **Omission and Over Enumeration:**
   - Some individuals may be missed (omission), while others might be counted more than once (over enumeration).
2. **Misreporting of Age:**
   - Memory lapses, preference for certain terminal digits, or over/under estimation can lead to inaccurate age data.
3. **Overstating Occupational Status:**
   - Respondents may exaggerate or misrepresent their occupation or employment status.
4. **Underreporting of Births:**
   - Issues with memory lapse or confusion over the reference period can lead to missing birth data.
5. **Underreporting of Deaths:**
   - Deaths, especially infant deaths, may be underreported due to memory lapse or reluctance to report.

Survey and Registration of Vital Events

Survey

- A technique based on **sampling methods** to gather specific information from a part of the population, which represents the whole.
- Conducted at a **specific moment** and **territory**.
- **Not periodic**, typically used for **in-depth study** of a problem.

Cont'd

Registration of Vital Events

- **Vital Events**: Includes **births, deaths**, and other key life events.
- Population changes are ongoing:
  - **Additions**: By births or arrivals.
  - **Reductions**: Due to deaths or people leaving.
- **Key Characteristics of Vital Registration**:
  1. **Comprehensive**: Covers all occurrences.
  2. **Compulsory by Law**: Legal requirement to report.
  3. **Central Compilation**: Data compiled at a central level.
  4. **Ongoing Process**: Continuous registration of events.

Basic demographic measures

Population Size

- The size of a population is determined by **counting** the individuals at a given time.

Key Characteristics Recorded:

- Age
- Sex
- Other demographic features

Describing and Comparing Populations:

- **Proportions and Ratios**: Calculated to describe and compare populations.

Cont'd

Combining Population Stock with Demographic Events:

- **Population Stock:** Refers to the size and structure of the population.
- **Demographic Events:** Includes births, deaths, migrations, etc.

Demographic Rates:

- By combining population stock and demographic events, we can calculate various **rates** (e.g., birth rate, death rate).

Proportions

- Proportions show the size of a sub-group of the population relative to the total population.
- They are calculated by dividing the number within the sub-group of interest by the total population

$$Proportion = \frac{Pi}{P}$$

Where:

- Pi is the sub-group of interest and P is the total population.

- A proportion will always be between 0 and 1.
- However, we can express it as a percentage by multiplying the value by 100, or per 1000 by multiplying by 1000 and so on.

Ratio

- Expresses the relation of one population sub-group to another sub-group**.**
- The numerator is **not** included in the denominator.

$$\text{Ratio} = \frac{P1}{P2}$$

Where:

- P1 is the size of one population group and P2 is the size of the other population group.

- When the population at risk is unavailable, we use a ratio than a rate.

Ratio: Examples

- Sex ratio at birth (male births per 100 female births)
- Child-woman ratio (number of children aged 0 to 4 years divided by the number of women aged 15 to 44 years or 15 to 49 years)
- Dependency ratio (population aged under 15 and over 64 divided by the population aged 15 to 64).

Rate

- Number of events of a specific type in a given time period divided by number of population at risk at that specific time period.
    - Occurrence / Exposure
- Population at risk : population that could potentially experience that type of event in the given time period.
- Strictly speaking, a rate is a measure of change in a population over a specified time period (usually a year).

Rate

$$\text{Rate} = \frac{\text{Number of Occurrences}}{\text{Person-years of Exposure to the Risk of Occurrence}}$$

- The population at risk is approximated by:
  - 'Person-years of exposure' or
  - most conviently 'mid year population'

Rate…..

- When a rate is applied to a certain time period(duration):

$$\text{Rate}[0,T] = \frac{\text{Number of Occurrences between Time 0 and } T}{\text{Person-years Lived in the Population between Time 0 and } T}$$

- When an interval is greater than 1 year, we evaluate exposure in terms of person years at risk.
- Time should always be included.

Person-years at risk

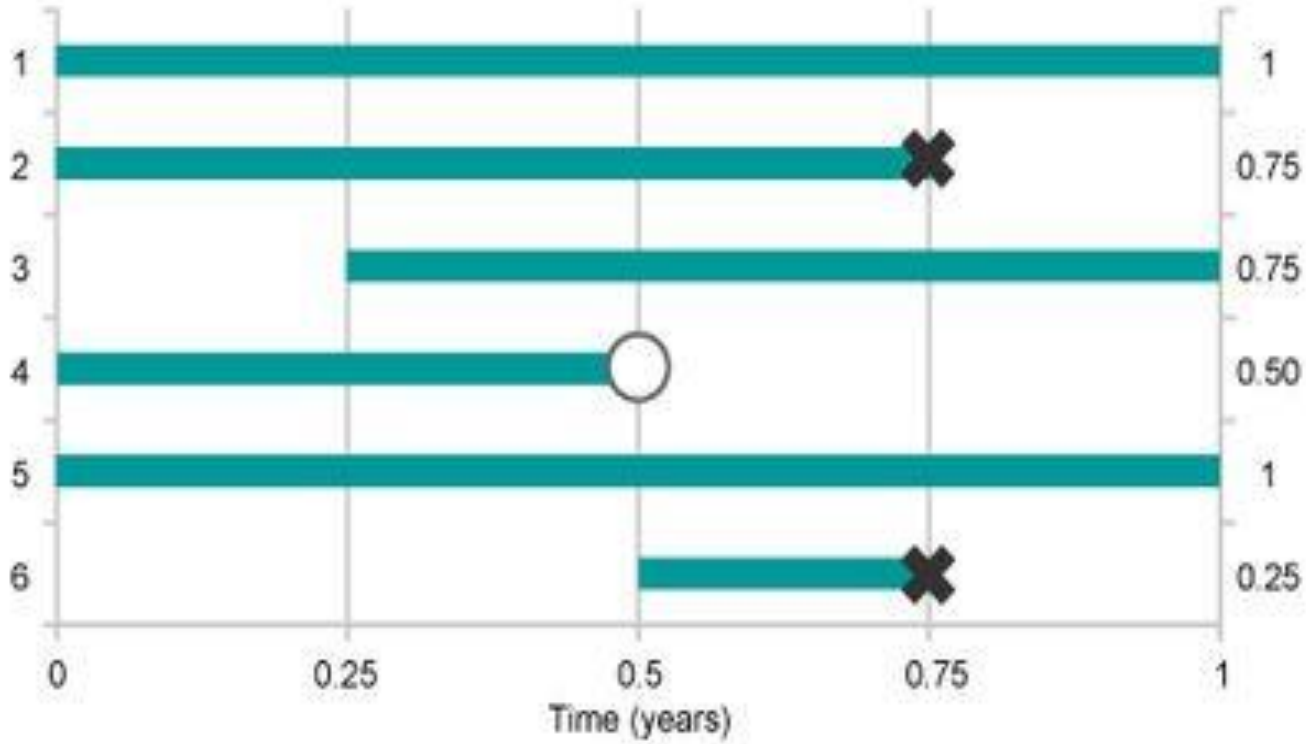
- When calculating rates we need to think about who is at risk and how long they are at risk for.
- The time spent during the specified period when the event of interest could happen to them is called person-years at risk.
- This can be calculated as any unit of time - days, weeks, months etc.

Exercise

- The chart in the next slide shows a population of 6 people and what happens to them over the course of one year.
  - Individual 1 was present at the beginning and not die or leave the population during the study.
  - Individual 2 was present at the beginning and die three-quarters of the way through the year.
  - Individual 3 was not present at the beginning and not die or leave the population during the study.

Exercise

- The chart in the next slide shows a population of 6 people and what happens to them over the course of one year.
  - Individual 4 was present at the beginning and lost to follow up after 6 months
  - Individual 5 was present at the beginning and not die or leave the population during the study
  - Individual 6 was enters the study halfway through the year and die at three-quarters of the way through the year.



Now we know how long each member of the population was at risk of dying.

- Based on the above information calculate the total person-years at risk for this population.

Types of Demographic Rates

- **Crude rates:** summary rates for total population.
  - But populations vary in composition (most notably, age) differences in crude rates are difficult to interpret.
- **Specific rates:** are useful because they are unaffected by differences in group composition.
  - compares homogeneous groups
  - such specific rates are useful for epidemiological and public health purposes
- **Adjusted rates:** eliminate effects of population composition by standardizing to a particular population.
  - Differences in the composition of population is "removed" allowing unbiased comparison

Rates, Ratios and Proportions

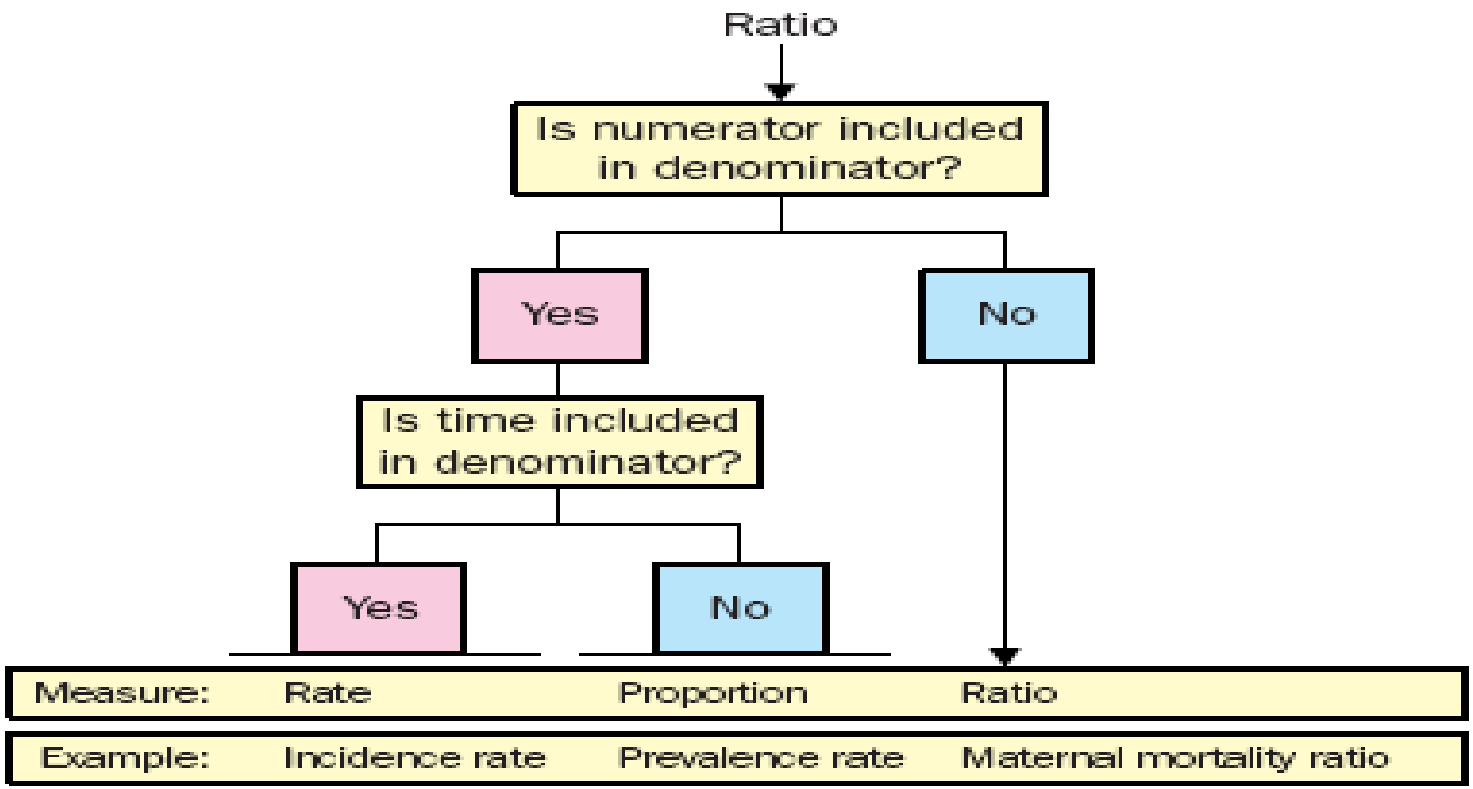


Figure 3: **Algorithm for distinguishing rates, proportions, and ratios**

Probability

- A probability expresses the chance that an event will occur within a defined time period.

- Is more meaningful in a cohort analysis and for a specific age or time interval

$$\text{Probability} = \frac{\text{Number of Occurrences}}{\text{Number of Preceding Events or Trials}}$$

Probability

- Is similar to a rate, but their important difference is;
  - The denominator is composed of all those persons in the population at the beginning of the period of observation
- Probabilities are based on initial population
  - Ex. the probability of dying at age 50 in a cohort is based on the number of people who celebrated their 50th birth day.

Period & Cohort measures

Period (cross-sectional) data

- Period data are events which occur during a specified period of time, often either one year or five.
- Period data contains the demographic experiences of people of many different ages (in other words, people belonging to many different birth cohorts).
- Many of the commonly used demographic rates are based on period data, such as crude birth rates.

Advantages of using period data

- Period data are easy to collect.
- Period analysis can use the most up-to-date information and thus be used to measure current trends.
- Period analysis is useful for when we wish to show rapid fluctuations which are a response to social, political or cultural events.

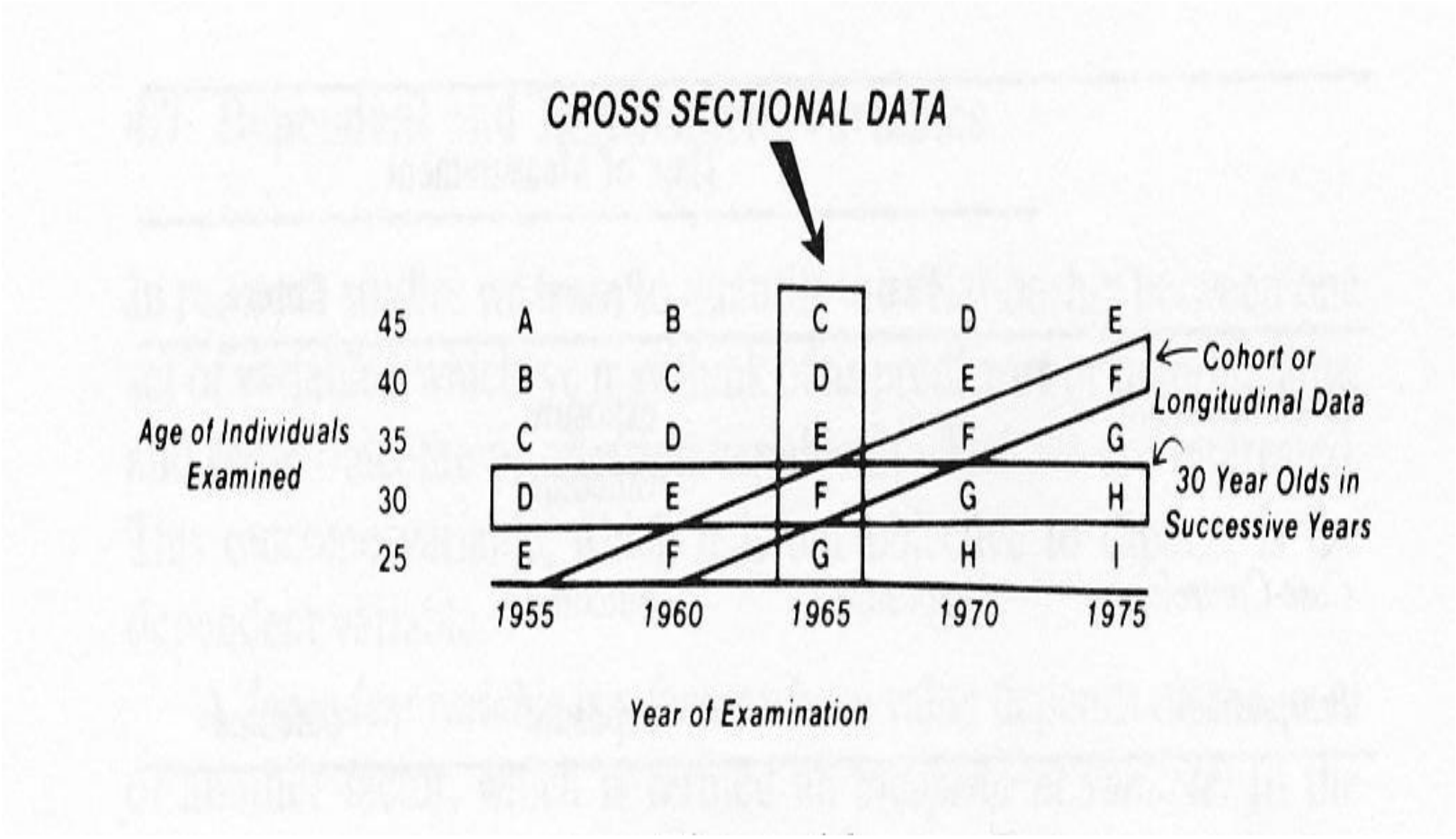
Disadvantages of using period data

- The summary rates calculated from period data do not represent real life experiences.
  - Over the course of an individual's life the rate of occurrence of demographic events will change.
- Period analysis is subject to tempo effects.

Cohort(Longitudinal) data

- A certain cohort is defined by a shared demographic event which occurred during a specified period of time.
  - This will most commonly refer to births during a set period of time, but may also refer to marriages, onset of puberty, migration, first birth etc.
- Cohort data is collected on members of the cohort longitudinally.
- Cohort data have the benefit of telling us about a group of people's real life experiences.
- However, they are often not feasible, as data collection takes decades.

Period (Cross-sectional) vs. Cohort (Longitudinal) Data



- In period and cohort data analysis we need to consider the following effects:

  »Cohort effects,

  »period effects and

  » age effects

Cohort effects

- A cohort effect refers to a distinctive formative experience that members of a cohort share that lasts and marks them throughout their lives.
- They are attributable to the singular experience of each cohort antecedent to the observation of the outcome under investigation.
  - Ex. Patterns of disease frequency due to an exposure occurring to a group of people at about the same time in their lives, such as famine, or war.

Period effects

- Period effects are environmental effects common to a given time period.
- Generations exposed to the same time period, overlap in experience but at different ages.
- Patterns of disease which are due to the period of time in which the people live.
  - e.g. epidemics, seasonal influences on disease.

Age effects

- Age effects are intrinsic effects depending on the differentiation and maturation of the individual.
- Patterns of disease which are due to the aging of the population.
  - Example: Rates of coronary artery disease increase by age.

Differences among 3 effects

- The three effects are interrelated.
- Once two are specified the third is determined by the other two
  - Period – Age = Cohort
- Cohort effects are often hard to detect because analysis is usually by age in a specific period.

- The differences between period and cohort measures can be easily conceptualized graphically.
- We will now see how this can be done by using **Lexis diagrams**.

The lexis diagram

Lexis diagram

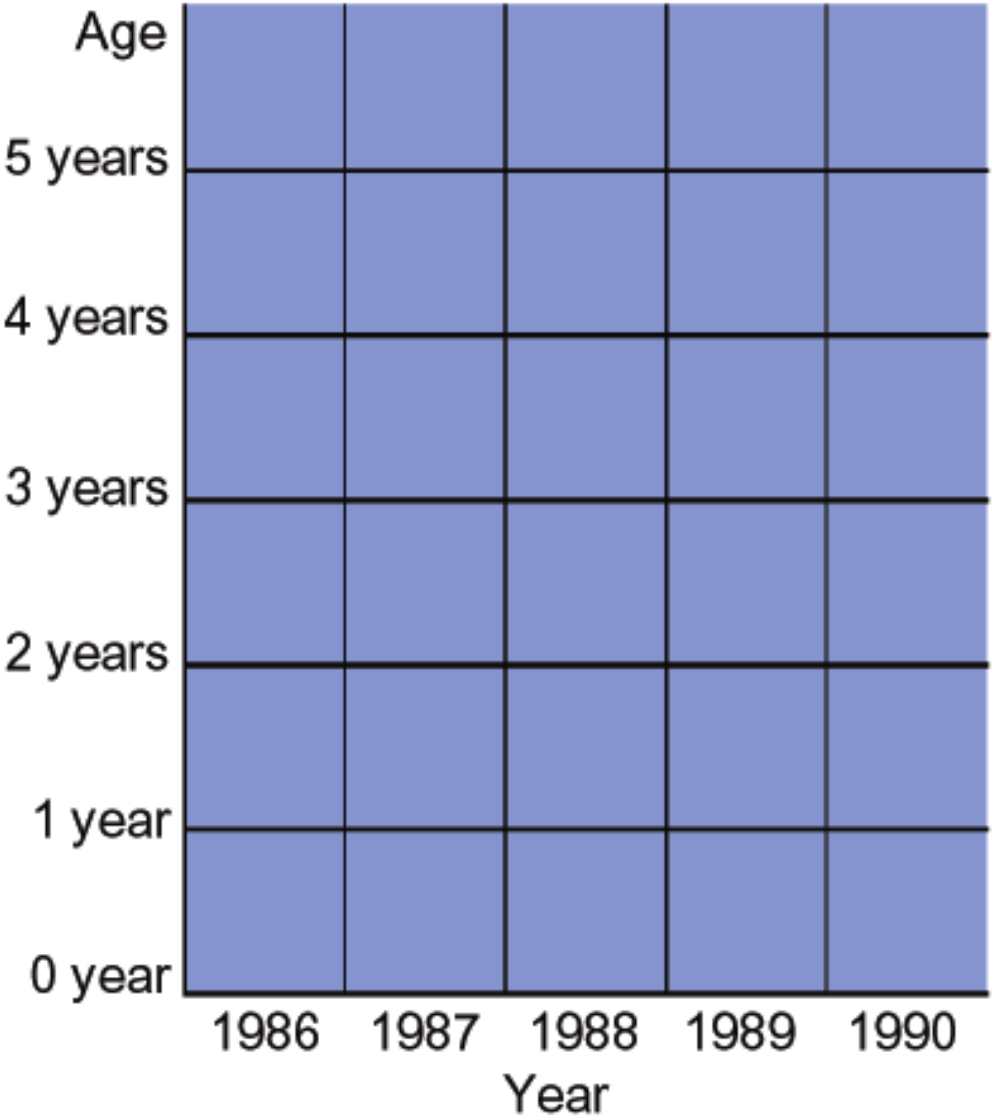
- Lexis diagrams provide a graphic representation of the relationship between demographic events in time and persons at risk.

- Age, period and cohort dynamics are illustrated by Lexis diagram.

Construction of Lexis diagram

- Every demographic event is characterized by two numbers:
  - the time (e.g., year) at which it occurs and
  - the age (or other duration measure) of the person to whom it occurs.

Construction



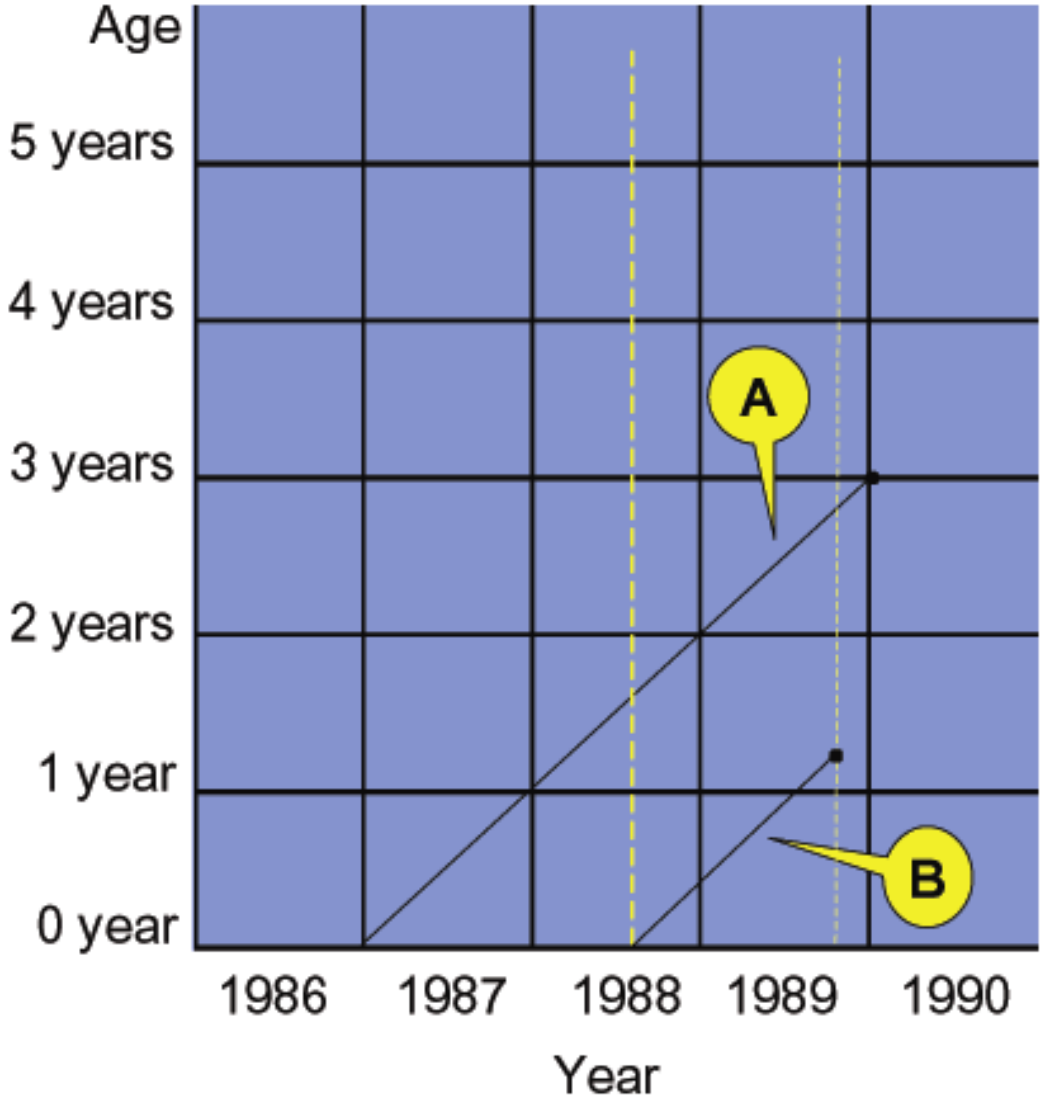
By convention

- Time is represented on the horizontal axis
- Age (or other duration measure) is represented on the vertical axis

Life Lines

- A person's life is represented in the Lexis diagram by a straight line called "life line"
  - The line begins on the time axis at the time of the person's birth
  - Continues diagonally upwards
  - Ends at the age/time point representing the person's death.

Life Lines



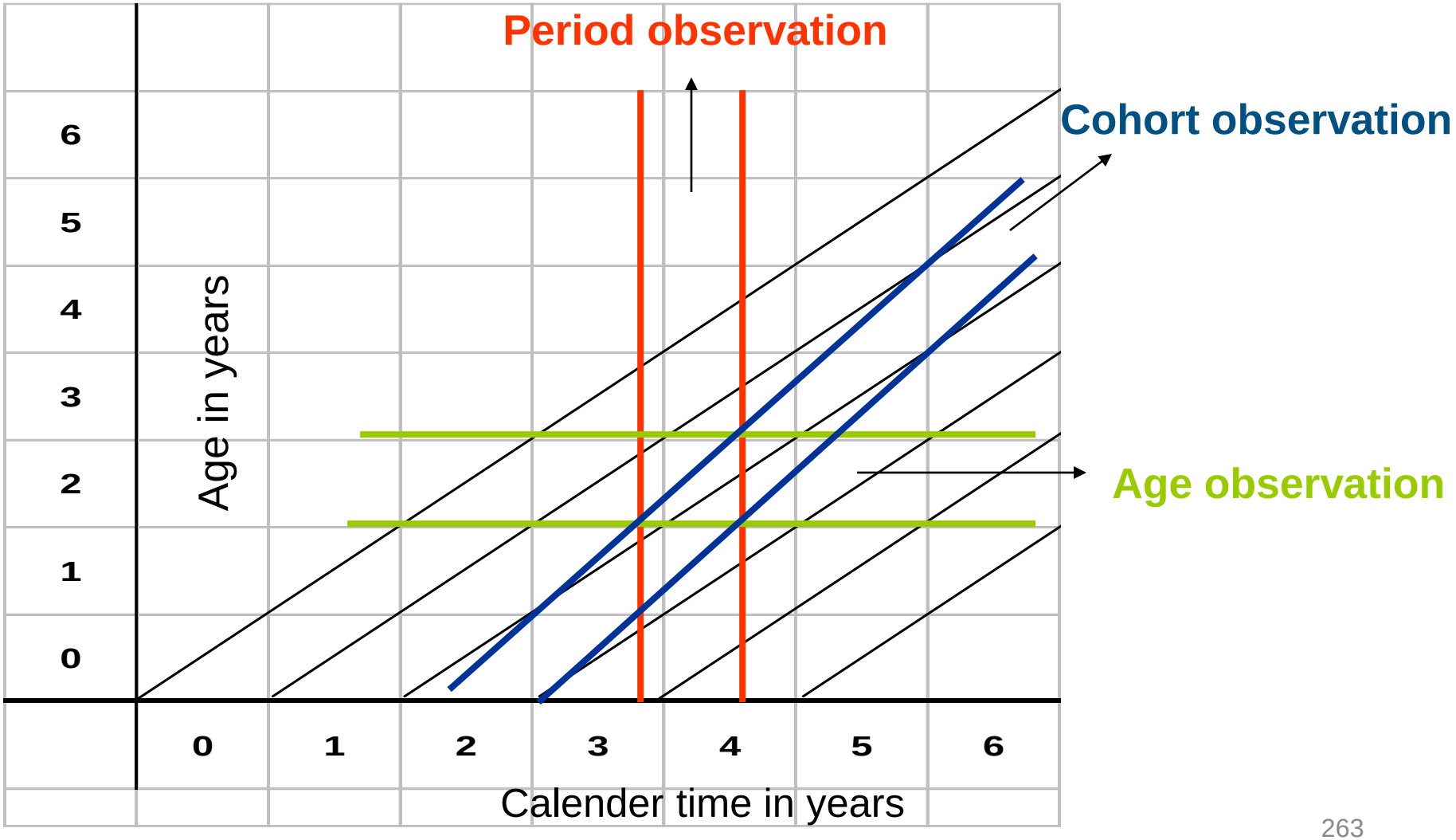
A: From birth on 1/1/1987 to age 3 on 1/1/1990

B: From birth on 7/15/1988 to 1 year and 3 months on 10/15/1989

Life Lines

- The sum of all the life line lengths in a particular portion of the diagram represents person-years lived or exposure in that area
- Life lines and events can be considered from a cohort or period perspective.

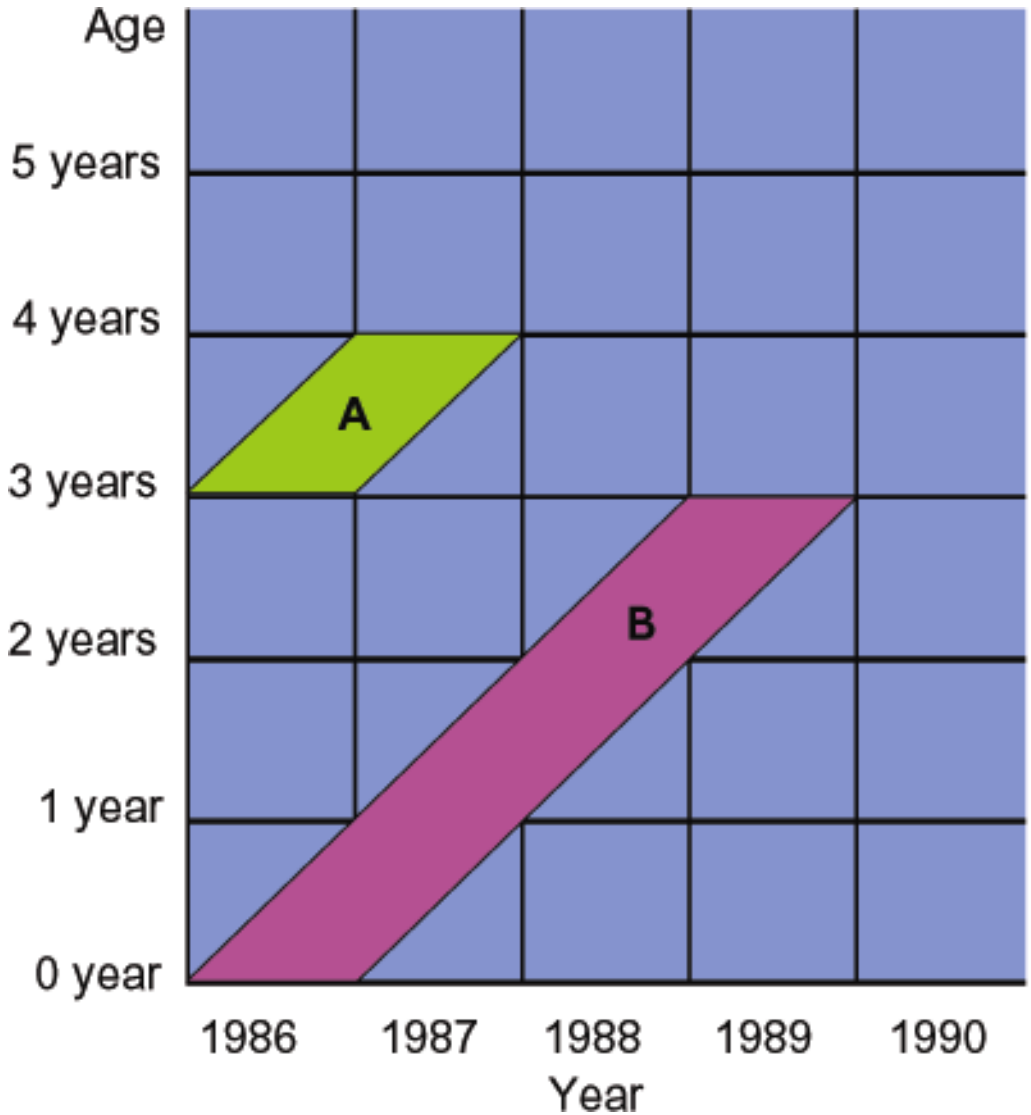
Lexis diagram



Cohort Measure

- As you can see from the picture in the next slide:
  - The Lexis diagram can show the experience of a cohort in a particular age interval (A).
  - It can also show the experience of a cohort as they move through life (B).

Cohort ….



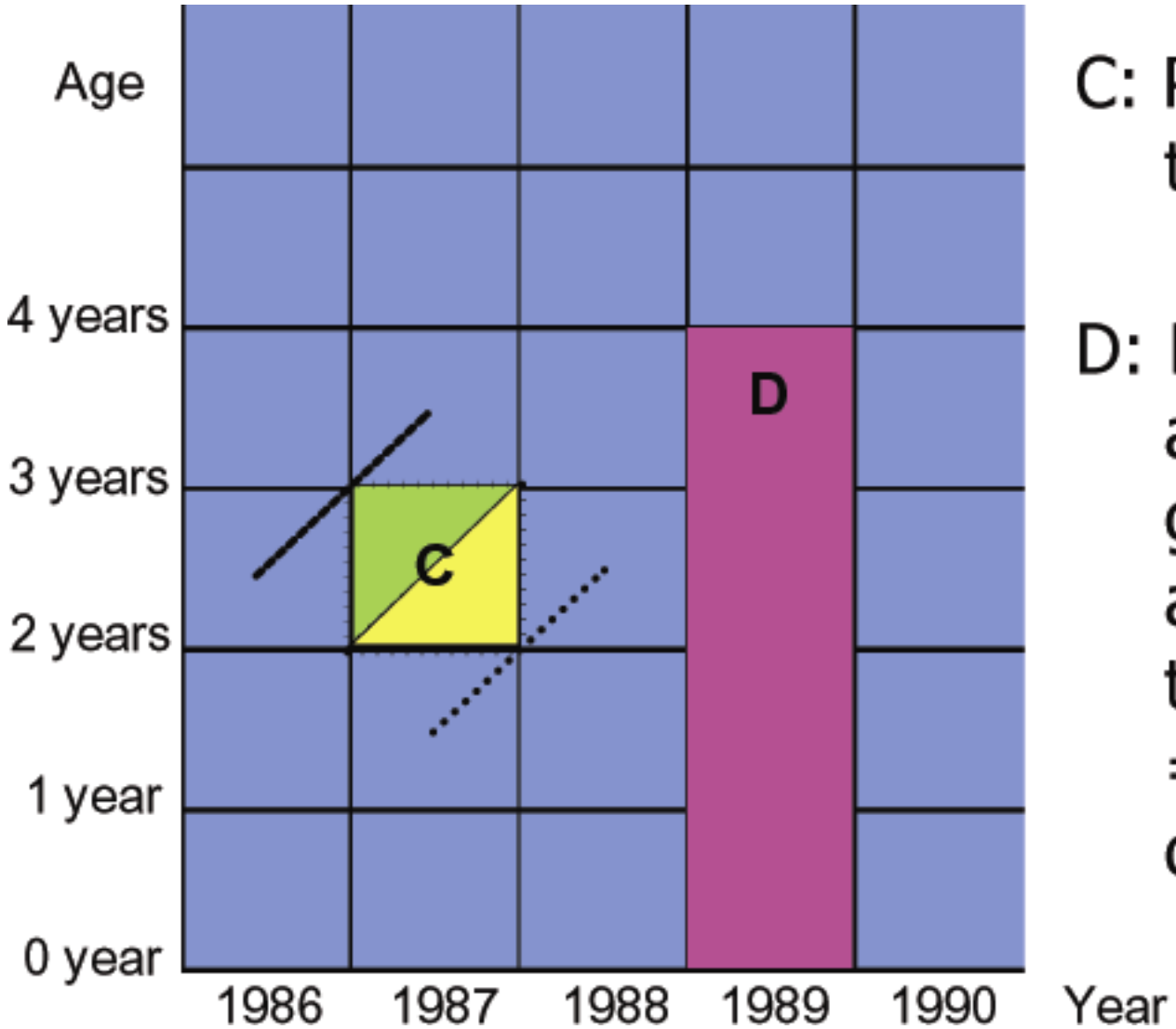
A: Person-years lived at age three by a cohort

B: Persons born in 1986, as they move through life until their 3rd birthday

Period Measure

- As you can see from the picture in the next slide:
  - Square (C) = person-time of exposure at a particular age and period
    - A square includes the life lines from experiences of two different cohorts
  - Column (D) = experience of all age groups of interest during a specific time period

Period …



C: Persons of age two in 1987

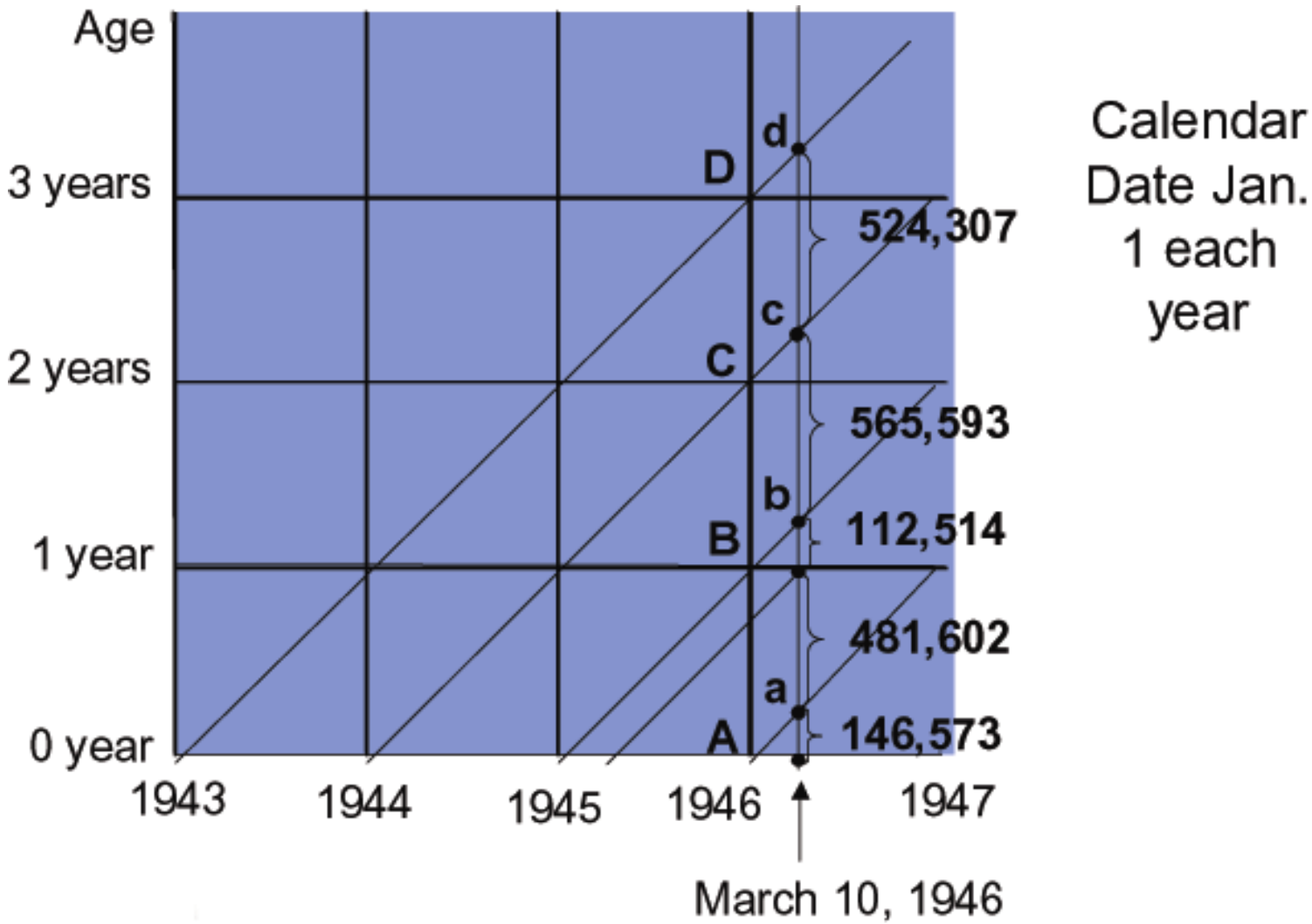
D: Experience of all the age groups under age four during the year 1989 = population during 1989

Example 1: display of census information by age and year of birth

France: Total Enumerated Population, March 10, 1946

Year of Birth	Age in Completed Years on 1 Jan 1946	Population
1946: January 1 to March 9	-	146,573
1945: March 10 to December 30	0	481,602
1945: January 1 to March 9	0	112,514
1944	1	565,593
1943	2	524,307

Example 1



Source: Pressat R. *Demographic Analysis: Methods, Results, and Applications.* 1972

Example 2

- Each square in a Lexis Diagram represents the experience of two cohorts (distribution of deaths during a calendar year between different birth cohorts)

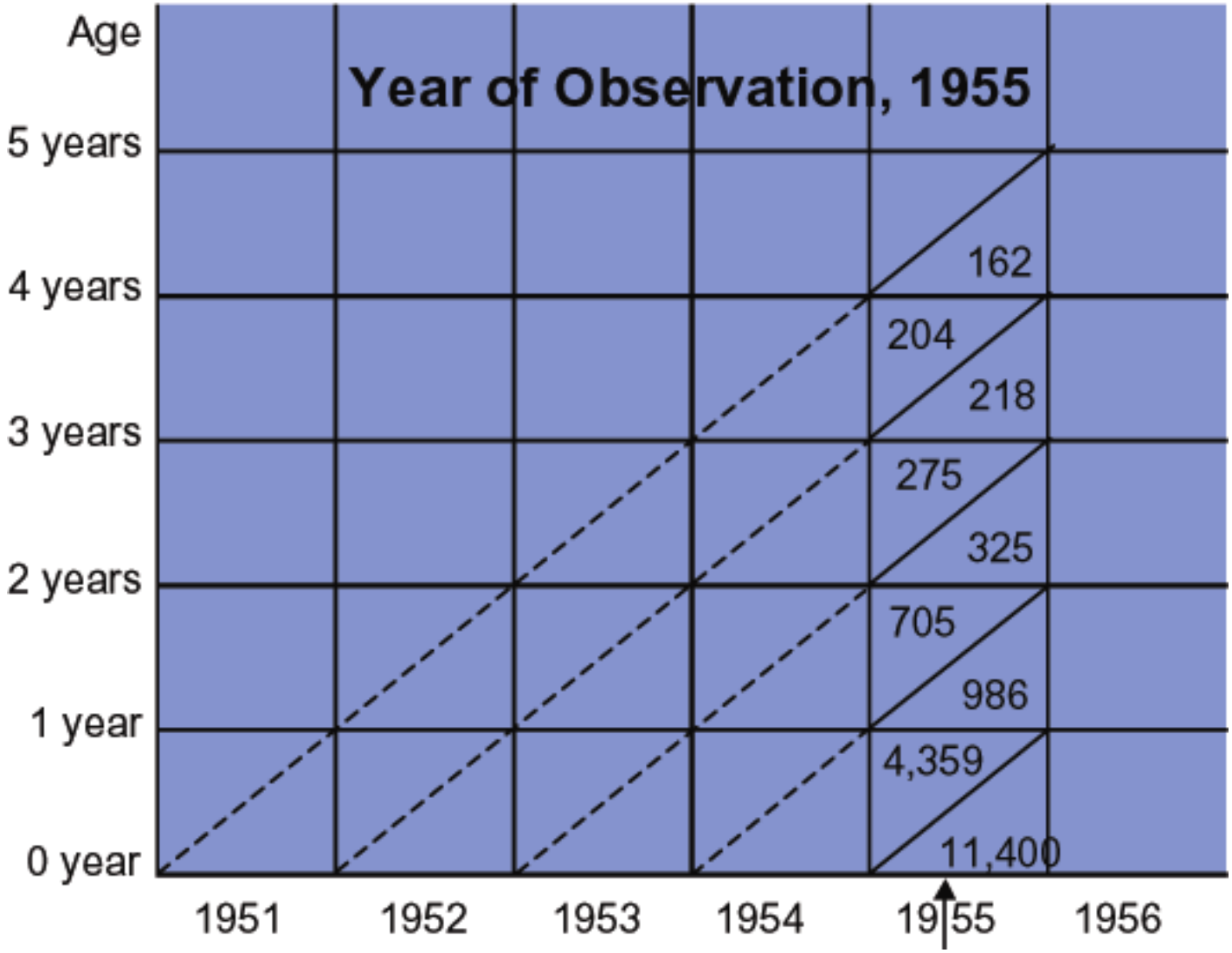
Example 2

France: Male Deaths in 1955, by Year of Birth and by Age

Year of Birth	Age in Completed Years	Number of Deaths
1955	0	11,400
1954	0	4,359
1954	1	986
1953	1	705
1953	2	325
1952	2	275
1952	3	218
1951	3	204
1951	4	162

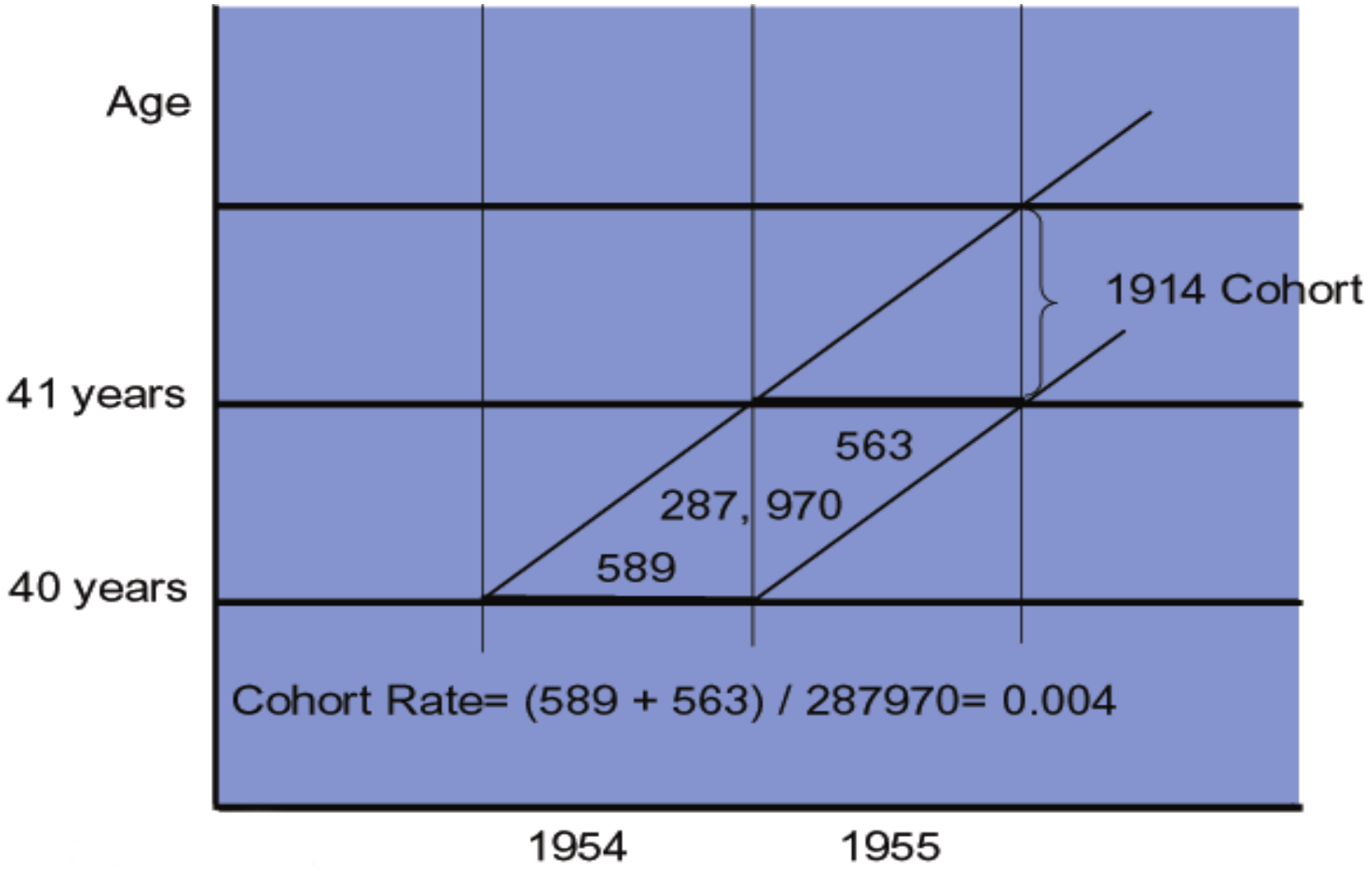
Source: Pressat, 1972

Example 2



Source: Pressat R. *Demographic Analysis: Methods, Results, and Applications.* 1972

Example 3:Representation of the deaths occurring to a specific cohort at a particular age



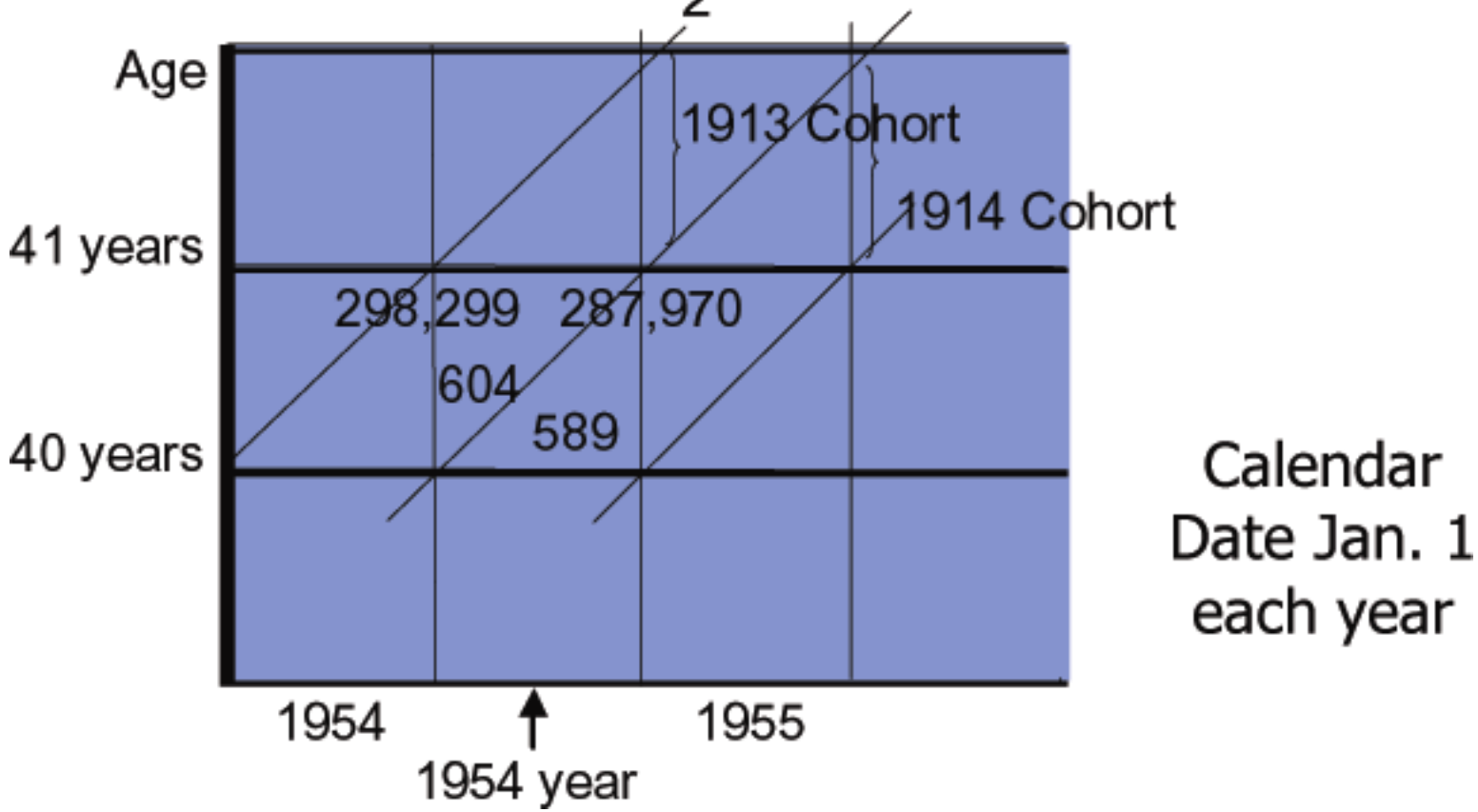
Source: Pressat R. *Demographic Analysis: Methods, Results, and Applications.* 1972

Example 4

- Each square in a Lexis Diagram represents the experience of two birth cohorts (distribution of the deaths occurring at a specific age during a calendar year)

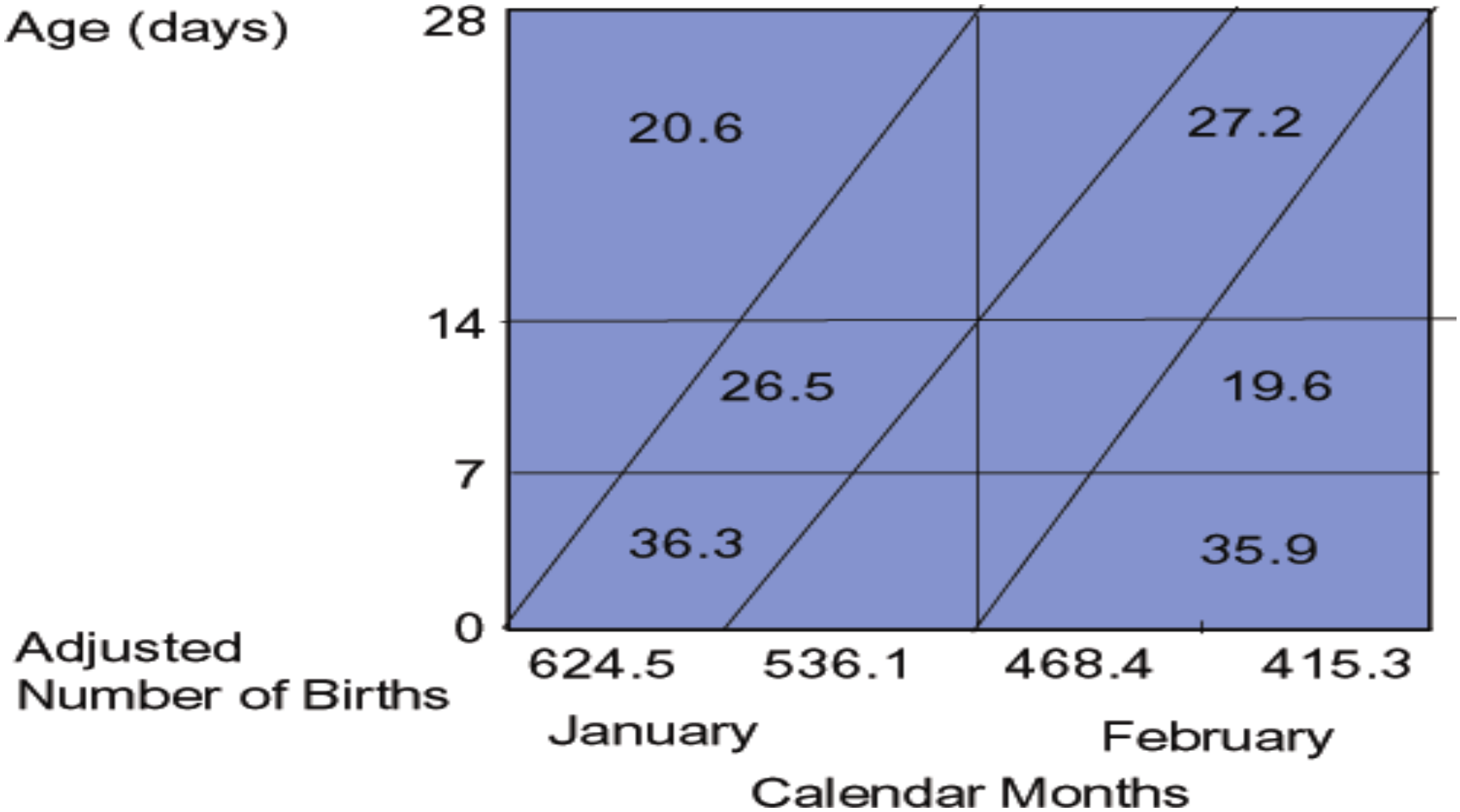
Example 4

$$\text{Period rate} = \frac{604 + 589}{\frac{298299 + 287970}{2}} = 0.008$$

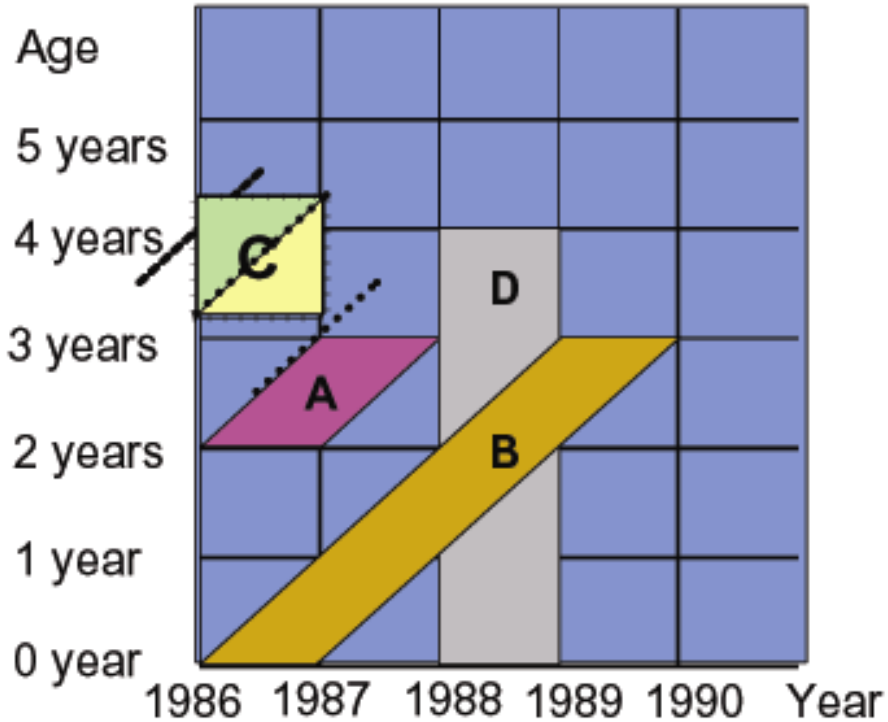


Source: Pressat R. *Demographic Analysis: Methods, Results, and Applications.* 1972

Example 5: Relating events in periods of time to cohorts of individuals (attributing period numbers of deaths to birth cohorts)



Summary



A: Experience of a cohort at age two

B: Experience of a cohort of persons born in 1986, as they move through the first three years of life

C: Persons of age three in 1986

D: Experience of all the age groups under four in 1988

Population Composition

- Population composition answer one of the demographic question: how many people, *of what kind*, are where?
- To classify population based on composition, characteristics that either do not change easily or change in a predictable fashion tend to be useful.
- Although age and sex are basic dimensions of population composition, many other social and economic characteristics also are used to ascribe social roles;
  - For example, ethnicity, literacy and education, wealth, occupation, and income.

Population Composition

- Societies construct roles and assign status on the basis of age and sex more than on any other characteristics.
- Consequently, nations uniformly include questions on age and sex in their census questionnaires.
- So important is age-sex composition, many demographers give it a special label: **population structure**.

Age and Sex

Age-sex structure

- The age-sex structure of a population is the distribution of the population by age and sex.
- It may be described in terms of absolute numbers or percentages.
- Age-sex structures are often depicted graphically in the form of a population pyramid.

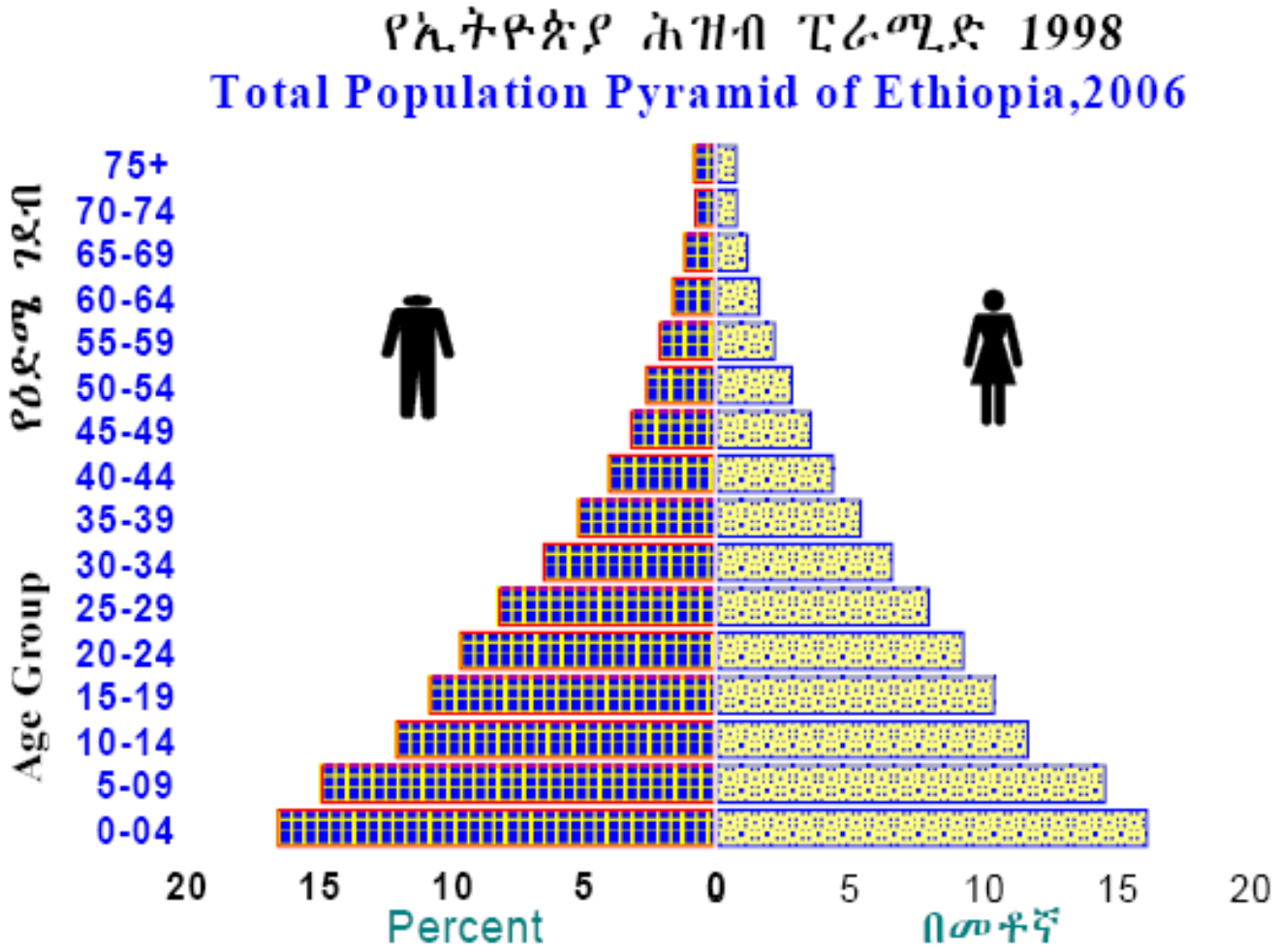
Population pyramid

- Population pyramid gives a detailed picture of the age-sex structure of a population
- Consists of bars representing age groups in ascending order from lowest to highest pyramided on one another
- Bars are generally by single or 5-year age Groups
- Males on left side; females on right side

Types of population pyramid

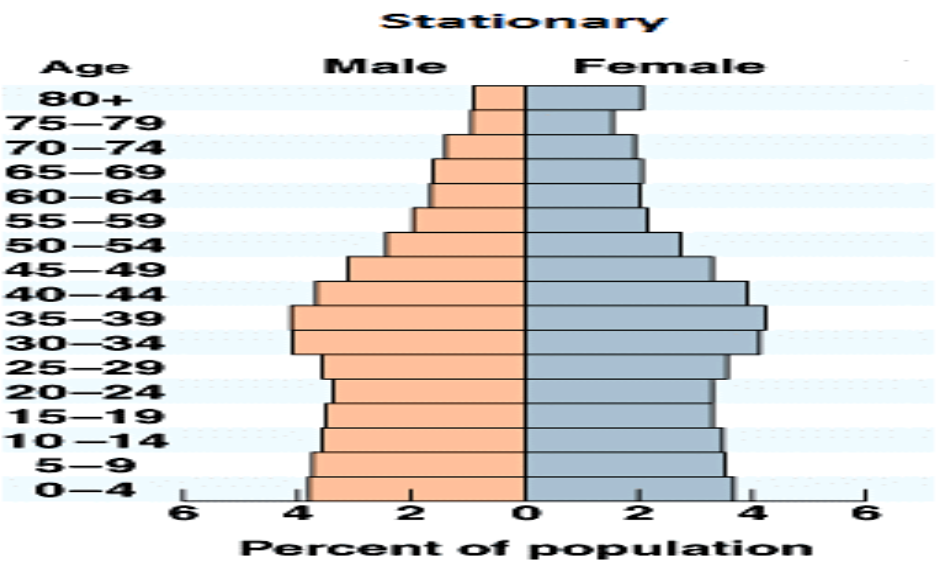
- **Expansive** - pyramid with a **wide base** (larger percentage of people in younger age groups, indicating high birth rates and high fertility rates) and narrow top (high death rate and lower life expectancies).
- It suggests a growing population.

Example: Ethiopia Population Pyramid


የኢትዮጵያ ሕዝብ ፒራሚድ 1998
Total Population Pyramid of Ethiopia,2006
Age Group
የዕድሜ ገደብ
75+
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-09
0-04
20
15
10
5
0
5
10
15
20
Percent
በመቶኛ

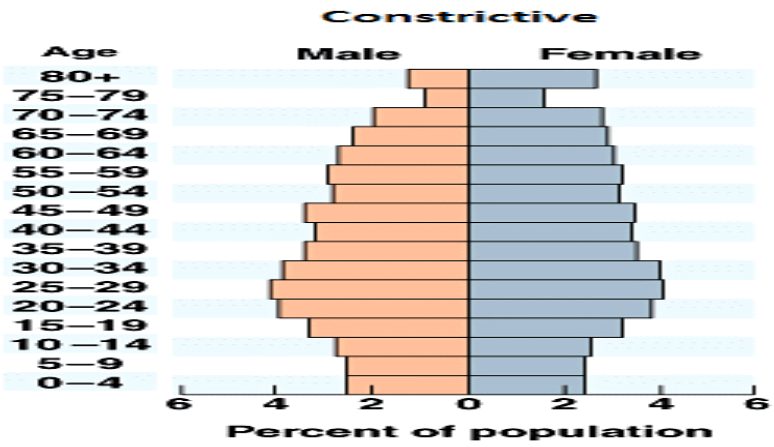
Types of population pyramid

- **Stationary** - with a somewhat **equal proportion** of the population in each age group.
- The population is stable, neither increasing nor decreasing

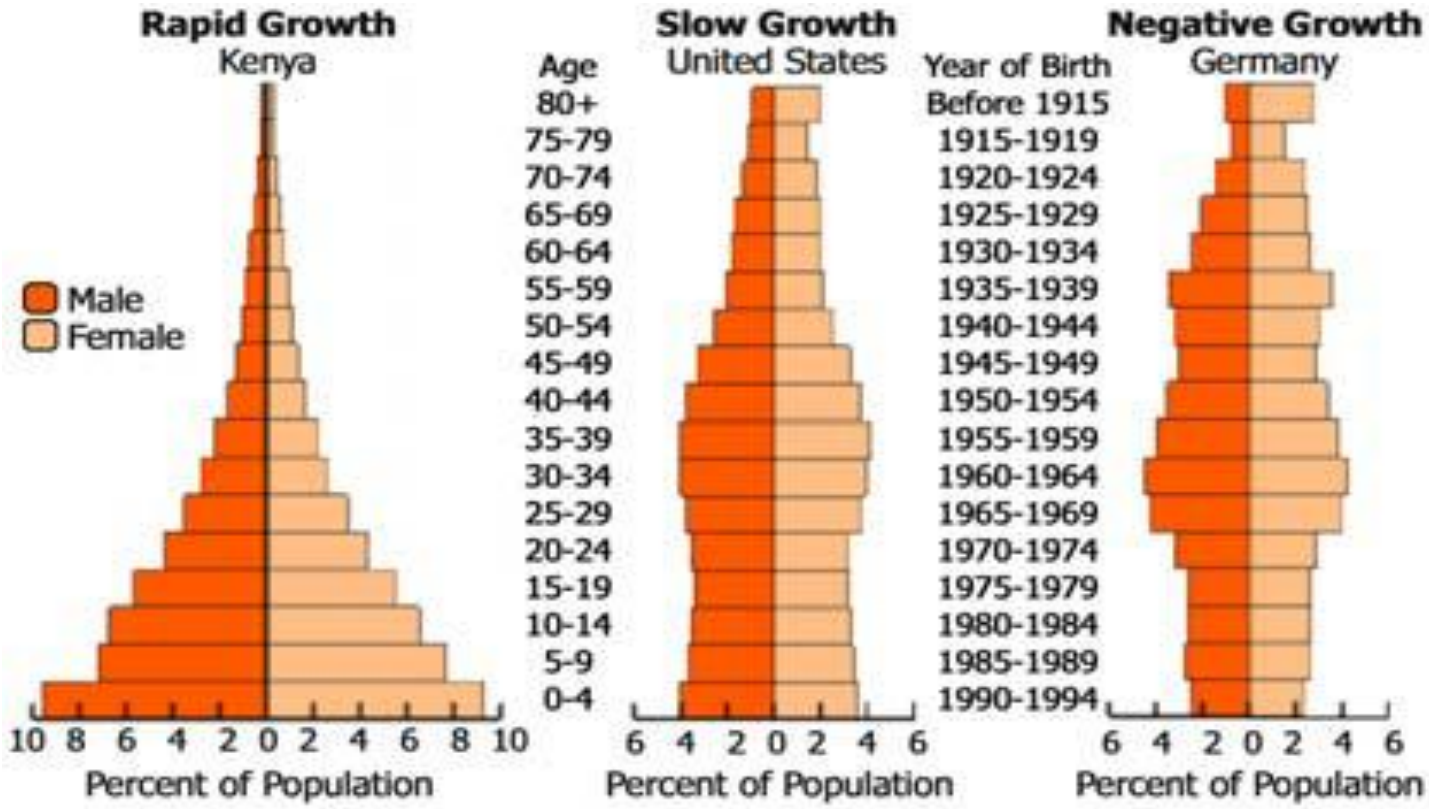


Types of population pyramid

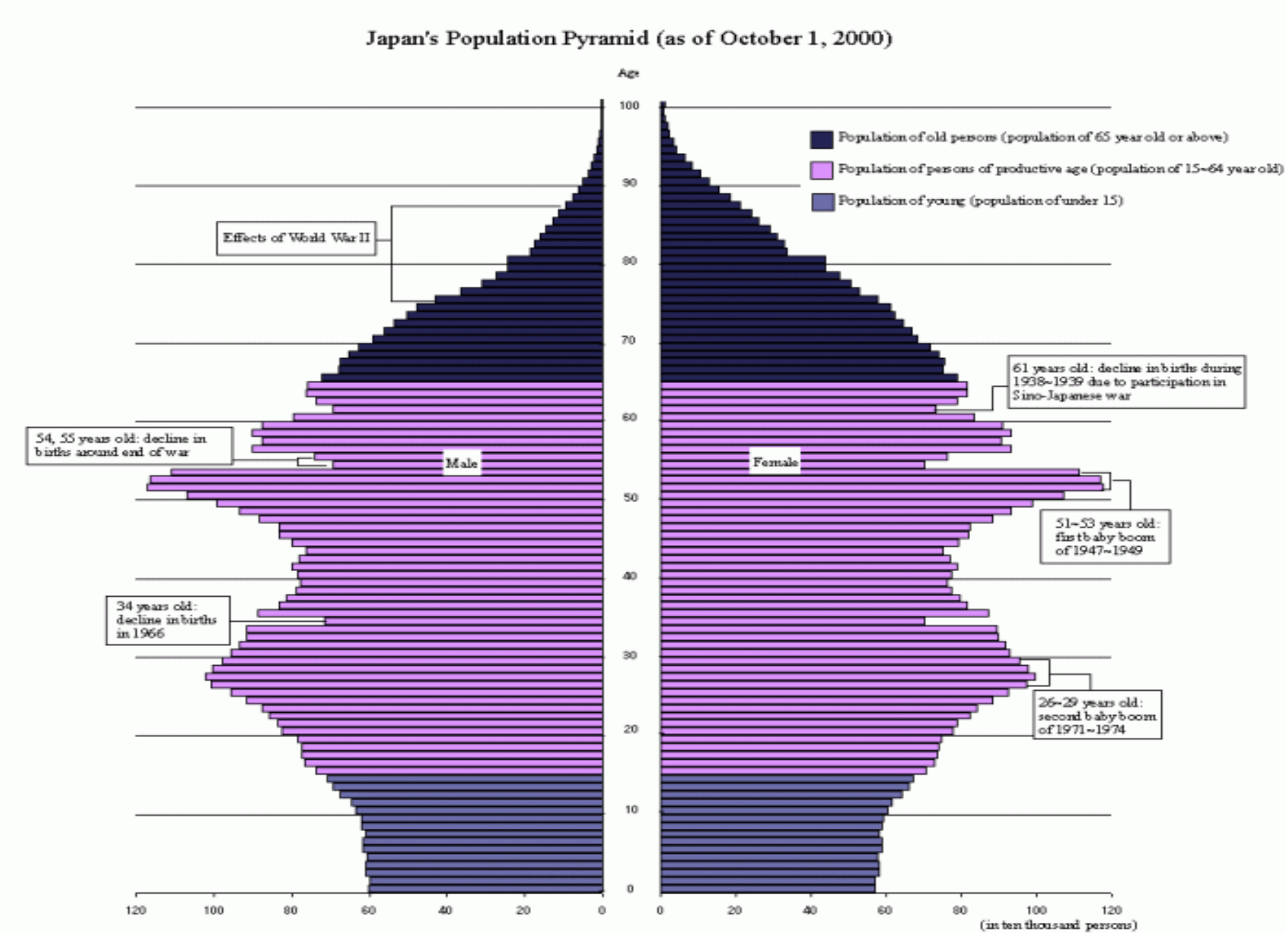
- **Constrictive** - pyramid with a **narrow base** (lower percentage of younger people, indicating declining birth rates with each succeeding age group getting smaller than the previous one).



Pyramids & Patterns of population growth



pyramids


Japan's Population Pyramid (as of October 1, 2000)
Age
Population of old persons (population of 65 year old or above)
Population of persons of productive age (population of 15~64 year old)
Population of young (population of under 15)
Effects of World War II
54, 55 years old: decline in births around end of war
34 years old: decline in births in 1966
61 years old: decline in births during 1938~1939 due to participation in Sino-Japanese war
51~53 years old: first baby boom of 1947~1949
26~29 years old: second baby boom of 1971~1974
Male
Female
(in ten thousand persons)

Age Composition

- Demographers' definition of age— Completed age,

i.e. age of an individual at last birthday

- Most important variable in demographic analyses
- Is most susceptible to reporting errors
- comparison of age distributions is a classic demographic analysis

Data Collection on Age

- Data on age may be secured by
  - Asking a direct question on age
  - Asking a question on date of birth, or month and year of birth
  - Or a combination of these

Age-Reporting Errors

- Content
  - Centenarians
    - Those close to 100 years tend to overestimate their age
  - Understatement
    - Women tend to understate their age

Age-Reporting Errors

- Overstatement
  - Mothers tend to round up the age of their children
- Heaping/Digit preference
  - People tend to report certain ages at the expense of others
  - Can occur at any digit but happens most often with 0 and 5

Age-Reporting Errors

- Distributing unknown and/or unreported age
- *Coverage*—Missed or counted twice
  - There is a tendency to miss the people in certain age groups (e.g. young men)
  - Some people are counted twice

Method for Detecting the Extent of Age Errors

- Age ratios used to estimate data quality
- Two formulas available:
  - Let a_x= number of persons of age x

$$\frac{a_x}{1/2\left(a_{x-1}+a_{x+1}\right)}*100$$

or

$$\frac{a_x}{1/3\left(a_{x-1}+a_x+a_{x+1}\right)}*100$$

Correction of Age Errors

- Grouping
  - Generally by 5 or 10 years
  - Avoid problem of fluctuations in single year data.
- Interpolation

Analysis of age

- When working with several populations, comparison of age distributions is a classic demographic analysis.
- We can analyze the population age using the following different functions.

Mean Age

There are two ways of calculating the mean of any distribution:

$$\sum xf(x) \quad \text{or} \quad \sum\left[1-F(x)\right]$$

Where:

- f(x) = Discrete density function of x = probability that a case has value x
- F(x) = Cumulative distribution function of x = Probability that a case has value at or below x

$$= \sum_{j=0}^{x} f(j)$$

Example of Calculation

x	f(x)	F(x)	xf(x)	1-F(x)
0	0.00	0.00	0.00	1.00
1	0.50	0.50	0.50	0.50
2	0.25	0.75	0.50	0.25
3	0.25	1.00	0.75	0.00
Mean			1.75	1.75

Note that F(0) = 0

Median Age

Median: Age which divides the population into two equal-size groups, one younger and the other older.

For class interval median

$$= \ell_{Md} + \left(\frac{\frac{N}{2} - \Sigma f_x}{f_{Md}} \right) * i$$

- **Where:**
  - l_{Md} = Lower limit of class containing the middle (N/2th) item
  - N= Sum of all frequencies, i.e. population size
  - Σf_x = Sum of frequencies in all classes preceding the one containing the N/2th item
  - f_{MD}= Frequency of class containing the N/2th item
  - i = Width of class interval containing the N/2th item

Median

- The median is preferred to the mean because of:
  - The marked skewness of the age distribution
  - The calculation of mean is often complicated by open-ended age groups (e.g. 80+)

Age-Dependency Ratios

- *Age-Dependency Ratios*—Proportion of children less than 15 and elderly 65+ relative to the population of "working ages"

$$= \frac{P_{0-14} + P_{65+}}{P_{15-64}} * 100$$

Child-Dependency Ratio

- *Child-Dependency Ratio*—Proportion of children less than 15 relative to the population of "working ages"

$$= \frac{P_{0-14}}{P_{15-64}} * 100$$

Old-Age Dependency Ratio

- *Old-Age Dependency Ratio*—Proportion of adults over age 65 relative to the population of "working ages"

$$= \frac{P_{65+}}{P_{15-64}} * 100$$

Sex Composition

- Sex is a variable of prime importance
- separate data for males and females are important for:

  - analysis of other types of data

  - evaluation of the completeness and accuracy of population counts

Sex Composition

- Sex composition is affected by;
  - Sex ratio at birth (usually greater than 100)
  - Differential patterns of mortality for males and females
  - Differential patterns of migration for males and females

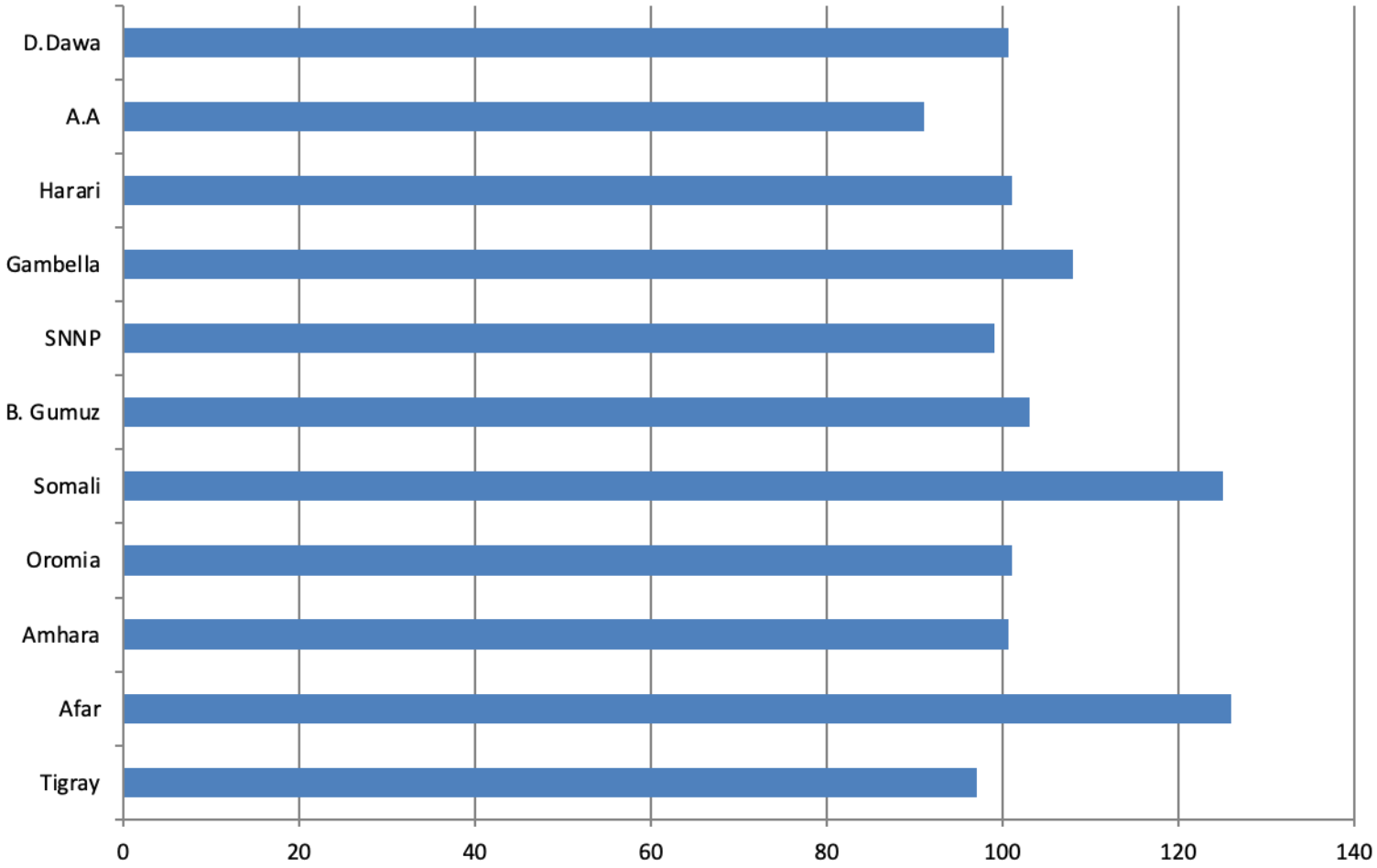
Sex ratio

- Proportion of males relative to females in a population

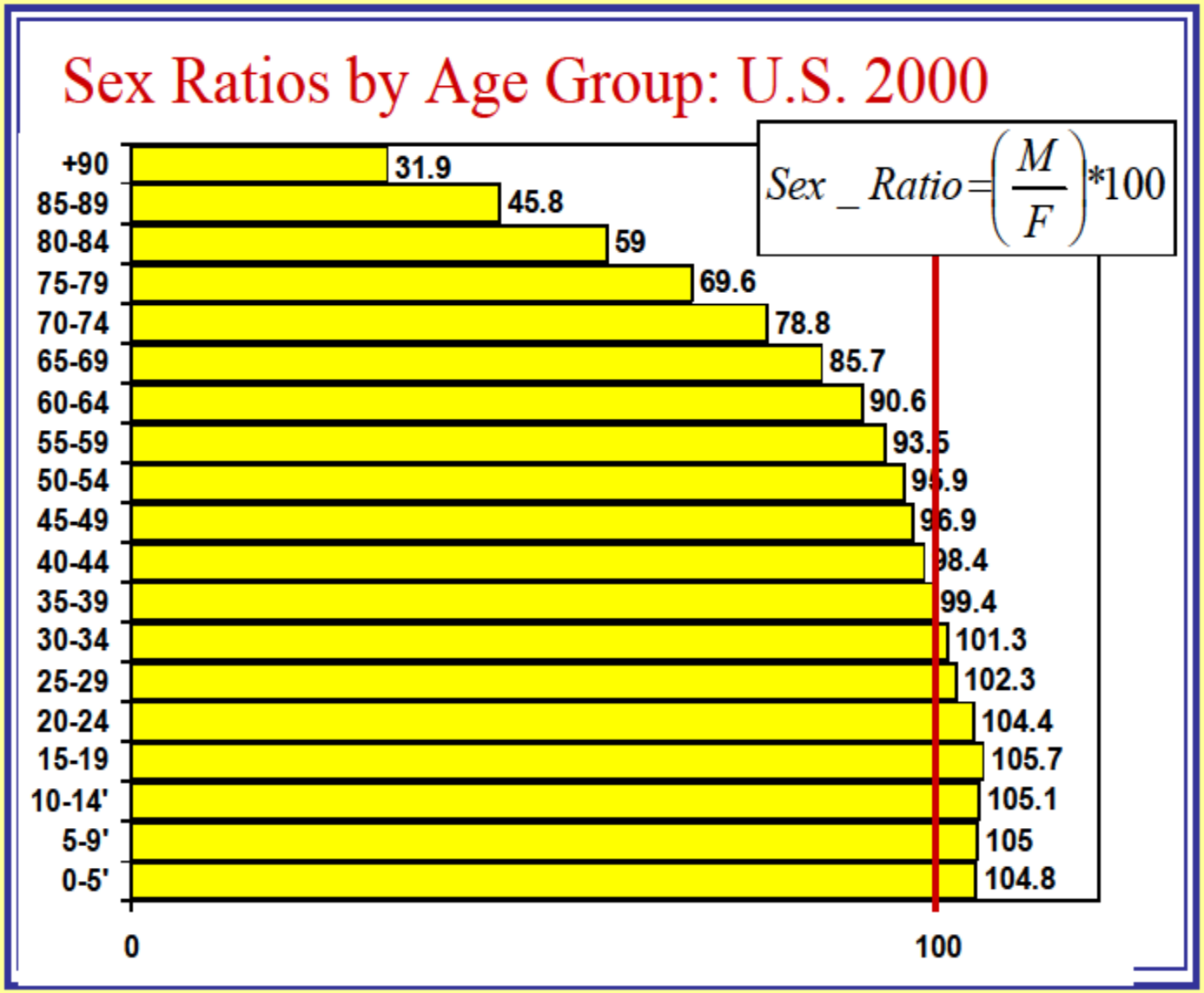
$$\text{Sex ratio} = \frac{\text{number of males}}{\text{number of females}} * 100$$

- Sex ratio pattern by age in a population:
  - Higher at very young ages
  - Gap narrows with increasing age
  - Goes below 100 by middle age
- Sex Ratio at Birth, is often higher than 100 & ranges from 104 to 107.

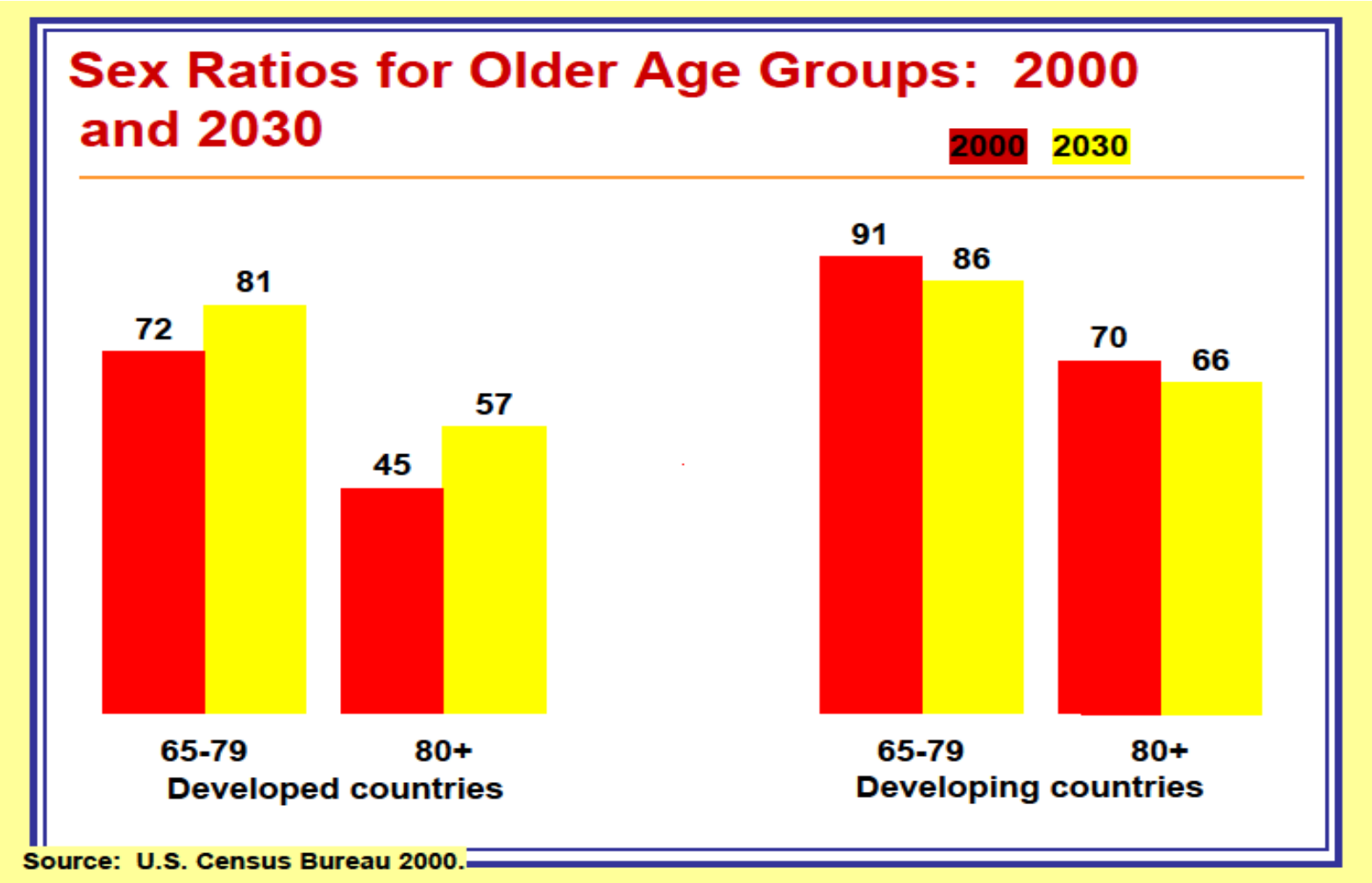
Sex ratio of regions in Ethiopia,2007



Sex ratio declines by age



Sex ratio


Sex Ratios for Older Age Groups: 2000 and 2030
2000
2030
72
81
45
57
91
86
70
66
65-79
80+
Developed countries
65-79
80+
Developing countries
Source: U.S. Census Bureau 2000.

- **Race**
- **Ethnic group**
- **Nationality**
- **Country of Birth**
- **Marital Status**

Race and Ethnic Group

- There are no standard definitions of race and ethnicity.
- *Ethnic Group*—Sub-Race.
  - Covers racial, national, cultural, or linguistic groups.
  - To varying extents, each such group will have a common descent, history, and habitat.

Nationality

- *Nationality* is
  - Country of present citizenship, or
  - Country or other area of origin, sometimes an extinct country or a country that once had very different boundaries

Country of Birth

- Used to identify and describe immigrant minorities and throw light on ethnic composition
- Inclusion in censuses recommended by UN.
- Ask for country of birth for foreign born

United States Practices

- *Native*—Persons born in the U.S., the Commonwealth of Puerto Rico, or a possession of the U.S.
- Also includes persons born in a foreign country or at sea who have at least one native American parent
- *Foreign Born*—Persons not classified as native i.e. persons who report a foreign country as place of birth

Citizenship

- *Citizenship*—Legal nationality
- *Alien*—Person included in census, register, etc… of a country but non-citizen thereof
  - An alien, however, is usually citizen of another country; that country is his/her legal nationality

Citizenship

- May be acquired by birth or naturalization
- Almost all people born in some countries are automatically citizens, or at least have that option; this is not so in some other countries.

Citizenship

- Useful in connection with problems of legal status and civil rights of immigrants, and for studies of naturalization and assimilation of alien populations.
- Limitations
  - Confusion in definition with country of birth
  - Also sensitive where minorities are insecure in their status

Language

- One of the most sensitive indices of ethnic origin because linguistic differences tend to persist until complete assimilation.

Types of Language Data Recommended by UN

- Language spoken at home in early childhood or language of parents (mother tongue)
  - Is an indicator of ethnic origin
- Language currently spoken, or most often spoken in present home (usual language)
  - Is an indicator of ethnic origin
  - Can also be used to see degree of assimilation and integration of foreign stock or of specific ethnic minorities
- Ability to speak various languages (designated languages)
  - Indicates linguistic skills of population, both native and foreign born

Religion

- Item mentioned by the United Nations in its consideration of ethnic origin
- If race or country of origin is also known, religion can be used to make further distinctions among ethnic groups
- Useful personal characteristic because it is associated with a variety of differences in attitudes, statuses, and behavior
- Of great interest for its hypothesized relationship with fertility, mortality, nuptiality, and migration

Marital Status

- Categories
  - Minimum list of groups recommended by the United Nations
    - Single (never married)
    - Married and not legally separated
    - Married but legally separated
    - Widowed and not remarried
    - Divorced and not remarried

Marital Status

- Note: Ever married = married + widowed + divorced + separated
- Additional options to take into account when appropriate:
  - *Consensual unions*—Common-law, extra-legal or de facto unions
  - Practices such as polygamy, concubinage and inherited widows
  - Same sex

Order of Marriage

- The variables most often used in demographic analysis of marriage and of the marital factor in fertility are:
- *Order of Marriage*—Number of times married, or whether married more than once

Duration of Marriage

- Also of interest are age at and date of marriage
- Can be asked for first marriage or most recent marriage
- Date of first marriage is used to identify 1st marriage cohorts, i.e. persons who entered marriage for the 1st time in a specified period such as a calendar year

Importance for Fertility

- Age at first marriage is probably the most useful fact about women's marital history for the study of their fertility
- Data on spacing of children in relation to date of first marriage, or number of years since first marriage, are highly useful for analysis of fertility

Education and Economic Status

Education

- Important variable in accounting for demographic behavior

School Enrollment

- *School Enrollment*—Enrollment at any regular educational institution, public or private, for systematic instruction at any level of education during a well-defined and recent time period (United Nations)

School Enrollment

- Data used to measure extent of the participation in the school systems of an area:
  - By persons of school age
  - The potential future participation of such persons
  - The relative participation of different segments of the population

School Enrollment

- Importance
  - Measures of school enrollment usually relate to a point in time or a very short period of time

School Attendance

- *School Attendance*—While enrollment data are available at yearly intervals, attendance data may be collected daily in school classrooms

Literacy

- *Literacy*—Ability of a person to both read and write, with understanding, a short statement on his everyday life (United Nations)

Literacy

- An illiterate person may not read and write at all, or may read and write only figures and his/her own name, or may only read and write a ritual phrase which has been memorized
- The language or languages in which a person can read and write are not a factor in determining literacy

Economic Participation

- Importance:
  - Many demographic characteristics are cross-tabulated with the economic characteristics of the individuals
  - Although all persons consume goods and services, only a part of the entire population of a country is engaged in producing such goods and services

Labor Force

- *Labor Force*—Involves the carrying on of an activity from which the person derives, or attempts to derive, pay or profit
- In censuses and surveys, the entire population can be subdivided into the employed and the unemployed

Employed

- *Employed*—Persons who worked for pay or profit during the time-reference period or who had a job or business that period but were absent from it because of vacation, illness, etc.

Unemployed

- *Unemployed*—Persons who, during the reference period, were not working but who were seeking work for pay or profit, including those who never worked before

Not in Labor Force

- *Not in Labor Force*—All persons who are neither employed nor unemployed
- Note:
  - The difference between the amount of work performed by persons in employment and the amount of work they would normally be able or willing to perform is a measure of underemployment

Status in Labor Force

- *Status in Labor Force*—Refers to the status of a person in the labor force with respect to his/her employment, that is, whether s/he is (or was, if unemployed) an employer, own-account worker, employee, unpaid family worker, or a member of a producers' cooperative (United Nations)

Status in Labor Force

- Note:
  - Categories can be combined or further subdivided to better reflect a country's reality

Occupation

- Refers to the kind of work done during the time-reference period established for data on economic characteristics by the person employed (or performed previously by the unemployed) irrespective of the industry or the status in which the person should be classified (United Nations)
  - Examples: Teacher, sales clerk, farmer, car washer, etc.

Personal Income

- *Personal Income*—One of the best measures of economic well-being; vies with educational attainment and occupation as a measure of socioeconomic status
- No international standard definition and sometimes difficult to collect
- In principle, income from all sources should be counted, both cash income and income in kind
- Reference period = usually one year

Household Income

- *Household Income*—Income data collected for each person in the household may be added to produce total for the household

Wealth

- *Wealth*—All goods and resources having value in terms of exchange or use
  - Land holdings, stocks, and bonds
  - Jewelry and other possessions, livestock

Measures of Fertility

Introduction

- The clinical concept of fertility is the potential of a woman to become pregnant and bear a live child.
- However, in demography fertility is entirely concerned with the number of living babies that women give birth to.
  - The reason for this demographic approach is that it is only live births that produce new members of the population.

Introduction…

- Fecundity: The physiological ability to have children—that is manifest roughly in the period between menarche and menopause in women
- Fecundability: the probability of becoming pregnant, or the likelihood of exposure to the possibility, that depends on the pattern of sexual and pregnancy preventive behaviors.

Introduction…

- Infertility(sterility): the inability to bear any children, either due to the inability to conceive or the inability to carry a pregnancy to a live birth.
  - WHO recommends that inability to conceive within two years of exposure is epidemiological definition, and one-year period is used for clinical definition of infertility.
- Infecundity: The inability to have a live birth after several years of exposure to the risk of pregnancy.

Fertility measures

- Fertility measurement is commonly about producing rates that relate numbers of births to numbers of women.
- Fertility is more easily measured for women because they, and not men, actually have the babies.
- The fertility measurement is based on the data from vital statistics, censuses and Surveys.

Fertility measures

- The most common fertility measures are:
  - Crude Birth Rate (CBR)
  - Child Women Ratio (C/W Ratio)
  - General Fertility Rate (GFR)
  - Age-specific Fertility Rates (ASFRs)
  - Total Fertility Rate (TFR)
  - Gross Reproduction Rate (GRR)
  - Net Reproduction Rate (NRT)

Crude Birth Rate (CBR)

- It is a simple ratio of the number of births in a particular period (usually a year) divided by the total population size.

$$\textbf{CBR} = \frac{\text{Number of live births in a stated period}}{\text{mean population over that period}}$$

- The denominator is technically the person-years lived in the year, but this is usually estimated as the **mid-period population**, which is itself estimated as the mean of the population at start and end of the period.
- The CBR is commonly quoted per 1,000 population.

...CBR

- It is not a true rate measure b/c, the denominator contains men and children and older people who are not at risk of childbearing, therefore it is a ratio rather than a rate.

➢ It is high for population of the developing countries and low for those of the developed ones.

➢ Based on CBR values levels of Fertility could be:

✓ ***High fertility Rate*** if CBR is > 30/1000

✓ ***Medium fertility rate*** if CBR is = 20-30/1000

✓ ***Low fertility rate*** if CBR is < 20/1000

...CBR

- The Crude Birth Rate is a useful measure to approximate numbers of births when you have limited information available.
- It is also an important measure because of its place in the balancing equation.
- However, it has limitations of:

✓Only a crude estimate of fertility.

✓All the population included in the denominator is not exposed to the risk of pregnancy.

✓ Not good for comparing fertility across populations

Exercise

Crude Birth Rate (CBR)

- Use the following data to calculate the CBR per 1,000

Island of Mauritius, 1985

Total Births: 18,247
Total female population: 491,310
Total male population: 493,900

Source: U.N. Demographic Yearbook, 1986

Child/Woman Ratio (C/W ratio)

- It is a measure of fertility and "youngness" in a population .

Child/Woman Ratio =	living children aged 0 - 4
	women aged 15 - 49

- So as a fertility measure it is rather rough-and-ready but it can still tell you about the burden of children in a community.
- It is normally a figure below one – considerably below one for low fertility countries, just under one for high fertility countries.

Child/Woman Ratio

- Its is not a true measure of fertility due to numerator misses children who have died.
- In a high mortality situation numbers of these could be substantial and fertility will therefore be underestimated.
- It can be constructed from census data alone; it does not require information about births.

General Fertility Rate (GFR)

The General Fertility rate is the number of live births per 1000 females aged 15-49 years (fertile age group) in a given year.

$$\textbf{GFR} = \frac{\text{Number of live births in a stated period}}{\text{Number of women aged 15-49 in the same period}} \times 1000$$

- It eliminates distortions that might arise due to different age and sex distributions among the total population.

General Fertility Rate (GFR)

- It can easily be constructed from vital registration, census or survey data and does not rely on exact ages (except at beginning and end of age range).
- It is reasonably accurate measure of fertility because the numerator and denominator are focused and well specified and the denominator is truly those at risk of producing the numerator.

General Fertility Rate (GFR)

- The major limitation of GFR is that ***not all women in the denominator are exposed to the risk of child birth***.

Exercise

General Fertility Rate (GFR)

- Use the following data to calculate the GFR per 1,000 women aged 15–44:

Island of Mauritius, 1985

Age Group	Women
15-19	52 013
20-24	54 307
25-29	46 990
30-34	40 211
35-39	30 401
40-44	23 496

Total births: 18 247

Source: U.N. Demographic Yearbook, 1986

Age-specific Fertility Rates (ASFRs)

- It is the number of children born alive to females in a specific age group.

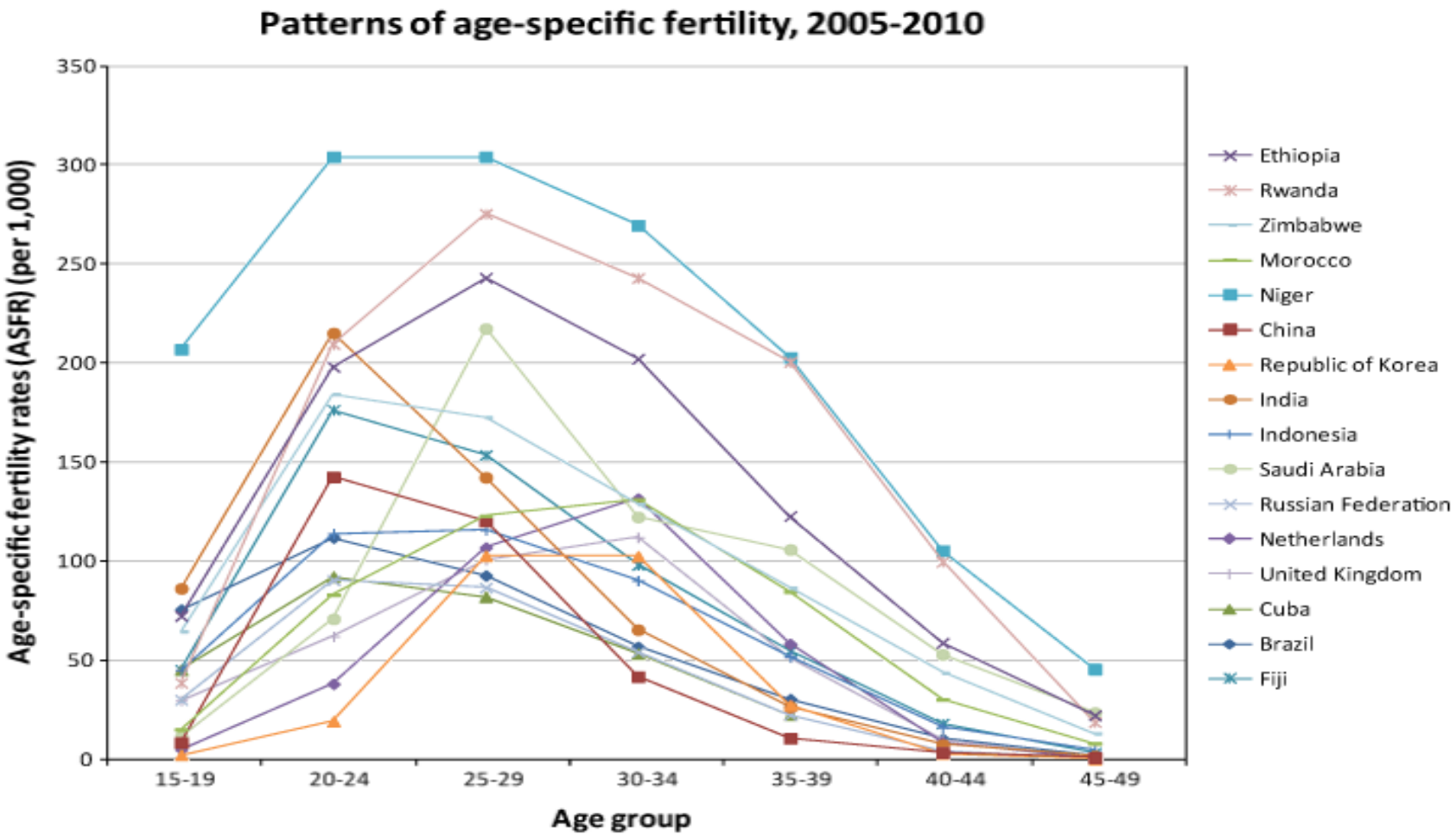
$$\textbf{ASFR} = \frac{\text{No. of live births to women in a specific age group in a stated period}}{\text{No. of women of the same age group in the same period}}$$

- Quite often they are expressed per thousand women *i.e.* multiplied by 1,000 but this is not essential.

Age-specific Fertility Rates (ASFRs)

- ASFRs allow us to see **patterns** of fertility, which give us a good idea of both traditional and modern/early and late patterns of fertility.
- The fertility pattern can give us some clues about the extent of fertility control although this is speculative.
  - to investigate this properly would need a contraceptive prevalence survey and/or the analysis of Parity Progression Ratios.
- The patterns are remarkably similar across populations although there are recognizable differences.

Patterns of age specific fertility



Patterns of age specific fertility in the above figure

- **Pattern A:** Very high fertility - Niger, Ethiopia, Rwanda.
  - This shows a broad pattern with a rapidly rising pace of fertility from teenage years and a considerable spread that remains substantial even at ages 39-44yrs.
- **Pattern B:** Low fertility - Netherlands, UK, Korea.
  - Childbearing is concentrated across a narrower age-range, later than the peak in Pattern A. There is little teenage childbearing and little after age 40.
- **Pattern C:** Medium level fertility – Zimbabwe, India, Fiji.
  - There is clearly some intentional control of fertility but a cultural preference for having children early and then limiting further children.
- **Pattern D:** Low fertility with early childbearing – Brazil, Cuba, Russia.
  - Low fertility comparable to Pattern B but in terms of timing more similar to Pattern C.

Exercise

Age-Specific Fertility Rate (ASFR[a, n])

- Use the following data to calculate the ASFR per 1,000 for women age 20–24 and 25–29

Island of Mauritius, 1985

Age Group of Mother	Women	Births
15-19	52 013	1884
20-24	54 307	6371
25-29	46 990	5362
30-34	40 211	2901
35-39	30 401	1170
40-44	23 496	268

Source: U.N. Demographic Yearbook, *1986*

The Total Fertility Rate (TFR)

- It is the number of children a woman would have if she experienced the age-specific fertility rates for the period in question throughout her reproductive life.
- The TFR is one of the most useful indicators of fertility, because it gives the best picture of how many children women are having currently.
- It is derived directly from ASFRs and is a standardised measure which allows international comparisons.

The Total Fertility Rate (TFR)

- The TFR sums up in a single number from the ASFRs of all women at a given point in time.

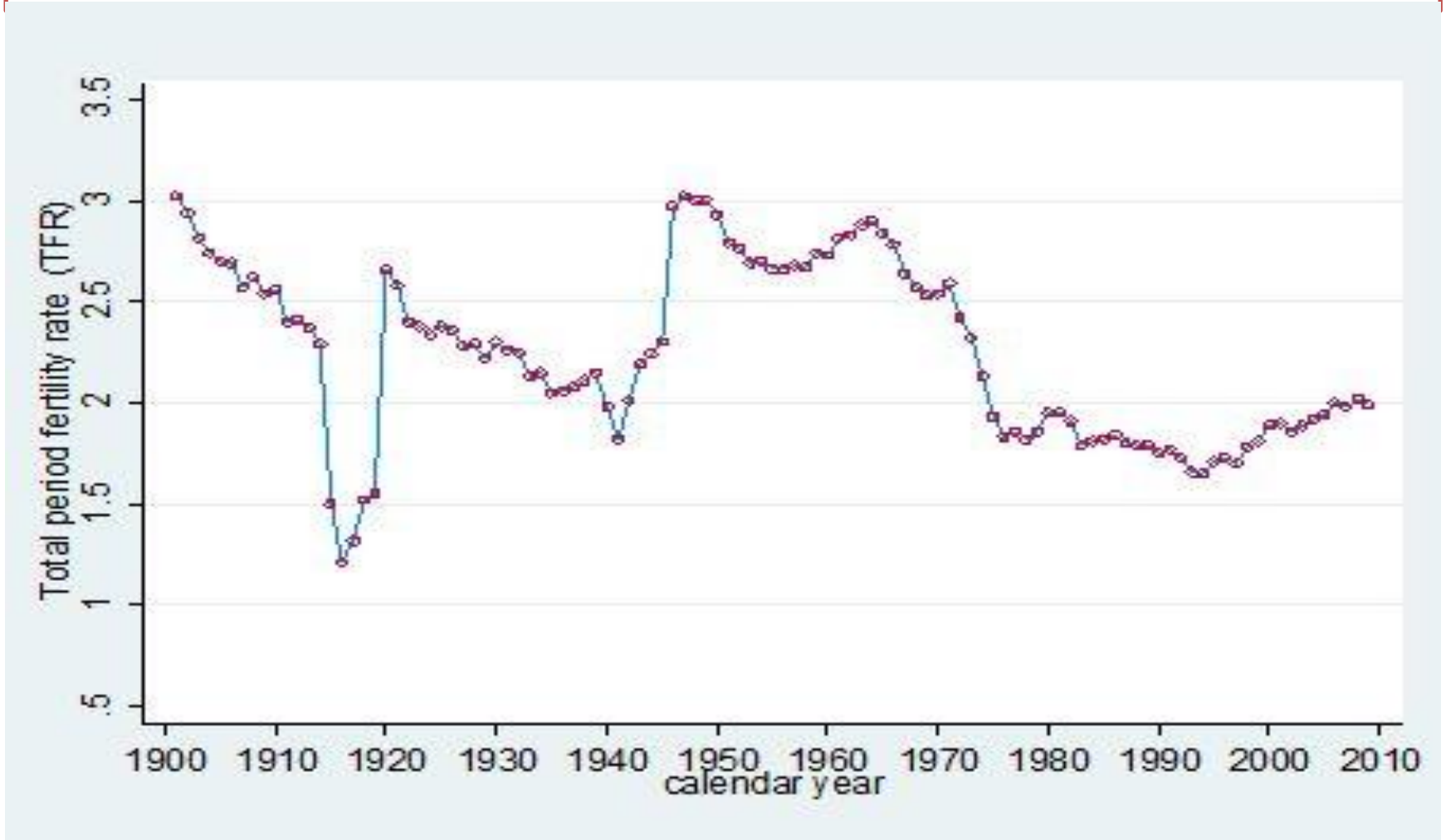
$$TFR = \sum \text{Age specific fertility rates} * x$$

  - If 5 – year age groups are used, the sum of the rates is multiplied by 5.
  - It is possible to use single year age groups, although this is not common.
  - In that case there would be 35 ASFRs to be calculated and summed – but there would be no need for multiplication.

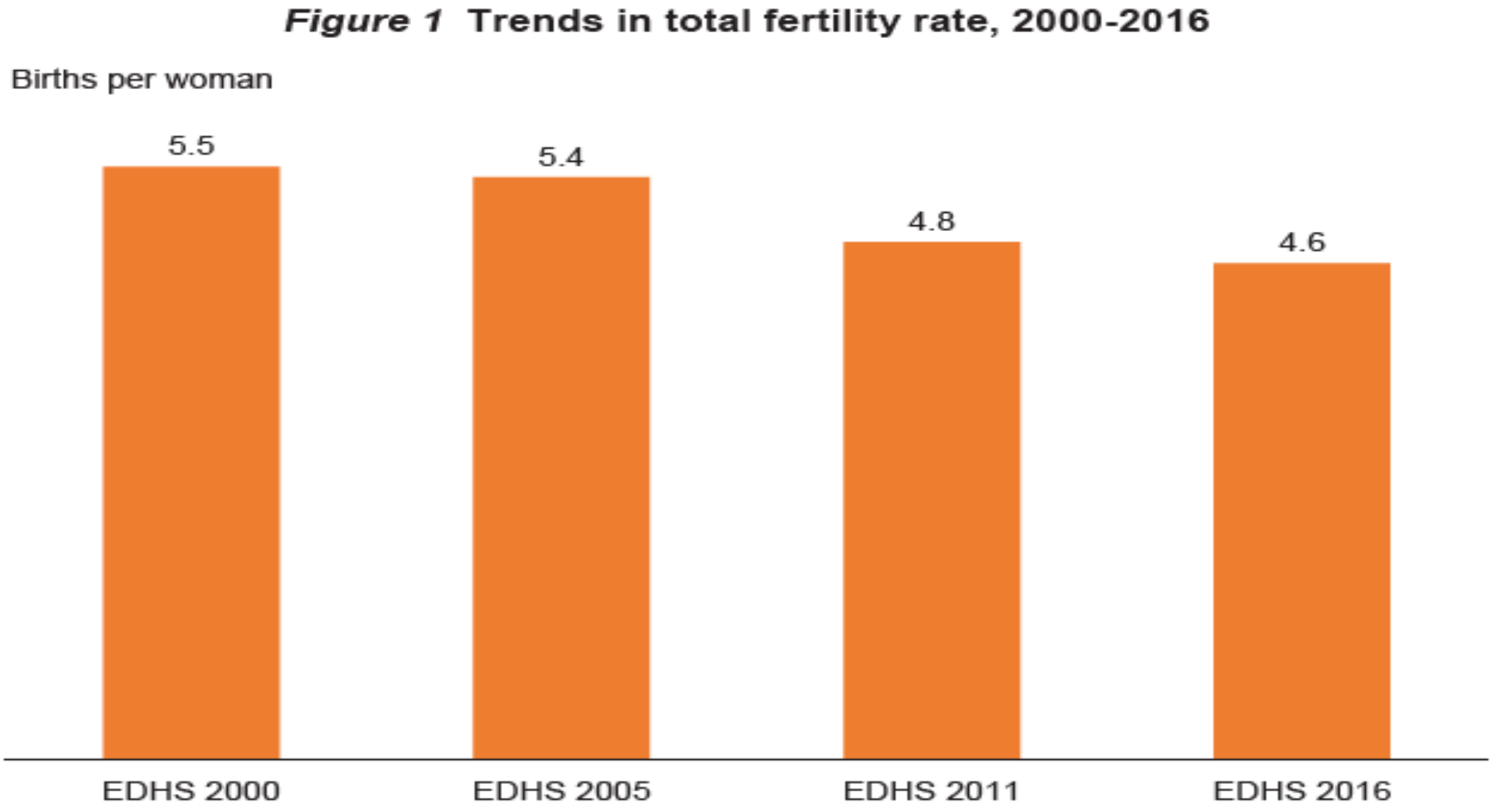
The Total Fertility Rate (TFR)

- TFR can fluctuate considerably from year to year – following economic trends, strife etc.
- If the TFR rises for a few years but then drops back the quantum of fertility may not be really changing – it is just the timing of births that has produced an apparent and temporary change.
  - Thus, this is a **tempo** effect

Figure that shows the France: total period fertility rate by year with tempo effect



Trends of TFR in Ethiopia



Period fertility vs. cohort fertility

Period Fertility

- Relates to short intervals of time
- Looks at births to women in all age groups
- Can produce very current estimates
- Data can be collected in a single time period or in a cross-sectional survey.
- Summary measure is Total Fertility Rate (TFR).

Cohort Fertility

- Relates to lifetime experience
- Follows real women over their reproductive life-times
- Estimates typically relate to an earlier period
- Requires either longitudinal data or retrospective questioning by means of a birth history from women aged 50+.
- Summary measure is Completed Family Size (CFS)

Replacement level

- This is a figure that represents the level of TFR at which the population, **in the long run**, would remain the same size – not grow or shrink.
- This is popularly taken to be a TFR of 2.1.
- A slight margin on the TFR (2.1 rather than 2) allows for the fact that some girls will not survive to become mothers.
- The commonly quoted replacement figure of TFR=2.1 is an approximation, but in general an acceptable one given modern levels of mortality.
- In very high mortality situations 2.1 would be an under-estimate.

Exercise

Total Fertility Rate (TFR)

- Use the following data to calculate the TFR per 1,000

Island of Mauritius, 1985

Age Group of Mother	Women	Births
15–19	52,013	1,884
20–24	54,307	6,371
25–29	46,990	5,362
30–34	40,211	2,901
35–39	30,401	1,170
40–44	23,496	268

Source: U.N. Demographic Yearbook, 1986

Gross Reproduction Rate (GRR)

- ✓ GRR is the average number of daughters that would be born to a woman throughout her child bearing age (15-49 years), if she were to pass through all her child bearing age.

Gross Reproduction Rate (GRR)

- ✓The GRR is calculated by multiplying the TFR by the proportion of female births to total birth.
- ✓This rate is like the TFR except that, it counts only daughters and literally measures “reproduction”; a woman reproducing herself by having a daughter.

Gross Reproduction Rate (GRR)

- If the sex ratio at birth is assumed constant across ages of women

$$GRR = \sum ASFR(a) * \frac{B^{f}(a)}{B^{m+f}(a)}$$

$$GRR = TFR * (\text{Proportion of female births})$$

Where B^f = Number of female births

B^{m+f} = Number of male and female births, i.e., all births

Exercise

Gross Reproduction Rate (GRR)

- Use the following data to calculate the GRR

United States, 1990

Age Group of Mother	Women	Total	Births
Males			
15-19	8 651	522	267
20-24	9 345	094	560
25-29	10 617	1277	653
30-34	10 986	886	454
35-39	10 061	318	163
40-44	8 924	49	25

Numbers are in 1,000s

Sources:

Exercise Answer

Gross Reproduction Rate (GRR)

- The correct answer for the gross reproduction rate is as follows:
  - **1.01 daughters per woman**

United States, 1990

Age Group of Mother	Women	Total	Births Males
15-19	8 651	522	267
20-24	9 345	094	560
25-29	10 617	1277	653
30-34	10 986	886	454
35-39	10 061	318	163
40-44	8 924	49	25

Numbers are in 1,000s

Net Reproduction Rate (NRR)

- It is the average number of daughters that would be born to a woman if she passed through her child bearing ages conforming to the age specific fertility and **mortality** rates of a given year.

Net Reproduction Rate (NRR)

- For single-year age groups

$$NRR = \sum ASFR * \frac{{}_1L_a^f}{l_0^f} * \frac{B^f}{B^{m+f}}$$

- For five-year age groups

$$NRR = \sum ASFR * \frac{{}_5L_a^f}{l_0^f} * \frac{B^f}{B^{m+f}}$$

Where:

L^f_a = Life table person-years lived by women in age group "a"

l_0 = Radix of life table

B^f = Number of female births

B^{m+f} = Number of male and female births

Net Reproduction Rate (NRR)

- **NRR** – measures the extent to which a cohort of newly born girls will replace their mothers under predetermined schedules of fertility & mortality.
  - **NRR** of 1 = replacement level
  - **NRR** <1 = below replacement(-ve growth)
  - **NRR** >1,implies more than replacement level

Net Reproduction Rate (NRR)

- NRR is always lower than GRR, because it takes into account the fact that some women will die before entering and completing their child-bearing years.
- Correspondingly NRR will be less than half the magnitude of the TFR.

Exercise

Net Reproduction Rate (NRR)

- Use the following data to calculate the NRR

United States, 1990

Age Group of Mother	Women	Births Total	Births Males	Stationary Population $_5L_x$
15-19	8 651	522	267	493 629
20-24	9 345	094	560	492 399
25-29	10 617	1277	653	490 989
30-34	10 986	886	454	489 203
35-39	10 061	318	163	486 812
40-44	8 924	49	25	483 465

Numbers are in 1,000s

Sources: U.S. Census *1990 and* Vital Statistics of the U.S. Vol. I

Exercise Answer

Net Reproduction Rate (NRR)

- The correct answer for the NRR is as follows:
  - **1.00 daughter per woman**

United States, 1990

Age Group of Mother	Women	Total	Births Males	Stationary Population ${}_5L_x$
15-19	8 651	522	267	493 629
20-24	9 345	094	560	492 399
25-29	10 617	1277	653	490 989
30-34	10 986	886	454	489 203
35-39	10 061	318	163	486 812
40-44	8 924	49	25	483 465

Numbers are in 1,000s

Determinants of Fertility

What is necessary to have a baby?

Determinants of Fertility

- Human fertility is governed by a number of immediate biological and behavioral factors which are themselves influenced by distal socio-economic, environmental and cultural factors.
- The proximate determinants of fertility are the biological and behavioral factors through which social, economic, cultural and environmental variables influence fertility.
- They have direct effects on fertility.

Concept of proximate determinants

- Davis and Blake (1956), identified a set of 11 'intermediate variables' which directly affect fertility which they grouped into the following three categories.
  - Factors affecting exposure to intercourse ('intercourse variables')
  - Factors affecting exposure to conception ('conception variables')
  - Factors affecting gestation and successful parturition ('gestation variables')

Intercourse variables

A. Those governing the formation and dissolution of unions in the reproductive period:

1. Age of entry into sexual unions
2. Permanent celibacy: proportions of women never entering sexual unions
3. Amount of reproductive period spent after or between unions
   - When unions are broken by divorce, separation, or desertion
   - When unions are broken by death of husband

...Intercourse variables

B. Those governing the exposure to intercourse within union

4. Voluntary abstinence
5. Involuntary abstinence (from impotence, illness, unavoidable but temporary separations)
6. Coital frequency (excluding periods of abstinence)

Conception variables:

7. Fecundity or infecundity, as affected by involuntary causes
8. Use or non-use of contraception
9. Fecundity or infecundity, as affected by voluntary causes

Gestation variables

10. Fetal mortality from involuntary causes
11. Fetal mortality from voluntary causes

- Measures of Mortality

- Life Table

Objectives

- At the end of this session you should be able to:
  - Calculate the basic measures of mortality
  - Evaluate the limitations of basic measures of mortality
  - Recognize different presentations of life tables and understand their utility as a form of survival analysis
  - Understand how the life table columns are constructed
  - Appreciate the utility of each life table column

Introduction

- Death is the permanent disappearance of all evidence of life at any time after live birth has taken place.
- All forms of prenatal mortality are excluded from this definition, irrespective of cause – since according to this definition, one cannot die prior to having live birth.
- In order to calculate any mortality rate, ratio or probability, two items of information are required:
  - a numerator representing a count of events
  - a denominator giving some (possibly approximate) measure of the population exposed to the risk of generating those events.

Basic measures of mortality

1. The crude death rate (CDR)

- It is the ratio of the numbers of deaths (*D*) observed in a population in a year to the population at risk of dying in that year (*N*), usually multiplied by 1 000:

$$CDR = \frac{D}{N} \times 1000$$

- Demographers tend to use a period of a year – typically a calendar year running from 1 January to 31 December.
- The numerator comes directly from a tabulation of the number of deaths observed.
- The denominator should reflect the person-years of exposure to the risk of dying in the population over the course of the year.

...measures of mortality

2. Specific mortality rates:

Age-specific mortality rates

- The crude death rate makes no allowance for the mortality pattern in a population by age.
- Thus, death rates can be calculated for specific age groups, in order to compare mortality at different ages.

$$ {}_{n}m_{x} = \frac{{}_{n}D_{x} \times 1000}{{}_{n}N_{x}} $$

where:

${}_{n}m_{x}$ = Age specific mortality rate *between ages x and x+n in the year*

${}_{n}D_{x}$ *= the number of deaths between ages x and x+n in the year*

${}_{n}N_{x}$ *= the (mid-year) population aged between x and x+n.*

For ease of interpretation, these rates are typically expressed per 1000, or sometime per 100 000.

...measures of mortality

Infant Mortality Rate (IMR)

- Infant Mortality Rate is the number of deaths of infants under one year of age (0-11 months of age) per 1000 live births in a given year.
- Infant (children under one year of age) are at highest risk of death than any other age group.

$$IMR = \frac{\text{No of death of children < 1 year in a year}}{\text{Total live births during that year}} \times 1000$$

Why Infant Mortality Rates ?

- The IMR is a good indicator of the overall health status of a population.
- It is a major determinant of life expectancy at birth
- The IMR is sensitive to levels and changes in socio-economic conditions of a population

...measures of mortality

Child Mortality Rate (CMR)

- It is the number of deaths of children 1-4 years of age per 1000 children 1-4 years of age.

$$CMR = \frac{\text{No of deaths of children 1-4 year in a year}}{\text{Total number of children 1-4 years of age}} \times 1000$$

...measures of mortality

Under Five Mortality Rate (<5MR)

- It is the number of deaths of children under five years of age in a year (0-4 years of age) per 1000 children under five years of age (0-4 year).

$$<5MR = \frac{\text{No of deaths of children <5 yrs in a year}}{\text{Total number of children < 5 years of age}} X\ 1000$$

…measures of mortality

Sex Specific Mortality Rates (SSMR):

- It is the number of deaths among a specific sex group (males or females) per 1000 population of the same sex group.

$$\textit{SSMR for males} = \frac{\textit{Number of deaths among males}}{\textit{Total number of males}} \times 1000$$

- Sex specific mortality rate is used to determine which sex group is at higher risk of death than the other.

…measures of mortality

Maternal Mortality Indicators

- A maternal death is defined as the death of a woman while pregnant or within 42 days of termination of pregnancy from any cause related to or aggravated by the pregnancy or its management but not from accidental or incidental causes.
- Commonly the following indicators are used to measure maternal death:
  - Maternal mortality *ratio*
  - Maternal mortality *rate*
  - Life-time risk of maternal mortality

Maternal Mortality Ratio

- It is the number of maternal deaths due to maternal causes in a given year per 100,000 live births in that year.
  - Deaths due to complications of spontaneous or induced abortions are included.

$$= \frac{D_{mc}}{B} * 100000$$

Let Dmc= Number of deaths due to maternal causes

B = Total live birth

- Represents the risk associated with each pregnancy, i.e., the obstetric risk

Maternal Mortality Rate

- Number of maternal deaths due to maternal causes per 1,000 women of reproductive ages

$$= \frac{D_{mc}}{W_{15-49}} * 1000$$

Where: Dmc= Number of deaths due to maternal causes
W15-49 = Number of women of reproductive age

- Represents both the obstetric risk and the frequency with which women are exposed to this risk.

Note: Maternal mortality ratio is more widely used

Life time risk of maternal mortality

- Life-time risk reflects the risk that a woman will die of maternal causes at some point of her reproductive span, given current rates of maternal mortality and fertility.

Life time risk of maternal mortality

=

[1-(1-maternal mortality rate)35]

…measures of mortality

3. Adjusted Mortality Rates:

- Populations may differ in regard to many characteristics & comparisons of crude rates can be misleading.
- Differences in determinants can distort comparisons between populations.
- Thus, mortality rates adjusted for major confounders like age and sex could be used to compare mortality in two or more populations or one population in different time periods.

…measures of mortality

- There are several techniques for adjusting mortality rates:

  - ✓ Standardization
  - ✓ Life table rates
  - ✓ Regression analysis…

Life Table

Introduction

- Life tables are, in essence, one form of combining mortality rates of a population at different ages into a single statistical model.
- They have a much wider application than mortality alone and are a particular form of survival analysis.
- As such the technique can be applied to any data which comes in the form of varying durations of time between a definable start point and end point.

Introduction

- It is used by public health workers, demographers, actuaries, and many others:
  - in studies of longevity, fertility, migration, and population growth.
  - in making projections of population size and characteristics .
  - in studies of widowhood, orphanhood, length of married life, length of working life, and length of disability free life.

Demographic Applications of Life Table

- One important method of assessing the health of a population is to ask how long people can expect to live.
- Life expectancy, usually reported at birth although it can be applied to other ages as well, is a commonly used summary measure which can also be used to compare against countries.
- Life expectancy is calculated using life tables.

Types of Life Tables

Life tables differ according to;

- reference year (time) of the table:
  - cohort life table
  - period life table
- number of factors comprehended by the table :
  - Single decrement life table
  - multiple decrement life table
  - Multi-state life table
- level of analysis: Macro vs micro
  - Macro: aggregate level
  - Micro: individual level

Period life table (L.T) (Assumptions)

- The period L.T follows a hypothetical cohort of some standard number of births (100,000, 10,000 or 1,000).
- The cohort is 'closed' against migration (in & out).
- People die at each age according to a schedule that is fixed in advance & does not change.

  - Age-specific mortality rates remain constant

- The ASDR in the life table population is the same as the ASDR in the actual population during a specified period.

Construction of Period Life Table

Prerequisites:

- Population living at all individual ages in a selected period.
- No. of deaths that occurred in these ages during the selected period.

Construction of Life Table

- There are many functions of life table but the Basic functions of a Period life table are :

Columns of a Life Table

The basic columns of a life table are:

(1) (x, x+n)	(2) ${}_nq_x$	(3) l_x	(4) ${}_nd_x$	(5) ${}_nL_x$	(6) T_x	(7) e_x
Age Interval	Probability of dying in interval	Survivors at exact age x	Number dying in interval	# years lived by population in interval	# years lived by population in interval & subsequent intervals	Average years of life remaining

Construction of Life Table

The following are the definition of L.T functions:

l_x – number of people left alive at age ***x***

${}_nd_x$ – number of people dying between age ***x*** and ***x+n***

${}_nq_x$ – probability of dying between age ***x*** and ***x+n***

${}_np_x$ – probability of surviving from age ***x*** and ***x+n***

${}_nL_x$ – person-years lived between age ***x*** and ***x+n***

T_x – person-years lived above age ***x***

e_x^o – expectation of life at age ***x***

${}_nm_x$ – age-specific death rates between age ***x*** and ***x+n***

${}_na_x$ – average person-years lived between age ***x*** and ***x+n*** for persons dying in the interval

Steps in Construction of period life table

1. Calculate ASDR (${}_nm_x$) from period data
2. Adopt a set of ${}_na_x$ values
3. Compute probability of dying(${}_nq_x$) x, x+n
4. Compute probability of surviving (${}_np_x$) x, x+n
5. Choose number of births (l_0) for hypothetical cohort
6. Calculate survivors at exact age x
7. Calculate number of dying x, x+n
8. Calculate person years lived x, x+n
9. Calculate person years lived above age x
10. Calculate life expectancy at age x

Age-specific death rates
($_nm_x$)

- Age specific death rate is calculated as:

$$ {}_n m_x = \frac{{}_nD_x \times 1000}{{}_nN_x} $$

where:

$_nm_x$ = Age specific mortality rate *between ages x and x+n in the year*

$_nD_x$ *= the number of deaths between ages x and x+n in the year*

$_nN_x$ *= the (mid-year) population aged between x and x+n.*

For ease of interpretation, these rates are typically expressed per 1000, or sometimes per 100 000.

Average person-years lived for persons dying in the interval($_n a_x$)

- This is the average proportion of the time lived in the interval x to x+n for those persons who die during the interval.

- The $_n a_x$ values, however, are not easily obtained

- But, there are several strategies for choosing a set of $_n a_x$ values.

Strategies for choosing ${}_n a_x$

- Borrow from another population
- Rule of the thumb:
  - If n is small, use ${}_n a_x = n / 2$.
  - For very young age groups, value of a_0=0.3 and ${}_4a_1$=0.4 is used for high mortality countries, and for low mortality countries a_0 value of 0.1 is used.
- Direct observation: if data on exact age at death are available for a population, it is possible to estimate ${}_n a_x$ values.

Probability of dying ($_nq_x$)

- It is the probability of dying between exact age x and x+n.
- The proportion of persons reaching exact age x who will die before reaching age x+n.

$$_nq_x = \frac{_nd_x}{l_x} = \frac{\text{Number of deaths in the cohort between ages } x \text{ and } x+n}{\text{Number of survivors to age } x \text{ in the cohort}}$$

- It could also derived from the age specific death rate by using the following formula.

Probability of dying (${}_nq_x$)

- Conversion of the ${}_nm_x$ to ${}_nq_x$ requires information on ${}_na_x$.
- If we have data on ${}_na_x$ values, then:

$$ {}_nq_x = \frac{n \cdot {}_nm_x}{1 + (n - {}_na_x)_n m_x} $$

- If we have no ${}_na_x$, *a simple way to estimate* ${}_nq_x$ from ${}_nm_x$ is;

$$ {}_nq_x = \frac{2*n*(nMx),}{2+n*(nMx)} $$

Probability of surviving ($_np_x$)

- This is the probability of surviving between exact ages x and x+n.

- It is the compliment of $_nq_x$.

$$_np_x = 1 - {_nq_x}$$

Or. $_np_x = {_nL_x}/l_o$

Survivors at exact age x(l_x)

- This is the number of persons alive at exact age x out of the original cohort.
- The hypothetical cohort (l_0), is usually set at around 1,000, 10,000 or 100,000.
- Then to calculate the remaining l_x values, the following formula is used.

$$l_{x+n} = l_x \cdot {}_np_x$$

Number of Death($_nd_x$)

- This is the number of persons dying between exact ages x and x+n.
- It is the difference between two subsequent l_x values.

$$_nd_x = l_x - l_{x+n}$$

- $_nd_x$ can also be calculated by using ;

$$_nd_x = l_x * {}_nq_x$$

Interval Person-years (${}_nL_x$)

- It is the number of person years lived between exact ages x and x+n.
- person years : is the sum of the time(expressed in years) spent in a given category by the individuals comprising the population.
- Each person surviving through the interval contributes n person years, while those who die during the interval will contribute only ${}_na_x$ years.

$$ {}_nL_x = n \cdot l_{x+n} + {}_na_x \cdot {}_nd_x $$

For open ended age interval

Since, $_n m_x = \frac{_n d_x}{_n L_x}$ when n = ∞

$$_\infty m_x = \frac{_\infty d_x}{_\infty L_x}$$

$$_\infty q_x = 1.00$$

$$_\infty L_x = \frac{_\infty d_x}{_\infty m_x}$$

$$_\infty p_x = 1 - {}_\infty q_x = 0.00$$

$$_\infty L_x = \frac{l_x}{_\infty m_x}$$

Total person years (T_x)

- This is the total number of person years lived after exact age x.
- It is thus the ${}_nL_x$ column cumulated from the bottom.
- The main purpose of this function is in the calculation of the expectation of life.
- It is given by the following formula;

$$T_x = \sum_{a=x}^{\infty} {}_nL_a$$

Life expectancy (e_x)

- This is the expectation of life at age x, or the average number of additional years a person aged x will live.

- Since the total number of years left to be lived by l_x people is T_x, then the expectation of life is obtained by;

$$e_x^o = \frac{T_x}{l_x}$$

Some general notes on L.T

- Some functions (l_x, T_x, e_x) refer to a single (exact) age while,
- Other functions (${}_nd_x$, ${}_np_x$, ${}_nq_x$, ${}_nm_x$, ${}_na_x$) refer to age intervals that begin with exact age x and extend for n years.
- The length of the interval may vary within the same life table.
  - Most conventional: 0,1,5,10,…,80,85+

life table

Age x	${}_nN_x$	${}_nD_x$	${}_nm_x$	${}_na_x$	${}_nq_x$	${}_np_x$	l_x	${}_nd_x$	${}_nL_x$	T_x	e_x^o
0	47,925	419	0.008743	0.068	0.008672	0.991328	100,000	867	99,192	7,288,901	72.889
1	189,127	70	0.000370	1.626	0.001479	0.998521	99,133	147	396,183	7,189,709	72.526
5	234,793	36	0.000153	2.500	0.000766	0.999234	98,986	76	494,741	6,793,526	68.631
10	238,790	46	0.000193	3.143	0.000963	0.999037	98,910	95	494,375	6,298,785	63.682
15	254,996	249	0.000976	2.724	0.004872	0.995128	98,815	481	492,980	5,804,410	58.740
20	326,831	420	0.001285	2.520	0.006405	0.993595	98,334	630	490,106	5,311,431	54.014
25	355,086	403	0.001135	2.481	0.005659	0.994341	97,704	553	487,127	4,821,324	49.346
30	324,222	441	0.001360	2.601	0.006779	0.993221	97,151	659	484,175	4,334,198	44.613
35	269,963	508	0.001882	2.701	0.009368	0.990632	96,492	904	480,384	3,850,023	39.900
40	261,971	769	0.002935	2.663	0.014577	0.985423	95,588	1,393	474,686	3,369,639	35.252
45	238,011	1,154	0.004849	2.698	0.023975	0.976025	94,195	2,258	465,777	2,894,953	30.734
50	261,612	1,866	0.007133	2.676	0.035082	0.964918	91,937	3,225	452,188	2,429,176	26.422
55	181,385	2,043	0.011263	2.645	0.054861	0.945139	88,711	4,867	432,096	1,976,988	22.286
60	187,962	3,496	0.018600	2.624	0.089062	0.910938	83,845	7,467	401,480	1,544,893	18.426
65	153,832	4,366	0.028382	2.619	0.132925	0.867075	76,377	10,152	357,713	1,143,412	14.971
70	105,169	4,337	0.041238	2.593	0.187573	0.812427	66,225	12,422	301,224	785,699	11.864
75	73,694	5,279	0.071634	2.518	0.304102	0.695898	53,803	16,362	228,404	484,475	9.005
80	57,512	6,460	0.112324	2.423	0.435548	0.564452	37,441	16,307	145,182	256,070	6.839
85+	32,248	6,146	0.190585	5.247	1.000000	0.000000	21,134	21,134	110,889	110,889	5.247

Lesson 4: Probability

Course Outline

- Introduction to Biostatistics
- Descriptive Biostatistics
- Probability
- Discrete Probability Distributions
- Continuous Probability Distributions
- Estimation
- Hypothesis Testing

Lecture Outline

- Introduction to probability
- Sample Space & Events
- Mutually Exclusive Events
- Union & Intersection
- Joint Probability
- Complementary Events
- Conditional Probability
- Multiplication Law of Probability
- Independent Events
- Dependent Events
- Addition Law of Probability
- Marginal Probability
- Sensitivity & Specificity
- Bayes' Law
- Bayesian Inference

Introduction

- The theory of probability provides the foundation for statistical inference.
- The concept of probability is not foreign to health workers and is frequently encountered in everyday communication.

For example;

i. A physician say that a patient has a *50–50 chance* of surviving a certain operation.

ii. A physician may say that she is *95 percent* certain that a patient has a particular disease.

iii. A public health nurse may say that *nine times out of ten* a certain client will break an appointment.

- As these examples suggest, most people express probabilities in terms of percentages.

Introduction

In dealing with probabilities mathematically, it is more convenient to express probabilities as fractions.

Thus, we measure the probability of the occurrence of some event by a number between zero and one.

The more likely the event, the closer the number is to one; and the more unlikely the event, the closer the number is to zero.

An event that cannot occur has a probability of zero, and an event that is certain to occur has a probability of one.

Cont'd

Probability: can be defined as a measure of the likelihood that a particular event will occur or it is a science of decision making with calculated risk in face of uncertainty.

In general, probability is the chance of an outcome of an experiment. It is the measure of how likely an outcome is to occur.

Sample Space (S)

- The ***sample space*** is the set of all possible outcomes.
- The sample space is represented by the symbol S.
- There are 2 possible outcomes with the sample space of tossing a coin;

$$S = \{\text{Head, Tail}\}$$

- There are 6 possible outcomes with the sample space of rolling a dice;

$$S = \{1, 2, 3, 4, 5, 6\}$$

- There are 52 possible outcomes with the sample space of drawing a card;

$$S = \{2\spadesuit, 2\clubsuit, 2\diamondsuit, 2\heartsuit, 3\spadesuit, 3\clubsuit, 3\diamondsuit, 3\heartsuit, \ldots, A\spadesuit, A\clubsuit, A\diamondsuit, A\heartsuit\}$$

Event

- An ***event*** is a subset of a sample space S.
- In referring to probabilities of events, an event is any set of outcomes of interest.
- The symbol { } is used as shorthand for the phrase “the event.”
- The probability of an event E, denoted by $P(E)$ or $Pr(E)$, always satisfies $0 \leq Pr(E) \leq 1$.

Probability

The ***probability*** of an event is the relative frequency of this set of outcomes over an indefinitely large (or infinite) number of trials i.e. sample space.

Probability (Example)

The primary aim of a study by Carter et al. was to investigate the effect of the age at onset of bipolar disorder on the course of the illness. One of the variables investigated was family history of mood disorders. Table shows the frequency of a family history of mood disorders in the two groups of interest (Early age at onset defined to be 18 years or younger and Later age at onset defined to be later than 18 years). Suppose we pick a person at random from this sample. What is the probability that this person will be 18 years old or younger ?

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

Source: Tasha D. Carter, Emanuela Mundo, Sagar V. Parkh, and James L. Kennedy, "Early Age at Onset as a Risk Factor for Poor Outcome of Bipolar Disorder," *Journal of Psychiatric Research, 37* (2003), 297–303.

$$P(E) = \text{number of Early subjects/total number of subjects}$$
$$= 141/318 = .4434$$

Mutually Exclusive

Two events A and B are ***mutually exclusive***, or ***disjoint***, if $A \cap B = \Phi$, that is, if A and B have no elements in common.

or

Two events A and B are ***mutually exclusive*** if they cannot both happen at the same time.

If outcomes A and B are two events that cannot both happen at the same time, then;

$$Pr(A \text{ or } B) = Pr(A) + Pr(B)$$

Example-1: Hypertension

- Let A be the event that a person has normotensive diastolic blood pressure (DBP) readings (DBP < 90), and
- Let B be the event that a person has borderline DBP readings (90 ≤ DBP < 95).
- Suppose that $Pr(A) = 0.7$, and $Pr(B) = 0.1$.
- Let Z be the event that a person has a DBP < 95.
- Then $\boldsymbol{Pr(Z) = Pr(A) + Pr(B) = 0.8}$

The events A and B are ***mutually exclusive*** because they cannot occur at the same time.

Example-2: Hypertension

- Let X be diastolic blood pressure (DBP),
- Let C be the event $X \geq 90$, and
- Let D be the event $75 \leq X \leq 100$.

Events C and D are ***not mutually exclusive***, because they both occur when $\mathbf{90 \leq X \leq 100}$.

Union

The ***union*** of the two events A and B, denoted by the symbol $A \cup B$, is the event containing all the elements that belong to A or B or both.

There are two special cases in the union.

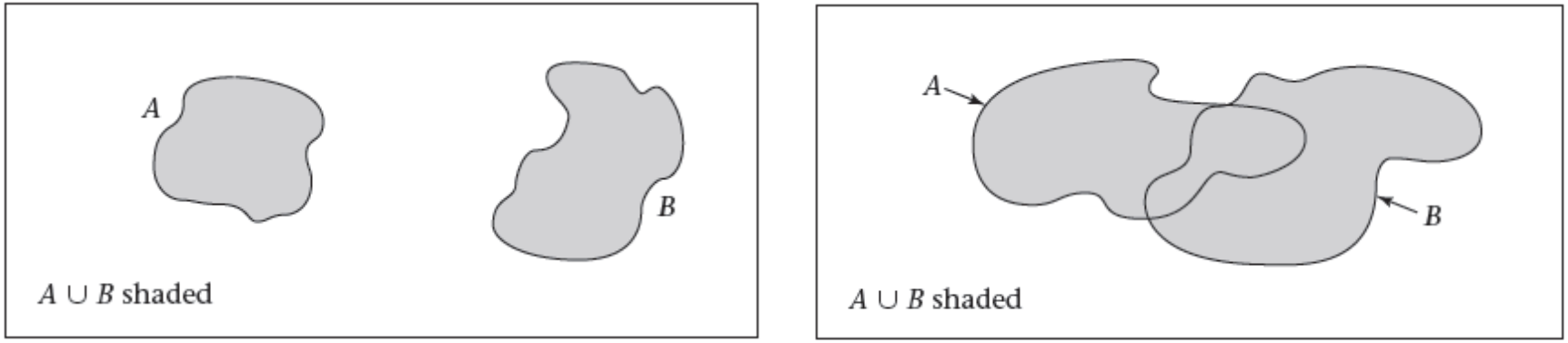
i. **$A \cup B$ can be mutually exclusive**

ii. **$A \cup B$ can be not mutually exclusive**

The given figure diagrammatically depicts $A \cup B$ both for the case in which A and B are

and are not mutually exclusive.

Diagrammatic representation of $A \cup B$:



(a) **A, B mutually exclusive**

(b) **A, B not mutually exclusive**

Union (Examples)

Example: when $A \cup B$ is mutually exclusive

Example Hypertension:

- Let events A is defined as $A = \{X < 90\}$,
- Let events B is defined as $B = \{90 \leq X < 95\}$,
- where X = diastolic blood pressure (DBP).
- Then $\boldsymbol{A \cup B = \{X < 95\}}$.

Example: when $A \cup B$ is not mutually exclusive

Example Hypertension:

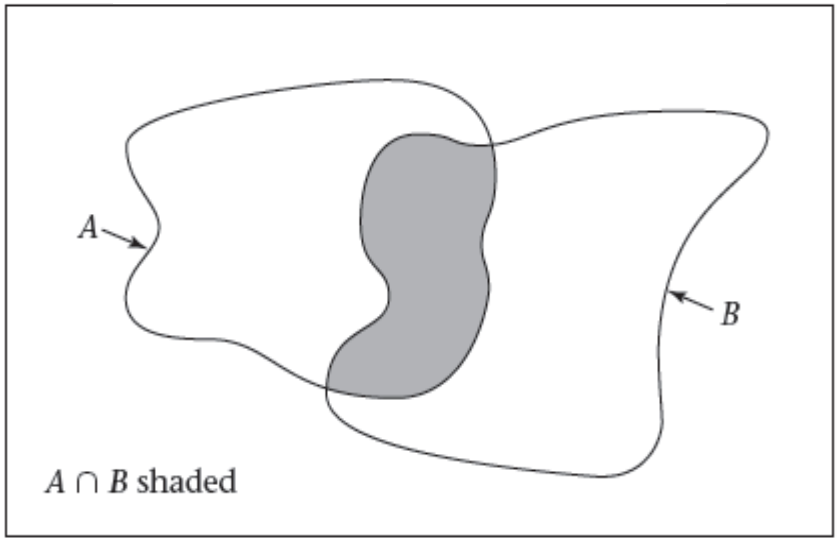
- Let events C is defined as $C = \{X \geq 90\}$,
- Let events D is defined as $D = \{75 \leq X \leq 100\}$,
- where X = diastolic blood pressure (DBP).
- Then $\boldsymbol{C \cup D = \{X \geq 75\}}$.

Joint Probability or Intersection

- Sometimes we want to find the probability that a subject picked at random from a group of subjects possesses two characteristics at the same time. Such a probability is referred to as a ***joint probability***.

- The ***intersection*** of two events A and B, denoted by the symbol $A \cap B$, is the event containing all elements that are common to A and B.
- $A \cap B$ is depicted diagrammatically in the figure.

Diagrammatic representation of $A \cap B$



Joint Probability or Intersection (Example-1)

Example-1: What is the probability that a person picked at random from the 318 subjects will be Early (*E*) and will be a person who has no family history of mood disorders (*A*)?

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Cont'd

- The joint probability may be written in symbolic notation as $P(E \cap A)$.
- The symbol $\cap$ is read either as "**intersection**" or "**and**".
- The statement indicates the joint occurrence of conditions *E* and *A*.
- The number of subjects satisfying both of the desired conditions is found at the intersection of the column labeled *E* and the row labeled *A* and is seen to be **28**.
- Since the selection will be made from the total set of subjects, the denominator is **318**.
- Thus, we may write the joint probability as $\boldsymbol{P(E \cap A) = 28/318 = 0.881}$

Example-2: Hypertension

- Let events C is defined as $C = \{X \geq 90\}$,
- Let events D is defined as $D = \{75 \leq X \leq 100\}$,
- where X = diastolic blood pressure (DBP).
- Then $\boldsymbol{C \cap D = \{90 \leq X \leq 100\}}$.

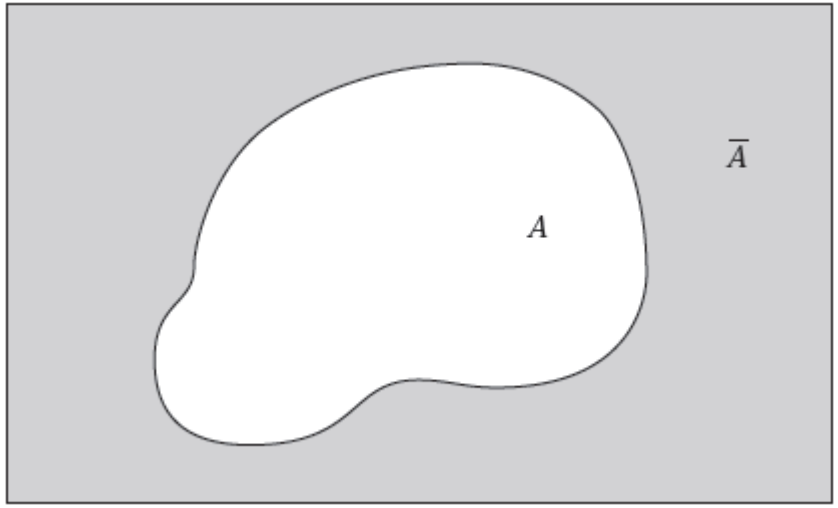
Complementary Events

- $\bar{A}$ is the event that event A does not occur. It is called the complement of A.

- The probability of an event A is equal to 1 minus the probability of its complement, which is written $\bar{A}$ and

$$P(\bar{A}) = 1 - P(A)$$

Diagrammatic representation of $\bar{A}$



Complementary Events (Example-1)

Example: Hypertension

- Let events A is defined as $A = \{X < 90\}$,
- Let events C is defined as $C = \{X \geq 90\}$,
- Suppose that $Pr(A) = 0.7$,
- where X = diastolic blood pressure (DBP).
- Then $\boldsymbol{C = \bar{A}}$, because C can only occur when A does not occur.
- Notice that $\boldsymbol{Pr(C) = Pr(\bar{A}) = 1 - 0.7 = 0.3}$
- Thus, if 70% of people have DBP < 90, then 30% of people must have DBP ≥ 90.

Complementary Events (Example-2)

Suppose that of 1200 admissions to a general hospital during a certain period of time, 750 are private admissions. If we designate these as set $\boldsymbol{A}$, then $\overline{\boldsymbol{A}}$ is equal to 1200 minus 750, or 450. We may compute

$$P(A) = 750/1200 = .625$$

and

$$P(\overline{A}) = 450/1200 = .375$$

and see that

$$P(\overline{A}) = 1 - P(A)$$
$$.375 = 1 - .625$$
$$.375 = .375$$

Independent Events

Two events *A* and *B* are called ***independent events*** if

$$Pr\,(A \cap B) = Pr\,(A) \times Pr\,(B)$$

If we want to compute the probability of two or several events occurring simultaneously, and If the events are independent, then we can use the multiplication law of probability to do so.

Conditional Probability

When probabilities are calculated with a subset of the total group as the denominator, the result is a ***conditional probability***.

- If two events are independent, then

$$Pr(A \cap B) = Pr(A) \times Pr(B)$$

- The conditional probability can be calculated by dividing both sides by $Pr(A)$ or $Pr(B)$ (depending what is the given condition should go in the denominator), then

$$Pr(B) = Pr(A \cap B)/Pr(A) = Pr(B \mid A)$$

Conditional Probability (Example)

Suppose we pick a subject at random from the 318 subjects and find that he is 18 years or younger (*E*). What is the probability that this subject will be one who has no family history of mood disorders (*A*)?

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Solution

- The total number of subjects is no longer of interest, since, with the selection of an Early subject, the Later subjects are eliminated.

Conditional Probability (Example)

- We may define the desired probability, then, as follows: What is the probability that a subject has no family history of mood disorders (*A*), given that the selected subject is Early (*E*)?

- This is a conditional probability and is written as $P(A \mid E)$ in which the vertical line is read "given."

- The 141 Early subjects become the denominator of this conditional probability, and 28, the number of Early subjects with no family history of mood disorders, becomes the numerator.

- Our desired probability, then, is

$$P(A \mid E) = 28/141 = .1986$$

The Multiplication Law of Probability

If $A_1, \ldots, A_k$ are mutually independent events, then

$$Pr\left(A_1 \cap A_2 \cap \ldots \cap A_k\right) = Pr\left(A_1\right) \times Pr\left(A_2\right) \times \ldots \times Pr\left(A_k\right)$$

Example: Hypertension, Genetics

Suppose we are conducting a hypertension-screening program in the home. Consider all possible pairs of DBP measurements of the mother and father within a given family, assuming that the mother and father are not genetically related. This sample space consists of all pairs of numbers of the form (X, Y) where $X > 0$, $Y > 0$. Certain specific events might be of interest in this context. In particular, we might be interested in whether the mother or father is hypertensive, which is described, respectively, by events ***A* = {mother's DBP ≥ 95}**, ***B* = {father's DBP ≥ 95}.**

Independent Events (Example-1)

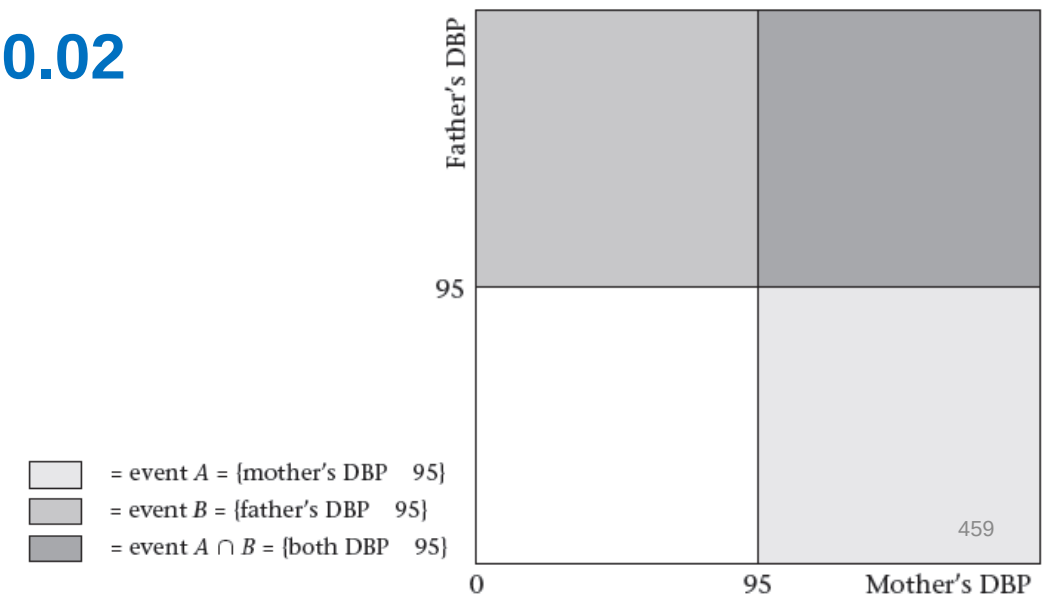
These events are diagrammed in the figure. Suppose that ***Pr*(*A*) = 0.1**, ***Pr*(*B*) = 0.2**. **Compute the probability that both mother and father are hypertensive if the events are independent.**

If *A* and *B* are independent events, then

$$Pr\,(A \cap B) = Pr\,(A) \times Pr\,(B)$$

$$\boldsymbol{Pr\,(A \cap B) = 0.1 \times 0.2 = 0.02}$$

Possible diastolic blood-pressure measurements of the mother and father within a given family



Independent Events (Example-2)

In a certain high school class, consisting of 60 girls and 40 boys, it is observed that 24 girls and 16 boys wear eyeglasses. If a student is picked at random from this class, the probability that the student wears eyeglasses, $P(E)$, is 40/100, or 0.4.

a) What is the probability that a student picked at random wears eyeglasses, given that the student is a boy?

$$P(E \mid B) = \frac{P(E \cap B)}{P(B)} = \frac{16/100}{40/100} = .4$$

- Thus the additional information that a student is a boy does not alter the probability that the student wears eyeglasses, and $P(E) = P(E \mid B)$.
- We say that the events being a boy and wearing eyeglasses for this group are independent.
- We may also show that the event of wearing eyeglasses, E, and *not* being a boy, $\bar{B}$ are also independent as follows:

$$P(E \mid \bar{B}) = \frac{P(E \cap \bar{B})}{P(\bar{B})} = \frac{24/100}{60/100} = \frac{24}{60} = .4$$

Independent Events (Example-2)

b) What is the probability of the joint occurrence of the events of wearing eyeglasses and being a boy?

$$P(E \cap B) = P(B)P(E \mid B)$$

- since we have shown that events E and B are independent, we may replace $P(E \mid B)$ by $P(E)$

$$\begin{aligned} P(E \cap B) &= P(B)P(E) \\ &= \left(\frac{40}{100}\right)\left(\frac{40}{100}\right) \\ &= .16 \end{aligned}$$

Dependent Events

- If two events are not independent, then they are said to be ***dependent events***.

- Two events A, and B are dependent if

$$Pr\,(A \cap B) \neq Pr\,(A) \times Pr\,(B)$$

Example: Hypertension, Genetics

Consider all possible diastolic blood pressure (DBP) measurements from a mother and her first-born child.

- Let A = {mother's DBP ≥ 95},
- Let B = {first-born child's DBP ≥ 80},
- Suppose $Pr(A \cap B) = 0.05$, $Pr(A) = 0.1$, $Pr(B) = 0.2$
- Then $\boldsymbol{Pr(A \cap B) = 0.05 > Pr(A) \times Pr(B) = 0.02}$
- The events A, B would be ***dependent***.

Cont'd

- This outcome would be expected because the mother and first-born child both share the same environment and are genetically related.

- In other words, the firstborn child is more likely to have elevated blood pressure in households where the mother is hypertensive than in households where the mother is not hypertensive.

The Addition Law of Probability

We know that if A and B are mutually exclusive events, then $\boldsymbol{Pr(A \cup B) = Pr(A) + Pr(B)}$.

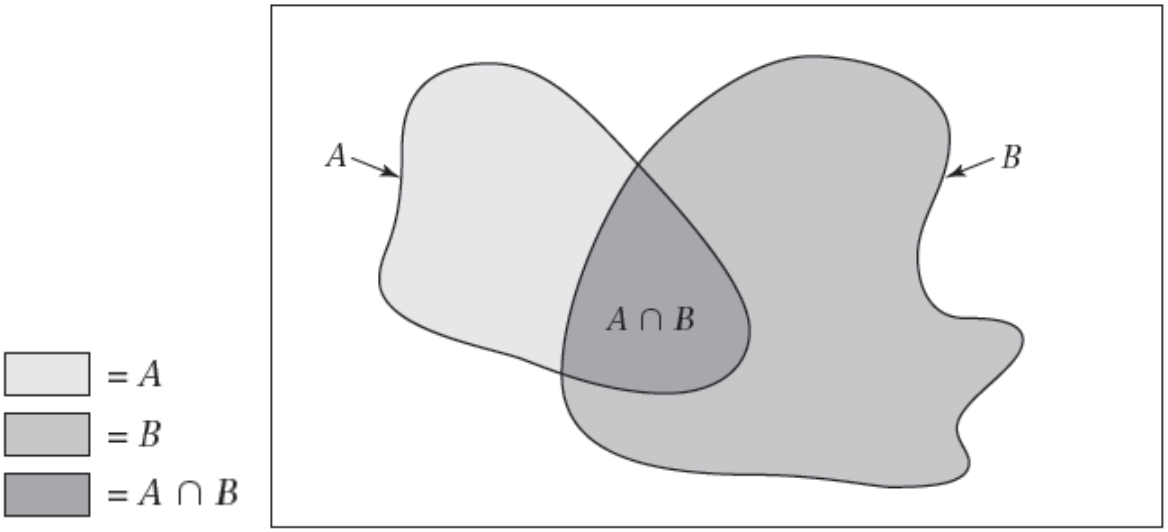
A more general formula for $Pr(A \cup B)$ can be developed when events A and B are not necessarily mutually exclusive.

The ***addition law of probability***, is stated as follows:

If A and B are any events, then

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

Diagrammatic representation of the addition law of probability



The Addition Law of Probability

The ***addition law of probability*** of two events A and B is

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A \cap B)$$

There are Two special cases of the addition law of probability.

First Case: If events *A* and *B* are mutually exclusive

Then, $Pr(A \cap B) = 0$ and the addition law reduces to

$$Pr(A \cup B) = Pr(A) + Pr(B)$$

Second Case: If events *A* and *B* are independent

Then, by definition $Pr(A \cap B) = Pr(A) \times Pr(B)$ and $Pr(A \cup B)$ can be rewritten as

$$Pr(A \cup B) = Pr(A) + Pr(B) - Pr(A) \times Pr(B)$$

- This leads to the following important special case of the addition law.

If two events *A* and *B* are independent, then

$$Pr(A \cup B) = Pr(A) + Pr(B) \times [1 - Pr(A)]$$

The Addition Law of Probability

$$Pr\,(A \cup B) = Pr\,(A) + Pr\,(B) \times [1 - Pr\,(A)]$$

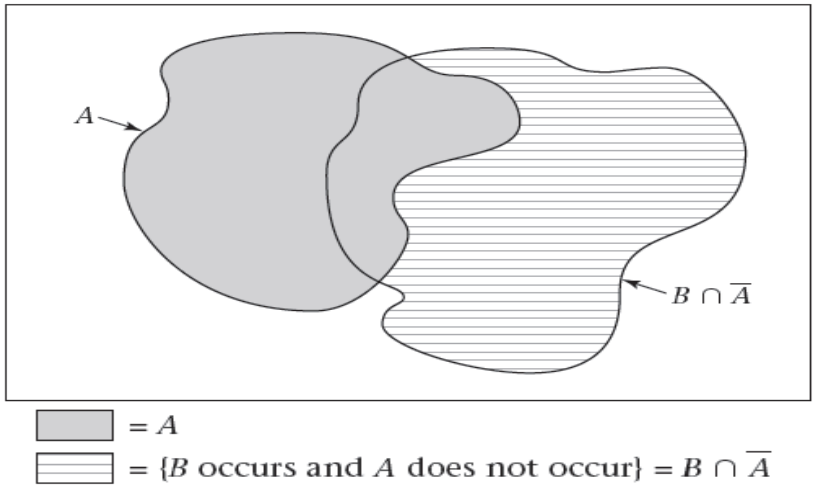
The above special case of the addition law can be interpreted as follows:

The event $A \cup B$ can be separated into two mutually exclusive events:

{*A* occurs} and **{*B* occurs and *A* does not occur}**

Furthermore, because of the independence of A and B, the probability of the latter event can be written as $Pr(B) \times [1 - Pr(A)]$.

Diagrammatic representation of the addition law of probability for independent events



The Addition Law of Probability

It is possible to extend the addition law to more than two events. In particular, if there are three events *A*, *B*, and *C*, then

$$Pr(A \cup B \cup C) = Pr(A) + Pr(B) + Pr(C) - Pr(A \cap B) - Pr(A \cap C) - Pr(B \cap C) + Pr(A \cap B \cap C)$$

This result can be generalized to an arbitrary number of events.

The Addition Law of Probability (Example-1)

Example: Hypertension, Genetics

Let event ***A* = {mother's DBP ≥ 95}**, ***B* = {father's DBP ≥ 95}**, ***Pr*(*A*) = 0.1**, and ***Pr*(*B*) = 0.2**. Assume *A* and *B* are independent events. Suppose a "hypertensive household" is defined as one in which either the mother or the father is hypertensive, with hypertension defined for the mother and father, respectively, in terms of events *A* and *B.* What is the probability of a hypertensive household?

Solution

Pr(hypertensive household) is

$Pr\,(A \cup B) = Pr\,(A) + Pr\,(B) \times [1 - Pr\,(A)]$

$Pr\,(A \cup B) = 0.1 + 0.2 \times [1 - 0.1]$

$\boldsymbol{Pr\,(A \cup B) = 0.28}$

Thus 28% of all households will be hypertensive.

The Addition Law of Probability (Example-2)

If we select a person at random from the 318 subjects represented in the Table, what is the probability that this person will be an Early age of onset subject (*E*) or will have no family history of mood disorders (*A*) or both?

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Solution

From the information in Table, we calculate;

- $P(E) = 141/318 = 0.4434$
- $P(A) = 63/318 = 0.1981$
- $P(E \cap A) = 28/318 = 0.0881$
- $P(E \cup A) = P(E) + P(A) - P(E \cap A)$
- $P(E \cup A) = 0.4434 + 0.1981 - 0.0881 = 0.5534$

Marginal Probability

We may define marginal probability as follows:

Given some variable that can be broken down into m categories designated by $A_1, A_2, \ldots, A_i, \ldots, A_m$ and another jointly occurring variable that is broken down into n categories designated by $B_1, B_2, \ldots, B_j, \ldots, B_n$, the marginal probability of A_i, $P(A_i)$, is equal to the sum of the joint probabilities of A_i with all the categories of B. That is,

$$P(A_i) = \sum P(A_i \cap B_j), \qquad \text{for all values of } j$$

Marginal Probability (Example)

Part-A: Compute the marginal probability *P(E)* of the data given in the table.

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (E)	Later > 18 (L)	Total
Negative (A)	28	35	63
Bipolar disorder (B)	19	38	57
Unipolar (C)	41	44	85
Unipolar and bipolar (D)	53	60	113
Total	141	177	318

The variable age at onset is broken down into two categories,

i. Early for onset 18 years or younger,

ii. Later for onset occurring at an age over 18 years.

Marginal Probability (Example)

- The variable family history of mood disorders is broken down into four categories:

i. Negative family history
ii. Bipolar disorder only
iii. Unipolar disorder only and
iv. Subjects with a history of both unipolar and bipolar disorder.

- The category Early occurs jointly with all four categories of the variable family history of mood disorders.
- The four joint probabilities that may be computed are;

$$P(E \cap A) = 28/318 = .0881$$
$$P(E \cap B) = 19/318 = .0597$$
$$P(E \cap C) = 41/318 = .1289$$
$$P(E \cap D) = 53/318 = .1667$$

Marginal Probability (Example)

The marginal probability $P(E)$ by adding these four joint probabilities as follows:

$$\begin{aligned} P(E) &= P(E \cap A) + P(E \cap B) + P(E \cap C) + P(E \cap D) \\ &= .0881 + .0597 + .1289 + .1667 \\ &= .4434 \end{aligned}$$

The result, as expected, is the same as the one obtained by using the marginal total for Early as the numerator and the total number of subjects as the denominator.

$$P(E) = 141/318 = .4434.$$

Marginal Probability (Example)

Part-B: Compute the marginal probability *P(L)* of the data given in the table. (Exercise)

Table: Frequency of Family History of Mood Disorder by Age Group Among Bipolar Subjects

Family History of Mood Disorders	Early = 18 (*E*)	Later > 18 (*L*)	Total
Negative (*A*)	28	35	63
Bipolar disorder (*B*)	19	38	57
Unipolar (*C*)	41	44	85
Unipolar and bipolar (*D*)	53	60	113
Total	141	177	318

Bayes' Rule and Screening Tests

- In the health sciences field a widely used application of probability laws and concepts is found in the evaluation of screening tests and diagnostic criteria.

- Of interest to clinicians is an enhanced ability to correctly predict the presence or absence of a particular disease from knowledge of test results (positive or negative) and/or the status of presenting symptoms (present or absent).

- Also of interest is information regarding the likelihood of positive and negative test results and the likelihood of the presence or absence of a particular symptom in patients with and without a particular disease.

False Positive

A ***false positive*** is defined as a positive test result when the disease or condition being tested for is not actually present.

False Negative

A ***false negative*** is defined as a negative test result when the disease or condition being tested for is actually present.

Sensitivity

The ***sensitivity*** of a symptom (or screening test) is the probability that the symptom is present given that the person has a disease.

Specificity

The ***specificity*** of a symptom (or screening test) is the probability that the symptom is *not* present given that the person does *not* have a disease.

Predictive Value Positive (PV^+)

The ***predictive value positive*** **(PV^+)** of a screening test (or symptom) is the probability that a person has a disease given that the screening test is positive (or has the symptom).

$$Pr(\text{disease}|\text{test}^+)$$

Predictive Value Negative (PV−)

The ***predictive value negative (PV⁻)*** of a screening test (or symptom) is the probability that a person does *not* have a disease given that the screening test is negative (or does not have the symptom).

$$Pr(\text{no disease} \mid \text{test}^{-})$$

Bayes' Rule

Let *A* = symptom and *B* = disease.

$$\text{Predictive value positive} = PV^{+} = Pr\left(B|A\right)$$

$$\text{Predictive value negative} = PV^{-} = Pr\left(\overline{B}|\overline{A}\right)$$

$$\text{Sensitivity} = Pr\left(A|B\right)$$

$$\text{Specificity} = Pr\left(\overline{A}|\overline{B}\right)$$

Bayes' Rule

The ***Bayes' rule*** allows to obtain the estimates of predictive value positive and predictive value negative of a test (or symptom) from knowledge of a test's (or symptom's) sensitivity and specificity and the probability of the relevant disease in the general population.

Bayes' Rule

Let A = symptom and B = disease.

$$PV^{+} = Pr\left(B|A\right) = \frac{Pr\left(A|B\right) \times Pr(B)}{Pr\left(A|B\right) \times Pr(B) + Pr\left(A|\overline{B}\right) \times Pr\left(\overline{B}\right)}$$

In words, this can be written as

$$PV^{+} = \frac{\text{Sensitivity} \times x}{\text{Sensitivity} \times x + \left(1 - \text{Specificity}\right) \times (1 - x)}$$

where $x = Pr(B)$ = prevalence of disease in the reference population. Similarly,

$$PV^{-} = \frac{\text{Specificity} \times (1 - x)}{\text{Specificity} \times (1 - x) + \left(1 - \text{Sensitivity}\right) \times x}$$

Bayes' Rule

To derive the equation of predictive value positive (PV^+), we have, from the definition of conditional probability,

$$PV^+ = Pr\left(B|A\right) = \frac{Pr(B \cap A)}{Pr(A)}$$

Also, from the definition of conditional probability,

$$Pr(B \cap A) = Pr\left(A|B\right) \times Pr(B)$$

Finally, from the total-probability rule,

$$Pr(A) = Pr\left(A|B\right) \times Pr(B) + Pr\left(A|\overline{B}\right) \times Pr\left(\overline{B}\right)$$

Cont'd

If the expressions for $Pr(B \cap A)$ and $Pr(A)$ are substituted into the equation for PV^+, we obtain,

$$PV^+ = Pr\left(B|A\right) = \frac{Pr\left(A|B\right) \times Pr(B)}{Pr\left(A|B\right) \times Pr(B) + Pr\left(A|\bar{B}\right) \times Pr\left(\bar{B}\right)}$$

That is, PV^+ can be expressed as a function of sensitivity, specificity, and the probability of disease in the reference population. A similar derivation can be used to obtain PV^-.

Example: Hypertension

Suppose 84% of hypertensives and 23% of normotensives are classified as hypertensive by an automated blood-pressure machine. What are the PV^+ and PV^- of the machine, assuming 20% of the adult population is hypertensive?

Bayes' Rule

Solution

The sensitivity = 0.84
The specificity = 1 - 0.23 = 0.77

Thus, from Bayes' rule, the PV^+ and PV^- can be calculated as

Similarly

$$PV^+ = (.84)(.2)/[(.84)(.2)+(.23)(.8)]$$
$$= .168/.352 = .48$$

$$PV^- = (.77)(.8)/[(.77)(.8)+(.16)(.2)]$$
$$= .616/.648 = .95$$

Bayes' Rule

- Thus a negative result from the machine is reasonably predictive because we are 95% sure a person with a negative result from the machine is normotensive.

- However, a positive result is not very predictive because we are only 48% sure a person with a positive result from the machine is hypertensive.

- The table shows for the *n* subjects and their status with regard to a disease and results from a screening test designed to identify subjects with the disease.

Bayes' Rule

- The cell entries represent the number of subjects falling into the categories defined by the row and column headings.

- For example, *a* is the number of subjects who have the disease and whose screening test result was positive.

Sample of n Subjects (Where *n* is Large) Cross-Classified According to Disease Status and Screening Test Result

	Disease		
Test Result	**Present (D)**	**Absent ($\overline{D}$)**	**Total**
Positive (T)	a	b	$a + b$
Negative ($\overline{T}$)	c	d	$c + d$
Total	$a + c$	$b + d$	n

Bayes' Rule

Sample of n Subjects (Where *n* is Large) Cross-Classified According to Disease Status and Screening Test Result

	Disease		
Test Result	Present (D)	Absent ($\overline{D}$)	Total
Positive (T)	a	b	$a + b$
Negative ($\overline{T}$)	c	d	$c + d$
Total	$a + c$	$b + d$	n

The ***sensitivity*** of the screening test is estimated to be $P(T \mid D) = a/(a + c)$.

The ***specificity*** of the screening test is estimated to be $P(\overline{T} \mid \overline{D}) = d/(b + d)$.

The ***predictive value positive PV⁺*** of a screening test is estimated to be

$$P(D \mid T) = \frac{P(T \mid D)\,P(D)}{P(T \mid D)\,P(D) + P(T \mid \overline{D})\,P(\overline{D})}$$

The ***predictive value negative PV⁻*** of a screening test is estimated to be

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})\,P(\overline{D})}{P(\overline{T} \mid \overline{D})\,P(\overline{D}) + P(\overline{T} \mid D)\,P(D)}$$

Bayes' Rule

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

	Alzheimer's Diagnosis?		
Test Result	**Yes (D)**	**No ($\overline{D}$)**	**Total**
Positive (T)	436	5	441
Negative ($\overline{T}$)	14	495	509
Total	450	500	950

Solution

The sensitivity of the test is estimated to be $P(T \mid D) = 436/450 = 0.97$

The specificity of the test is estimated to be $P(\overline{T}|\overline{D}) = 495/500 = 0.99$

Continue on next slide…

Bayes' Rule

we wish to estimate the probability that a subject who is positive on the test has Alzheimer's disease. From the tabulated data we compute

$$P(T \mid D) = 436/450 = .9689$$

and

$$P(T \mid \overline{D}) = 5/500 = .01$$

The predictive value positive of a screening test (or symptom) can be computed as:

$$P(D \mid T) = \frac{P(T \mid D)\, P(D)}{P(T \mid D)\, P(D) + P(T \mid \overline{D})\, P(\overline{D})}$$

Cont'd

- Substitution of the results into above Equation gives

$$P(D \mid T) = \frac{(.9689)\,P(D)}{(.9689)\,P(D) + (.01)\,P(\overline{D})}$$

The predictive value positive of the test depends on the rate of the disease in the relevant population in general.

In this case the relevant population consists of subjects who are 65 years of age or older.

Bayes' Rule

For *P*(*D*) that is not given in the data table, is given separately. Evans et al. (A-5) estimated that 11.3 percent of the U.S. population aged 65 and over have Alzheimer's disease, so *P*(*D*) = 0.113.

When we substitute this estimate of *P*(*D*) into Equation we obtain

$$P(D \mid T) = \frac{(.9689)(.113)}{(.9689)(.113) + (.01)(1 - .113)} = .93$$

As we see, in this case, the predictive value of the test is very high.

Similarly, let us now consider the predictive value negative of the test.

$$P(\overline{T} \mid D) = 14/450 = .0311$$

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{T} \mid \overline{D})\, P(\overline{D})}{P(\overline{T} \mid \overline{D})\, P(\overline{D}) + P(\overline{T} \mid D)\, P(D)}$$

$$P(\overline{D} \mid \overline{T}) = \frac{(.99)(1 - .113)}{(.99)(1 - .113) + (.0311)(.113)} = .996$$

As we see, the predictive value negative is also quite high.

Lesson 5: Probability Distributions

Course Outline

- Introduction to Biostatistics
- Descriptive Biostatistics
- Probability
- Discrete Probability Distributions
- Continuous Probability Distributions
- Estimation
- Hypothesis Testing

Lecture Outline

- Random Variable and its types
- Probability distribution
- Frequency Distribution
- Relative Frequency Distribution
- Cumulative Frequency Distribution
- Binomial Distribution
- Poisson Distribution

Random Variable

A ***random variable*** is a function that assigns numeric values to different events in a sample space.

Categories of a Random Variable

There are two categories of a random variable.

1. **Discrete Random Variable**
2. **Continuous Random Variable**

Discrete Random Variable

A random variable for which there exists a discrete set of numeric values is a ***discrete random variable***.

Continuous Random Variable

A random variable whose possible values cannot be enumerated is a ***continuous random variable***.

Probability Distribution

A ***probability distribution*** (or ***probability mass function***) indicates the values that a random variable may have.

Types of Probability Distribution

There are two types of probability distribution.

1. **Discrete Probability Distribution**
2. **Continuous Probability Distribution**

Probability Distribution of a Discrete Random Variable

The ***probability distribution of a discrete random variable*** ***or*** ***discrete probability distribution*** is a table, graph, formula, or other device used to specify all possible values of a discrete random variable along with their respective probabilities.

Discrete Probability Distributions

Types of discrete probability distributions are;

1. **The Binomial Distribution**
2. **The Poisson Distribution**

Discrete Probability Distribution

Let the discrete probability distribution be represented by P(x), then

$P(x) = P(X = x)$ is the probability of the discrete random variable X to assume a value x.

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

In an article appearing in the *Journal of the American Dietetic Association,* Holben et al. looked at food security status in families in the Appalachian region of southern Ohio. The purpose of the study was to examine hunger rates of families with children in a local Head Start program in Athens, Ohio. The survey instrument included the 18-question U.S. Household Food Security Survey Module for measuring hunger and food security. In addition, participants were asked how many food assistance programs they had used in the last 12 months. Table shows the number of food assistance programs used by subjects in this sample.

Table: Number of Assistance Programs Utilized by Families with Children in

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

We wish to construct the probability distribution of the discrete variable X, where X = of food assistance programs used by the study subjects.

Solution

- The values of X are $x_1 = 1$, $x_2 = 2$, …, $x_7 = 7$ and $x_8 = 8$.
- We compute the probabilities for these values by dividing their respective frequencies by the total, 297.
- Thus, for example, $p(x_1) = P(X = x_1) = 62/297 = .2088$.

Table: Number of Assistance Programs Utilized by Families with Children in

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

- The desired probability distribution is shown in the Table.
- Note that each of the probabilities in the Table is the *relative frequency of occurrence* of the corresponding value of *X.*

Table: Probability Distribution of Programs Utilized by Families Among the Subjects.

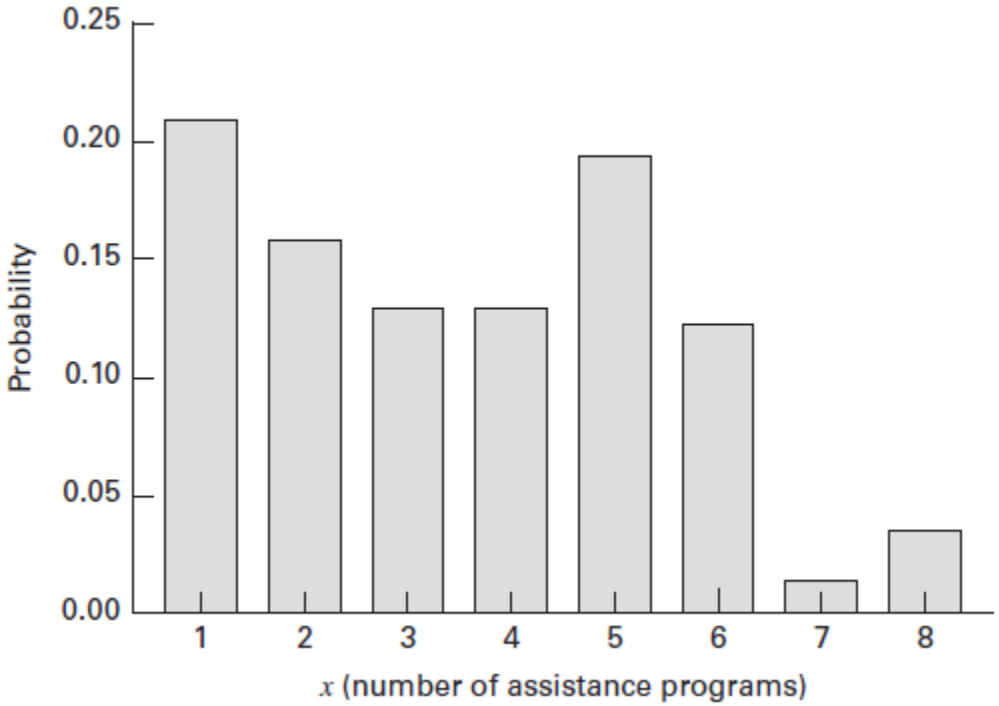
Number of Programs	Frequency	$P(X = x)$
1	62	.2088
2	47	.1582
3	39	.1313
4	39	.1313
5	58	.1953
6	37	.1246
7	4	.0135
8	11	.0370
Total	297	1.0000

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

- We can also present the probability distribution in the form of a graph, as shown in the figure.
- In the figure, the length of each vertical bar indicates the probability for the corresponding value of *x.*

Figure: Graphical representation of the probability distribution.



Discrete Probability Distribution (Properties)

- If $x_1, x_2, x_3, \ldots, x_k$ are all possible values of the discrete random variable X.

- Then the two essential properties of a probability distribution of a discrete variable are:

$$(1) \quad 0 \leq P(X = x) \leq 1$$

$$(2) \quad \sum P(X = x) = 1, \quad \text{for all } x$$

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a randomly selected family used three assistance programs?

Solution: We may write the desired probability as

$$p(3) = P(X = 3).$$

We see in the Table that the answer is **.1313**.

Table: Probability Distribution of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	$P(X = x)$
1	.2088
2	.1582
3	.1313
4	.1313
5	.1953
6	.1246
7	.0135
8	.0370
Total	1.0000

Discrete Probability Distribution (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a randomly selected family used either one or two programs?

Solution

we can use the addition rule for mutually exclusive events.

$$P(1 \cup 2) = P(1) + P(2) = .2088 + .1582 = .3670.$$

Table: Probability Distribution of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	$P(X = x)$
1	.2088
2	.1582
3	.1313
4	.1313
5	.1953
6	.1246
7	.0135
8	.0370
Total	1.0000

Cumulative Probability Distribution

The ***cumulative probability distributions*** or ***cumulative distribution function (cdf)*** of a random variable X is denoted by $F(X)$ and, for a specific value x of randam variable X, is defined by $Pr(X \leq x)$ and denoted by $F(x)$.

Cumulative Probability Distribution

Food Assistance Programs in Southern Ohio

- The ***cumulative probability distribution*** for the discrete variable X whose probability distribution is given in the Table below may be obtained by successively adding the probabilities $P(X = x_i)$.
- The cumulative probability for x_i is written as
- It gives the probability that X is less than or equal to a specified value, x_i.

$$F(x_i) = P(X \leq x_i).$$

Table: Probability Distribution of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	$P(X = x)$
1	.2088
2	.1582
3	.1313
4	.1313
5	.1953
6	.1246
7	.0135
8	.0370
Total	1.0000

Cumulative Probability Distributions

Food Assistance Programs in Southern Ohio

The resulting ***cumulative probability distribution*** is shown in the table

Table: Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects.

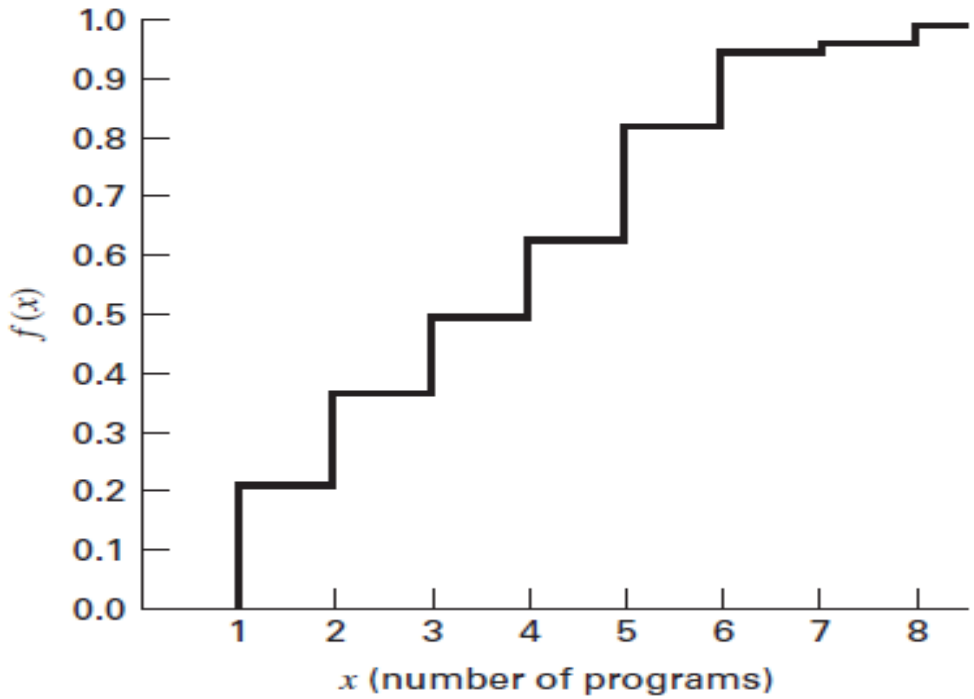
Number of Programs (x)	$P(X = x)$	Cumulative Frequency $P(X \leq x)$
1	.2088	.2088
2	.1582	.3670
3	.1313	.4983
4	.1313	.6296
5	.1953	.8249
6	.1246	.9495
7	.0135	.9630
8	.0370	1.0000

Food Assistance Programs in Southern Ohio

- The graph of the ***cumulative probability distribution*** is shown in the figure.
- The graph of a cumulative probability distribution is called an ***ogive***.
- The graph of $F(x)$ consists solely of the horizontal lines.
- The length of each vertical line represents the same probability as that of the corresponding line in graph of probability distribution.
- For example, the length of the vertical line at $X = 3$ is .1313.

Cont'd

Figure: Cumulative probability distribution of number of assistance programs among the subjects.



Cumulative Probability Distributions

- Another way to distinguish between a discrete and continuous random variable is by each variable's cumulative distribution function (cdf).
- For a discrete random variable, the cdf looks like a series of steps and is sometimes called a *step function.*
- For a continuous random variable, the cdf is a smooth curve.
- As the number of values increases, the cdf for a discrete random variable approaches that of a smooth curve.

Cumulative Probability Distributions

Otolaryngology

Otitis media, a disease of the middle ear, is one of the most common reasons for visiting a doctor in the first 2 years of life other than a routine wellbaby visit. Let X be the random variable that represents the number of episodes of otitis media in the first 2 years of life. Then X is a discrete random variable, which takes on the values 0, 1, 2, and so on.

Consider the random variable representing the number of episodes of otitis media in the first 2 years of life. Suppose this random variable has a probability-mass function as given in the table .

Probability-mass function for the number of episodes of otitis media in the first 2 years of life

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

Cumulative Probability Distributions

Otolaryngology

Compute the cdf for the otitis-media random variable and display it graphically.

Probability-mass function for the number of episodes of otitis media in the first 2 years of life

r	0	1	2	3	4	5	6
$Pr(X = r)$	.129	.264	.271	.185	.095	.039	.017

Solution

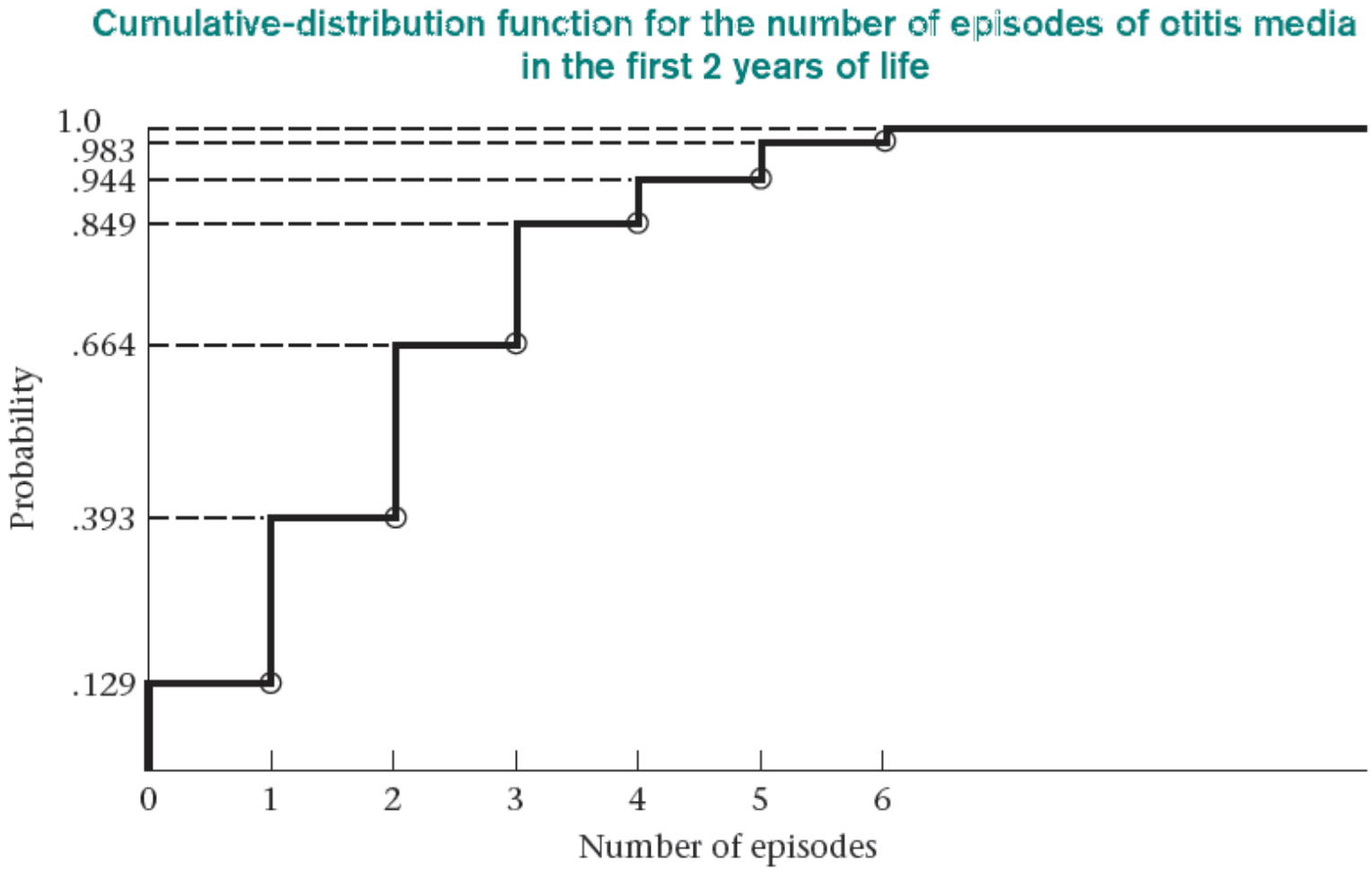
The cdf is given by

$$
\begin{aligned}
F(x) &= 0 && \text{if} \quad x < 0 \\
F(x) &= .129 && \text{if} \quad 0 \le x < 1 \\
F(x) &= .393 && \text{if} \quad 1 \le x < 2 \\
F(x) &= .664 && \text{if} \quad 2 \le x < 3 \\
F(x) &= .849 && \text{if} \quad 3 \le x < 4 \\
F(x) &= .944 && \text{if} \quad 4 \le x < 5 \\
F(x) &= .983 && \text{if} \quad 5 \le x < 6 \\
F(x) &= 1.0 && \text{if} \quad x \ge 6
\end{aligned}
$$

Cumulative Probability Distributions

Otolaryngology

The ***cumulative distribution function (cdf)*** is displayed in the figure.



Cumulative Probability Distributions (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a family picked at random used two or fewer assistance programs?

Table: Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	Cumulative Frequency $P(X \leq x)$
1	.2088
2	.3670
3	.4983
4	.6296
5	.8249
6	.9495
7	.9630
8	1.0000

Solution

The probability we seek may be found directly in the Table by reading the cumulative probability opposite $x = 2$. $P(X \leq 2) = .3670.$

Cumulative Probability Distributions (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a randomly selected family used fewer than four programs?

Table: Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	Cumulative Frequency $P(X \leq x)$
1	.2088
2	.3670
3	.4983
4	.6296
5	.8249
6	.9495
7	.9630
8	1.0000

Solution

Since a family that used fewer than four programs used either one, two, or three programs, the answer is the cumulative probability for 3. That is,

$$P(X < 4) = P(X \leq 3) = .4983$$

Cumulative Probability Distributions (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a randomly selected family used five or more programs?

Table: Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	Cumulative Frequency $P(X \leq x)$
1	.2088
2	.3670
3	.4983
4	.6296
5	.8249
6	.9495
7	.9630
8	1.0000

Cont'd

Solution

- we make use of the concept of complementary probabilities.
- The set of families that used five or more programs is the complement of the set of families that used fewer than five (that is, four or fewer) programs.
- The sum of the two probabilities associated with these sets is equal to 1.
- We write this relationship in probability notation as
- There,

$$P(X \geq 5) + P(X \leq 4) = 1.$$

$$P(X \geq 5) = 1 - P(X \leq 4) = 1 - .6296 = .3704$$

Cumulative Probability Distributions (Example)

Food Assistance Programs in Southern Ohio

What is the probability that a randomly selected family used between three and five programs, inclusive?

Table: Cumulative Probability Distribution of Number of Programs Utilized by Families Among the Subjects.

Number of Programs (x)	Cumulative Frequency $P(X \leq x)$
1	.2088
2	.3670
3	.4983
4	.6296
5	.8249
6	.9495
7	.9630
8	1.0000

Solution

- The probability that a family used between one and five programs, inclusive is;

$$P(X \leq 5) = .8249$$

- To get the probability of between three and five programs, we subtract, from .8249, the probability of two or fewer.
- So the answer is, $P(3 \leq X \leq 5) = P(X \leq 5) - P(X \leq 2) = .8249 - .3670 = .4579$

Binomial Distribution

- The ***binomial distribution*** is one of the most widely encountered probability distributions in applied statistics.

- The distribution is derived from a process known as a *Bernoulli trial,* named in honor of the Swiss mathematician James Bernoulli (1654–1705).

- James Bernoulli made significant contributions in the field of probability, including, in particular, the binomial distribution.

- When a random process or experiment, called a trial, can result in only one of two mutually exclusive outcomes, such as dead or alive, sick or well, full-term or premature, the trial is called a Bernoulli trial.

Bernoulli Trial

A binomial trial or experiment is a probability experiment with the following properties.

1. Each trial can have only two outcomes which can be considered

success (p) or failure (q).

2. There must be a fixed number of trials.

3. The outcomes of each trial must be independent of each other.

4. The probability of success must remain the same in each trial.

The outcomes of a binomial experiment are called a *binomial distribution*. These properties (or conditions) can be summarized as:

A binomial experiment has a fixed number of independent trials, each of which can only have two possible outcomes.

Examples of Binomial Trials

- Tossing a coin 5 times to see how many tails occur.
- Asking 10 patients if they like chicken soup.
- Rolling a die to see if a 5 appears.
- Asking 30 people if they watch Geo news.
- Asking 100 people if they would vote for the PTI candidate.
  - » The 3 conditions are met - there's a fixed number [100], the trials are independent [what one person does doesn't affect the next person], and there's only two outcomes [yes or no]).

Binomial Trials

Example: A game in which a true die is thrown and if a 4, 5, or 6 comes up you win; and if a 1, 2, or 3 comes up you lose.

1. 2 possible outcomes: win, lose.
2. The probability of winning remains constant at 0.5 from one trial to the next.
3. The chances of winning are not increased from previous trials because the dice have no memory.

So all 3 conditions are met in this example.

Examples which are not Binomial Trials

- Rolling a die until a 6 appears (not a fixed number of trials)
- Asking 20 patients how old they are (not two outcomes)
- Drawing 5 cards from a deck for a poker hand (done without replacement, so not independent)

Example: What is the probability of rolling exactly two sixes in 6 rolls of a die?

There are five steps you need to do to work a binomial problem.

Binomial Trials

1. Define Success for a single trial. **Success = "Rolling a 6 on a single die"**
2. Define the probability of success (*p*): **p = 1/6**
3. Find the probability of failure (*q*): **q = 1 – p = 5/6**
4. Define the number of trials (*n*): **n = 6**
5. Define the number of successes out of those trials (*x*): **x = 2**

- Anytime a 6 appears, it is a success (denoted S)
- Anytime something else appears, it is a failure (denoted F).

The total number of combinations of getting exactly two sixes in 6 rolls of a die can be computed as

$$_6C_2 = 6!/[2!(6 - 2)!]$$

$$\boxed{_6C_2 = 15}$$

Binomial Trials

- Also, note that the 1/6 is the probability of success and you needed 2 successes.
- The 5/6 is the probability of failure, and if 2 of the 6 trials were success, then 4 of the 6 must be failures.
- Note that 2 is the value of x and 4 is the value of $n - x$.
- Further note that there are fifteen ways this can occur.
- This is the number of ways 2 successes can be occur in 6 trials *without repetition* and *order not being important*, or a combination of 6 things, 2 at a time.

Binomial Trials

The probability of getting exactly ***x*** success in ***n*** trials, with the probability of success on a single trial being ***p*** is:

$$
\begin{aligned}
f(x) &= {}_nC_x q^{n-x} p^x = {}_nC_x\, p^x q^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n \\
&= 0, \qquad \text{elsewhere}
\end{aligned}
$$

Binomial Trials

So the probability of rolling exactly two sixes in 6 rolls of a die can be computed as

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(2) = P(X = 2) = {}_6C_2\ (1/6)^2\ (5/6)^4$$

$$f(2) = P(X = 2) = 15\ (0.02777)\ (0.48225)$$

$$\mathbf{f(2) = P(X = 2) = 0.2009}$$

Binomial Trials

Example: A coin is tossed 10 times. What is the probability that exactly 6 heads will occur.

Solution

1. Success = "A head is flipped on a single coin"
2. The probability of success ***p = 0.5***
3. The probability of failure ***q = 1 – p = 0.5***
4. The number of trials ***n = 10***
5. The number of successes out of those trials ***x = 6***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(6) = P(X = 6) = {}_{10}C_6\ (0.5)^6\ (0.5)^4$$

$$f(6) = P(X = 6) = 210\ (0.015625)\ (0.0625)$$

$$\boldsymbol{f(6) = P(X = 6) = 0.205078125}$$

Binomial Trials

Example: 60% of the students in 11BM at MUET are male. What is the probability that a randomly selected group of 25, there will be exactly 15 males?

Solution

1. Success = “Male students from 11 BM at MUET"
2. The probability of success ***p = 0.60***
3. The probability of failure ***q = 1 – p = 0.40***
4. The number of trials ***n = 25***
5. The number of successes out of those trials ***x = 15***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(15) = P(X = 15) = {}_{25}C_{15}\ (0.60)^{15}\ (0.40)^{10}$$

$$f(15) = P(X = 15) = 0.1612$$

Binomial Trials

Example: A marksman hits a target 80% of the time. He fires 5 shots at the target. What is the probability that exactly 3 shots hit the target?

Solution

1. Success = “A marksman hits a target "
2. The probability of success ***p = 0.80***
3. The probability of failure ***q = 1 – p = 0.20***
4. The number of trials ***n = 5***
5. The number of successes out of those trials ***x = 3***

$$f(3) = f(x) = {}_nC_x\, p^x q^{n-x} (0.80)^3 \ (0.20)^2$$

$$f(3) = P(X = 3) = 0.2048$$

Binomial Trials

Example: A marksman hits a target 80% of the time. He fires 5 shots at the target. What is the probability that more than 3 shots hit the target?

Solution

1. Success = “A marksman hits a target "
2. The probability of success ***p = 0.80***
3. The probability of failure ***q = 1 – p = 0.20***
4. The number of trials ***n = 5***
5. The number of successes out of those trials ***x = 4***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(4) = P(X > 3) = {}_5C_4\ p^4\ q^{5-4} + {}_5C_5\ p^5\ q^{5-5}$$

$$f(4) = P(X > 3) = {}_5C_4\ (0.80)^4\ (0.20)^{5-4} + {}_5C_5\ (0.80)^5\ (0.20)^{5-5}$$

$$f(4) = P(X > 3) = 0.7373$$

Binomial Trials

Example: What is the probability of obtaining 2 boys out of 5 children if the probability of a boy is 0.51 at each birth and the sexes of successive children are considered independent random variables?

Solution

1. Success = "A boy at each birth"
2. The probability of success ***p = 0.51***
3. The probability of failure ***q = 1 – p = 0.49***
4. The number of trials ***n = 5***
5. The number of successes out of those trials ***x = 2***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(2) = P(X = 2) = {}_5C_2\ (0.51)^2\ (0.49)^3$$

$$f(2) = P(X = 2) = 10\ (0.2601)\ (0.117649)$$

$$\boldsymbol{f(2) = P(X = 2) = 0.306}$$

Binomial Trials

Example: Suppose the probability that a person who is 50 years old will die before s/he reaches the age of 55 is .001. What is the likelihood that in a group of one thousand 50-year-olds, exactly 2 will die before reaching the age of 55?

Solution

1. Success = “50-year-olds die before reaching the age of 55"
2. The probability of success ***p = 0.001***
3. The probability of failure ***q = 1 – p = 0.999***
4. The number of trials ***n = 1000***
5. The number of successes out of those trials ***x = 2***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(2) = P(X = 2) = {}_{1000}C_2\ (0.001)^2\ (0.999)^{998}$$

$$f(2) = P(X = 2) = 0.1840$$

Binomial Trials

Example: An air conditioner made by a certain company is made of 20 distinct parts. Each part has a .004 probability of being defective. What is the probability that a randomly selected air conditioner will not work perfectly?

Solution

1. Success = “An air conditioner will not work perfectly "
2. The probability of success ***p = 0.004***
3. The probability of failure ***q = 1 – p = 0.996***
4. The number of trials ***n = 20***
5. The number of successes out of those trials ***x = 0***

Probability that the air conditioner will work is: $f(x) = {}_nC_x\, p^x q^{n-x}$

$$f(0) = P(X = 0) = {}_{20}C_0\ (0.004)^0\ (0.996)^{20}$$

$$f(2) = P(X = 2) = 0.923$$

Probability that the air conditioner will not work is **1 - .923 = .077**

Hence, there is a **7.7%** chance that a randomly selected air conditioner made by this company will not work.

Binomial Trials

Example: A hybrid car made by a certain company consists of 400 distinct components. Each component has a .001 chance of being defective. What is the probability that a car made by this company will not work perfectly?

Solution

1. Success = “A car made by the company will not work perfectly "
2. The probability of success ***p = 0.001***
3. The probability of failure ***q = 1 – p = 0.999***
4. The number of trials ***n = 400***
5. The number of successes out of those trials ***x = 0***

Probability that the car will work perfectly is:

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(0) = P(X = 0) = {}_{400}C_0\ (0.001)^0\ (0.999)^{400}$$

$$f(2) = P(X = 2) = 0.6702$$

Probability that it will not work perfectly is **1 - .6702 = .3298**

There is almost a **33%** chance that something will be wrong with the car.

Binomial Trials

Example: if we examine all birth records from the North Carolina State Center for Health Statistics for the calendar year 2001, we find that 85.8 percent of the pregnancies had delivery in week 37 or later. We will refer to this as a full-term birth. With that percentage, we can interpret the probability of a recorded birth in week 37 or later as .858. If we randomly select five birth records from this population, what is the probability that exactly three of the records will be for full-term births?

Solution

1. Success = “A record of the full term births"
2. The probability of success ***p = 0.858***
3. The probability of failure ***q = 1 – p = 0.142***
4. The number of trials ***n = 5***
5. The number of successes out of those trials ***x = 3***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(3) = P(X = 3) = {}_5C_3\ (0.858)^3\ (0.142)^2$$

$$f(3) = P(X = 3) = 10\ (0.6316)\ (0.0202)$$

$$f(3) = P(X = 3) = 0.1276$$

Binomial Trials

Example: The data from the North Carolina State Center for Health Statistics show that 14 percent of mothers admitted to smoking one or more cigarettes per day during pregnancy. If a random sample of size 10 is selected from this population, what is the probability that it will contain exactly four mothers who admitted to smoking during pregnancy?

Solution

1. Success = " A mother admitted to smoking cigarettes per day during pregnancy "
2. The probability of success ***p = 0.14***
3. The probability of failure ***q = 1 – P = 0.86***
4. The number of trials ***n = 10***
5. The number of successes out of those trials ***x = 4***

$$f(x) = {}_nC_x\, p^x q^{n-x}$$

$$f(4) = P(X = 4) = {}_{10}C_4\,(0.14)^4\,(0.86)^6$$

$$f(4) = P(X = 4) = 210\,(0.0003842)\,(0.4045672)$$

$$f(4) = P(X = 4) = 0.0326$$

Poisson Distribution

- The ***Poisson distribution*** named after the French mathematician Simeon Denis Poisson (1781–1840).

- The Poisson distribution is useful when there are a large number of independent trials with a small probability of success on a single trial and the variables occur over a period of time.

- It can also be used when a density of items is distributed over a given area or volume.

Poisson Random Variable

The ***Poisson random variable*** *x* is often a model for data that represent the number of occurrences of a specified event in a given unit of time or space.

Examples:

- The number of calls received by a receptionist during a given period of time.
- The number of medical instruments breakdowns in a day
- The number of traffic accidents at a given intersection during a given time period.

The Poisson Process

The Poisson distribution is based on the Poisson process.

The following statements describe what is known as the ***Poisson process***.

1. The occurrences of the events are independent in an interval.

2. An infinite number of occurrences of the event are possible in the interval.

3. The probability of a single event in the interval is proportional to the length of the interval.

4. In an infinitely small portion of the interval, the probability of more than one occurrence of the event is negligible.

Poisson Distribution

If ***x*** is the number of occurrences of some random event in an interval of time or space (or some volume of matter), the probability that ***x*** will occur is given by

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \qquad x = 0, 1, 2, \ldots$$

- The Greek letter **λ** (lambda) is called the parameter of the distribution.
- The **λ** is the average number of occurrences of the random event in the interval (or volume).
- The symbol ***e*** is the constant (to four decimals) 2.7183.

Poisson Distribution

- The Poisson distribution depends on a single parameter **λ**.

- Note that the parameter **λ** represents the *expected number of events over time period t.*

- One important difference between the Poisson and binomial distributions concerns the numbers of trials and events.

- For a binomial distribution there are a finite number of trials ***n***, and the number of events can be no larger than ***n***.

- For a Poisson distribution, the number of trials is essentially infinite and the number of events (or number of deaths) can be indefinitely large, although the probability of ***x*** events becomes very small as ***x*** increases.

Poisson Distribution

Example: If there are 500 customers per eight-hour day in a check out lane, what is the probability that there will be exactly 3 in line during any five-minute period?

Solution

- There are **(60 minutes /5 minuets) × 8 hours = 96** five-minute periods in eight hours.
- So, the expected value during any one five-minute period would be **λ = 500 / 96 = 5.2083**.
- So, you expect about 5.2 customers in 5 minutes and want to know the probability of getting exactly 3.

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$F(3) = p(3;500/96) = e^{-500/96} * (500/96)^{3} / 3!$$

$$F(3) = p(3;500/96) = 0.1288$$

Poisson Distribution

Example: In a study of drug-induced anaphylaxis among patients taking rocuronium bromide as part of their anesthesia, Laake and Rottingen found that the occurrence of anaphylaxis followed a Poisson model with λ = 12 incidents per year in Norway. Find the probability that in the next year, among patients receiving rocuronium, exactly three will experience anaphylaxis.

Solution

λ = 12

x = 3

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

$$P(X = 3) = \frac{e^{-12}12^{3}}{3!} = .00177$$

Poisson Distribution

Example: In a study of drug-induced anaphylaxis among patients taking rocuronium bromide as part of their anesthesia, Laake and Rottingen found that the occurrence of anaphylaxis followed a Poisson model with λ = 12 incidents per year in Norway. What is the probability that at least three patients in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

Solution

We can u $f(x) = \dfrac{e^{-\lambda}\lambda^x}{x!}$ of complementary events in this case.

$$P(X \geq 3) = 1 - P(X \leq 2) = 1 - [P(X = 0) + P(X = 1) + P(X = 2)]$$

$$= 1 - \left[\frac{e^{-12}12^0}{0!} + \frac{e^{-12}12^1}{1!} + \frac{e^{-12}12^2}{2!}\right]$$

$$= 1 - [.00000614 + .00007373 + .00044238]$$

$$= 1 - .00052225$$

$$P(X \geq 3) = .99947775$$

Infectious Disease

Example: Consider the distribution of number of deaths attributed to typhoid fever over a long period of time, for example, 1 year. Assuming the probability of a new death from typhoid fever in any one day is very small and the number of cases reported in any two distinct periods of time are independent random variables, then the number of deaths over a 1-year period will follow a Poisson distribution. Suppose the number of deaths from typhoid fever over a 1-year period is Poisson distributed with parameter $\lambda = 4.6$. What is the probability distribution of the number of deaths over a 6-month period?

Solution

- Let X = the number of deaths in 6 months.
- For a 1-year period, the parameter $\lambda = 4.6$,
- For a 6-month period, the parameter $\lambda = 4.6/2 = 2.3$

Poisson Distribution

$$Pr(X=0)=e^{-2.3}=.100$$

$$Pr(X=1)=\frac{2.3}{1!}e^{-2.3}=.231$$

$$Pr(X=2)=\frac{2.3^2}{2!}e^{-2.3}=.265$$

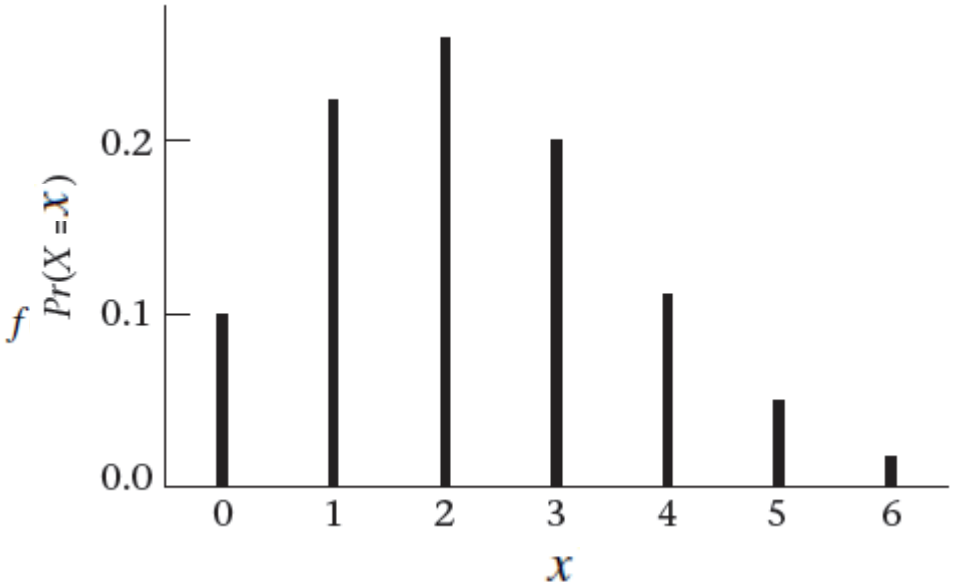
$$Pr(X=3)=\frac{2.3^3}{3!}e^{-2.3}=.203$$

$$Pr(X=4)=\frac{2.3^4}{4!}e^{-2.3}=.117$$

$$Pr(X=5)=\frac{2.3^5}{5!}e^{-2.3}=.054$$

$$Pr(X\geq 6)=1-(.100+.231+.265+.203+.117+.054)=.030$$

Distribution of the number of deaths attributable to typhoid fever over various time intervals



Poisson Distribution

Infectious Disease

Example: Suppose the number of deaths from typhoid fever over a 1-year period is Poisson distributed with parameter λ = 4.6. What is the probability distribution of the number of deaths over a 3-month period?

Solution

Let Y = the number of deaths in 3 months.

For a 1-year period, the parameter λ = 4.6,

For a 3-month period, the parameter λ = 4.6/4 = 1.15

$$Pr(Y=0)=e^{-1.15}=.317$$

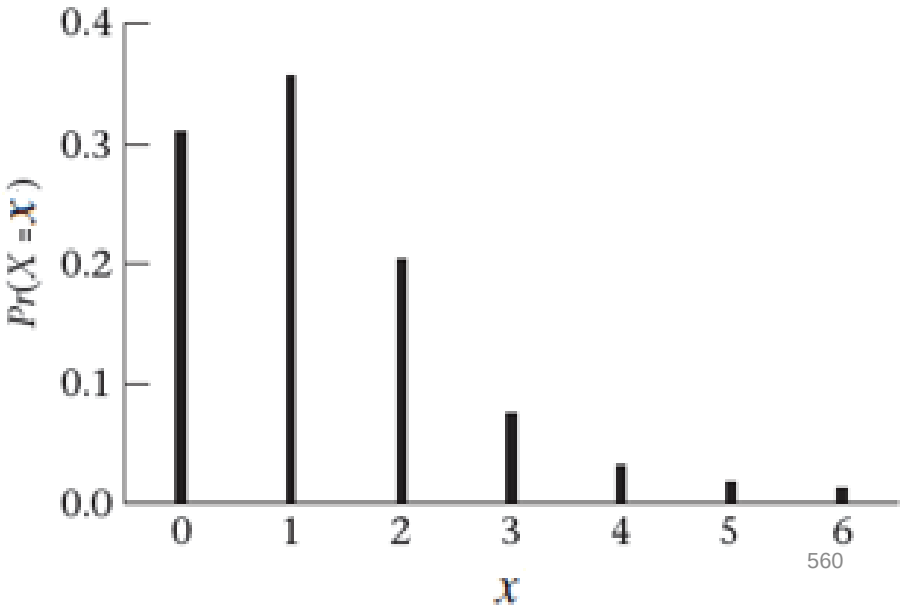
$$Pr(Y=1)=\frac{1.15}{1!}e^{-1.15}=.364$$

$$Pr(Y=2)=\frac{1.15^2}{2!}e^{-1.15}=.209$$

$$Pr(Y=3)=\frac{1.15^3}{3!}e^{-1.15}=.080$$

$$Pr(Y\geq 4)=1-(.317+.364+.209+.080)=.030$$

Distribution of the number of deaths attributable to typhoid fever over various time intervals



Poisson Distribution

Bacteriology

Example: Suppose we have a 100 cm^2 agar plate. The probability of finding any bacterial colonies at anyone-point *a* (or more precisely in a small area around *a*) is very small, and the events of finding bacterial colonies at any two points a_1, a_2 are independent. The number of bacterial colonies over the entire agar plate will follow a Poisson distribution. If λ = 2, calculate the probability distribution of at least five number of bacterial colonies.

Solution

Let *X* = the number of colonies.

$$Pr(X=0)=e^{-2}=.135$$

$$Pr(X=1)=e^{-2}2^1/1!=2e^{-2}=.271$$

$$Pr(X=2)=e^{-2}2^2/2!=2e^{-2}=.271$$

$$Pr(X=3)=e^{-2}2^3/3!=\frac{4}{3}e^{-2}=.180$$

$$Pr(X=4)=e^{-2}2^4/4!=\frac{2}{3}e^{-2}=.090$$

$$Pr(X\geq 5)=1-(.135+.271+.271+.180+.090)=.053$$

CONTINUOUS PROBABILITY DISTRIBUTIONS

- A continuous variable is one that can assume any value within a specified interval of values assumed by the variable. Consequently, between any two values assumed by a continuous variable, there exist an infinite number of values.

FIGURE 4.5.2 Graphical representation of a continuous distribution.

Probability distribution of c.r.v.

A nonnegative function $f(x)$ is called a probability distribution (some- times called a probability density function) of the continuous random variable X if the total area bounded by its curve and the x-axis is equal to 1 and if the subarea under the curve bounded by the curve, the x-axis, and perpendiculars erected at any two points a and b give the probability that X is between the points a and b.

Properties of continuous probability Distributions:

1- Area under the curve = 1.

2- $P(X = a) = 0$, where a is a constant.

3- Area between two points a *and* b is $P(a<x<b)$.

The normal distribution:

- It is one of the most important probability distributions in statistics.

- The normal density is given by;

$$f(x) = \frac{1}{\sqrt{2ps}} e^{-\frac{(x-m)^2}{2s^2}}$$

$-\infty < x < \infty$, $-\infty < \boldsymbol{\mu} < \infty$, $\boldsymbol{\sigma} > 0$

- π and e : constants, μ: population mean.

σ : Population standard deviation.

Characteristics of the normal distribution

- The following are some important characteristics of the normal distribution:

- It is symmetrical about its mean, μ.
- The mean, the median, and the mode are all equal.

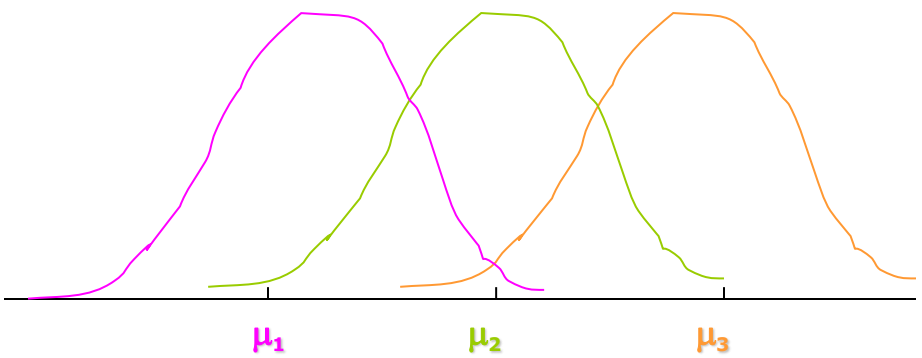
- The total area under the curve above the x-axis is one.
- The normal distribution is completely determined by the parameters μ and σ

Cont'd

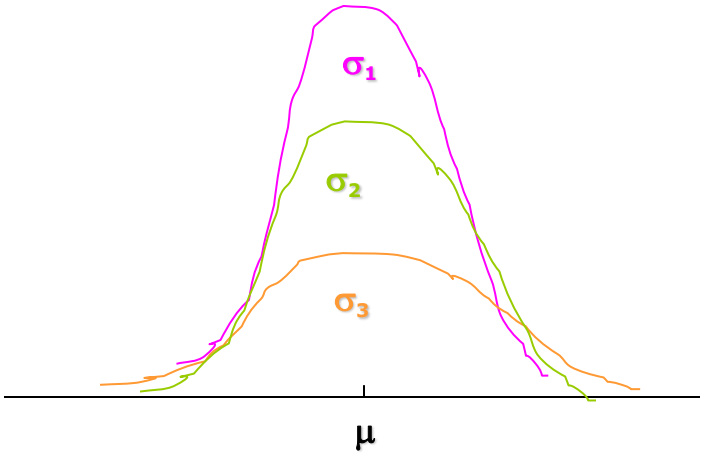
- **The normal distribution depends on the two parameters μ and σ.**

μ determines the location of the curve.

But, σ determines the scale of the curve, i.e. the degree of flatness or peakedness of the curve.



$\mu_1 < \mu_2 < \mu_3$



$\sigma_1 < \sigma_2 < \sigma_3$

Note that

1. $P(\mu - \sigma < x < \mu + \sigma) = 0.68$

2. $P(\mu - 2\sigma < x < \mu + 2\sigma) = 0.95$

3. $P(\mu - 3\sigma < x < \mu + 3\sigma) = 0.997$

The Standard normal distribution

- Is a special case of normal distribution with mean equal 0 and a standard deviation of 1.

- The equation for the standard normal distribution is written as

$$z = \frac{x - \mu,}{\sigma} \qquad -\infty < z < \infty$$

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

Characteristics of the standard normal distribution

1- It is symmetrical about 0.

2- The total area under the curve above the x-axis is one.

3- We can use normal distribution tables to find the probabilities and areas.

“How to use tables of Z”

Note that

The cumulative probabilities $P(Z \leq z)$ are given in tables for $-3.49 < z < 3.49$. Thus, $P(-3.49 < Z < 3.49) \cong 1$.

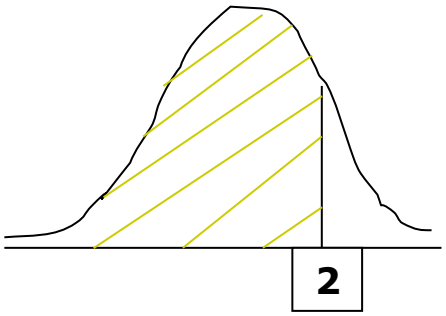
For standard normal distribution,

$P(Z > 0) = P(Z < 0) = 0.5$

<u>Example 4.6.1:</u>

If Z is a standard normal distribution, then

1) $P(Z < 2) = 0.9772$ is the area to the left to 2 and it equals 0.9772.



Example

- Given the standard normal distribution, find the area under the curve, above the z-axis between z=- infinity and z=2.

FIGURE 4.6.7 The standard normal distribution showing area between $z = -\infty$ and $z = 2$.

LE D

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
0	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.20
0	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.30
0	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879	0.40
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224	0.50
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549	0.60
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852	0.70
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133	0.80
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389	0.90
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621	1.00
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830	1.10
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015	1.20
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177	1.30
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319	1.40
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441	1.50
1.60	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545	1.60
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633	1.70
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706	1.80
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767	1.90
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817	2.00
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857	2.10
2.20	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890	2.20

Area under curve is = 0.9772

Cont'd

We may interpret this as

1. the probability that a z picked at random from the population of z's will have a value between – infinity and 2.

2. the relative frequency of occurrence (or proportion) of values of z between –infinity and 2.

3. We may interpret that 97.72 percent of the z's have a value between – infinity and 2.

Example

What is the probability that a z picked at random from the population of z's will have a value between -2.55 and +2.55.

$$P(-2.55 < z < 2.55) = .9946 - .0054 = .9892$$

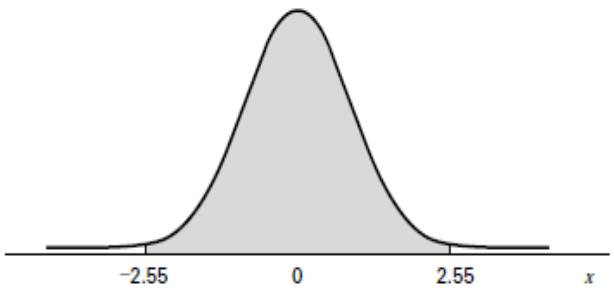


FIGURE 4.6.8 Standard normal curve showing $P(-2.55 < z < 2.55)$.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480
0.40	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844
0.50	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190
0.60	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517
0.70	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823
0.80	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106
0.90	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365
1.00	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599
1.10	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810
1.20	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997
1.30	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162
1.40	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306
1.50	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429
1.60	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535
1.70	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625
1.80	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699
1.90	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761
2.00	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812
2.10	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854
2.20	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963

Cont'd

EXAMPLE 4.6.3

What proportion of z values are between -2.74 and 1.53?

Solution: Figure 4.6.9 shows the area desired. We find in Table D that the area between $-\infty$ and 1.53 is .9370, and the area between $-\infty$ and -2.74 is .0031. To obtain the desired probability we subtract .0031 from .9370. That is,

$$P(-2.74 \leq z \leq 1.53) = .9370 - .0031 = .9339$$

FIGURE 4.6.9 Standard normal curve showing proportion of z values between $z = -2.74$ and $z = 1.53$.

Cont'd

EXAMPLE 4.6.4

Given the standard normal distribution, find $P(z \geq 2.71)$.

Solution: The area desired is shown in Figure 4.6.10. We obtain the area to the right of $z = 2.71$ by subtracting the area between $-\infty$ and 2.71 from 1. Thus,

$$\begin{aligned} P(z \geq 2.71) &= 1 - P(z \leq 2.71) \\ &= 1 - .9966 \\ &= .0034 \end{aligned}$$

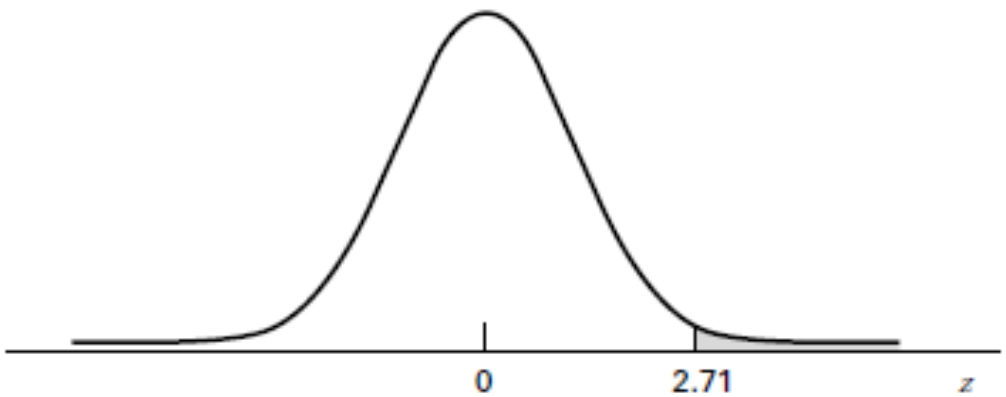


FIGURE 4.6.10 Standard normal distribution showing $P(z \geq 2.71)$.

■

cont'd

EXAMPLE 4.6.5

Given the standard normal distribution, find $P(.84 \leq z \leq 2.45)$.

Solution: The area we are looking for is shown in Figure 4.6.11. We first obtain the area between $-\infty$ and 2.45 and from that subtract the area between $-\infty$ and .84. In other words,

$$\begin{aligned} P(.84 \leq z \leq 2.45) &= P(z \leq 2.45) - P(z \leq .84) \\ &= .9929 - .7995 \\ &= .1934 \end{aligned}$$

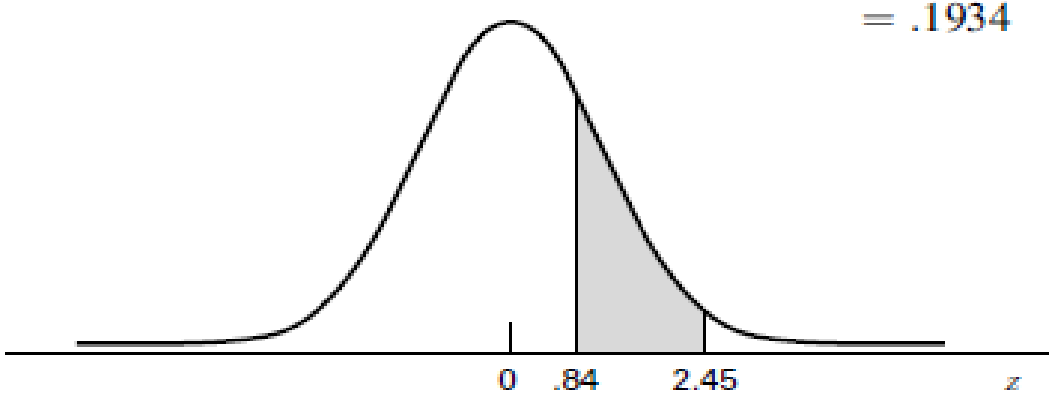


FIGURE 4.6.11 Standard normal curve showing $P(.84 \leq z \leq 2.45)$.

■

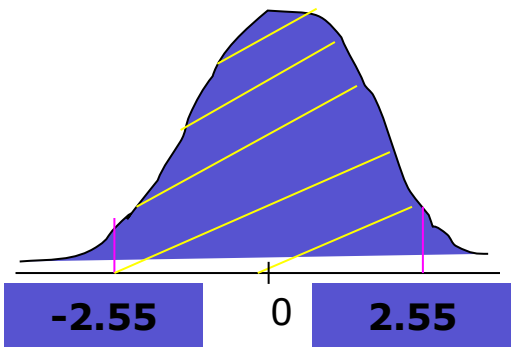
Cont'd

Example 4.6.2:

P(-2.55 < Z < 2.55) is the area between -2.55 and 2.55, Then it equals

P(-2.55 < Z < 2.55) =0.9946 – 0.0054

= 0.9892.

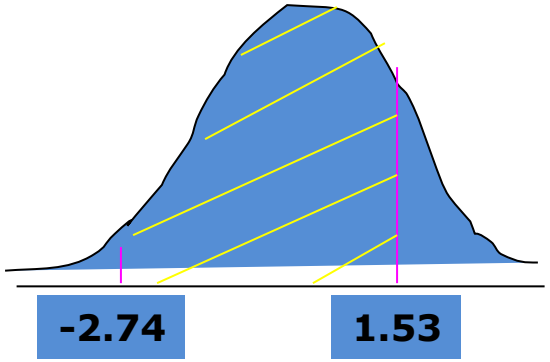


Example 4.6.2:

P(-2.74 < Z < 1.53) is the area between -2.74 and 1.53.

P(-2.74 < Z < 1.53) =0.9370 – 0.0031

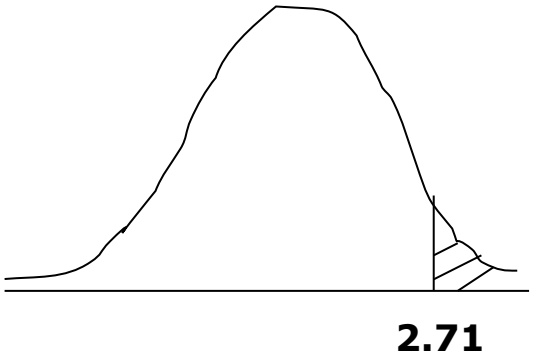
= 0.9339.



Cont'd

Example:

P(Z > 2.71) is the area to the right of 2.71.

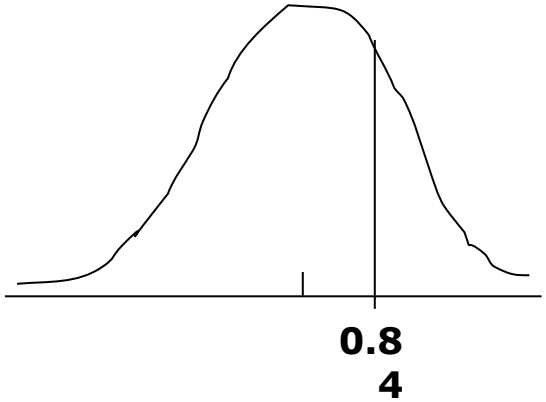


And so,

P(Z > 2.71) =1 – 0.9966 = 0.0034.

Example :

P(Z = 0.84) is the area at z = 0.84.



And so,

P(Z = 0.84) =1 – 0.9966 = 0.0034

How to transform normal distribution (X) to standard normal distribution (Z)?

- This is done by the following formula:

$$z = \frac{x - \mu}{\sigma}$$

Example:

- If X is normal with μ = 3, σ = 2. Find the value of standard normal Z, If X= 6?

Answer:

$$z = \frac{x - \mu}{\sigma} = \frac{6 - 3}{2} = 1.5$$

Normal Distribution Applications

The normal distribution can be used to model the distribution of many variables that are of interest. This allow us to answer probability questions about these random variables.

Example:

The ‘Uptime ’is a custom-made light weight battery-operated activity monitor that records the amount of time an individual spend the upright position. In a study of children ages 8 to 15 years. The researchers found that the amount of time children spend in the upright position followed a normal distribution with mean of 5.4 hours and standard deviation of 1.3.Find

If a child selected at random ,then

1-The probability that the child spend less than 3 hours in the upright position 24-hour period

$$P(X < 3) = P\left(Z = \frac{X-\mu}{\sigma} < \frac{3-5.4}{1.3}\right) = P(Z < -1.85) = 0.0322$$

2-The probability that the child spend more than 5 hours in the upright position 24-hour period

$$P(X > 5) = P\left(\frac{X-\mu}{\sigma} > \frac{5-5.4}{1.3}\right) = P(Z > -0.31)$$

$$= 1 - P(Z < -0.31) = 1 - 0.3520 = 0.648$$

3-The probability that the child spend exactly 6.2 hours in the upright position 24-hour period $P(X = 6.2) = 0$

4-The probability that the child spend from 4.5 to 7.3 hours in the upright position 24-hour period

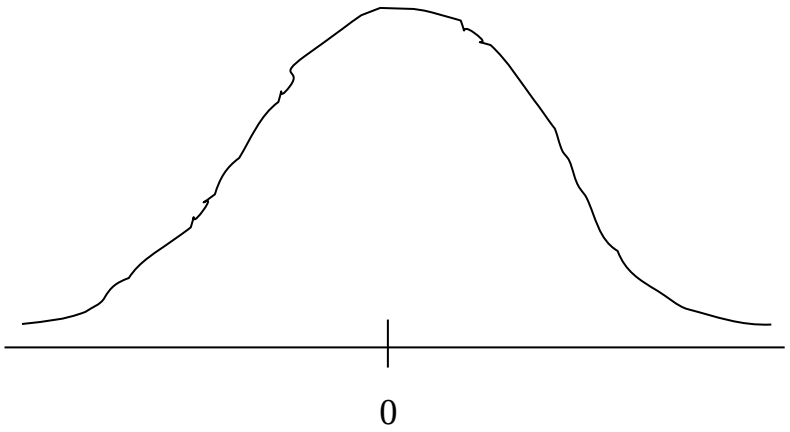
$$P(4.5 < X < 7.3) = P\left(\frac{4.5-5.4}{1.3} < \frac{X-\mu}{\sigma} < \frac{7.3-5.4}{1.3}\right)$$

$$= P(-0.69 < Z < 1.46) = P(Z<1.46) - P(Z< -0.69)$$

$$= 0.9279 - 0.2451 = 0.6828$$

The T Distribution:

1- It has mean of zero.
2- It is symmetric about the mean.
3- It ranges from -∞ to ∞.



Cont'd

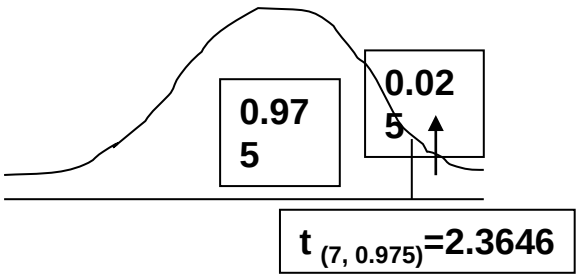
4- compared to the normal distribution, the t distribution is less peaked in the center and has higher tails.

5- It depends on the degrees of freedom (n-1).

6- The t distribution approaches the standard normal distribution as (n-1) approaches ∞.

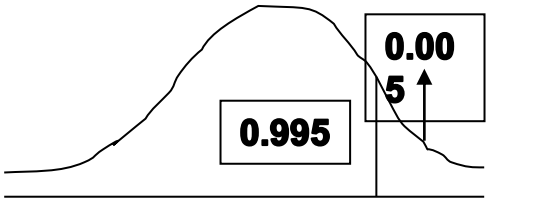
Examples

t (7, 0.975) = 2.3646

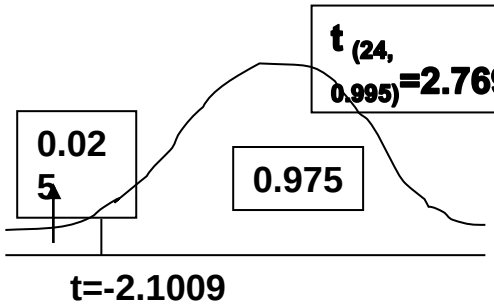


---

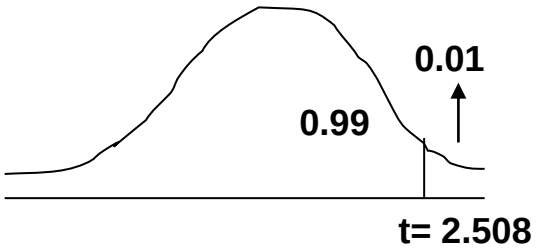
t (24, 0.995) = 2.7696



If P ($T_{(18)}$ > t) = 0.975,
then t = -2.1009



If P ($T_{(22)}$ < t) = 0.99,
then t = 2.508



Lesson 6: Sampling Theory

Introduction

- **Focus**
  - ✓To answer the following three questions
    - ❖What is the group of people from which we want to draw the sample?
    - ❖How many people do we need in our sample?
    - ❖How will this people be selected?

The Group of People

- **Population**
  - ✓ Is a complete set of elements that possess some common characteristic
  - ✓ Should be clearly and explicitly defined in terms of place, time, and other relevant criteria

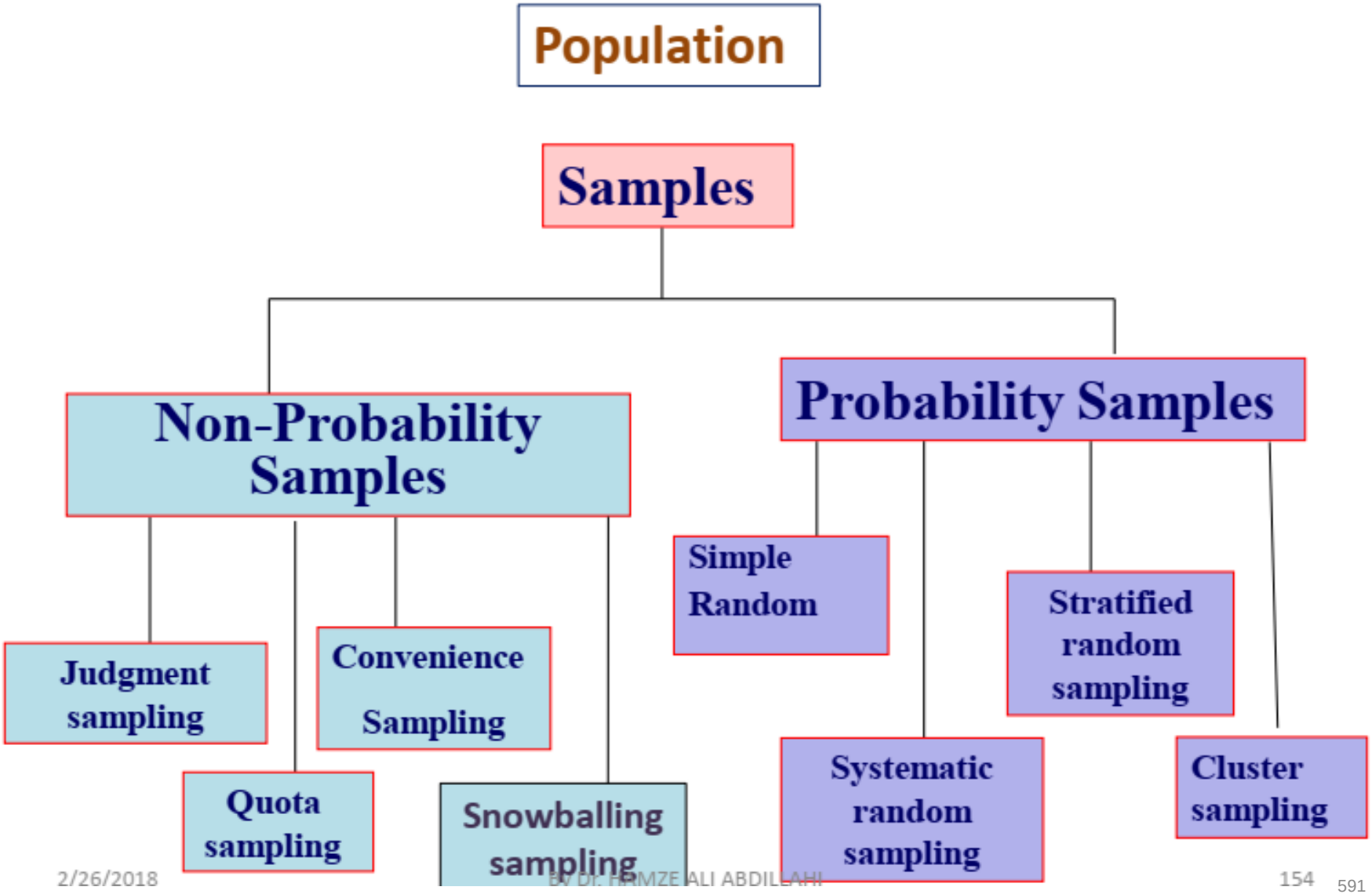
- **Target Population**
  - ✓ The entire group of people or objects to which the researcher wishes to generalize the study findings
    - All people with AIDS

Cont'd

- **Source Population**
  - ✓ A subset of the target population
  - ✓ May be limited to region, state, city, or institution
    - ❖ All people with AIDS in A.A.

- **Study Population**
  - ✓ A subset of the source population
  - ✓ Population from which the sample actually was drawn and about which a conclusion can be made.
    - ❖ All people with AIDS in A.A. satisfying the inclusion criteria

Types of Sampling Methods



Cont'd

Probability: means the chance of an occurrence. To compute the chance of occurrence, we need to know all the items in the population.

Sampling frame refers to complete list of all the items in the population.

Random means that every item in the population has an equal chance of being picked.

Why sampling?

- Investigation entire population by a census

  - is costly
  - Time consuming
  - Requires large manpower

- Sampling is a more cost-effective and convenient means of collecting information.

Simple Random Sampling

- Every individual or item from the frame has an equal chance of being selected

- **Samples obtained from:**
  - table of random numbers or
  - computer random number generators.

Cont'd

- Advantages of SR
  - minimal knowledge of population needed
  - statistical estimation of error
  - Easy to analyze data
- Disadvantages
  - High cost; low frequency of use
  - Requires sampling frame
  - Does not use researchers' expertise
  - Larger risk of random error than stratified

Cont'd

Table of random numbers

6 8 4 20, 5 7 9 57, 4 1 82 5, 6 3 29 1,
5 8 2 10, 3 62 1 5, 4 07 8 5, 9 6 02 0,
3 6 25 3, 3 34 2 5, 4 77 8 9, 1 22 0 3,
9 8 56 2, 6 31 0 1, 7 84 2 4, 5 05 3 6

- Locate one row and one column in the table.
- Close the eyes and use pencil to choose any number.
- Say the number is 5821.
- Read the digits horizontally, can also be read vertically down.
- Split the digits into two-digit numbers : **example** 58, 21, 03 ...
- Remove the repeat numbers and rearrange the selected numbers

Fore example in a class of 40 students, each students has a 1/40 (0.025) chance of being picked.

Systematic random sampling

- Decide on sample size: n
- Divide frame of N individuals into groups of k individuals: k=N/n
- Randomly select one individual from the 1st group
- Select every k-th individual thereafter.
- First number that is within the range 1 – 8 is 3
- Then the next number is 3+8 = 11 and third is 11 + 8 = 19 and so on.....

N = 64

n = 8

k = 8

First Group

Cont'd

Advantages: Systematic Sampling

- Moderate cost; moderate usage
- statistical estimation of error
- Simple to draw sample; easy to verify

Disadvantages

- Requires sampling frame
- Potential for bias if there are underlying patterns to the sampling frame

Stratified Samples

- Population divided into two or more groups according to some common characteristic with similar groups in each strata.
- Simple random sample selected from each group
- The two or more samples are combined into one.

Advantages

- minimal knowledge of population needed
- Allows calculation statistical estimation of error
- Easy to analyze data

Cont'd

Disadvantages

- High cost
- Requires sampling frame
- Does not use researchers' expertise
- Larger risk of random error than stratified
- Unhelpful if there are no homogenous groups

Cont'd

- **For example**: we have 16 boys and 24 girls in a class, and we wand to stratify the class by gender.

- First divide class list into two (boys and girls lists). •We want select 5 from the sampling frame.

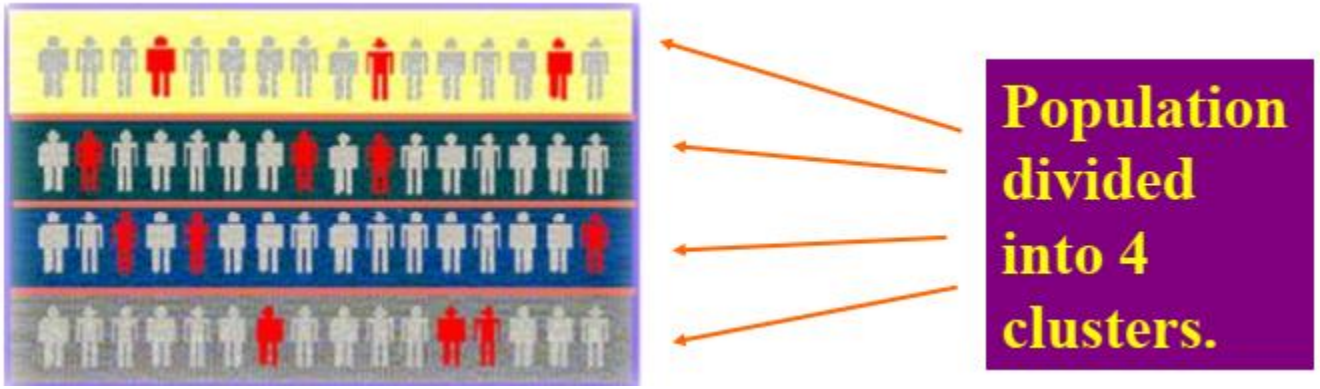
- Subjects from each stratum is usually proportionate to the population size within each stratum.

- n = 5/40 *100 = 12.5% . The number of boys will be 16*12.5/100 = 2, we select two boys from sampling frame using simple random sampling.

- The number of girls = 24 *12.5/100 = 3 we select 3 girls from the sampling frame using simple random sampling

Cluster Samples

- Population divided into several “clusters,” each representative of the population
- Simple random sample selected from each
- The samples are combined into one



Cont'd

- is useful when it is difficult or costly to develop a complete list of the population members or when the population elements are widely dispersed geographically.
- Cluster sampling may increase sampling error due to similarities among cluster members.

Advantages

- Low cost
- Requires list of all clusters
- Can estimate characteristics of both cluster and population

Disadvantages

- Increase sampling error

Stratification vs. Clustering

Stratification

- Divide population into groups different from each other: sexes, races, ages
- Sample randomly from each group
- Less error compared to simple random
- More expensive to obtain stratification information before sampling

Clustering

- Divide population into comparable groups: schools, cities
- Randomly sample some of the groups
- More error compared to simple random
- Reduces costs to sample only some areas or organizations

Cont'd

We use when the sampling frame is absent .

1. **Convenience sampling**
2. **Quota sampling**
3. **Judgment sampling**
4. **Snowballing sampling**

Convenience Sample

- Subjects are selected on basis of being readily available.
- Target population is defined, and the required sample size is determined.
- Subjects are selected until we reach the required sample size.

Cont'd

Advantages

- Very low cost
- Extensively used/understood
- No need for list of population elements

Disadvantages

- Variability and bias cannot be measured or controlled- volunteer bias

Quota Sampling

1. Select demographic characteristics of interest (e.g. age, sex, ethnicity).

2. After selecting the target population into homogenous groups , the number of subjects in each group will not be the same.

3. So, we find the percentage composition of each group in the population, similar to the first stage of stratified sampling method.

4. Then we choose the subjects using convenient procedure , on first-come-first serve basis

Cont'd

Advantages

- moderate cost
- Very extensively used/understood
- No need for list of population elements
- Introduces some elements of stratification
- Representative with regard to known characteristics

Disadvantages

- Variability and bias cannot be measured or controlled –volunteer bias

Cont'd

- In a study on perception of outpatients on services provided at a hospital, the patients may be sub- divided into various age groups .

- Target population is (patients between 21 to 60 years old seeking services at the particular hospital. Age groups are (21,30) (31,40) (41,50) (51, 60) .

- The percentage of the patients taken from hospital records were 10%, 30%, 40%, 20% respectively.

- If the overall sample size is 50 , then the 50*10/100 = 5 patients will be choosing from the first group interval (21,30) …also 15, 20 and 10 from other group respectively.

Judgment sampling

- Subjects chosen purposively on the basis of having particular features
- Used by specialists or authorities in a specific area.
- Most case studies are done in this manner.
- Sample size may not be large but an in- depth study of the cases is the main focus.
- Also used when choosing controls for epidemiological studies.
- Useful for rare characteristics

Cont'd

Advantages

- Moderate cost
- Commonly used/understood
- Sample will meet a specific objective
- Useful for qualitative research
- Useful for rare characteristics

Disadvantages

- Bias

Snowballing sampling

- Researchers move from one known case to another just by referrals.
- Used in rare events(sentinel events) .
- Enables researcher to reach groups that are otherwise hard to reach.
- For example; when studying rare behaviors in the population such as drug abuse

Cont'd

Advantages

- low cost
- Useful in specific circumstances
- Useful for locating rare populations

Disadvantages

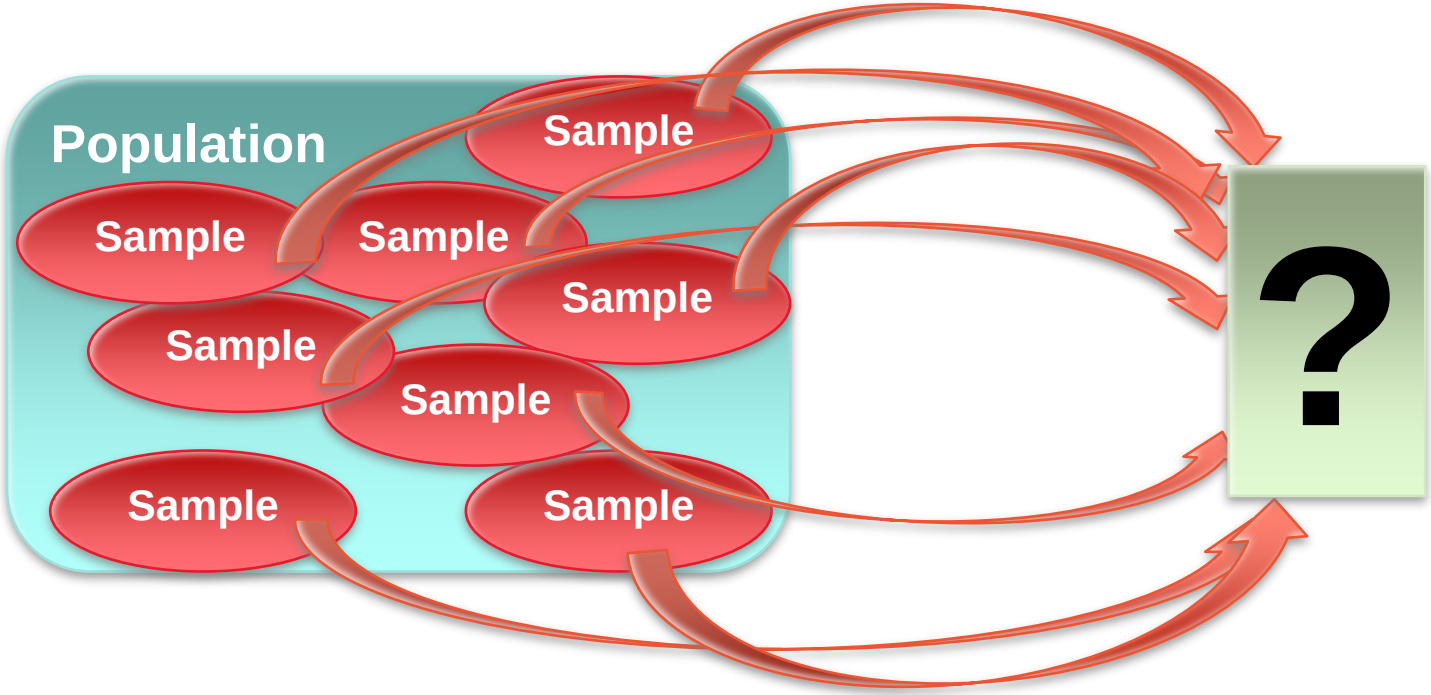
- Bias because sampling units not independent

Sampling Variability

How can $\bar{x}$ be an accurate estimate of μ? After all, different random samples would produce different values of $\bar{x}$.

This basic fact is called **sampling variability**: the value of a statistic varies in repeated random sampling.

To make sense of sampling variability, we ask, "What would happen if we took many samples?"



Sampling Distribution

- **Sampling Distribution**: the distribution of all possible values that can be assumed by some statistic, computed from samples of the same size randomly drawn from the same population, is called the sampling distribution of that statistic

Sampling Distributions: Construction Sampling distributions may be constructed empirically when sampling from a discrete, finite population. To construct a sampling distribution, we proceed as follows:

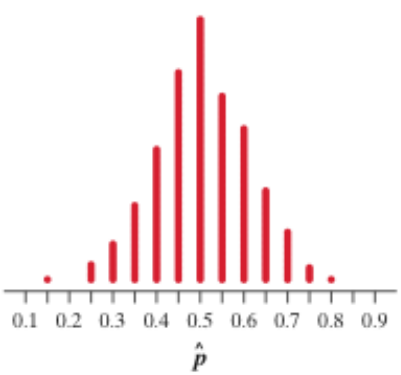
1. From a finite population of size N, randomly draw all possible samples of size n.
2. Compute the statistic of interest for each sample.
3. List in one column the different distinct observed values of the statistic, and in another column list the corresponding frequency of occurrence of each distinct observed value of the statistic

Describing Sampling Distributions

The fact that statistics from random samples have definite sampling distributions allows us to answer the question, "How trustworthy is a statistic as an estimator of the parameter?" To get a complete answer, we consider the center, spread, and shape.

Center: Biased and unbiased estimators

In the chips example, we collected many samples of size 20 and calculated the sample proportion of red chips. How well does the sample proportion estimate the true proportion of red chips, $p = 0.5$?



Note that the center of the approximate sampling distribution is close to 0.5. In fact, if we took ALL possible samples of size 20 and found the mean of those sample proportions, we'd get *exactly* 0.5.

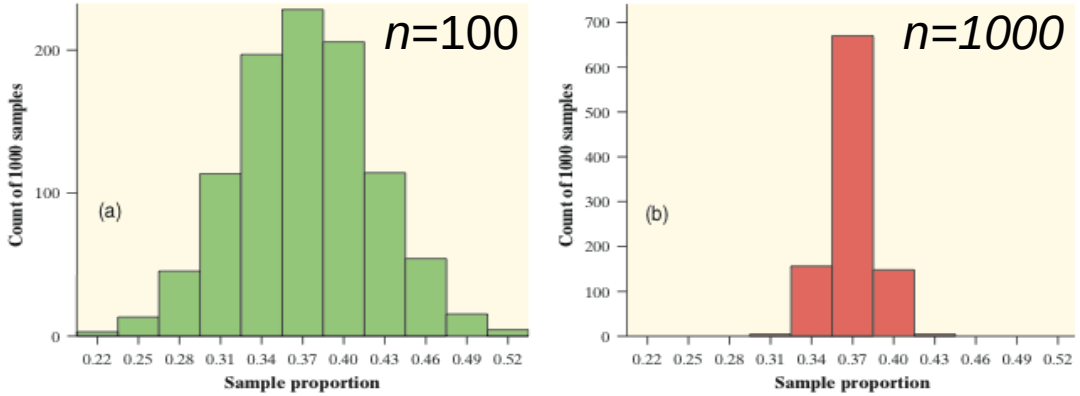
Definition:

A statistic used to estimate a parameter is an **unbiased estimator** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Describing Sampling Distributions

Spread: Low variability is better!

To get a trustworthy estimate of an unknown population parameter, start by using a statistic that's an unbiased estimator. This ensures that you won't tend to overestimate or underestimate. Unfortunately, using an unbiased estimator doesn't guarantee that the value of your statistic will be close to the actual parameter value.



Larger samples have a clear advantage over smaller samples. They are much more likely to produce an estimate close to the true value of the parameter.

Variability of a Statistic

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined primarily by the size of the random sample. Larger samples give smaller spread. The spread of the sampling distribution does not depend on the size of the population, as long as the population is at least 10 times larger than the sample.

Distribution of the Sample Mean

- Statistics such as $\bar{x}$ are random variables since their value varies from sample to sample.

- As such, they have probability distributions associated with them. In this chapter we focus on the shape, center and spread of statistics such as $\bar{x}$.

- The **sampling distribution of the sample mean** $\bar{x}$ is the probability distribution of all possible values of the random variable $\bar{x}$ computed from a sample of size *n* from a population with mean μ and standard deviation σ.

DISTRIBUTION OF THE SAMPLE MEAN

- Let's consider a population consisting of {4, 5, 9} -→ ages of someone's children
- If two ages are randomly selected with replacement from the population, identify the sampling distribution of the sample mean.

Sample	Sample Mean	Probability
4,4	4.0	1/9
4,5	4.5	1/9
4,9	6.5	1/9
5,4	4.5	1/9
5,5	5.0	1/9
5,9	7.0	1/9
9,4	6.5	1/9
9,5	7.0	1/9
9,9	9	1/9

Sampling Distribution of Mean(condensed

Sample Mean	Probability
4.0	1/9
4.5	2/9
5.0	1/9
6.5	2/9
7.0	2/9
9.0	1/9

Using this table, let's calculate the mean of the sample mean: 4(1/9)+4.5(2/9)+5(1/9)+…+9(1/9)=6.0

The mean of the population {4,5,9} is also 6.0!

Cont'd

- Because the mean of the sample means is equal to the mean of the population, we conclude that the values of the sample mean do *target* the value of the population mean.
- *The mean of the sample mean is equal to the population mean μ*

Illustrating Sampling Distributions

- **Step 1:** Obtain a simple random sample of size *n*.
- **Step 2:** Compute the sample mean.
- **Step 3:** Assuming we are sampling from a finite population, repeat Steps 1 and 2 until all simple random samples of size *n* have been obtained.

Sampling Distribution of the Sample Mean-Normal Population

The weights of pennies minted after 1982 are approximately normally distributed with mean 2.46 grams and standard deviation 0.02 grams.

Approximate the sampling distribution of the sample mean by obtaining 200 simple random samples of size n = 5 from this population

- The data on the following slide represent the sample means for the 200 simple random samples of size n = 5.
- For example, the first sample of n = 5 had the following data:
  2.493 2.466 2.473 2.492 2.471
- Note: $\bar{x}$ = 2.479 for this sample

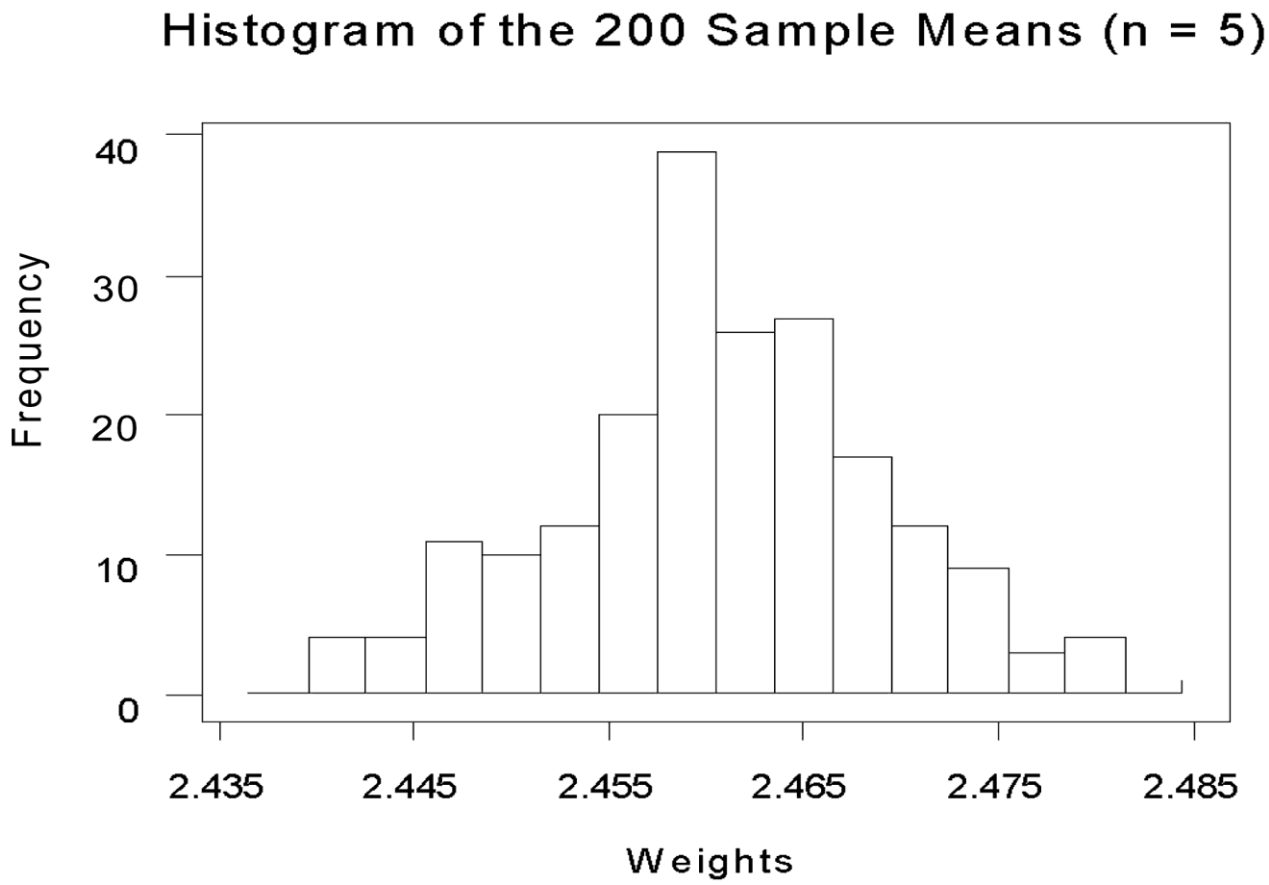
Cont'd

Sample Means for Samples of Size *n* = 5

2.479	2.463	2.459	2.455	2.468	2.467	2.464	2.480	2.467	2.468
2.465	2.463	2.463	2.459	2.465	2.461	2.459	2.467	2.475	2.455
2.466	2.475	2.465	2.470	2.449	2.474	2.458	2.471	2.461	2.463
2.451	2.446	2.470	2.465	2.461	2.460	2.460	2.454	2.466	2.466
2.458	2.454	2.467	2.471	2.453	2.468	2.467	2.476	2.456	2.473
2.446	2.459	2.460	2.472	2.467	2.471	2.466	2.459	2.454	2.459
2.471	2.468	2.460	2.461	2.468	2.463	2.467	2.446	2.469	2.458
2.472	2.457	2.444	2.480	2.443	2.459	2.452	2.458	2.460	2.460
2.466	2.465	2.442	2.461	2.466	2.446	2.454	2.466	2.463	2.455
2.459	2.451	2.451	2.458	2.461	2.459	2.457	2.464	2.459	2.459
2.445	2.459	2.474	2.458	2.447	2.451	2.478	2.465	2.470	2.438
2.469	2.456	2.445	2.450	2.473	2.449	2.450	2.463	2.447	2.453
2.457	2.457	2.446	2.447	2.456	2.465	2.453	2.458	2.455	2.460
2.460	2.460	2.465	2.466	2.475	2.460	2.450	2.472	2.455	2.462
2.454	2.467	2.463	2.474	2.460	2.465	2.465	2.460	2.466	2.483
2.456	2.453	2.453	2.446	2.465	2.461	2.466	2.463	2.457	2.442
2.457	2.459	2.461	2.442	2.471	2.465	2.459	2.448	2.466	2.460
2.479	2.461	2.462	2.459	2.453	2.473	2.463	2.461	2.442	2.478
2.446	2.467	2.455	2.449	2.462	2.461	2.464	2.457	2.460	2.469
2.461	2.462	2.460	2.456	2.458	2.471	2.459	2.455	2.465	2.456

Cont'd

- The mean of the 200 sample means is 2.46, the same as the mean of the population.
- The standard deviation of the sample means is 0.0086, which is smaller than the standard deviation of the population.



Cont'd

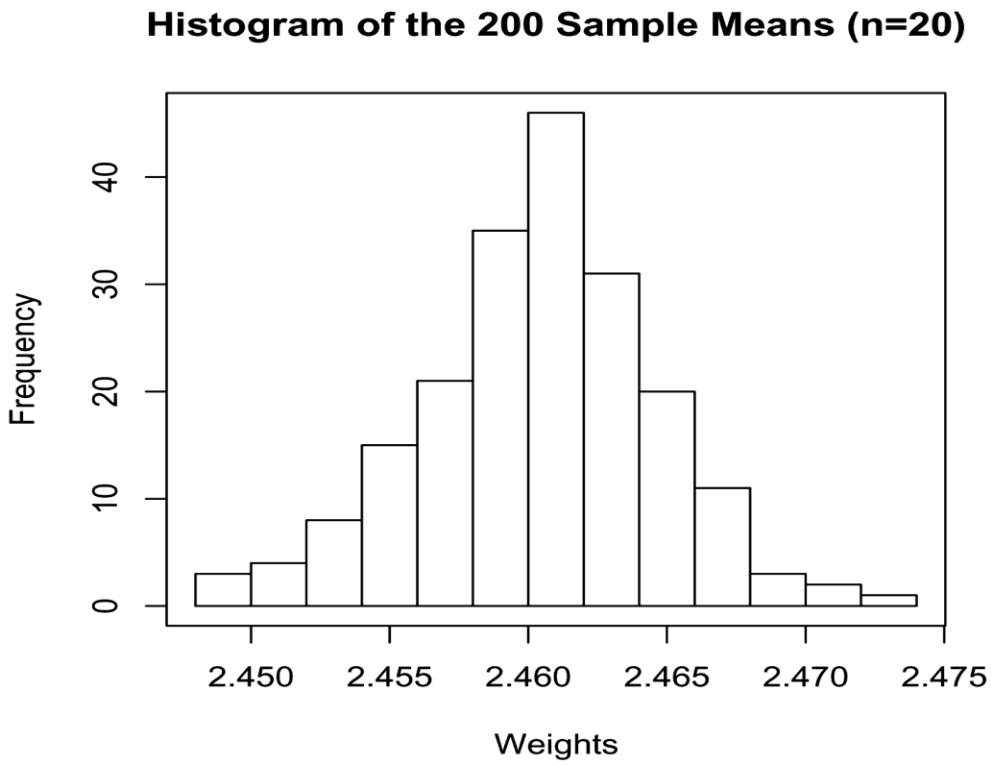
- What role does *n*, the sample size, play in the standard deviation of the distribution of the sample mean?

- As the size of the sample increases, the standard deviation of the distribution of the sample mean decreases.

The Impact of Sample Size on Sampling Variability

- Approximate the sampling distribution of the sample mean by obtaining 200 simple random samples of size $n = 20$ from the population of weights of pennies minted after 1982 ($\mu = 2.46$ grams and $\sigma = 0.02$ grams)

Cont'd



- The mean of the 200 sample means for n = 20 is still 2.46, but the standard deviation is now 0.0045 (0.0086 for n = 5). As expected, there is less variability in the distribution of the sample mean with n =20 than with n = 5.

The Mean and Standard Deviation of the Sampling Distribution of $\bar{x}$

Suppose that a simple random sample of size n is drawn from a large population with mean μ and standard deviation σ. The sampling distribution of $\bar{x}$ will have mean $\mu_{\bar{x}} = \mu$ and standard deviation

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

The standard deviation of the sampling distribution of $\bar{x}$ is called the **standard error of the mean** and is denoted $\sigma_{\bar{x}}$.

The Shape of the Sampling Distribution of $\bar{x}$ If X is Normal

- If a random variable X is normally distributed, the distribution of the sample mean $\bar{x}$ is normally distributed.

Describing the Distribution of the Sample Mean

The weights of pennies minted after 1982 are approximately normally distributed with mean 2.46 grams and standard deviation 0.02 grams.

What is the probability that in a simple random sample of 10 pennies minted after 1982, we obtain a sample mean of at least 2.465 grams?

This is not an individual value!

Cont'd

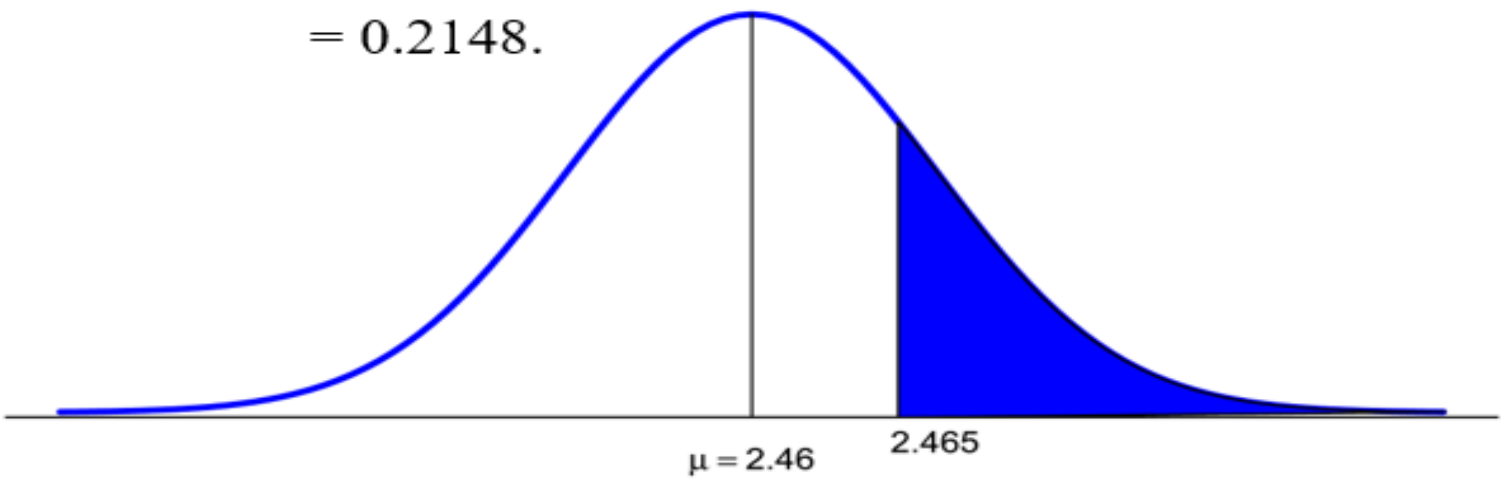
Solution

$\bar{x}$ is normally distributed with $\mu_{\bar{x}} = 2.46$ and $\sigma_{\bar{x}} = \frac{0.02}{\sqrt{10}} = 0.0063$.

$$Z = \frac{2.465 - 2.46}{0.0063} = 0.79 .$$

$$P(Z > 0.79) = 1 - 0.7852$$

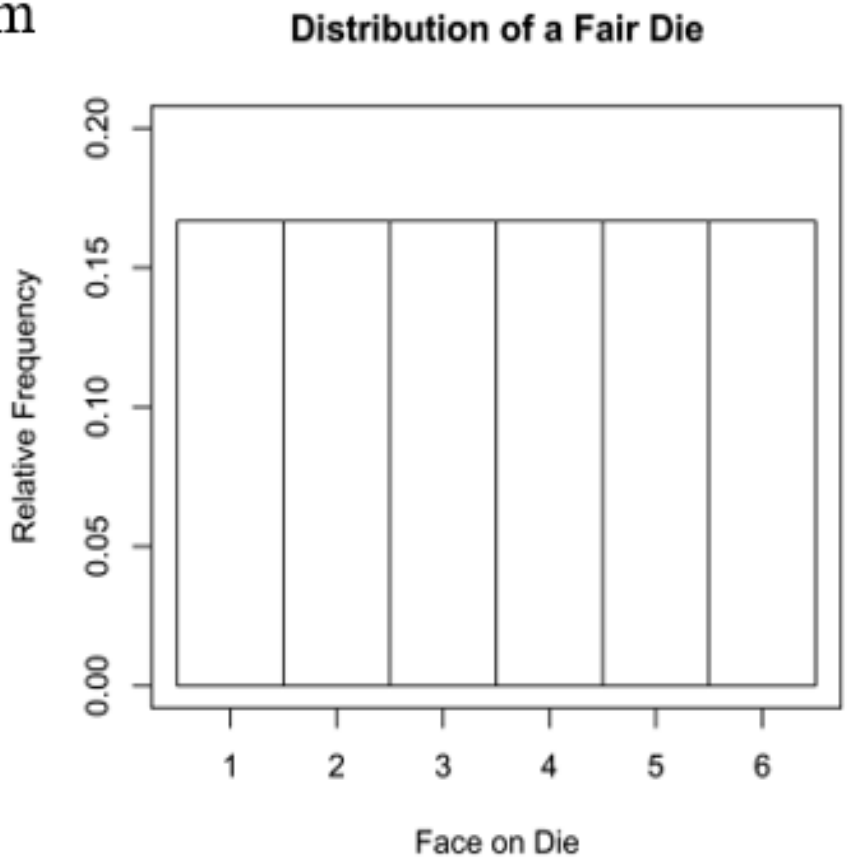
$$= 0.2148.$$



Sampling from a Population that is Not Normal

The following table and histogram give the probability distribution for rolling a fair die:

Face on Die	Relative Frequency
1	0.1667
2	0.1667
3	0.1667
4	0.1667
5	0.1667
6	0.1667

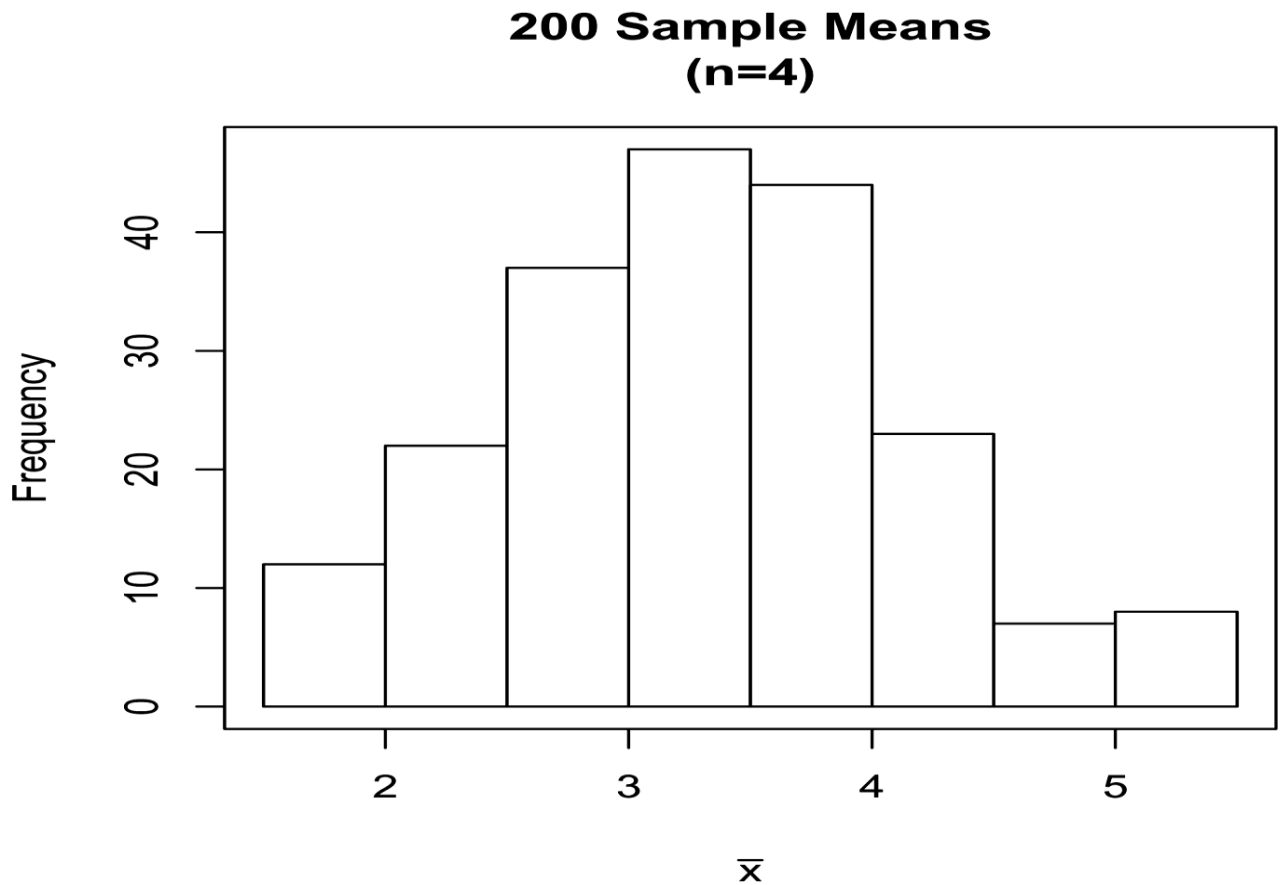


$\mu = 3.5, \ \sigma = 1.708$

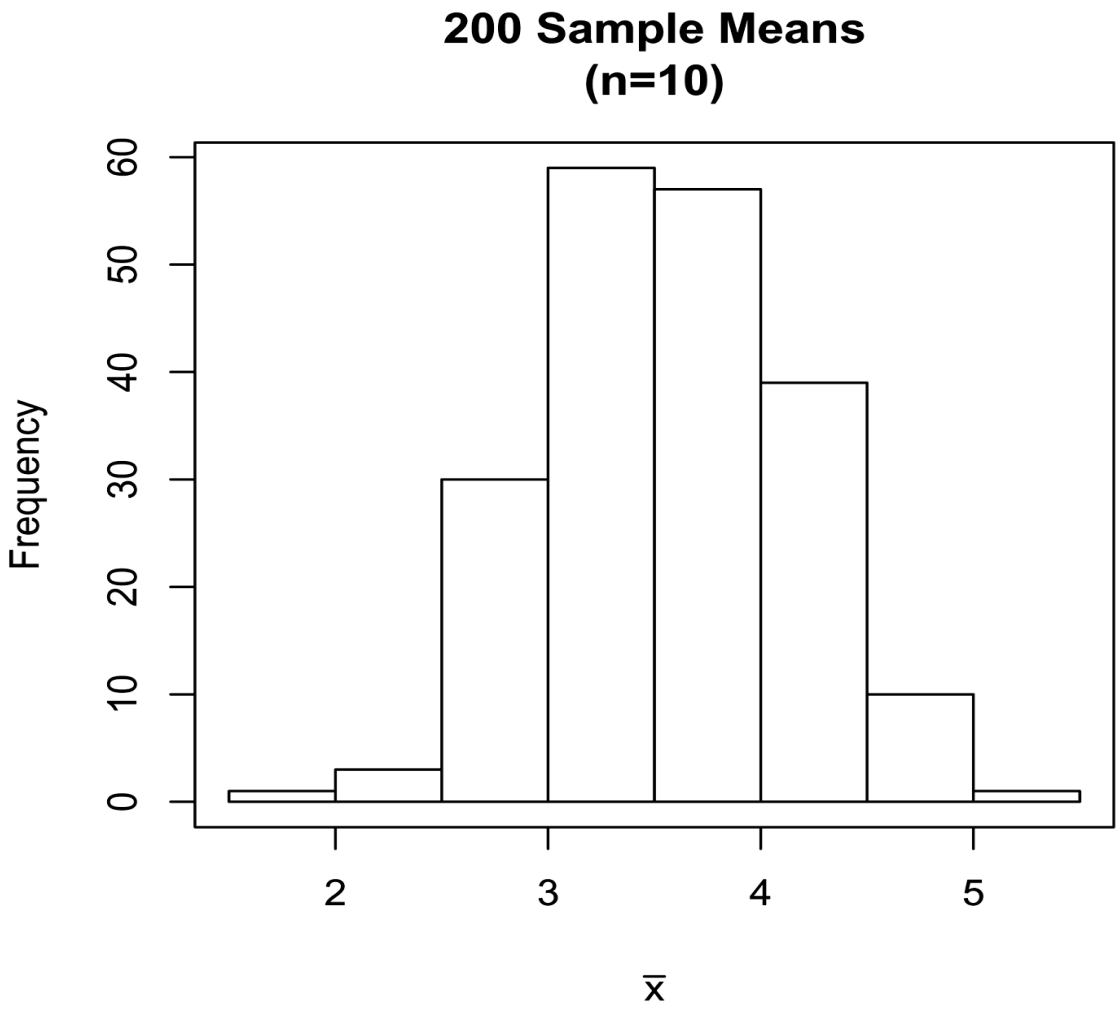
Note that the population distribution is NOT normal

Cont'd

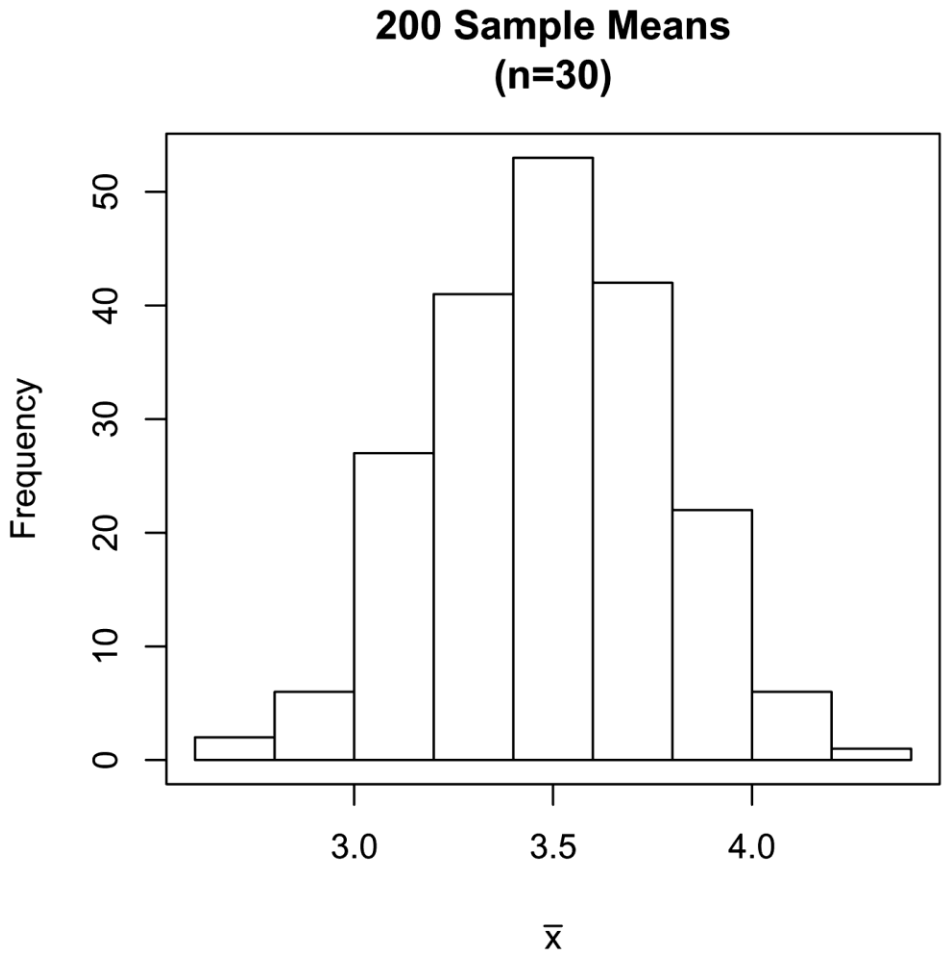
- Estimate the sampling distribution of $\bar{x}$ by obtaining 200 simple random samples of size n = 4 and calculating the sample mean for each of the 200 samples. Repeat for n = 10 and 30.



Cont'd



Cont'd



Cont'd

- The mean of the sampling distribution is equal to the mean of the parent population and the standard deviation of the sampling distribution of the sample mean is $\frac{\sigma}{\sqrt{n}}$ regardless of the sample size.
- **The Central Limit Theorem:** the shape of the distribution of the sample mean becomes approximately normal as the sample size n increases, regardless of the shape of the population.

Using the Central Limit Theorem

- Suppose that the mean time for an oil change at a "10-minute oil change joint" is 11.4 minutes with a standard deviation of 3.2 minutes.

(a) If a random sample of n = 35 oil changes is selected, describe the sampling distribution of the sample mean.

Solution: $\bar{x}$ **is approximately normally distributed with mean = 11.4 and std. dev. =** $\frac{3.2}{\sqrt{35}} = 0.5409$.

(b) If a random sample of n = 35 oil changes is selected, what is the probability the mean oil change time is less than 11 minutes?

Solution: $Z = \frac{11 - 11.4}{0.5409} = -0.74$, $\mathbf{P(Z < -0.74) = 0.23.}$

Distribution of the Sample Proportion

Suppose that a random sample of size n is obtained from a population in which each individual either does or does not have a certain characteristic. The **sample proportion**, denoted $\hat{p}$ (read "p-hat") is given by

$$\hat{p} = \frac{x}{n}$$

where x is the number of individuals in the sample with the specified characteristic. The sample proportion $\hat{p}$ is a statistic that estimates the population proportion, p.

Cont'd

In a Quinnipiac University Poll conducted in May of 2008 1745 registered voters nationwide were asked whether they approved of the way George W. Bush is handling the economy. 349 responded "yes". Obtain a point estimate for the proportion of registered voters who approve of the way George W. Bush is handling the economy.

Solution: $\hat{p} = \frac{349}{1745} = 0.2$

Cont'd

Sampling Distribution of $\hat{p}$

For a simple random sample of size n with population proportion p:

- The shape of the sampling distribution of $\hat{p}$ is approximately normal provided $np(1-p) \geq 10$.
- The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$
- The standard deviation of the sampling distribution of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Cont'd

Sampling Distribution of $\hat{p}$

- The model on the previous slide requires that the sampled values are independent. When sampling from finite populations, this assumption is verified by checking that the sample size n is no more than 5% of the population size N ($n \leq 0.05N$).

- Regardless of whether $np(1-p) \geq 10$ or not, the mean of the sampling distribution of $\hat{p}$ is p, and the standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Lesson 7 : Estimation

Lesson Outline

- Introduction
- Confidence interval for a population mean
- The t distribution
- Confidence interval for the difference between two population means
- Confidence interval for a population proportion
- Confidence interval for the difference between two population proportions
- Determination of sample size for estimating means
- Determination of sample size for estimating proportions

Introduction

Statistical inference is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

The process of estimation entails calculating, from the data of a sample, some statistic that is offered as an approximation of the corresponding parameter of the population from which the sample was drawn.

The rationale behind estimation in the health sciences field rests on the assumption that workers in this field have an interest in the parameters, such as means and proportions, of various populations.

Example

- Suppose the administrator of a large hospital is interested in the mean age of patients admitted to his hospital during a given year. He may consider it too expensive to go through the records of all patients admitted during that particular year and, consequently, elect to examine a sample of the records from which he can compute an estimate of the mean age of patients admitted that year.

- A physician in general practice may be interested in knowing what proportion of a certain type of individual, treated with a particular drug, suffers undesirable side effects. No doubt, her concept of the population consists of all those persons who ever have been or ever will be treated with this drug. Deferring a conclusion until the entire population has been observed could have an adverse effect on her practice.

Cont'd

- Statistical inference can be further subdivided into the two main areas of estimation and hypothesis testing.

- Estimation is concerned with estimating the values of specific population parameters; hypothesis testing is concerned with testing whether the value of a population parameter is equal to some specific value.

- To any parameters, we can compute two types of estimate: a point estimate and an interval estimate.

- A point estimate is a single numerical value used to estimate the corresponding population parameter.
- An interval estimate consists of two numerical values defining a range of values that, with a specified degree of confidence, most likely includes the parameter being estimated.

Properties of Good Estimator

Consistent Estimator: An estimator which gets closer to the value of the parameter as the sample size increases. Let be an estimator of $\widehat{\boldsymbol{\theta}}$, i.e. gets closer to θ as the sample size increases.

Unbiased Estimator: An estimator whose expected value is the value of the parameter being estimated. Let be an estimator of $\widehat{\boldsymbol{\theta}}$, i.e. $E(\widehat{\boldsymbol{\theta}}) = \theta$

Relatively Efficient Estimator: The estimator for a parameter with the smallest variance.

Estimation (Point estimation and Interval estimation) of population mean

Point Estimation

- Another term for statistic is point estimate, since we are estimating the parameter value. A point estimator is the mathematical way we compute the point estimate. For instance, sum of

x_i over n is the point estimator used to compute the estimate of the population means, μ

.That is $\bar{X} = \dfrac{\sum x_i}{n}$ is a point estimator of the population mean.

- **Point estimate for (μ) is $\bar{X}$**
- **Point estimate for (σ) is S**

Cont'd

Point Estimation of $\mu_1 - \mu_2$:

Result:

$\bar{X}_1 - \bar{X}_2$ is a "good" point estimate for $\mu_1 - \mu_2$.

(i) Point Estimate for (p):

Result:

A good point estimate for the population proportion (p) is the sample proportion ($\hat{p}$).

Point Estimation for $p_1 - p_2$:

Result:

A good point estimator for the difference between the two proportions, $p_1 - p_2$, is:

$$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$$

Confidence Interval for a Population Mean

An ***interval estimate*** consists of two numerical values defining a range of values that, with a specified degree of confidence, we feel includes the parameter being estimated.

Confidence Interval: An interval estimate with a specific level of confidence

Confidence Level: The percent of the time the true value will lie in the interval estimate given.

Degrees of Freedom: The number of data values which are allowed to vary once a statistic has been determined.

Confidence Interval for a Population Mean

Suppose researchers wish to estimate the mean of some normally distributed population.

- They draw a random sample of size n from the population and compute , which they use as a point estimate of μ.
- Because random sampling involves chance, then $\bar{x}$ can't be expected to be equal to μ.
- The value of $\bar{x}$ may be greater than or less than μ.
- It would be much more meaningful to estimate μ by an interval.

The 1-α percent confidence interval (C.I.) for μ

- We want to find two values L :Lower bound and U:Upper bound between which μ lies with high probability, i.e.

  $P(L \leq \mu \leq U) = 1-\alpha$

Example: when,

- $\alpha = 0.01$,

  then $1- \alpha = 1-0.01= 0.99$
- $\alpha = 0.05$,

  then $1- \alpha = 1-0.05= 0.95$
- $\alpha = 0.1$,

  then $1- \alpha = 1-0.1= 0.90$

(1- α)100% confident interval for the mean μ

- **When the value of sample size (n):**

population is normal or not normal

(n ≥ 30)

σ is known

$$\overline{X} \pm Z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

σ is not known

$$\overline{X} \pm Z_{1-\alpha} \frac{S}{\sqrt{n}}$$

population is normal

(n< 30)

σ is known

$$\overline{X} \pm Z_{1-\alpha} \frac{\sigma}{\sqrt{n}}$$

σ is not known

$$\overline{X} \pm t_{1-\alpha,n-1} \frac{S}{\sqrt{n}}$$

Cont'd

- Example: Suppose a researcher , interested in obtaining an estimate of the average level of some enzyme in a certain human population, takes a sample of 10 individuals, determines the level of the enzyme in each, and computes a sample mean of approximately $\bar{x}$=22. Suppose further it is known that the variable of interest is approximately normally distributed with a variance of 45. We wish to estimate μ. **(α=0.05)**

Cont'd

- 1- α=0.95→ α=0.05→ $\alpha/2$=0.025,
- variance = σ^2 = 45 → $\sigma=\sqrt{45}$,n=10
- 95%confidence interval for μ is given by:

$$\bar{x} = 22$$

- $P\left(\bar{x} - Z_{1-\alpha/2}\,{}^{\sigma}/_{\sqrt{n}} < \mu < \bar{x} + Z_{1-\alpha/2}\,{}^{\sigma}/_{\sqrt{n}}\right) = 1 - \alpha$

- $Z_{(1-\alpha/2)} = Z_{0.975} = 1.96$ (refer to table D)
- $Z_{0.975}(\sigma/\sqrt{n}) = 1.96\,(\sqrt{45}\,/\,\sqrt{10}) = 4.1578$
- $22 \pm 1.96\,(\sqrt{45}\,/\,\sqrt{10})$ →
- (22-4.1578, 22+4.1578) → (17.84, 26.16)

Cont'd

The activity values of a certain enzyme measured in normal gastric tissue of 35 patients with gastric carcinoma has a mean of 0.718 and a standard deviation of 0.511.We want to construct a 90 % confidence interval for the population mean.

- **Solution:**

- Note that the population is not normal,

- n=35 (n>30) n is large and σ is unknown , s=0.511

- 1- α=0.90$\rightarrow$ α=0.1, $\rightarrow$ $\alpha/2$=0.05$\rightarrow$ 1-$\alpha/2$=0.95,

Cont'd

Then 90% confident interval for μ is given by :

- $P\left(\bar{x} - Z_{1-\alpha/2}\, {}^{s}/_{\sqrt{n}} < \mu < \bar{x} + Z_{1-\alpha/2}\, {}^{s}/_{\sqrt{n}}\right) = 1 - \alpha$

- $Z_{(1-\alpha/2)} = Z_{0.95} = 1.645$ (refer to table D)
- $Z_{0.95}(s/\sqrt{n}) = 1.645\ (0.511/\sqrt{35}) = 0.1421$

  $0.718 \pm 1.645\ (0.511)\ /\ \sqrt{35} \rightarrow$

  $(0.718-0.1421,\ 0.718+0.1421) \rightarrow$

  $(0.576, 0.860)$.

Confidence Interval for the difference between two Population Means

If we draw two samples from two independent population and we want to get the confident interval for the difference between two population means , then we have the following cases :

a) When the population is normal

1) When the variance is known and the sample sizes is large or small, the C.I. has the form:

$$(\bar{x}_1 - \bar{x}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Cont'd

2) When variances are unknown but equal, and the sample size is small, the C.I. has the form

$$(\bar{x}_1 - \bar{x}_2) - t_{1-\frac{\alpha}{2},(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{1-\frac{\alpha}{2},(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Example

The researcher team interested in the difference between serum uric and acid level in a patient with and without Down's syndrome .In a large hospital for the treatment of the mentally retarded, a sample of 12 individual with Down's Syndrome yielded a mean of $\bar{x}_1 = 4.5$ mg/100 ml. In a general hospital a sample of 15 normal individual of the same age and sex were found to have a mean value of $\bar{x}_2 = 3.4$. If it is reasonable to assume that the two population of values are normally distributed with variances equal to 1 and 1.5,find the 95% C.I for $\mu_1 - \mu_2$

<u>Solution:</u>

$1-\alpha=0.95 \rightarrow \alpha=0.05 \rightarrow \alpha/2=0.025 \rightarrow Z_{(1-\alpha/2)} = Z_{0.975} = 1.96$

$$(\bar{x}_1 - \bar{x}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = (4.5 - 3.4) \pm 1.96 \sqrt{\frac{1}{12} + \frac{1.5}{15}}$$

- $1.1 \pm 1.96(0.4282) = 1.1 \pm 0.84 = (0.26, 1.94)$

Cont'd

- We say that we are 95 percent confident that the true difference, $\mu_1 - \mu_2$, is somewhere between .26 and 1.94 because, in repeated sampling, 95 percent of the intervals constructed in this manner would include the difference between the true means.

- Since the interval does not include zero, we conclude that the two population means are not equal.

Example 2

The purpose of the study was to determine the effectiveness of an integrated outpatient dual-diagnosis treatment program for mentally ill subject. The authors were addressing the problem of substance abuse issues among people with sever mental disorder. A retrospective chart review was carried out on 50 patient ,the recherché was interested in the number of inpatient treatment days for physics disorder during a year following the end of the program. Among 18 patient with schizophrenia, The mean number of treatment days was 4.7 with standard deviation of 9.3. For 10 subject with bipolar disorder, the mean number of treatment days was 8.8 with standard deviation of 11.5. We wish to construct 99% C.I for the difference between the means of the populations Represented by the two samples

Cont'd

- $1-\alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha/2 = 0.005 \rightarrow 1-\alpha/2 = 0.995$
- $n_2 - 2 = 18 + 10 - 2 = 26 + n_1$
- $t_{(1-\alpha/2),(n_1+n_2-2)} = t_{0.995,26} = 2.7787$, then 99% C.I for $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{1-\frac{\alpha}{2},(n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

- where
- $$S_p = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{(17x9.3^2) + (9x11.5^2)}{18+10-2} = 102.33$$

- then

$(4.7-8.8) \pm 2.7787 \sqrt{102.33} \sqrt{(1/18)+(1/10)}$

$-4.1 \pm 11.086 = (-15.186, 6.986)$

Confidence Interval for a Population proportion (P)

- Many questions of interest to the health worker relate to population proportions.
- What proportion of patients who receive a particular type of treatment recover?
- What proportion of some population has a certain disease?
- What proportion of a population is immune to a certain disease?
- To estimate a population proportion, we proceed in the same manner as when estimating a population mean.

Confidence Interval for a Population proportion (P)

- A sample is drawn from the population of interest ,then compute the sample proportion $\hat{p}$ such as

$$\hat{p} = \frac{\text{no. of element in the sample with some charachtaristic}}{\text{Total no. of element in the sample}} = \frac{a}{n}$$

- This sample proportion is used as the point estimator of the population proportion . A confident interval is obtained by the following formula

$$\hat{P} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}}$$

Example 1

The Pew internet life project reported in 2003 that 18% of internet users have used the internet to search for information regarding experimental treatments or medicine. The sample consist of 1220 adult internet users, and information was collected from telephone interview. We wish to construct 98% C.I for the proportion of internet users who have search for information about experimental treatments or medicine

Cont'd

Solution :

$1-\alpha = 0.98 \longrightarrow \alpha = 0.02 \longrightarrow \alpha/2 = 0.01 \longrightarrow 1-\alpha/2 = 0.99$

$Z_{1-\alpha/2} = Z_{0.99} = 2.33$, n=1220, $\hat{p} = \frac{18}{100} = 0.18$

The 98% C. I is

$$\hat{P} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}(1-\hat{P})}{n}} = 0.18 \pm 2.33 \sqrt{\frac{0.18(1-0.18)}{1220}}$$

$0.18 \pm 0.0256 = (0.1544 , 0.2056)$

Confidence Interval for the difference between two Population proportions

- Two samples is drawn from two independent population of interest ,then compute the sample proportion for each sample for the characteristic of interest. An unbiased point estimator for the difference between two population proportions $\hat{P}_1 - \hat{P}_2$

- A 100(1-α)% confident interval for P_1 - P_2 is given by

$$(\hat{P}_1 - \hat{P}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$

Example

Connor investigated gender differences in proactive and reactive aggression in a sample of 323 adults (68 female and 255 males). In the sample ,31 of the female and 53 of the males were using internet in the internet café. We wish to construct 99 % confident interval for the difference between the proportions of adults go to internet café in the two sampled population .

Cont'd

Solution :

$1-\alpha = 0.99 \rightarrow \alpha = 0.01 \rightarrow \alpha/2 = 0.005 \rightarrow 1-\alpha/2 = 0.995$

$Z_{1-\alpha/2} = Z_{0.995} = 2.58$, $n_F = 68$, $n_M = 255$,

$$\hat{p}_F = \frac{a_F}{n_F} = \frac{31}{68} = 0.4559, \hat{p}_M = \frac{a_M}{n_M} = \frac{53}{255} = 0.2078$$

The 99% C. I is

$$(\hat{P}_F - \hat{P}_M) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{P}_F(1-\hat{P}_F)}{n_F} + \frac{\hat{P}_M(1-\hat{P}_M)}{n_M}}$$

$$(0.4559 - 0.2078) \pm 2.58 \sqrt{\frac{0.4559\,(1-0.4559)}{68} + \frac{0.2078\,(1-0.2078)}{255}}$$

$0.2481 \pm 2.58(0.0655) = (0.07914, 0.4171)$

Determination of Sample Size for Estimating Means

- The question of how large a sample to take arises early in the planning of any survey or experiment. This is an important question that should not be treated lightly.

- To take a larger sample than is needed to achieve the desired results is wasteful of resources, whereas very small samples often lead to results that are of no practical use.

- The objectives in interval estimation are to obtain narrow intervals with high reliability. If we look at the components of a confidence interval, we see that the width of the interval is determined by the magnitude of the quantity.

$$(\text{reliability coefficient}) \times (\text{standard error of the estimator})$$

- since the total width of the interval is twice this amount.

Determination of Sample Size for Estimating Means

- This quantity is usually called the precision of the estimate or the margin of error.
- For a given standard error, increasing reliability means a larger reliability coefficient.
- But a larger reliability coefficient for a fixed standard error makes for a wider interval.
- On the other hand, if we fix the reliability coefficient, the only way to reduce the width of the interval is to reduce the standard error.
- Since the standard error is equal to $\sigma/\sqrt{n}$ and since σ is a constant, the only way to obtain a small standard error is to take a large sample. How large a sample?

Determination of Sample Size for Estimating Means

- That depends on the size of σ , the population standard deviation, the desired degree of reliability, and the desired interval width.
- Let us suppose we want an interval that extends d units on either side of the estimator. We can write

$$d = (\text{reliability coefficient}) \times (\text{standard error of the estimator})$$

- If sampling is to be with replacement, from an infinite population, or from a population that is sufficiently large to warrant our ignoring the finite population correction,

$$d = z\frac{\sigma}{\sqrt{n}}$$

which, when solved for n, gives

$$n = \frac{z^2\sigma^2}{d^2}$$

Determination of Sample Size for Estimating Means

When sampling is without replacement from a small finite population, the finite population correction is required

$$d = z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

which, when solved for n, gives

$$n = \frac{Nz^2\sigma^2}{d^2(N-1) + z^2\sigma^2}$$

Determination of Sample Size for Estimating Means

Estimating σ^2 The formulas for sample size require knowledge of σ^2 but, as has been pointed out, the population variance is, as a rule, unknown. As a result, σ^2 has to be estimated. The most frequently used sources of estimates for σ^2 are the following:

1. A *pilot* or preliminary sample may be drawn from the population, and the variance computed from this sample may be used as an estimate of σ^2. Observations used in the pilot sample may be counted as part of the final sample, so that n (the computed sample size) $-n_1$ (the pilot sample size) $= n_2$ (the number of observations needed to satisfy the total sample size requirement).
2. Estimates of σ^2 may be available from previous or similar studies.
3. If it is thought that the population from which the sample is to be drawn is approximately normally distributed, one may use the fact that the range is approximately equal to six standard deviations and compute $\sigma \approx R/6$. This method requires some knowledge of the smallest and largest value of the variable in the population.

Example

- A health department nutritionist, wishing to conduct a survey among a population of teenage girls to determine their average daily protein intake (measured in grams), is seeking the advice of a biostatistician relative to the sample size that should be taken.

- What procedure does the biostatistician follow in providing assistance to the nutritionist? Before the statistician can be of help to the nutritionist, the latter must provide three items of information: (1) the desired width of the confidence interval, (2) the level of confidence desired, and (3) the magnitude of the population variance.

Cont'd

- **Solution:** Let us assume that the nutritionist would like an interval about 10 grams wide; that is, the estimate should be within about 5 grams of the population mean in either direction. In other words, a margin of error of 5 grams is desired. Let us also assume that a confidence coefficient of .95 is decided on and that, from past experience, the nutritionist feels that the population standard deviation is probably about 20 grams. The statistician now has the necessary information to compute the sample size $Z = 1.96, \sigma = 20 \; and \; d = 5.$ Let us assume that the population of interest is large so that the statistician may ignore the finite population correction and use the above formula

$$n = \frac{(1.96)^2(20)^2}{(5)^2}$$
$$= 61.47$$

- The nutritionist is ac of size 62.

Determination of Sample Size for Estimating Proportions

- The method of sample size determination when a population proportion is to be estimated is essentially the same as that described for estimating a population mean.

$$n = \frac{z^2 pq}{d^2}$$

where q= 1-p

- If the finite population correction cannot be disregarded, the proper formula for n is

$$n = \frac{Nz^2 pq}{d^2(N-1) + z^2 pq}$$

Estimating *p*

- One solution to this problem is to take a pilot sample and compute an estimate to be used in place of p in the formula for n.

- Sometimes an investigator will have some notion of an upper bound for p that can be used in the formula. For example, if it is desired to estimate the proportion of some population who have a certain disability, we may feel that the true proportion cannot be greater than, say, 0.30. We then substitute 0.30 for p in the formula for n.

- If it is impossible to come up with a better estimate, one may set p equal to 0.5 and solve for n. Since p=0.5 in the formula yields the maximum value of n, this procedure will give a large enough sample for the desired reliability and interval width.

- It may, however, be larger than needed and result in a more expensive sample than if a better estimate of p had been available. This procedure should be used only if one is unable to arrive at a better estimate of p.

Example

- A survey is being planned to determine what proportion of families in a certain area are medically indigent. It is believed that the proportion cannot be greater than .35. A 95 percent confidence interval is desired with d=0.05. What size sample of families should be selected?

- **Solution**: If the finite population correction can be ignored, we have

$$n = \frac{(1.96)^2(.35)(.65)}{(.05)^2} = 349.59$$

- The necessary sample size, then, is 350.

2. Sample Size: Two Samples

A. Estimation of the difference between two population means

B. Estimation of the difference between two population proportions

A. Sample size for estimating a difference in two means

▶ **Aim: Estimate**

$$n = \frac{z^2 \cdot (\sigma_1^2 + \sigma_2^2)}{d^2}$$

If equal sample size in both groups is required

B. Sample size for estimating a difference in two proportions

- **Aim: Estimate p_1-p_2**

$$n = \frac{z^2 \cdot (p_1 q_1 + p_2 q_2)}{d^2}$$

If equal sample size in both groups is required

- But the population Values are most of the time unknown

As a result, they has to be estimated from:

- *Pilot or preliminary sample:*
  - Select a pilot sample and estimate the population values
- *Previous or similar studies*

3. Sample Size Based on Hypothesis Testing

- **Type I error (α)** = The probability of rejecting H_0 when it is true

$$\alpha = P(\text{reject } H_0 \mid H_0 \text{ true})$$

- **Type II error (β)** = The probability of not rejecting H_0 when it is false

$$\beta = P(\text{do not reject } H_0 \mid H_0 \text{ false})$$

- **Power (1-β)** = the probability H_0 is rejected given that it is false

  = P (rejecting H_0/H_0 is false)

Cont'd

- If the power of a test is low, then there is little chance of detecting a difference even if one really exists
- Power is an important part of the design of a study
  - Power $(1 - \beta) = 50\%$, $Z_\beta = 0.00$
  - Power $(1 - \beta) = 75\%$, $Z_\beta = 0.67$
  - Power $(1 - \beta) = 80\%$, $Z_\beta = 0.84$
  - Power $(1 - \beta) = 90\%$, $Z_\beta = 1.28$
- Most of the studies recommend power of at least 80%.
- **A. Comparison between two means (Equal sample sizes)**

$$n_1 = n_2 = \frac{(z_{\alpha/2} + z_\beta)^2 (\sigma_1^2 + \sigma_2^2)}{\Delta^2}$$

Cont'd

$\Delta = /\mu_1 - \mu_2/$

The means and variances of the two respective groups are (μ_1, σ^2_1) and (μ_2, σ^2_2).

- Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort studies .
- Example: 1. Determine the sample sizes required to detect a difference of 5 mm in mean blood pressure between individuals receiving placebo and those receiving drug with α =5% and power of 0.80

Assume $\sigma_1 = \sigma_2$ = 15 mm in each group.

- We are interested in testing:

Cont'd

- H_0: $\mu_1 - \mu_2 = 5$, H_A: $\mu_1 - \mu_2 \neq 5$

$$n = \frac{(z_\alpha + z_\beta)^2(\sigma_1^2 + \sigma_2^2)}{\Delta^2} = \frac{(1.96+0.84)^2(15^2+15^2)}{5^2} = 141.1$$

We would need 142 individuals in each group

Final sample size???

B. Comparison between two means (Unequal sample sizes)

$$n_1 = \frac{(r+1)}{r} \frac{\sigma^2 (Z_\beta + Z_{\alpha/2})^2}{\text{difference}^2}$$

$r = n_2/n_1$

Cont'd

Example: Suppose we anticipate twice as many non-OC users as OC users entering the study. Determine the sample size to achieve an 80% power in the study using α=0.05. r = 2.

$n_1 = 1.5*((15.34^2+18.23^2)(1.96+0.84)^2/(5.42)^2)$

= 228 OC users and n_2 = 2(228) = 456 non-OC users.

Final sample size???

C. Comparison between two proportions (Equal sample sizes)

$$n = \frac{2(\bar{p})(1-\bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

$$\bar{p} = \frac{p_1 + p_2}{2}$$

Cont'd

Let p_1=0.35, p_2=0.25, and $\Delta=p_1-p_2$=0.35-0.25 =0.10
We would need approximately 330 subjects in each group. Final Sample Size???

Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort

D. Comparison between two proportions (Unequal sample sizes)

$$n = \frac{r+1}{r} \frac{\bar{p}(1-\bar{p})(Z_\beta + Z_{\alpha/2})^2}{(p_1 - p_2)^2}$$

Were

$r=n_2/n_1$

$$\bar{q} = 1 - \bar{p}.$$

$\bar{p}$ $=(p_1+ rp_2)/(1+r)$

Cont'd

- Note: This formula is quite general, and applies to comparative cross sectional, case-control as well as cohort studies
- Example: A study is proposed to study the link b/n anticoagulant therapy and bleeding. Patients are divided into two groups: as cases and controls. Suppose that 5% of cases and 22% of controls are anticipated to have exposure for the treatment. How large sample should such a study be to have an 80% chance of finding a significance association at a ratio of 1:2 for cases and control at α =5%.
- Solution

p_1=0.05, p_2=0.22, = (0.05+2*0.22)/(1+2) = 0.16, Δ = 0.22-0.05 = 0.17

We would need approximately 165 subjects in total

$\mathbf{n_1}$= 55 cases and n_2 = 2(55) = 110 control.

Final Sample Size??

Comparison between two proportions

(Equal sample sizes)

$$n_1 = n_2 = \frac{\left(z_{\alpha/2}\sqrt{2\overline{p}\overline{q}} + z_{\beta}\sqrt{p_1q_1 + p_2q_2}\right)^2}{\Delta^2}$$

Where

$$\overline{p} = \frac{p_1 + p_2}{2} \qquad \overline{q} = 1 - \overline{p}.$$

Δ = p1-p2

Note: This formula is quite general, and applies to cohort case-control as well as comparative Cross sectional Studies.

Cont'd

- **Let p_1=0.35, p_2=0.25, and $\Delta=p_1-p_2$=0.35-0.25 =0.10**

$$n_1 = n_2 = \frac{\left(1.96\sqrt{2(0.30)(.70)} + 0.84\sqrt{(0.35)(0.65)+(0.25)(0.75)}\right)^2}{(0.10)^2}$$

$$n_1 = n_2 = 328.1$$

- **We would need approximately 329 subjects in each group**
- Final Sample Size??

Comparison between two proportions (Unequal sample sizes)

$$n_1 = \frac{\left[Z_{\alpha/2}\sqrt{\bar{p}\bar{q}\,(1+1/\lambda)} + Z_{\beta}\sqrt{p_1q_1+p_2q_2/\lambda}\right]^2}{\Delta^2}$$

Where $n_2=n_1\lambda$, $\bar{p} = (p_1+\lambda p_2)/(1+\lambda)$

- A study is proposed to study the effect of a new anticoagulant therapy. Patients are to be randomly divided into two groups: one receives the anticoagulant, and the other placebo. The groups are then followed for the incidence of major bleeding events over 3 years. Suppose that 5% of treated patients and 22% of controls are anticipated to experience a major event over 3 years. How large sample should such a study be to have an 80% chance of finding a significance difference at a ratio of 1:2 for treated and control at α =5%.

Cont'd

▶ Solution

p_1=0.05, p_2=0.22, = (0.05+2*0.22)/(1+2) = 0.16

q_1=0.95, q_2=0.78, Δ = 0.22-0.05 = 0.17

We would need approximately 147 subjects in total

n_1= 49 treated and n_2 = 2(49) = 98 control.

$$n_1 = \frac{\left[1.96\sqrt{0.16*0.84(1+1/2)} + 0.84\sqrt{0.05*0.95+0.22*0.78/2}\right]^2}{(0.17)^2}$$

$$= 49$$

$$n_2 = 2*n_1 = 2*49 = 98$$

We would approximately need **147** subjects

Cont'd

- But the population Values are most of the time unknown

As a result, they has to be estimated from:

- *Pilot or preliminary sample:*
  - Select a pilot sample and estimate the population values
- *Previous or similar studies*

Lesson 8: Hypotheses Testing

Lesson Outline

Introduction

Hypothesis testing: a single population mean

Hypothesis testing: the difference between two population means

Paired comparisons

Hypothesis testing: a single population proportion

Hypothesis testing: the difference between two population proportions

Key words

Null hypothesis $H_{0,}$ Alternative hypothesis H_A , testing hypothesis , test statistic , P-value

Definition of a hypothesis

- It is a statement about one or more populations .
- It is usually concerned with the parameters of the population. e.g. the hospital administrator may want to test the hypothesis that the average length of stay of patients admitted to the hospital is 5 days

Definition of Statistical hypotheses

- They are hypotheses that are stated in such a way that they may be evaluated by appropriate statistical techniques.
- There are two hypotheses involved in hypothesis testing
- **Null hypothesis H_0**: It is the hypothesis to be tested .
- **Alternative hypothesis H_A** : It is a statement of what we believe is true if our sample data cause us to reject the null hypothesis

Testing a hypothesis about the mean of a population

- We have the following steps:

1.Data: determine variable, sample size (n), sample mean($\bar{x}$) , population standard deviation or sample standard deviation (s) if is unknown

2. Assumptions : We have two cases:

- Case1: Population is normally or approximately normally distributed with known or unknown variance (sample size n may be small or large),
- Case 2: Population is not normal with known or unknown variance (n is large i.e. n≥30).

Cont'd

3.Hypotheses:

- we have three cases
- Case I : H_0: $\mu=\mu_0$

  H_A: $\mu \neq \mu_0$
- e.g. we want to test that the population mean is different than 50
- Case II : H_0: $\mu = \mu_0$

  H_A: $\mu > \mu_0$
- e.g. we want to test that the population mean is greater than 50
- Case III : $H_{0:}$ $\mu = \mu_0$

  H_A: $\mu < \mu_0$
- e.g. we want to test that the population mean is less than 50

Testing hypothesis for the mean μ

- **When the value of sample size (n):**

- **population is normal or not normal**
  (n ≥ 30)

population is normal
(n< 30)

σ is known

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

σ is not known

$$Z = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

σ is known

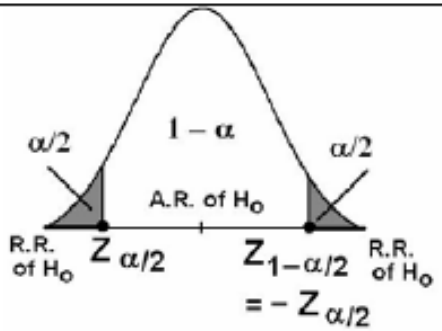
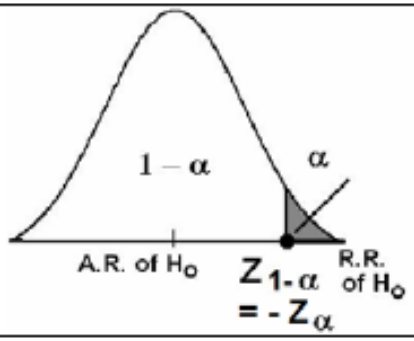
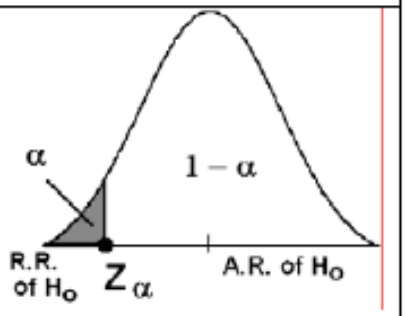
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

σ is not known

$$T = \frac{\bar{X} - \mu_0}{S / \sqrt{n}}$$

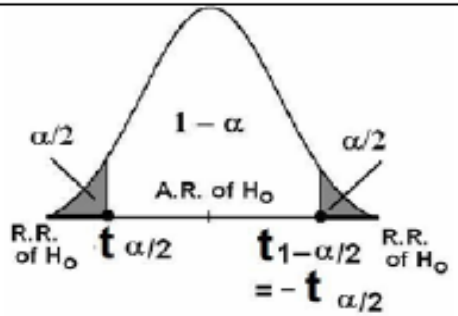
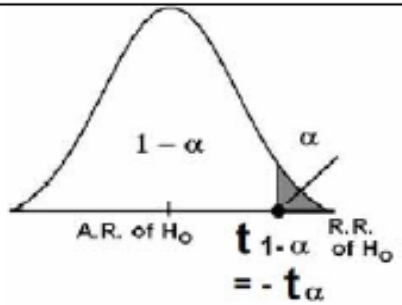
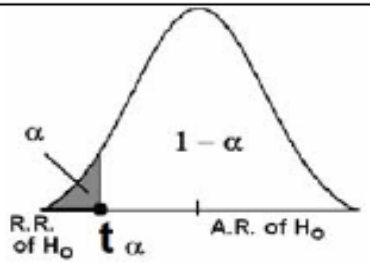
Cont'd

Test Procedures:

Hypotheses	$H_o: \mu = \mu_o$ $H_A: \mu \neq \mu_o$	$H_o: \mu \leq \mu_o$ $H_A: \mu > \mu_o$	$H_o: \mu \geq \mu_o$ $H_A: \mu < \mu_o$
Test Statistic (T.S.)	Calculate the value of: $Z = \dfrac{\bar{X} - \mu_o}{\sigma / \sqrt{n}} \sim N(0,1)$		
R.R. & A.R. of H_o			
Critical value (s)	$Z_{\alpha/2}$ and $-Z_{\alpha/2}$	$Z_{1-\alpha} = -Z_{\alpha}$	Z_{α}
Decision:	We reject H_o (and accept H_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_{\alpha}$ One-Sided Test	$Z < Z_{\alpha}$ One-Sided Test

Cont'd

Test Procedures:

Hypotheses	$H_o: \mu = \mu_o$ $H_A\ \mu \neq \mu_o$	$H_o: \mu \leq \mu_o$ $H_A: \mu > \mu_o$	$H_o: \mu \geq \mu_o$ $H_A: \mu < \mu_o$
Test Statistic (T.S.)	Calculate the value of: $t = \dfrac{\bar{X} - \mu_o}{S/\sqrt{n}} \sim t(n-1)$ $(df = \nu = n-1)$		
R.R. & A.R. of H_o			
Critical value (s)	$t_{\alpha/2}$ and $-t_{\alpha/2}$	$t_{1-\alpha} = -t_\alpha$	t_α
Decision:	We reject H_o (and accept H_A) at the significance level α if:		
	$t < t_{\alpha/2}$ or $t > t_{1-\alpha/2} = -t_{\alpha/2}$ Two-Sided Test	$t > t_{1-\alpha} = -t_\alpha$ One-Sided Test	$t < t_\alpha$ One-Sided Test

The Use of P – Values in Decision Definition Making

Note: Using P- Value as a decision tool:

P-value is the smallest value of α for which we can reject the null hypothesis H_o.

Calculating P-value:

* Calculating P-value depends on the alternative hypothesis H_A.
* Suppose that $z_c = \frac{\bar{X} - \mu_o}{\sigma / \sqrt{n}}$ is the computed value of the test Statistic.
* The following table illustrates how to compute P-value, and how to use P-value for testing the null hypothesis:

Cont'd

Alternative Hypothesis:	H_A: $\mu \neq \mu_o$	H_A: $\mu > \mu_o$	H_A: $\mu < \mu_o$
P-Value =	$2\times P(Z > \lvert z_c \rvert)$	$P(Z > z_c)$	$P(Z < z_c)$
Significance Level =	α		
Decision:	Reject H_o if P-value $< \alpha$.		

Example:

For the previous example, we have found that:

$$Z_C = \frac{\bar{X} - \mu_o}{\sigma / \sqrt{n}} = 2.02$$

The alternative hypothesis was H_A: $\mu > 70$.

$$P - Value = P(Z > Z_C)$$

$$= P(Z > 2.02) = 1 - P(Z < 2.02) = 1 - 0.9783 = 0.0217$$

The level of significance was $\alpha = 0.05$.

Since P-value $< \alpha$, we reject H_o.

6.Decision :

- If we reject H_0, we can conclude that H_A is true.
- If ,however ,we do not reject H_0, we may conclude that H_0 is true.

An Alternative Decision Rule using the p - value Definition

- The p-value is defined as the smallest value of α for which the null hypothesis can be rejected.
- If the p-value is less than or equal to α ,we reject the null hypothesis ($p \leq \alpha$)
- If the p-value is greater than α ,we do not reject the null hypothesis ($p > \alpha$)

Example

- Researchers are interested in the mean age of a certain population.
- A random sample of 10 individuals drawn from the population of interest has a mean of 27.
- Assuming that the population is approximately normally distributed with variance 20,can we conclude that the mean is different from 30 years ? (α=0.05) .
- If the p - value is 0.0340 how can we use it in making a decision?

Cont'd

Solution

1-Data: variable is age, n=10, =27 , σ^2=20,α=0.05

2-Assumptions: the population is approximately normally distributed with variance 20

3-Hypotheses:

- **H_0 : μ=30**
- **H_A: $\mu \neq 30$**

4-Test Statistic:

- **Z** = -2.12

5.Decision Rule

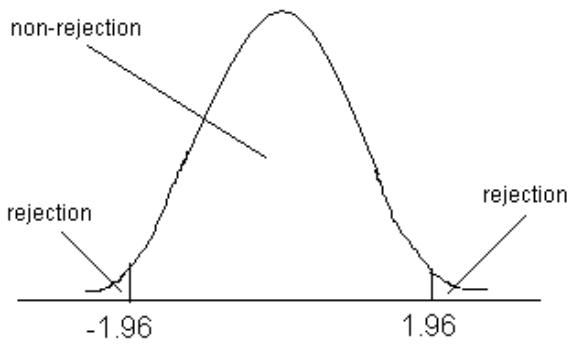
- **The alternative hypothesis is**

H_A: $\mu \neq 30$

- Hence we reject H_0 if $Z > Z_{1-0.025} = Z_{0.975}$

or $Z < -Z_{1-0.025} = -Z_{0.975}$

- $Z_{0.975}$=1.96



Cont'd

- **6.Decision:**
- **We reject H_0 ,since -2.12 is in the rejection region .**
- We can conclude that μ is not equal to 30
- Using the p value ,we note that p-value =0.0340< 0.05,therefore we reject H0

Example

- Among 157 African-American men ,the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. We wish to know if on the basis of these data, we may conclude that the mean systolic blood pressure for a population of African-American is greater than 140. Use $\alpha=0.01$.
- Solution

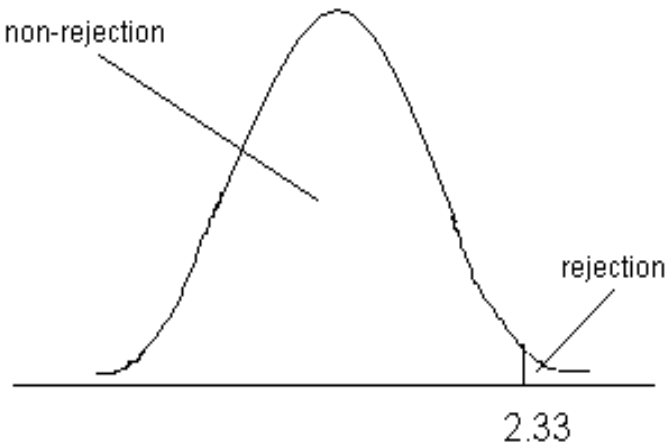
1. **Data:** Variable is systolic blood pressure, n=157 , =146, s=27, $\alpha=0.01$.
2. **Assumption:** population is not normal, σ^2 is unknown
3. **Hypotheses:** $H_0 : \mu=140$

   $H_A : \mu>140$

Cont'd

4.Test Statistic:

$$z = \frac{146 - 140}{27/\sqrt{157}} = \frac{6}{2.1548} = 2.78$$



5. Decision Rule:

we reject H_0 if $Z > Z_{1-\alpha}$

$= Z_{0.99} = 2.33$

6. Decision: We reject H_0.

Hence we may conclude that the mean systolic blood pressure for a population of African-American is greater than 140.

Hypothesis Testing :The Difference between two population mean

- We have the following steps:

1. **Data**: determine variable, sample size (n), sample means, population standard deviation or samples standard deviation (s) if is unknown for two population.

2. **Assumptions :** We have two cases:

- Case1: Population is normally or approximately normally distributed with known or unknown variance (sample size n may be small or large),
- Case 2: Population is not normal with known variances (n is large i.e. n≥30).

Cont'd

- **3.Hypotheses**:
- we have three cases
- Case I : H_0: $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

  H_A: $\mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$
- e.g. we want to test that the mean for first population is different from second population mean.
- Case II : H_0: $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

  H_A: $\mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$
- e.g. we want to test that the mean for first population is greater than second population mean.
- Case III : H_0: $\mu_1 = \mu_2 \rightarrow \mu_1 - \mu_2 = 0$

  H_A: $\mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0$
- e.g. we want to test that the mean for first population is greater than second population mean.

Cont'd

4.Test Statistic:

- **Case 1: Two population is normal or approximately normal**

σ^2 is known
(n_1 ,n_2 large or small)

σ^2 is unknown if
(n_1 ,n_2 small)

Population variances equal

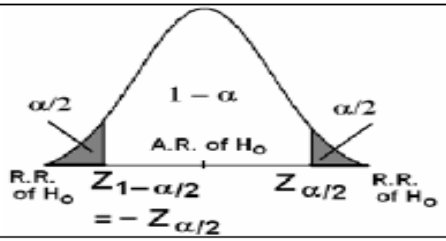
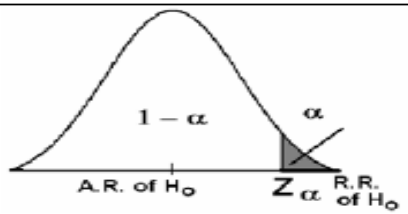
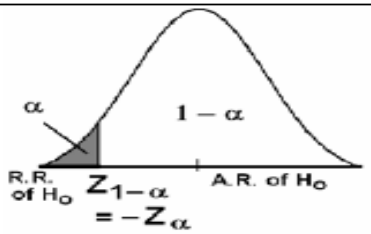
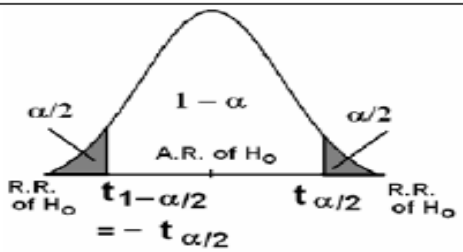
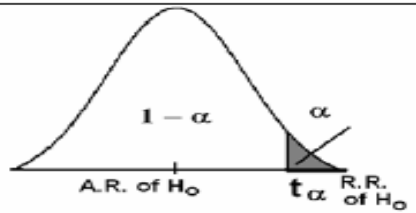
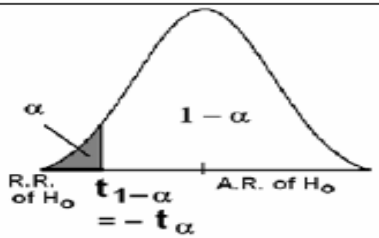
$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where $\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$

Cont'd

Summary of Testing Procedure:

Hypotheses	$H_o: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$	$H_o: \mu_1 - \mu_2 \leq 0$ $H_A: \mu_1 - \mu_2 > 0$	$H_o: \mu_1 - \mu_2 \geq 0$ $H_A: \mu_1 - \mu_2 < 0$
Test Statistic For the First Case:	$Z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$ {if σ_1^2 and σ_2^2 are known}		
R.R. and A.R. of H_o (For the First Case)			
Test Statistic For the Second Case:	$T = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}} \sim t(n_1+n_2-2)$ {if $\sigma_1^2=\sigma_2^2=\sigma^2$ is unknown}		
R.R. and A.R. of H_o (For the Second Case)			
Decision:	Reject H_o (and accept H_A) at the significance level α if:		
	T.S. $\in$ R.R. Two-Sided Test	T.S. $\in$ R.R. One-Sided Test	T.S. $\in$ R.R. One-Sided Test

Example

Researchers wish to know if the data have collected provide sufficient evidence to indicate a difference in mean serum uric acid levels between normal individuals and individual with Down's syndrome. The data consist of serum uric reading on 12 individuals with Down's syndrome from normal distribution with variance 1 and 15 normal individuals from normal distribution with variance 1.5 . The mean are and α=0.05.

Solution: $\bar{X}_1 = 4.5mg/100$ $\bar{X}_2 = 3.4mg/100$

1. Data: Variable is serum uric acid levels, $n_1=12$, $n_2=15$, $\sigma^2_1=1$, $\sigma^2_2=1.5$,$\alpha=0.05$.

Cont'd

2. Assumption: Two population are normal, σ^2_1 , σ^2_2 are known

3. Hypotheses:

H_0: $\mu 1 = \mu 2 \rightarrow \mu_1 - \mu_2 = 0$

H_A: $\mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$

4.Test Statistic:

$$z = \frac{(4.5 - 3.4) - 0}{\sqrt{1/12 + 1.5/15}} = \frac{1.1}{.4282} = 2.57$$

Cont'd

5. Decision Rule:

Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$, $Z_{1-\alpha/2} = Z_{1-0.05/2} = Z_{0.975} = 1.96$

6-**Conclusion:** Reject H_0 since 2.57 > 1.96

Or if p-value =0.102→ reject H_0 if $p < \alpha$ → then reject H_0

Example: The purpose of a study by Tam, was to investigate wheelchair Maneuvering in individuals with over-level spinal cord injury (SCI) And healthy control (C). Subjects used a modified a wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. The data for measurements of the left ischial tuerosity for SCI and control C are shown below

TABLE 7.3.1 Pressures (mm Hg) Under the Pelvis during Static Conditions for Example 7.3.2

Control	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

Cont'd

We wish to know if we can conclude, on the basis of the above data that the mean of left ischial tuberosity for control C lower than mean of left ischial tuerosity for SCI, Assume normal populations equal variances. α=0.05, p-value = -1.33

Solution:

1. Data:, $n_C=10$, $n_{SCI}=10$, $S_C=21.8$, $S_{SCI}=133.1$,$\alpha=0.05$.

- $\bar{X}_C = 126.1$ $\bar{X}_{SCI} = 133.1$
- (calculated from data)

2. Assumption: Two population are normal, σ^2_1 , σ^2_2 are unknown but equal

3. Hypotheses:

H_0: $\mu_C = \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} = 0$

H_A: $\mu_C < \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} < 0$

Cont'd

4.Test Statistic:

$$s_p^2 = \frac{9(21.8)^2 + 9(32.2)^2}{9+9} = 756.04$$

$$t = \frac{(126.1 - 133.1) - 0}{\sqrt{\frac{756.04}{10} + \frac{756.04}{10}}} = -.569$$

5. Decision Rule:

Reject H_0 if $T < -T_{1-\alpha,(n_1+n_2-2)}$

$T_{1-\alpha,(n_1+n_2-2)} = T_{0.95,18} = 1.7341$
(from table E)

6-**Conclusion:** Fail to reject H_0
since $-0.569 < -1.7341$
Or Fail to reject H_0 since $p = -1.33 > \alpha = 0.05$

Cont'd

7.4 Paired Comparisons:

- In this section, we are interested in comparing the means of two related (non-independent/dependent) normal populations.
- In other words, we wish to make statistical inference for the difference between the means of two related normal populations.
- Paired t-Test concerns about testing the equality of the means of two related normal populations.

Examples of related populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Example: (effectiveness of a diet program)
Suppose that we are interested in studying the effectiveness of a certain diet program. Let the random variables X and Y are as follows:

Cont'd

X = the weight of the individual before the diet program

Y= the weight of the same individual after the diet program

We assume that the distributions of these random variables are normal with means μ_1 and μ_2, respectively.

These two variables are related (dependent/non-independent) because they are measured on the same individual.

Populations:

1-st population (X): weights before a diet program

mean = μ_1

2-nd population (Y): weights after the diet program

mean = μ_2

Cont'd

where:

$$\mu_D = \mu_1 - \mu_2$$

- We calculate the following quantities:
  - The differences (D-observations):

$$D_i = X_i - Y_i \quad (i=1, 2, \ldots, n)$$

  - Sample mean of the D-observations (differences):

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \ldots + D_n}{n}$$

  - Sample variance of the D-observations (differences):

$$S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \overline{D})^2}{n-1} = \frac{(D_1 - \overline{D})^2 + (D_2 - \overline{D})^2 + \cdots + (D_n - \overline{D})^2}{n-1}$$

  - Sample standard deviation of the D-observations:

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

Cont'd

Confidence Interval for the Difference between the Means of Two Related Normal Populations ($\mu_D = \mu_1 - \mu_2$):

In this section, we consider constructing a confidence interval for the difference between the means of two related (non-independent) normal populations. As before, let us define the difference between the two means as follows:

$$\mu_D = \mu_1 - \mu_2$$

where μ_1 is the mean of the first population and μ_2 is the mean of the second population. We assume that the two normal populations are not independent.

Result:

A $(1-\alpha)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is:

Example

- John M. Morton et al., examined gallbladder function before and after fundoplication—a surgery used to stop stomach contents from flowing back into the esophagus (reflux)—in patients with gastroesophageal reflux disease. The authors measured gallbladder functionality by calculating the gallbladder ejection fraction (GBEF) before and after fundoplication. The goal of fundoplication is to increase GBEF, which is measured as a percent. The data are shown in Table 7.4.1. We wish to know if these data provide sufficient evidence to allow us to conclude that fundoplication increases GBEF functioning.

TABLE 7.4.1 Gallbladder Function in Patients with Presentations of Gastroesophageal Reflux Disease Before and After Treatment

Preop (%)	22	63.3	96	9.2	3.1	50	33	69	64	18.8	0	34
Postop (%)	63.5	91.5	59	37.8	10.1	19.6	41	87.8	86	55	88	40

Cont'd

- **Solution**: We will say that sufficient evidence is provided for us to conclude that the fundoplication is effective if we can reject the null hypothesis that the population mean change md is different from zero in the appropriate direction. We may reach a conclusion by means of the ten-step hypothesis testing procedure.

1. **Data.** The data consist of the GBEF for 12 individuals, before and after fundoplication. We shall perform the statistical analysis on the differences in preop and postop GBEF. We may obtain the differences in one of two ways: by subtracting the preop percents from the postop percents or by subtracting the postop percents from the preop percents. Let us obtain the differences by subtracting the preop percents from the postop percents. The di= postop -preop differences are: 41.5, 28.2, -37.0, 28.6, 7.0, -30.4, 8.0, 18.8, 22.0, 36.2, 88.0, 6.0.

Cont'd

2. Assumptions. The observed differences constitute a simple random sample from a normally distributed population of differences that could be generated under the same circumstances.

3. **Hypotheses.** The way we state our null and alternative hypotheses must be consistent with the way in which we subtract measurements to obtain the differences. In the present example, we want to know if we can conclude that the fundoplication is useful in increasing GBEF percentage. If it is effective in improving GBEF, we would expect the postop percents to tend to be higher than the preop percents. If, therefore, we subtract the preop percents from the postop percents (postop – preop), we would expect the differences to tend to be positive. Furthermore, we would expect the mean of a population of such differences to be positive. So, under these conditions, asking if we can conclude that the fundoplication is effective is the same as asking if we can conclude that the population mean difference is positive (greater than zero).

Cont'd

The null and alternative hypotheses are as follows:

$$H_0: \mu_d \leq 0$$
$$H_A: \mu_d > 0$$

If we had obtained the differences by subtracting the postop percents from the preop weights (preop − postop), our hypotheses would have been

$$H_0: \mu_d \geq 0$$
$$H_A: \mu_d < 0$$

If the question had been such that a two-sided test was indicated, the hypotheses would have been

$$H_0: \mu_d = 0$$
$$H_A: \mu_d \neq 0$$

regardless of the way we subtracted to obtain the differences.

4. **Test statistic.** The appropriate test statistic is given by Equation 7.4.1.
5. **Distribution of test statistic.** If the null hypothesis is true, the test statistic is distributed as Student's t with $n - 1$ degrees of freedom.
6. **Decision rule.** Let $\alpha = .05$. The critical value of t is 1.7959. Reject H_0 if computed t is greater than or equal to the critical value. The rejection and nonrejection regions are shown in Figure 7.4.1.

Cont'd

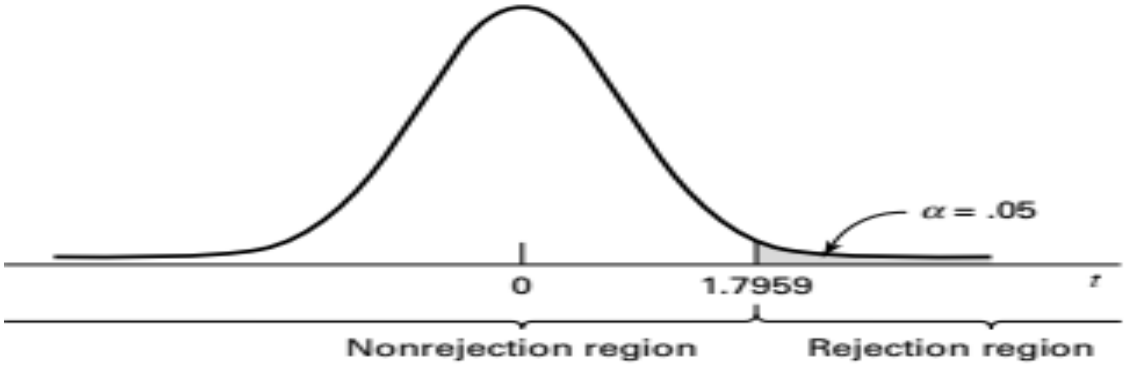


FIGURE 7.4.1 Rejection and nonrejection regions for Example 7.4.1.

7. **Calculation of test statistic.** From the $n = 12$ differences d_i, we compute the following descriptive measures:

$$\bar{d} = \frac{\sum d_i}{n} = \frac{(41.5) + (28.2) + (-37.0) + \cdots + (6.0)}{12} = \frac{216.9}{12} = 18.075$$

$$s_d^2 = \frac{\sum (d_i - \bar{d})^2}{n-1} = \frac{n\sum d_i^2 - (\sum d_i)^2}{n(n-1)} = \frac{12(15669.49) - (216.9)^2}{(12)(11)} = 1068.0930$$

$$t = \frac{18.075 - 0}{\sqrt{1068.0930/12}} = \frac{18.075}{9.4344} = 1.9159$$

8. **Statistical decision.** Reject H_0, since 1.9159 is in the rejection region.

Hypothesis Testing: A Single Population Proportion

- Testing hypotheses about population proportions is carried out in much the same way as for means when the conditions necessary for using the normal curve are met.

- One-sided or two-sided tests may be made, depending on the question being asked.

- When a sample sufficiently large for application of the central limit theorem as discussed in Section 5.5 is available for analysis, the test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}$$

which, when H_0 is true, is distributed approximately as the standard normal.

Example

- Wagenknecht et al., collected data on a sample of 301 Hispanic women living in San Antonio, Texas. One variable of interest was the percentage of subjects with impaired fasting glucose (IFG). IFG refers to a metabolic stage intermediate between normal glucose homeostasis and diabetes. In the study, 24 women were classified in the IFG stage. The article cites population estimates for IFG among Hispanic women in Texas as 6.3 percent. Is there sufficient evidence to indicate that the population of Hispanic women in San Antonio has a prevalence of IFG higher than 6.3 percent?

cont'd

Solution:

1. **Data.** The data are obtained from the responses of 301 individuals of which 24 possessed the characteristic of interest; that is, $\hat{p} = 24/301 = .080$.
2. **Assumptions.** The study subjects may be treated as a simple random sample from a population of similar subjects, and the sampling distribution of $\hat{p}$ is approximately normally distributed in accordance with the central limit theorem.
3. **Hypotheses.**

$$H_0: p \leq .063$$
$$H_A: p > .063$$

We conduct the test at the point of equality. The conclusion we reach will be the same as we would reach if we conducted the test using any other hypothesized value of p greater than .063. If H_0 is true, $p = .063$ and the standard error $\sigma_{\hat{p}} = \sqrt{(.063)(.937)/301}$. Note that we use the hypothesized value of p in computing $\sigma_{\hat{p}}$. We do this because the entire test is based on the assumption that the null hypothesis is true. To use the sample proportion, $\hat{p}$, in computing $\sigma_{\hat{p}}$ would not be consistent with this concept.

Cont'd

4. **Test statistic.** The test statistic is given by Equation 7.5.1.
5. **Distribution of test statistic.** If the null hypothesis is true, the test statistic is approximately normally distributed with a mean of zero.
6. **Decision rule.** Let $\alpha = .05$. The critical value of z is 1.645. Reject H_0 if the computed z is ≥ 1.645.
7. **Calculation of test statistic.**

$$z = \frac{.080 - .063}{\sqrt{\frac{(.063)(.937)}{301}}} = 1.21$$

8. **Statistical decision.** Do not reject H_0 since $1.21 < 1.645$.
9. **Conclusion.** We cannot conclude that in the sampled population the proportion who are IFG is higher than 6.3 percent.
10. ***p* value.** $p = .1131$.