(6/45)*30=(6*30)/45=((12/45)*30)/2=((12*30)/45)/2 ?

BIOSTATISTICS Module (BSTA 2422)

Mr. Mutayomba Sylvestre
Catholic University of Rwanda (CUR)

OUTLINE

1. **Introduction to Statistics**
2. **Summarizing data**
3. **Elementary Probability and probability distribution**
4. **Sampling methods**
5. **Estimation**
6. **Hypothesis Testing**
7. **Correlation and Regression**
8. **Demographic Methods and Health Services Statistics.**

I. Introduction to Statistics

After completing this chapter, the student will be able to:

i. Define Statistics and Biostatistics

ii. Enumerate the **importance** and limitations of statistics

iii. Define and identify the different types of data and understand **why we need to classifying variables**.

Limitations of statistics:

i. Statistics deals with **only** those subjects of inquiry that are capable of being quantitatively measured and numerically expressed.

ii. It deals on aggregates of facts and **no importance is attached to individual items**–suited only if their group characteristics are desired to be studied.

iii. Statistical data are **only approximately** and not mathematically **correct**.

- This course is about **information**—**how** it is **obtained**, **how** it is **analyzed**, and **how** it is **interpreted**.
- The information about which we are concerned is called **data**, and the data are available to us **in the form of numbers**.
- The principle objectives of this course are twofold:

(1) to learn how **to organize and summarize data**, and

(2) to learn **how to reach decisions** about a large body of data by **examining only a small part of the data**.

- Like all fields of learning, statistics has **its own vocabulary**. Some of the words and phrases encountered in the study of statistics will be new to those not previously exposed to the subject.
- Other terms, **though appearing to be familiar**, may have **specialized meanings** that are different from the meanings that we are accustomed to associating with these terms.

- **The tools of statistics are employed in many fields:**
  business, education, psychology, agriculture, economics, … etc.
- **When the data analyzed are derived from the biological science and medicine, we use the term biostatistics to distinguish this particular application of statistical tools and concepts.**

Data:
- **The raw material of Statistics is data.**
- **We may define data as figures/numbers.**
- **Figures result from the process of counting or from taking a measurement.**

For example:

- **When a hospital administrator counts the number of patients (counting).**
- **When a nurse weighs a patient (measurement)**

- Statistics: a field of study concerned with

(1) **The collection, organization**, **summarization**, and **analysis** of data; and

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

- Simply put, we may say that data are **numbers**, **numbers contain information**, and the purpose of statistics is to **investigate** and evaluate the nature and meaning of this information.

- The performance of statistical activities is motivated by **the need to answer a question**.

- When we determine that **the appropriate approach to seeking an answer to a question will require the use of statistics**, we begin to search for **suitable data to serve as the raw material for our investigation.**

CHARACTERISTICS OF STATISTICAL DATA

In order that numerical descriptions may be called statistics ***they must possess the following characteristics****:*

(i)They must **be in aggregates** – This means that statistics are **'number of facts**.' A single fact, even though numerically stated, cannot be called statistics.

(ii)They must be **affected to a marked extent by a multiplicity of causes**.

(iii) They must be numerically expressed

(iv) **They must be enumerated or estimated accurately**

(v) They must have been **collected in a systematic manner**

(vi) for a predetermined purpose.

(vii) They must be **placed in relation to each other**. That is, they must be comparable.

https://hemantmore.org.in/management/statistics-management/introduction-statistics/3913/ 18/7/2018

Sources of Data:

1- Routinely kept records.

For example:

- **Hospital medical records contain immense amounts of information on patients.**
- **Hospital accounting records contain a wealth of data on the facility's business activities.**
- **When the need for data arises, we should look for them first among routinely kept records.**

2- External sources.

The data needed to answer a question may already exist in the form of

published reports, commercially available data banks, or the research literature, i.e. someone else has already asked the same question and the answer obtained may be applicable to our present situation.

3- Surveys:

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

For example:

If the administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic, then a survey may be conducted among patients to obtain this information.

4- Experiments.

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

For example:

If a nurse wishes to know which of several strategies is best for maximizing patient compliance,

she might conduct an experiment in which the different strategies of motivating compliance are tried with different patients.

A VARIABLE:

It is a characteristic that takes on different values in different persons, places, or things i.e. the characteristic is not the same when observed in different possessors of it.

For example: - 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

Quantitative Variables

It can be measured in the usual sense.

For example:

- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a dental clinic.

Qualitative Variables

Many characteristics are **not capable of being measured**. Some of them can be ordered or ranked.

For example:

- classification of people into socio-economic groups,
- social classes based on income, education, etc.

Measurements made on quantitative variables convey information regarding amount.

Measurements made on qualitative variables convey information regarding attribute.

Types of quantitative variables

A discrete variable

Is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

Can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Because of the **limitations of available measuring instruments**, however, observations on variables that are inherently continuous are recorded as if they were discrete.

A POPULATION:

- We define a population of entities as the largest collection of entities for which we have an interest at a particular time.
- If we take a measurement of some variable on each of the entities in a population, we generate a population of values of that variable.
- We may, therefore, define a population of values as the largest collection of values of a random variable for which we have an interest at a particular time.
- Populations are determined or defined by our sphere of interest.

For example:

The weights of all the children enrolled in a certain elementary school.

Populations may be finite or infinite.

A SAMPLE:

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

DESCRIPTIVE STATISTICS

- Data generally consist of an extensive number of measurements or observations that are **too numerous or complicated to be understood through simple observation**.
- There are a number of **ways to condense and organize information into a set of descriptive measures and visual devices that enhance the understanding of complex data**.
- Measurements that have not been organized, summarized, or otherwise **manipulated** are called **raw data**.
- There are **several techniques for organizing and summarizing data** so that we may more easily determine what information they contain.
- The ultimate in summarization of data is the **calculation of a single number that in some way conveys important information about the data from which it was calculated**.

THE **ORDERED ARRAY**

- An ordered array is a **listing of the values of a collection** (either population or sample) **in order of magnitude from the smallest value to the largest value**.
- An ordered array **enables one to determine quickly the value of the smallest measurement**, **the value of the largest measurement**, and **other facts about the arrayed data that might be needed in a hurry**
- This unordered table (see next slide) requires **considerable searching for us to ascertain such elementary information as the age of the youngest and oldest subjects**.
- By referring to the ordered array (see slide 18) we are able to determine quickly the **age of the youngest subject** and **the age of the oldest subject**. We also readily note that about one-third of the subjects are 50 years of age or younger.

UNORDERED ARRAY

TABLE 1.4.1 Ages of 189 Subjects Who Participated in a Study on Smoking Cessation

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
1	48	49	38	97	51	145	52
2	35	50	44	98	50	146	53
3	46	51	43	99	50	147	61
4	44	52	47	100	55	148	60
5	43	53	46	101	63	149	53
6	42	54	57	102	50	150	53
7	39	55	52	103	59	151	50
8	44	56	54	104	54	152	53
9	49	57	56	105	60	153	54
10	49	58	53	106	50	154	61
11	44	59	64	107	56	155	61
12	39	60	53	108	68	156	61
13	38	61	58	109	66	157	64
14	49	62	54	110	71	158	53
15	49	63	59	111	82	159	53
16	53	64	56	112	68	160	54
17	56	65	62	113	78	161	61
18	57	66	50	114	66	162	60
19	51	67	64	115	70	163	51
20	61	68	53	116	66	164	50
21	53	69	61	117	78	165	53
22	66	70	53	118	69	166	64
23	71	71	62	119	71	167	64
24	75	72	57	120	69	168	53

(Continued)

1 3 5 7

Subject No.	Age	Subject No.	Age	Subject No.	Age	Subject No.	Age
25	72	73	52	121	78	169	60
26	65	74	54	122	66	170	54
27	67	75	61	123	68	171	55
28	38	76	59	124	71	172	58
29	37	77	57	125	69	173	62
30	46	78	52	126	77	174	62
31	44	79	54	127	76	175	54
32	44	80	53	128	71	176	53
33	48	81	62	129	43	177	61
34	49	82	52	130	47	178	54
35	30	83	62	131	48	179	51
36	45	84	57	132	37	180	62
37	47	85	59	133	40	181	57
38	45	86	59	134	42	182	50
39	48	87	56	135	38	183	64
40	47	88	57	136	49	184	63
41	47	89	53	137	43	185	65
42	44	90	59	138	46	186	71
43	48	91	61	139	34	187	71
44	43	92	55	140	46	188	73
45	45	93	61	141	46	189	66
46	40	94	56	142	48		
47	48	95	52	143	47		
48	49	96	54	144	43		

2 4 6 8

TABLE 2.2.1 Ordered Array of Ages of Subjects from Table 1.4.1

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						

GROUPED DATA: THE FREQUENCY DISTRIBUTION

- **Further useful** summarization may be achieved by grouping the data.
- One must bear in mind that **data contain information and that summarization is a way of making it easier to determine the nature of this information**.
- To group a set of observations, we select **a set of contiguous, non overlapping intervals such that each value in the set of observations can be PLACED IN ONE, AND ONLY ONE, OF THE INTERVALS**. These intervals are usually referred to as ***class intervals***.
- A commonly followed rule of thumb states that there should be **no fewer than five intervals and no more than 15**.
- If there are fewer than five intervals, the data have been **summarized too much** and the information they contain has been lost. If there are more than 15 intervals, the data have **not been summarized enough**.

Frequency Distribution of Ages of 189 Subjects Shown in Tables 1.4.1 and 2.2.1

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

SOME DEFINITIONS

- **Frequency**: The **number of times** a particular value occurs in the set of values.
- **Cumulative Frequency**: Cumulative frequency of **a particular** value in a table can be defined as **the sum of all the frequencies up to that value** (including the value itself).
- **Relative Frequency**: The ratio of the number of times a value of the data occurs in the set of all outcomes to the number of all outcomes
- **Cumulative Relative Frequency**: is **the sum of the relative frequencies for all values that are less than or equal to the given value**.
- **Range:** Range is the **difference** between the highest and the lowest values in a set of data.

Frequency Distribution for Discrete Random Variables

Example:

Suppose that we take a sample of size 16 from children in a primary school and get the following data about the number of their decayed teeth,

3,5,2,4,0,1,3,5,2,3,2,3,3,2,4,1

To construct a frequency table:

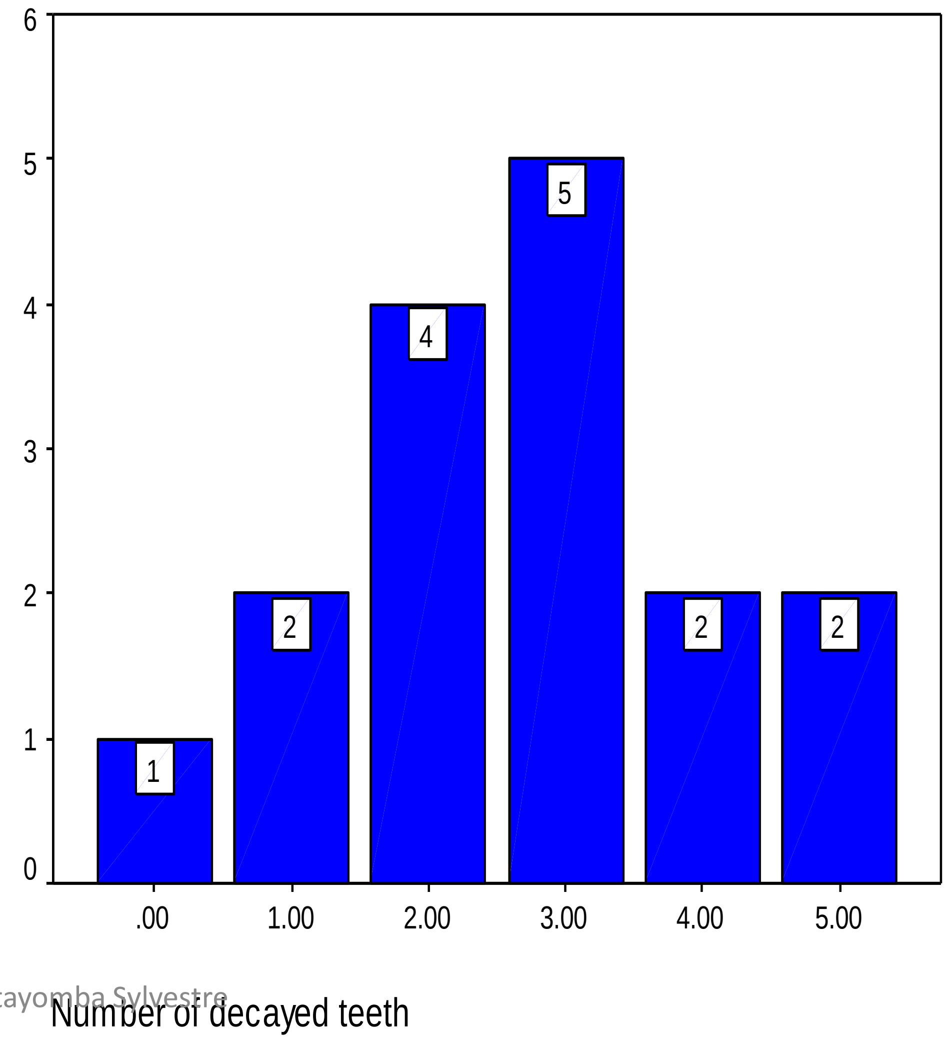
1- Order the values from the smallest to the largest.

0,1,1,2,2,2,2,3,3,3,3,3,4,4,5,5

2- Count how many numbers are the same.

Representing the simple frequency table using the bar chart

We can represent the above simple frequency table using the bar chart.



FREQUENCY DISTRIBUTION FOR CONTINUOUS RANDOM VARIABLES

For large samples, we can't use the simple frequency table to represent the data.

We need to divide the data into groups or intervals or classes.

So, we need to determine:

1- The number of intervals (k).

Too few intervals are not good because information will be lost.

Too many intervals are not helpful to summarize the data.

A commonly followed rule is that $5 \leq k \leq 15$,

or the following formula may be used,

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

2- The range (R).

It is the difference between the largest and the smallest observation in the data set.

3- The Width of the interval (w).

Class intervals generally should be of the same width. Thus, if we want k intervals, then w is chosen such that $w \geq R / k$.

Example:

Assume that the number of observations equal 100, then

$k = 1+3.322(\log 100)$

$= 1 + 3.3222 (2) = 7.6 \cong 8.$

Assume that the smallest value = 5 and the largest one of the data = 61, then

$R = 61 – 5 = 56$ and

$w = 56 / 8 = 7.$

To make the summarization more comprehensible, the class width may be 5 or 10 or the multiples of 10.

EXAMPLE

- We wish to know how many class interval to have in the frequency distribution of the data where the number of observation is 189, the largest value 82 and the smallest value 30 (the case of 189 subjects who Participated in a study on smoking cessation)

Solution :

- **Since the number of observations equal 189, then**
- ✓ **k = 1+3.322(log 189)**
  = 1 + 3.3222 (2.276)=8.6 ≅ 9,
- **R = 82 – 30 = 52 and**
- **w = 52 / 9 = 5.778**

- **It is better to let w = 10, then make a table of intervals along with their frequencies.**

w ≥ R / k

The Cumulative Frequency:

It can be computed by adding successive frequencies.

The Cumulative Relative Frequency:

It can be computed by adding successive relative frequencies.

The Mid-interval:

It can be computed by adding the lower bound of the interval plus the upper bound of it and then divide over 2.

For the above example, the following table represents the cumulative frequency, the relative frequency, the cumulative relative frequency and the mid-interval.

R.f= freq/n

Class interval	Mid – interval	Frequency Freq (f)	Cumulative Frequency	Relative Frequency R.f	Cumulative Relative Frequency
30 – 39	34.5	**11**	11	0.0582	0.0582
40 – 49	44.5	**46**	57	0.2434	
50 – 59	54.5		127		0.6720
60 – 69		**45**		0.2381	0.9101
70 – 79	74.5	**16**	188	0.0847	0.9948
80 – 89	84.5	**1**	189	0.0053	1
Total		189		1	

Example :

- From the above frequency table, complete the table then answer the following questions:
- 1-The number of objects with age **less than 50 years** ?
- 2-The number of objects with age **between 40-69 years** ?
- 3-Relative frequency of objects with age **between 70-79 years** ?
- 4-Relative frequency of objects with age more than 69 years ?
- 5-The percentage of objects with age between 40-49 years ?
- 6- The percentage of objects with age less than 60 years ?
- 7-The Range (R) ?
- 8- Number of intervals (K)?
- 9- The width of the interval (W) ?

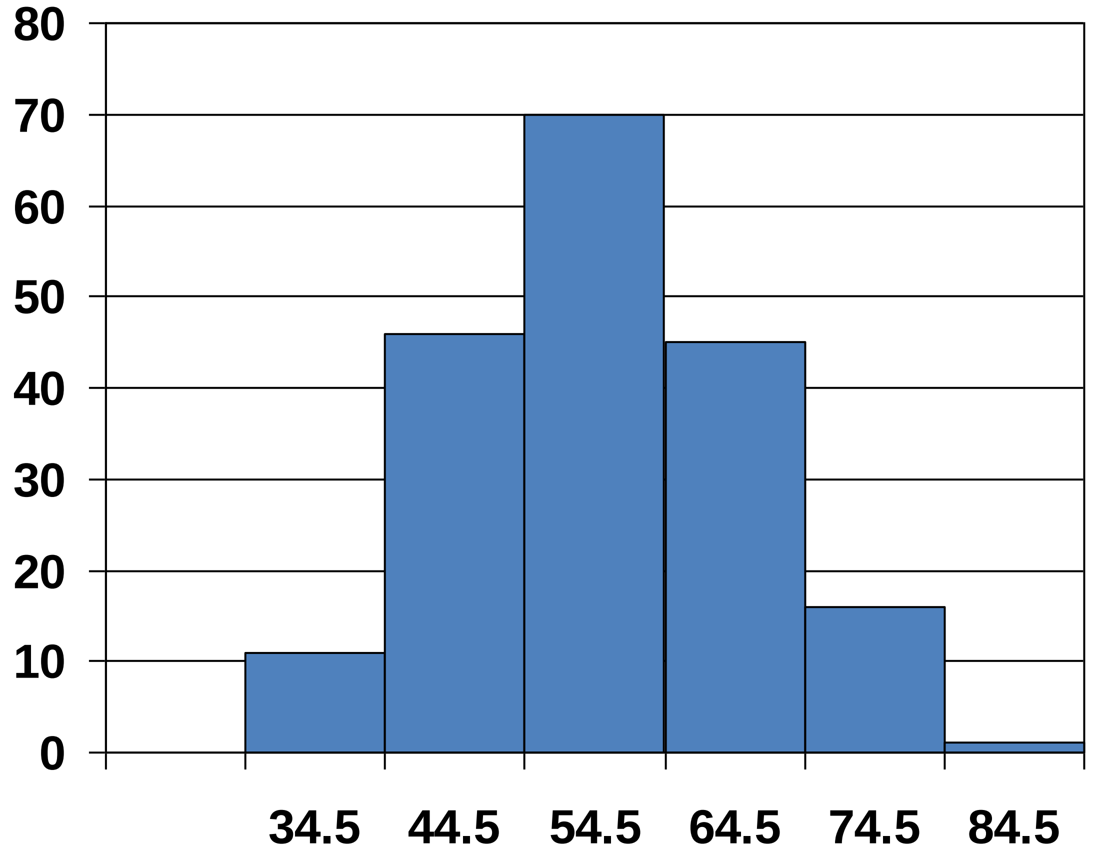
The Histogram

- Histogram is **a bar graph** which shows **the frequencies of data in a certain interval**.
- When we construct a histogram **the values of the variable** under consideration are represented by the **horizontal axis**, while the **vertical axis** has as its **scale the frequency** (or relative frequency if desired) of occurrence.
- Above each class interval on the horizontal axis a rectangular bar, or **cell**, as it is sometimes called, is erected so that **the height** corresponds to the respective **frequency** **when** the **class intervals** are of equal **width**.
- The **cells of a histogram must be joined** and, to accomplish this, we must take into account the **true boundaries** of the class intervals **to prevent gaps** from occurring between the cells of our graph.
- The class interval limits usually reflect **the degree of precision** of the raw data.

- **Some** of the values falling in the second class interval (See slide 20 and 32), for example, **when measured precisely**, would probably be **a little less** than 40 and **some** would be a **little greater than** 49.
- Considering the underlying continuity of our variable, and assuming that the data were rounded to the nearest whole number, we find it **convenient to think of 39.5 and 49.5 as the true limits** of this second interval.
- Each cell contains a certain **proportion of the total area**, depending on the frequency. The second cell, for example, contains 46/189 of the area. **This is the relative frequency of occurrence of values between 39.5 and 49.5**.
- From this we see that "***subareas of the histogram defined by the cells correspond to the frequencies of occurrence of values between the horizontal scale boundaries of the areas***".
- The **ratio of a particular subarea to the total area of the histogram is equal to *the relative frequency of occurrence* of values between the corresponding points on the horizontal axis**.

REPRESENTING THE GROUPED FREQUENCY TABLE USING THE HISTOGRAM

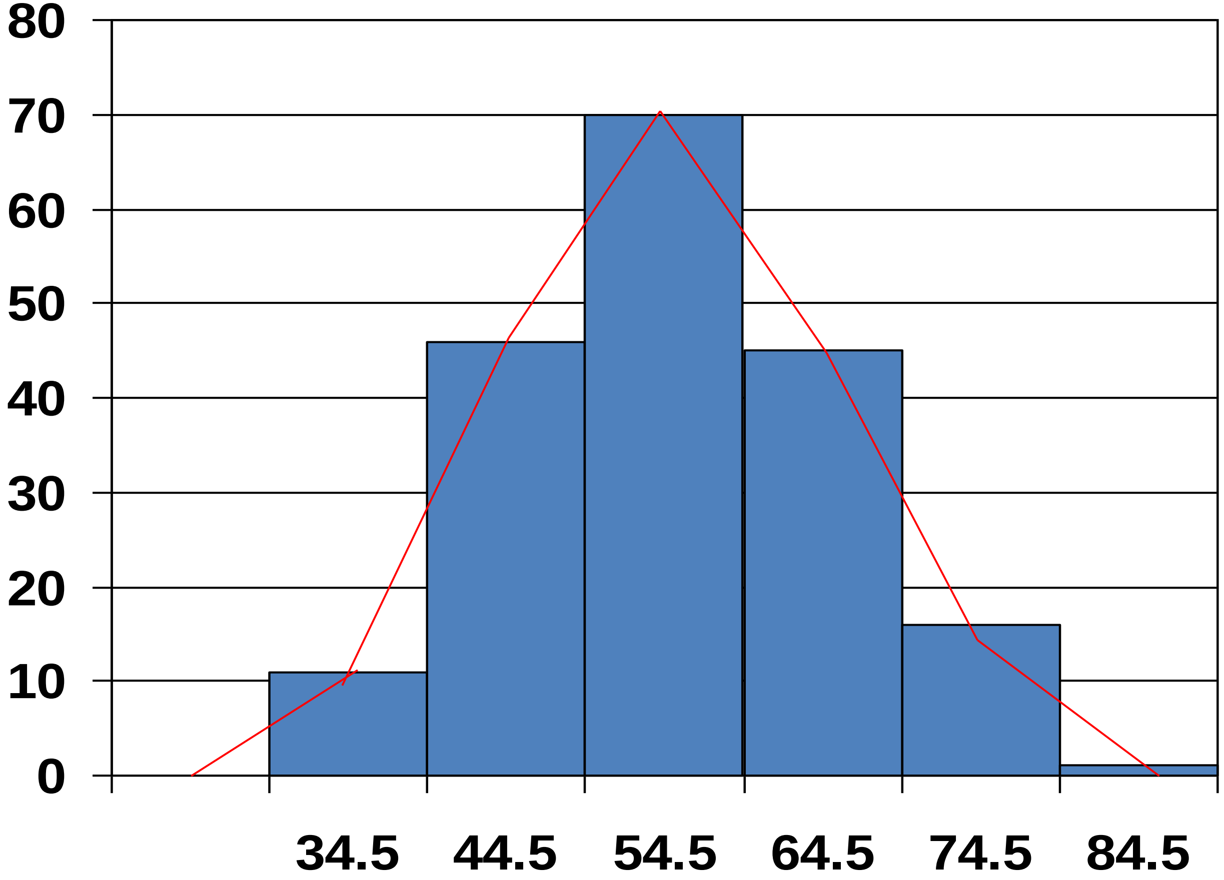
True class limits	Frequency
29.5 – <39.5	**11**
39.5 – < 49.5	**46**
49.5 – < 59.5	**70**
59.5 – < 69.5	**45**
69.5 – < 79.5	**16**
79.5 – < 89.5	**1**
Total	189



THE FREQUENCY POLYGON

- A frequency distribution can be portrayed graphically in yet another way by means of a frequency polygon, which 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**.
- The polygon is **brought down to the horizontal axis at the ends at points that would be the midpoints if there were an additional cell at each end of the corresponding histogram**. This allows for the total area to be enclosed.

REPRESENTING THE GROUPED FREQUENCY TABLE USING THE POLYGON



Descriptive Statistics Measures of Central Tendency

Location parameter

STATISTIC, PARAMETER, MEAN (M) ,MEDIAN, MODE.

The Statistic and The Parameter

A Statistic:

It is a descriptive measure computed from the data of a sample.

A Parameter:

It is a a descriptive measure computed from the data of a population.

Since it is difficult to measure a parameter from the population, a sample is drawn of size n, whose values are $\chi_1, \chi_2, \ldots, \chi_n$. From this data, we measure the statistic.

Measures of Central Tendency

A measure of central tendency is a measure which indicates where the middle of the data is.

The three most commonly used measures of central tendency are:

The Mean, the Median, and the Mode.

The Mean:

It is the average of the data.

The Population Mean:

$$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$$ which is usually unknown, then we use the sample mean to estimate or approximate it.

The Sample Mean:

Example:

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

Here is a random sample of size 10 of ages, where

$\chi_1 = 42$, $\chi_2 = 28$, $\chi_3 = 28$, $\chi_4 = 61$, $\chi_5 = 31$,
$\chi_6 = 23$, $\chi_7 = 50$, $\chi_8 = 34$, $\chi_9 = 32$, $\chi_{10} = 37$.

$\bar{\chi}$ **=?**

arithmetic mean, weighted mean, geometric mean (GM) and harmonic mean (HM) are different types of mean. If mentioned without an adjective (as mean), it generally refers to the arithmetic mean.

Properties of the Mean:

- **Uniqueness. For a given set of data there is one and only one mean.**
- **Simplicity. It is easy to understand and to compute.**
- The sum of the deviations from the mean is 0
- **Affected by extreme values. Since all values enter into the computation.**

Example: Assume the values are 115, 110, 119, 117, 121 and 126. The mean = 118.

But assume that the values are 75, 75, 80, 80 and 280. The mean = 118, a value that is not representative of the set of data as a whole. The single atypical value had the effect of inflating the mean.

The Median:

- **When ordering the data, it is the observation that divide the set of observations into two equal parts such that half of the data are before it and the other are after it.**
- **If n is odd, the median will be the middle of observations. It will be the $(n+1)/2$ th ordered observation.**

When n = 11, then the median is the 6^{th} observation.

- **If n is even, there are two middle observations. The median will be *the mean of these two middle observations*. It will be the $(n+1)/2$ th ordered observation.**

When n = 12, then the median is the 6.5^{th} observation, which is an observation halfway between the 6^{th} and 7^{th} ordered observation.

Example:

For the same random sample, the ordered observations will be as:

23, 28, 28, 31, 32, 34, 37, 42, 50, 61.

Since n = 10, then the median is the 5.5th observation, i.e. = (32+34)/2 = 33.

Properties of the Median:

- **Uniqueness. For a given set of data there is one and only one median.**
- **Simplicity. It is easy to calculate.**
- **It is not affected by extreme values as is the mean.**

The Mode:

It is the value which occurs most frequently.

If all values are different there is no mode.

Sometimes, there are more than one mode.

Example:

For the same random sample (on the previous slide), the value 28 is repeated two times, so it is the mode.

Properties of the Mode:

- **Sometimes, it is not unique.**
- **It may be used for describing qualitative data.**

An attractive property of a data distribution occurs when the **mean, median, and mode are all equal**. The well-known "bell-shaped curve" is a graphical representation of a distribution for which the mean, median, and mode are all equal. Much statistical inference is based on this distribution

Measures of Dispersion

Range ,variance, Standard deviation, coefficient of variation (C.V)

- A measure of dispersion conveys information regarding **the amount of variability present in a set of data**.

Note:

1. If all the values are the **same**

→ There is no dispersion .

2. If all the values are **different**

→ There is a dispersion:

3. If the values **close to each other**

→The amount of Dispersion small.

4. If the values are **widely scattered**

→ The Dispersion is greater.

The Range

- The range is **the difference between the largest and smallest value in a set of observations**.
- Since the range, expressed as a single measure, imparts minimal information about a data set and therefore, **is of limited use**, it is often preferable to express the range as a number pair. $[x_S, x_L]$
- Although this is not the traditional expression for the range, **it is intuitive to imagine that knowledge of the minimum and maximum values in this data set would convey more information**.
- An infinite number of distributions, each with quite different minimum and maximum values, may have a range of 52.

  - Range =Largest value- Smallest value =

$$x_L - x_S$$

  - Range concern only onto two values

The Variance

- When the values of a set of observations lie **close to their mean**, the dispersion is less than when they are **scattered over a wide range**.
- Since this is true, it would be intuitively appealing if we could measure **dispersion relative to the scatter of the values about their mean**.
- Such a measure is realized in what is known as **the variance**.
- In computing the variance of a sample of values, for example, we **subtract the mean from each of the values, square the resulting differences, and then add up the squared differences**. This sum of the squared deviations of the values from their mean is divided by the sample size, minus 1, to obtain the sample variance.

- **a) Sample Variance (S^2):** $S^2 = \dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$, **where $\bar{X}$ is sample mean**
- **B) population variance(σ^2)** $\sigma^2 = \dfrac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$ **,where µ is the population mean**

Standard Deviation

- The variance represents squared units and, therefore, is **not an appropriate measure of dispersion when we wish to express this concept in terms of the original units**.
- To obtain a measure of dispersion in original units, we merely take the square root of the variance.
- The result is called **the standard deviation**.

The Standard Deviation:

- **is the square root of variance=** $\sqrt{Varince}$

a) Sample Standard Deviation = S = $\sqrt{S^2}$

b) Population Standard Deviation = σ = $\sqrt{\sigma^2}$

THE COEFFICIENT OF VARIATION

- 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 fallacious 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 ***body weight***, measured in pounds.
- Furthermore, although the same unit of measurement is used, the two means may be quite different.
- 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.

- 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.**

  - It is a measure use **to compare the dispersion in two sets of data** which is independent of the unit of the measurement .

    $C.V = \frac{S}{\overline{X}}(100)$ where S: Sample standard deviation.

    $\overline{X}$: Sample mean.

- Suppose two samples of human males yield the following data:

	Sampe1	Sample2
Age	25-year-olds	11year-olds
Mean weight	145 pound	80 pound
Standard deviation	10 pound	10 pound

- We wish to know **which is more variable**.

Solution:

- ✓ c.v (Sample1)= (10/145)*100= 6.9
- ✓ c.v (Sample2)= (10/80)*100= 12.5

- Then age of 11-years old(sample2) has more variation

- x_i values are given below:
- ✓ $X_1, X_2 \ldots X_{62}=1; X_{63}, X_{64} \ldots X_{109}=2; X_{110}, X_{111} \ldots . X_{148}=3; X_{149}, X_{150}, X_{187}=4; X_{188}, X_{189} \ldots, X_{245}=5; X_{246}, X_{247} \ldots X_{282}=6; X_{283}, X_{284}, X_{286}=7; X_{287}, X_{288} \ldots X_{297}=8$
- Calculate

1. The mean 2 marks
2. The median 2 marks
3. The mode 2 marks
4. The range 2 marks
5. The variance 3 marks
6. The standard deviation 2 marks
7. and the coefficient of variation 2marks

8 $\sum_{i=100}^{i=200} x_i$ 2 marks

9. $\sum_{i=150}^{i=155} x_i^2$ 3 marks

HOMEWORK

GROUP ASSIGNMENTS

Measurement and measurement scales

1. **Simple random sampling, interval scale, weighted mean**
2. **Systematic sampling, Ratio scale, mean of grouped data**
3. **Stratified random sampling, nominal scale, median of grouped data**
4. **Convenience Sampling, ordinal and nominal scales, geometric mean**
5. **Multistage sampling**

Probability, the Basis of the Statistical inference

INTRODUCTION

- The concept of probability is **frequently encountered in everyday communication**.
- **For example**, a physician may say that a patient has a **50-50 chance of surviving a certain operation.**
- Another physician may say that **she is 95 percent certain that a patient has a particular disease**.
- Most people express probabilities in terms of percentages.
- But, **it is more convenient to express probabilities as fractions**.
- Thus, we may measure the probability of the occurrence of some event **by a number between 0 and 1**.
- **The more likely** the event, the closer the number is to one.
- An event that **can't occur** has a probability of zero, and an event that is **certain to occur** has a probability of one.

Two views of Probability: objective and subjective

Objective Probability: Classical and Relative

Some definitions:

1. Equally likely outcomes:

Are the outcomes that **have the same chance of occurring**.

2. Mutually exclusive:

Two events are said to be mutually exclusive if **they cannot occur simultaneously** such that $A \cap B = \Phi$.

3. The universal Set (S): The set **all possible outcomes**.

4. The empty set Φ : Contain no elements.

5. The event, E : is a set of outcomes in S which **has a certain characteristic.**

- **Classical Probability (a priori, probability)** : If an event **can occur in N mutually exclusive and equally likely ways**, and if **m of these possess a trait, E**, the probability of the occurrence of event E is equal to m/ N .

If we read $P(E)$ as "the probability of E," we may express this definition as

$$P(E) = \frac{m}{N}$$

- **For Example:** in the rolling of the die , each of the six sides is equally likely to be observed . So, the probability that a 4 will be observed is equal to 1/6.

➢ Relative Frequency Probability (a posteriori):

❖ The relative frequency approach to probability depends on **the repeatability** of some process and the **ability to count the number of repetitions**, as well as **the number of times that some event of interest occurs**.

Def: If some process is **repeated a large number of times**, **n**, and if **some resulting event E occurs m times** , the relative frequency of occurrence of E , m/n will be approximately equal to probability of E . P(E) = m/n .

To express this definition in compact form, we write

$$P(E) = \frac{m}{n}$$

Subjective Probability :

Probability measures the confidence that a particular individual has in the truth of a particular proposition.

This concept **does not rely on the repeatability of any process**.

In fact, by applying this concept of probability, one may evaluate **the probability of an event that can only happen once**,

For Example : the probability that a cure for cancer will be discovered within the next 10 years.

Although the subjective view of probability has enjoyed increased attention over the years, it has not been fully accepted by statisticians who have traditional orientations.

Elementary Properties of Probability:

- Given some process (or experiment) with n mutually exclusive events E_1, E_2, E_3,…………, E_n, then

1. $P(E_i) \geq 0$, i= 1,2,3,……n
2. $P(E_1) + P(E_2) + \ldots\ldots + P(E_n) = 1$
3. $P(E_i + E_J) = P(E_i) + P(E_J)$, E_i, E_J are mutually exclusive

RULES OF PROBABILITY

1). Addition Rule

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

2). If A and B are mutually exclusive (disjoint) ,then

$P(A \cap B) = 0$

Then , addition rule is $P(A \cup B) = P(A) + P(B)$.

3. Complementary Rule

$P(A') = 1 - P(A)$

where, A' = complement event

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

(**Early age at onset** defined to be **18 years or younger** and **Later age at onset** defined to be **later than 18 years**).

Answer the following questions:

Suppose we pick **a person at random from this sample**.

1.The probability that this person will be **18-years old or younger**?

2-The probability that this person has family history of mood orders Unipolar(C)?

3-The probability that this person has no family history of mood orders Unipolar($\overline{C}$)?

4-The probability that this person is 18-years old or younger or has no family history of mood orders :Negative (A)?

5-The probability that this person is more than18-years old and has family history of mood orders Unipolar and Bipolar(D)?

CONDITIONAL PROBABILITY

- **The set of "all possible outcomes" may constitute a subset of the total group**. In other words, **the size of the group of interest may be reduced by conditions not applicable to the total group**.
- When probabilities are calculated **with a subset of the total group as the denominator**, the result is **a conditional probability**.

❖ E.g.: suppose we pick a person at random and find **he is 18 years or younger** (E),what is the probability that this person will be one who has no family history of mood disorders (A)?

Solution

The total number of subjects is no longer of interest, since, with the selection of an Early subject, **the Later subjects are eliminated**. We may define the desired probability, then, as follows: **What is the probability that a subject has no family history of mood disorders , given that the selected subject is Early ?**

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. Answer:?????????????????????????

CONDITIONAL PROBABILITY:

P(A\B) is the probability of A assuming that B has happened.

$$P(A\backslash B)=\frac{P(A\cap B)}{P(B)}\ ,\ P(B)\neq 0$$

$$P(B\backslash A)=\frac{P(A\cap B)}{P(A)}\ ,\ P(A)\neq 0$$

Suppose we pick a person at random and find he has family history of mood (D). what is the probability that this person will be 18 years or younger (E)?

CALCULATING A JOINT PROBABILITY :

- 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**.
- E.g. Suppose we pick a person at random from the 318 subjects. Find the probability that he will early (E) and has no family history of mood disorders (A).

Solution

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. **Answer:**

THE MULTIPLICATION RULE

- A probability may be **computed from other probabilities**. For example, **a joint probability may be computed as the product of an appropriate marginal probability and an appropriate conditional probability**.
- This relationship is known as the multiplication rule of probability.

E.g. We wish to compute the joint probability of **Early age at onset** and **a negative family history of mood disorders A** from knowledge of an appropriate marginal probability and an appropriate conditional probability.

- Solution: The The probability we seek is P (E∩A)

$P(E) = 141/318 = .4434$ and $P(A \mid E) = 28/141 = .1986$

$$P(E \cap A) = P(E)P(A \mid E) = (.4434)(.1986) = .0881$$

- $P(A \cap B) = P(A \backslash B)P(B)$
- $P(A \cap B) = P(B \backslash A)P(A)$

Where,

- P(A): marginal probability of A.
- P(B): marginal probability of B.
- P(B\A):The conditional probability.

The *conditional probability* of *A* given *B* is equal to the probability of $A \cap B$ divided by the probability of *B*, provided the probability of *B* is not zero.

INDEPENDENT EVENTS:

- If **A has no effect on B**, we said that A and B are independent events.
- Then,

  1- $P(A\cap B)= P(B)P(A)$

  2- $P(A\backslash B)=P(A)$

  3- $P(B\backslash A)=P(B)$

Two events are not independent unless all these statements are true.

- If two events are independent, the probability of their joint occurrence is equal to the product of the probabilities of their individual occurrences.

E.g. 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 .

1.What is the probability that a student picked at random wears eyeglasses given that the student is a boy?

2.What is the probability of the joint occurrence of the events of wearing eye glasses and being a boy?

COMPLEMENTARY EVENTS

- The probability of an event A is equal to 1 minus the probability of its complement, which is written $\overline{A}$ and $P(\overline{A}) = 1 - P(A)$

MARGINAL PROBABILITY

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,

$$P(A_i) = \sum P(A_i \cap B_j), \qquad \text{for all values of } j$$

$$P(E \cap A) = 28/318 = .0881$$
$$P(E \cap B) = 19/318 = .0597$$
$$P(E \cap C) = 41/318 = .1289$$
$$P(E \cap D) = 53/318 = .1667$$

$$\begin{aligned} P(E) &= P(E \cap A) + P(E \cap B) + P(E \cap C) + P(E \cap D) \\ &= .0881 + .0597 + .1289 + .1667 \\ &= .4434 \end{aligned}$$

Exercise

It is known that a student who does his online homework on a regular basis has a chance of 83 percent to get a good grade (A or B) but the chance drops to 58 percent if he doesn't do the homework regularly. John has been very busy with other courses and an evening job and figures that he has only a 69 percent chance of doing the homework regularly. **What is his chance of not getting a good grade in the course?**

E:Doing online homework on a regular basis.

F:Getting a good grade A or B

BAYE'S THEOREM

- 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**.
- In the consideration of screening tests, one must be aware of the fact that they are **not always infallible**. That is, a testing procedure may yield **a false positive** or **a false negative**.

Definition

1. A **false positive results:** a test indicates a positive status when the true status is negative.
2. A **false negative** results: a test indicates a negative status when the true status is positive.

In summary, the following questions must be answered in order to evaluate **the usefulness of test results and symptom status** in determining whether or not a subject has some disease:

1. Given that a subject has the disease, **what is the probability of a positive test result** (or the presence of a symptom)? sensitivity

2. Given that a subject does not have the disease, **what is the probability of a negative test result** (or the absence of a symptom)? specificity

3. Given a positive screening test (or the presence of a symptom), **what is the probability that the subject has the disease?**

4. Given a negative screening test result (or the absence of a symptom), **what is the probability that the subject does not have the disease?**

- Suppose we have for a sample of n subjects (where n is a large number) the information shown in table below:

	Disease		
Test Result	Present (D)	Absent ($\bar{D}$)	Total
Positive (T)	a	b	$a + b$
Negative ($\bar{T}$)	c	d	$c + d$
Total	$a + c$	$b + d$	n

- The table shows for these n subjects, **their status with regard to a disease and results from a screening test** designed to identify subjects with the disease.
- 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.

Definition.1

The sensitivity of a test (or symptom) is the probability of a positive test result (or presence of the symptom) given the presence of the disease.

Definition.2

The **specificity of a test** (or symptom) is the probability of a negative test result (or absence of the symptom) given the absence of the disease.

Definition.3

The positive predictive value of a screening test (or symptom) is the probability that a subject has the disease given that the subject has a positive screening test result (or has the symptom).

Definition 4

The negative predictive value of a screening test (or symptom) is the probability that a subject does not have the disease, given that the subject has a negative screening test result (or does not have the symptom).

- **Estimates of the positive predictive value and negative predictive value of a test** (or symptom) may be obtained from knowledge of a test's (or symptom's) ***sensitivity*** **and** ***specificity*** and **the** ***probability of the relevant disease*** **in the general population**.
- To obtain these predictive value estimates, we make use of Bayes's theorem.

$$P(D \mid T) = \frac{P(T \mid D)\, P(D)}{P(T \mid D)\, P(D) + P(T \mid \overline{D})\, P(\overline{D})}$$

- To understand the logic of Bayes's theorem, we must recognize that the numerator of the Equation above represents **P(D∩T)** "multiplication rule" and that the denominator represents **P(T)** "We know that event T is the result of a **subject's being classified as positive with respect to a screening test** (or classified as having the symptom)".
- A subject classified as positive **may have the disease or may not have the disease.**

- Therefore, the occurrence of T is **the result of a subject having the disease and being positive or not having the disease and being positive.**
- These two events are mutually exclusive (their intersection is zero), and consequently, by the addition rule we may write:

$$P(T) = P(D \cap T) + P(\overline{D} \cap T)$$

Since, by the multiplication rule, $P(D \cap T) = P(T \mid D)\, P(D)$ and $P(\overline{D} \cap T) = P(T \mid \overline{D})\, P(\overline{D})$, we may rewrite

$$P(T) = P(T \mid D)\, P(D) + P(T \mid \overline{D})\, P(\overline{D})$$

Note, also, that the numerator of equation (on the previous slide) is equal to **the sensitivity times the rate** (prevalence) of the disease and the denominator is equal to **the sensitivity times the rate of the disease** plus **the term 1 minus the specificity** times **the term 1 minus the rate of the disease**. Thus, we see that the predictive value positive can be calculated from knowledge of the **sensitivity, specificity, and the rate of the disease**.

To answer question 4 (see slide 19) we follow a now familiar line of reasoning. the probability that a subject **does not have the disease given that the subject has a negative screening test result** is calculated using Bayes Theorem through the following formula

$$P(\overline{D}\mid\overline{T})=\frac{P(\overline{T}\mid\overline{D})P(\overline{D})}{P(\overline{T}\mid\overline{D})P(\overline{D})+P(\overline{T}\mid D)P(D)}$$

where, $p(\overline{T}\mid D)=1-P(T\mid D)$

Example

❖ 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 or older. The results are as follows.

Test Result	Alzheimer's Diagnosis? Yes (D)	No ($\overline{D}$)	Total
Positive (T)	436	5	441
Negative ($\overline{T}$)	14	495	509
Total	450	500	950

In the context of this example

a)What is a false positive?

A false positive is when the test indicates a positive result (T) when the person does not have the disease $\overline{D}$

b) What is the false negative?

A false negative is when a test indicates a negative result $(\overline{T})$ when the person has the disease (D).

c) Compute the sensitivity of the symptom.

d) Compute the specificity of the symptom.

We see that the positive predictive value of the test **depends on the rate of the disease in the relevant population in general**.

Suppose it is known that the rate of the disease in the general population is 11.3%.

What is the predictive value positive of the symptom and the predictive value negative of the symptom. The predictive value positive of the symptom is calculated as

The predictive value negative of the symptom is calculated as

PROBABILITY DISTRIBUTIONS

- Probability distributions of random variables **assume powerful roles in statistical analyses**.
- Since they **show all possible values of a random variable and the probabilities associated with these values**, probability distributions may be **summarized in ways that enable researchers to easily make objective decisions based on samples drawn from the populations** that the distributions represent.
- We shall see that the **relationship** between the values of a random variable and the probabilities of their occurrence may be summarized by means of a device called a probability distribution.
- Knowledge of the probability distribution of a random variable provides the clinician and researcher with a powerful tool for summarizing and describing a set of data and for reaching conclusions about a population of data on the basis of a sample of data drawn from the population.

PROBABILITY DISTRIBUTIONS OF DISCRETE VARIABLES

The Random Variable (X):

- When the values of a variable (height, weight, or age) **can't be predicted in advance**, the variable is called a random variable.
- An example is the adult height: when a child is born, we can't predict exactly his or her height at maturity.

Definition:

- The probability distribution of a discrete random variable is a **table**, **graph**, **formula**, or **other device** used to specify **all possible values of a discrete random variable along with their respective probabilities**.
- If we 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

- Table on next slide shows the **number of food assistance programs** used by subjects (family) in a given sample.

Number of Programs	Frequency
1	62
2	47
3	39
4	39
5	58
6	37
7	4
8	11
Total	297

- We wish to construct the probability distribution of the discrete variable X, where X= number of food assistance programs used by the study subjects.
- The values of X are $x_1=1$, $x_2=2$, $x_3=3$, $x_4=4$, $x_5=5$, $x_6=6$ and $x_8=8$ and We compute the probabilities for these values by dividing their respective frequencies by the total, 297. Thus, for example, $p(x_1) = P(X=x_1) = 62/297=0.2088$

- Probability distribution of programs utilized by families among the subjects described in table on previous slide, which is the desired probability distribution is shown **down here**

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

- The values of $p(x) = P(X = x)$ are **all positive**, they are **all less than 1**, and **their sum is equal to 1**. These are not phenomena peculiar to this particular example, but are **characteristics of all probability distributions of discrete variables**.

- If x_1, x_2, $x_{3,}$ x_4………., x_k are all possible values of the discrete random variable X, then we may give the following **two essential properties of a probability distribution of a discrete variable**

1. $0 \leq P(X = x) \leq 1$
2. $\sum P(X = x) = 1$

- ✓ What is the probability that a randomly selected family will be **one who** used three assistance programs?
- ✓ What is the probability that a randomly selected family used either one or two programs?

The Cumulative Probability Distribution of X, F(x):

- It shows **the probability that the variable X is less than or equal to a certain value**, **$P(X \leq x)$**.

Number of Programs	frequency	P(X=x)	F(x)= P(X≤ x)
1	62	0.2088	0.2088
2	47	0.1582	0.3670
3	39	0.1313	0.4983
4	39	0.1313	0.6296
5	58	0.1953	0.8249
6	37	0.1246	0.9495
7	4	0.0135	0.9630
8	11	0.0370	1.0000
Total	297	1.0000	

1. What is the probability that a family picked at random will be one who used **two or fewer** assistance programs?
2. What is the probability that a randomly selected family will be one who used **fewer than four programs**?
3. What is the probability that a randomly selected family used **five or more programs**?
4. What is the probability that a randomly selected family is one who used between **three and five** programs, inclusive?

- **Properties of probability distribution of discrete random variable.**

1. $0 \leq P(X = x) \leq 1$
2. $\sum P(X = x) = 1$
3. $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a-1)$
4. $P(X < b) = P(X \leq b-1)$

Mean and Variance of discrete probability distributions

The mean and variance of a discrete probability distribution can easily be found using the formulae below:

$$\mu = \sum xp(x)$$
$$\sigma^2 = \sum (x-\mu)^2 p(x) = \sum x^2 p(x) - \mu^2$$

where $p(x)$ is the relative frequency of a given random variable X. **The standard deviation** is simply the positive square root of the variance.

The Binomial Distribution:

- The binomial distribution is one of the most widely encountered probability distributions in applied statistics. It is derived from a process known as a **Bernoulli trial**.
- **Bernoulli trial is** :

✓ 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, the trial is called a Bernoulli trial.

The Bernoulli Process

- A sequence of Bernoulli trials forms a Bernoulli process under the following conditions

1- **Each trial** results in one of two possible, mutually exclusive, outcomes. One of the possible outcomes is denoted (***arbitrarily***) as a **success**, and the other is denoted a **failure**.

2- The probability of a success, denoted by **p**, **remains constant** from trial to trial. The probability of a failure, 1-p, is denoted by **q**.

3- The trials are **independent**, that is the outcome of any particular trial is not affected by the outcome of any other trial

Example

- ❖ If we examine all birth records from the North Carolina State Center for Health statistics for year 2001, we find that 85.8 percent of the pregnancies had **delivery in week 37 or later** (full- term birth).
- ✓ If we randomly selected five birth records from this population what is the probability that exactly three of the records will be for full-term births?
- ❑ Assign the number **1 to a success** (record for a full-term birth) and the **number 0 to a failure** (record of a premature birth).
- ❑ The process that eventually results in a birth record is considered to be a Bernoulli process.
- ❖ Suppose the five birth records selected resulted in this sequence of full-term births: $P(1,0,1,1,0) = pqppq = q^2p^3$

The **multiplication rule** is appropriate for computing this probability since we are seeking the probability of a full-term, and a premature, and a full term, and a full-term, and a premature, ***in that order*** or, in other words, the joint probability of the five events.

- Three successes and two failures could occur in any one of the following additional sequences as well:

Number	Sequence
2	11100
3	10011
4	11010
5	11001
6	10101
7	01110
8	00111
9	01011
10	01101

When we draw a single sample of size five from the population specified, **we obtain only one sequence of successes and failures**.

The question now becomes, **what is the probability of getting sequence number 1 or sequence number 2 . . . or sequence number 10?**

- From the addition rule we know that this probability is equal to **the sum of the individual probabilities**.
- In the present example we need to sum them or, equivalently, multiply by 10. We would have:

$$10(.142)^2(.858)^3 = 10(.0202)(.6316) = .1276$$

- We can easily anticipate that, as the size of the sample increases, **listing the number of sequences becomes more and more difficult and tedious**.
- What is needed is **an easy method of counting the number of sequences**.
- When the order of the objects in a subset is immaterial, the subset is called a **combination of objects**.

- If **a set** consists of **n objects**, and we wish to form **a subset** of **x objects** from these n objects, ***without regard to the order*** of the objects in the subset, **the result is called a combination**.

- The number of combinations of n objects that can be formed by taking **x of them at a time** is given by $\binom{n}{x} = \frac{n!}{x!(n-x)!}$

Which can also be written as $${}_nC_x = \frac{n!}{x!(n-x)!}$$

- The probability distribution of the binomial random variable **X**, the number of successes in **n** independent trials is:

$$f(x) = P(X = x) = \binom{n}{x} p^X q^{n-X} \quad , \quad x = 0,1,2,....,n$$

- Where $\binom{n}{x}$ is the number of combinations of **n** distinct objects taken **x** of them at **a** time.

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

Note: 0! =1

$$x! = x(x-1)(x-2)....(1)$$

Properties of the binomial distribution

1. $f(x) \geq 0$

2. $\sum f(x) = 1$

3. The parameters of the binomial distribution are *n* and *p*

4. $\boxed{\mu = E(X) = np}$

5. $\boxed{\sigma^2 = \text{var}(X) = np(1-p)}$

- 14 percent of pregnant mothers admitted to smok 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 smok during pregnancy**?

Probabilities for different values of n, p, and x have been **tabulated**, so that **we need only to consult an appropriate table to obtain the desired probability**.

It gives the **probability that X is less than or equal to some specified value**. That is, the *table gives the* ***cumulative probabilities*** **from $x = 0$ up through some specified positive number of successes**.

- Suppose it is known that in a certain population **10 percent** of the population is color blind. If a random sample of **25 people** is drawn from this population, find the probability that:

a) Five or fewer will be color blind?

b) Six or more will be color blind?

c) Between **six and nine inclusive** will be color blind?

d) Two, three, or four will be color blind?

- The table **does not give probabilities for values of p greater than 0.5**.
- We may obtain probabilities from the table, however, by **restating the problem in terms of the probability of a failure**, rather than in terms of the probability of a success, p.

As part of the restatement, we must also think in terms of the **number of failures**, rather than the number of successes, x.

$$P(X = x \mid n, p > .50) = P(X = n - x \mid n, 1 - p)$$

- In words, “The probability that X is equal to **some specified value given the sample size** and a probability of success greater than 0.5 is equal to the probability that X is equal to $\boxed{n - x}$ given the sample size and the probability of a failure of $\boxed{1 - p}$.”
- For purposes of using the binomial table we **treat the probability of a failure as though it were the probability of a success**.
- When p is greater than 0.5, we may obtain cumulative probabilities from the table by using the following relationship

$$P(X \le x \mid n, p > .50) = P(X \ge n - x \mid n, 1 - p)$$

Finally, to use Table B to find the **probability that X is greater than** or equal to some x when we use the following relationship.

$$P(X \ge x \mid n, p > .50) = P(X \le n - x \mid n, 1 - p)$$

e.g. According to a June 2003 poll conducted by the Massachusetts Health Benchmarks project (A-4), **approximately 55 percent** of residents answered "**serious problem**" to the question, "Some people think that childhood obesity is a national health problem. What do you think? Is it a very *serious problem*, *somewhat of a problem*, *not much of a problem*, or *not a problem at all*?" Assuming that the probability of giving this answer to the question **is 0.55** for any Massachusetts resident, use the table to find the probability that if 12 residents are chosen at random:

1. **Exactly** seven will answer "serious problem."
2. **Five or fewer** households will answer "serious problem."
3. **Eight or more** households will answer "serious problem."

- The binomial distribution has two parameters, n and p.
- They are parameters in the sense that **they are sufficient to specify a binomial distribution**.
- The binomial distribution is really **a family of distributions with each possible value of n and p designating a different member of the family**.
- Strictly speaking, the binomial distribution is applicable in situations **where sampling is from *an infinite population* or from *a finite population with replacement***.
- Since in actual practice samples are usually drawn without replacement from finite populations, the question arises as to **the appropriateness of the binomial distribution under these circumstances**.
- Whether or not the binomial is appropriate depends on how drastic the effect of these conditions is on the constancy of p from trial to trial.
- It is generally agreed that **when *n is small relative to N*, the binomial model is appropriate**. Some writers say that n is small relative to N **if N is at least 10 times as large as n**.

THE POISSON DISTRIBUTION

- If the random variable X is **the number of occurrences of some random event** in a certain period of **time** or **space** (or some **volume** of matter).
- The probability distribution of X is given by:

f (x) =P(X=x) = $\dfrac{e^{-\lambda}\lambda^{x}}{x!}$,x = 0,1,…..

The symbol **e** is the constant equal to 2.7183. λ (Lambda) is called the parameter of the distribution and is **the average number of occurrences of the random event in the interval** (or volume)

The following statements describe what is known as the Poisson process.

1. **The occurrences of the events are independent**. The occurrence of an event in an interval of space or time has no effect on the probability of a second occurrence of the event in the same, or any other, interval.

2. Theoretically, **an infinite number of occurrences of the event must be possible in the interval**.

3. The probability of the single occurrence of the event in a given interval is **proportional to the length of the interval**.

4. **In any infinitesimally small portion of the interval**, the probability of **more than one occurrence** of the event is negligible.

- An interesting feature of the Poisson distribution is the fact that **the mean and variance are equal**.

The Poisson distribution is employed as a model **when counts are made of events or entities that are distributed at random in space or time.**

- An additional use of the Poisson distribution in practice **occurs when n is large and p is small**.
- In this case, the Poisson distribution **can be used to approximate the binomial distribution**.
- We may, however, use the table C, which gives cumulative probabilities for various values of λ and X.

Properties of the Poisson distribution

1. $f(x) \geq 0$
2. $\sum f(x) = 1$
3. $\mu = E(X) = \lambda$
4. $\sigma^2 = \text{var}(X) = \lambda$

Exercise

- In a study of a 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:

1- The probability that in the next year, among patients receiving rocuronium, **exactly three** will experience anaphylaxis?

2- The probability that **less than two patients** receiving rocuronium, in the next year will experience anaphylaxis?

3- The probability that **more than two patients** receiving rocuronium, in the next year will experience anaphylaxis?

4- The **expected value** of patients receiving rocuronium, **in the next year** who will experience anaphylaxis.

5- **The variance** of patients receiving rocuronium, **in the next** year who will experience anaphylaxis

6- The **standard deviation** of patients receiving rocuronium, **in the next year** who will experience anaphylaxis

7-What is the probability that **at least three patients** in the **next year** will experience anaphylaxis if rocuronium is administered with anesthesia?

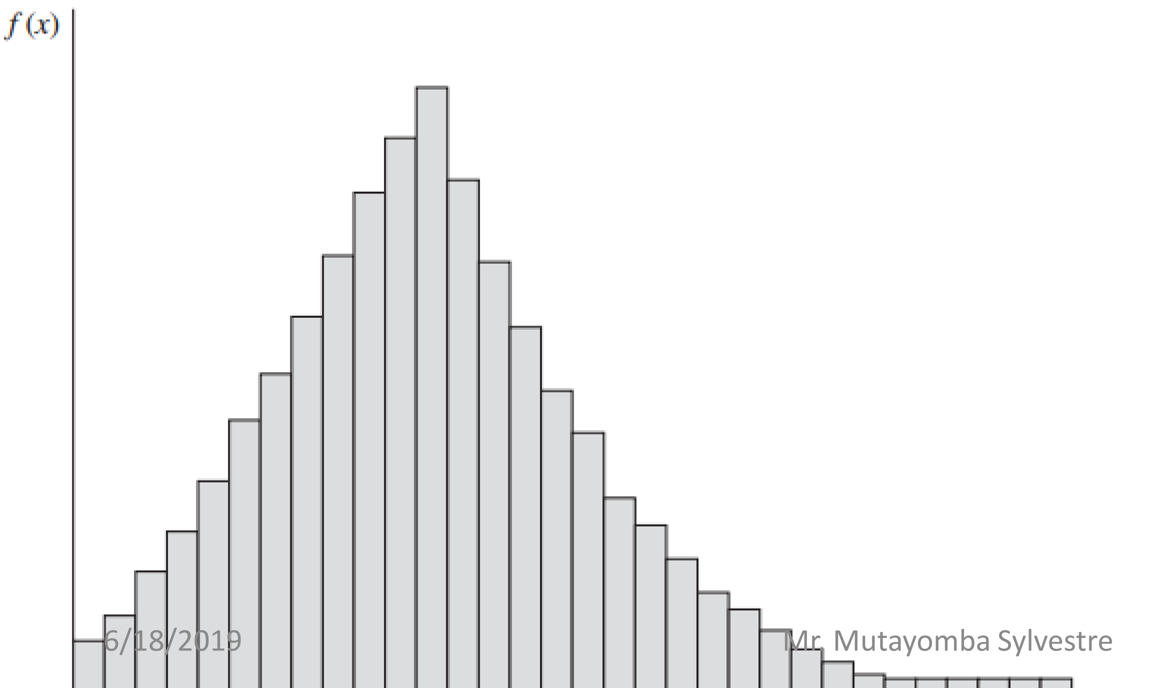
8-What is the probability that **exactly one patient** in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

9-What is the probability that **none of the patients** in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

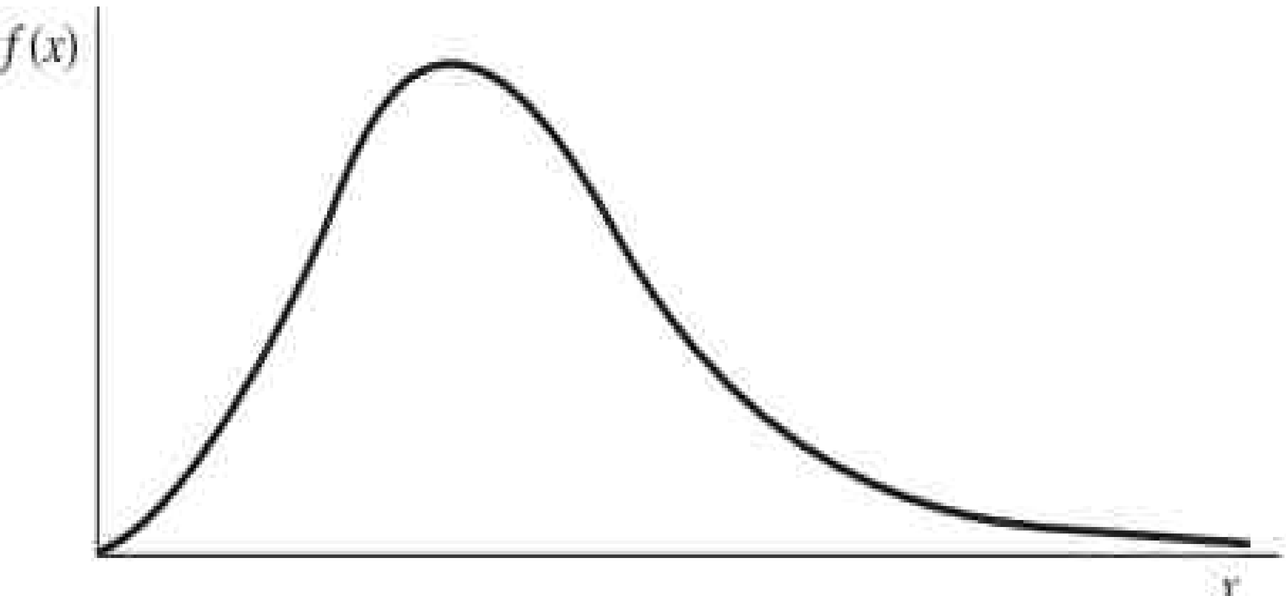
10-What is the probability that **at most two** patients in the next year will experience anaphylaxis if rocuronium is administered with anesthesia?

CONTINUOUS PROBABILITY DISTRIBUTION

- The binomial and the Poisson, are distributions of discrete variables.
- Let us now consider distributions of continuous random variables.
- **Between any two values assumed by a continuous variable**, there exist an infinite number of values.
- Imagine the situation where **the number of values of our random variable is very large and the width of our class intervals is made very small**.
- The resulting histogram could look like the one shown below:

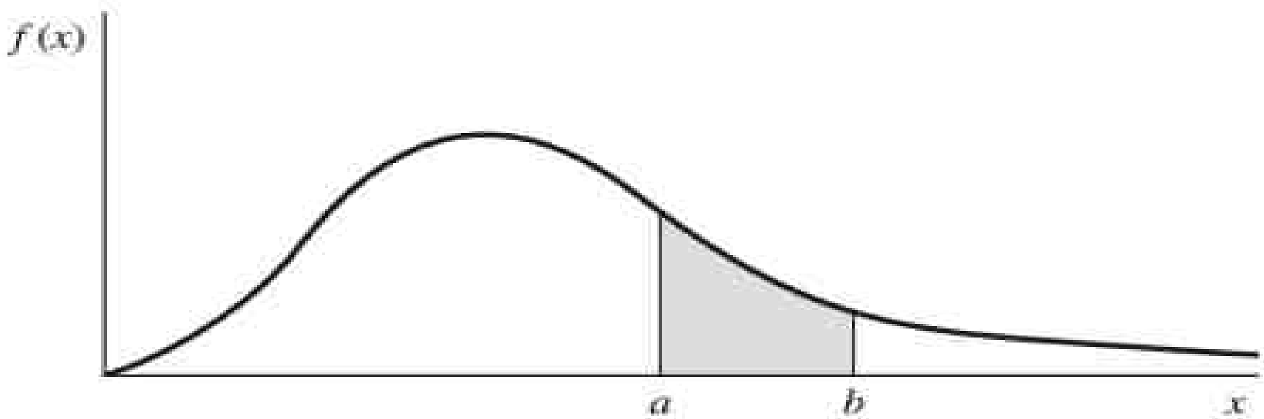


- If we were **to connect the midpoints of the cells of the histogram** in previous slide to form a frequency polygon, clearly we would have **a much smoother figure than the frequency polygon on slide 33** (In notes of part I)
- In general, as the number of observations, **n, approaches infinity**, and the **width of the class intervals approaches zero**, the frequency polygon approaches a smooth curve such as the one below.



Such smooth curves are used to represent graphically the distributions of continuous random variables.

- The total area under the curve **is equal to one**, as was true with the histogram, and the relative frequency of occurrence of values between any two points on the x-axis is equal to **the total area bounded by the curve, the x-axis, and perpendicular lines erected at the two points on the x-axis**. See figure below



Graph of a continuous distribution showing area between *a* and *b*.

- **The probability of any specific value of the random variable is zero**. This seems logical, since a specific value is represented by a point on the x-axis and **the area above a point is zero**.

The probability of a continuous random variable to assume values between a and b is denoted by $P(a < X < b)$

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 , $b = P(a<x<b)$.

THE NORMAL DISTRIBUTION, THE GAUSSIAN DISTRIBUTION:

- It is one of *the most important probability distributions* in statistics.
- The normal density is given by: $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Where

- ✓ $-\infty < x < \infty$, $-\infty < \boldsymbol{\mu} < \infty$, $\boldsymbol{\sigma} > 0$
- ✓ π, e : constants
- ✓ μ: population mean.
- ✓ σ : Population standard deviation.

CHARACTERISTICS OF THE NORMAL DISTRIBUTION:

- The following are some important characteristics of the normal distribution:

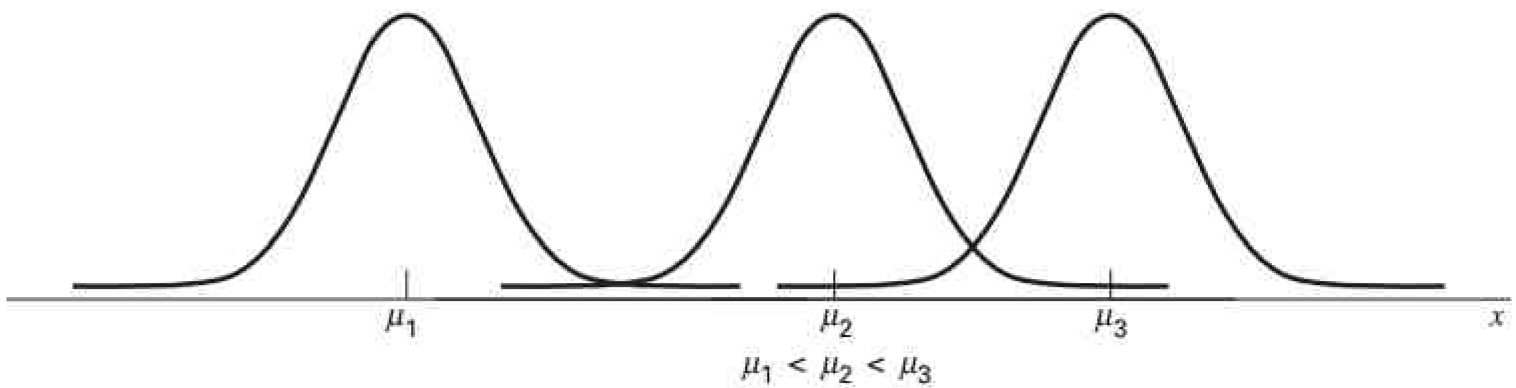
1- It is **symmetrical about its mean**, μ.

2- The mean, the median, and the mode are all equal.

3- The **total area** under the curve above the x-axis is one.

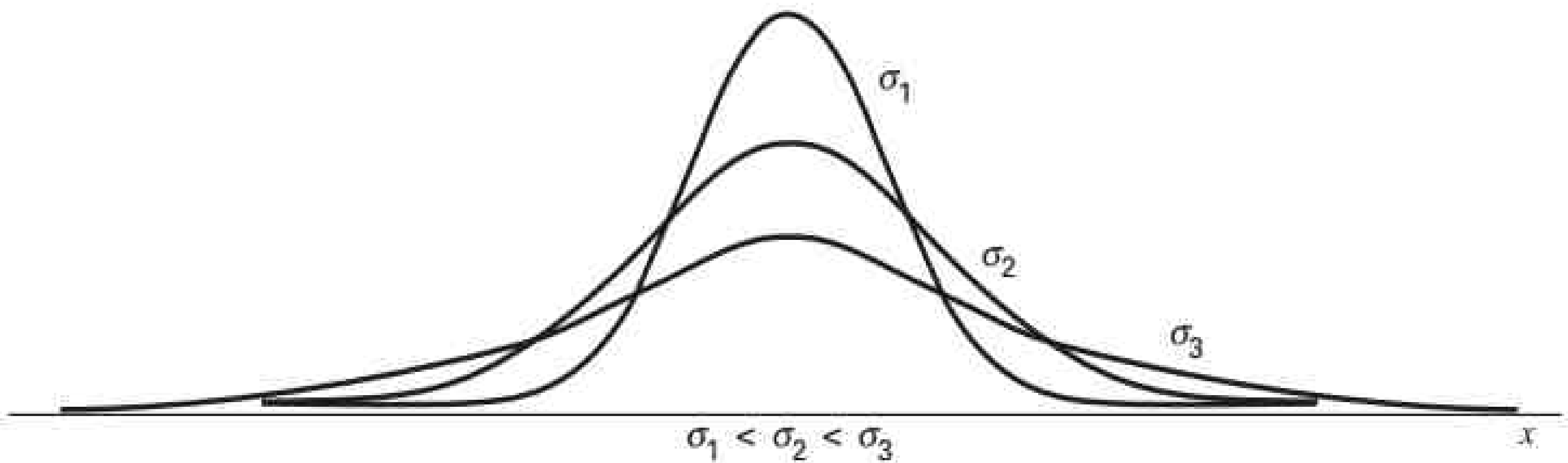
4-The normal distribution is **completely determined by the parameters μ and σ**. In other words, a different normal distribution is specified for each different value of μ and σ.

- Different values of **μ shift the graph of the distribution along the x-axis** as is shown on the figure below:



Three normal distributions with different means but the same amount of variability.

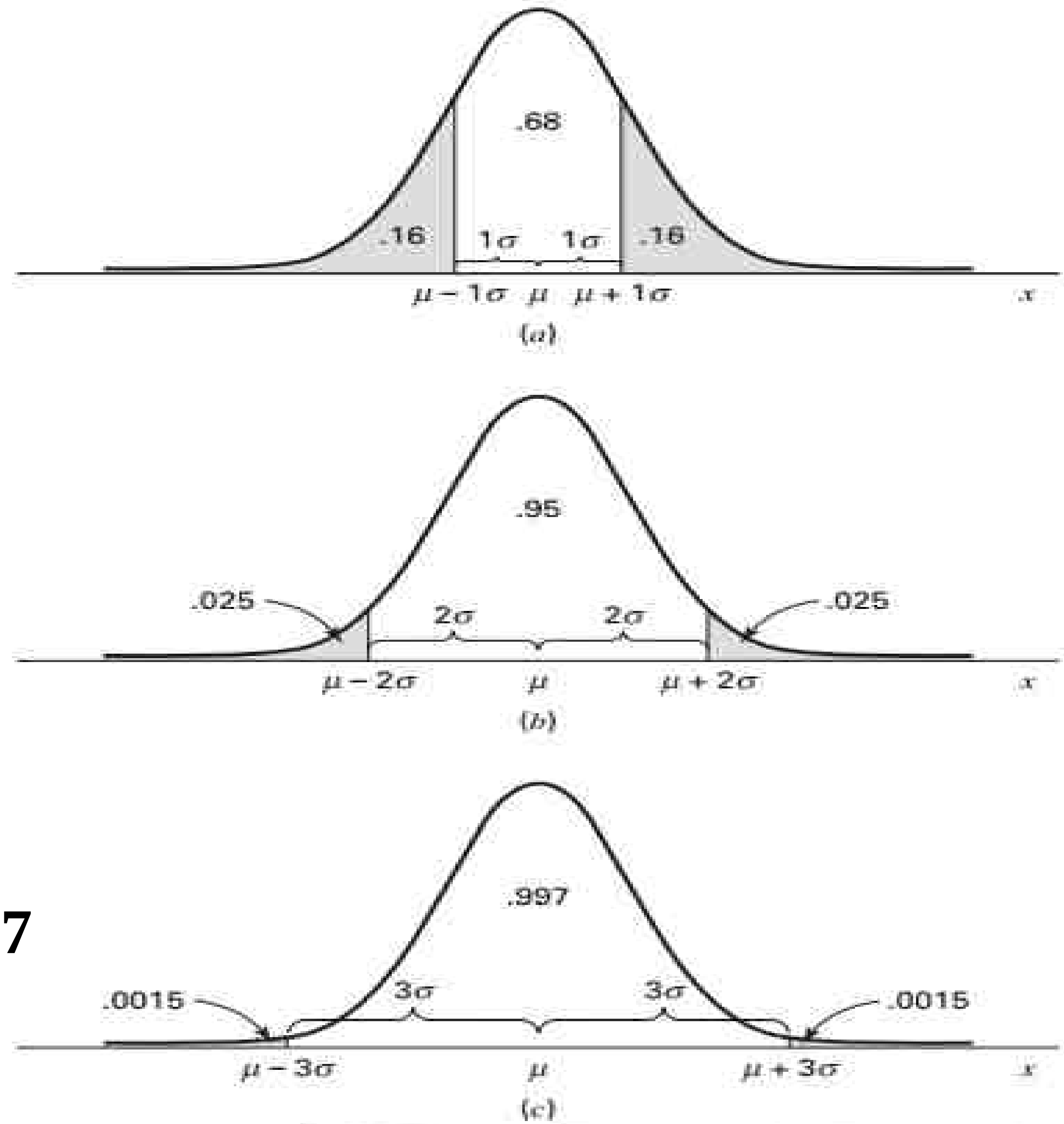
- Different values of σ determine the degree of **flatness** or peakedness of the graph of the distribution as shown on the figure below:



Three normal distributions with different standard deviations but the same mean.

- If we erect perpendiculars **a distance of 1 standard deviation from the mean in both directions**, the area enclosed by these perpendiculars, the x-axis, and the curve will be **approximately 68 percent of the total area**.
- If we extend these lateral boundaries **a distance of two standard deviations on either side of the mean**, approximately **95 percent** of the area will be enclosed, and

- Extending them **a distance of three standard deviations** will cause approximately 99.7 percent of the total area to be enclosed. See figures below



Subdivision of the area under the normal curve (areas are approximate).

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

$$f(z) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{z^2}{2}} \quad , \qquad -\infty < z < \infty$$

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 table (D) 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 $-3.89 < z < 3.89$. Thus,

$P(-3.89 < Z < 3.89) \cong 1$.

For standard normal distribution,

$P(Z > 0) = P(Z < 0) = 0.5$

Example :

If Z is a standard normal distribution, then

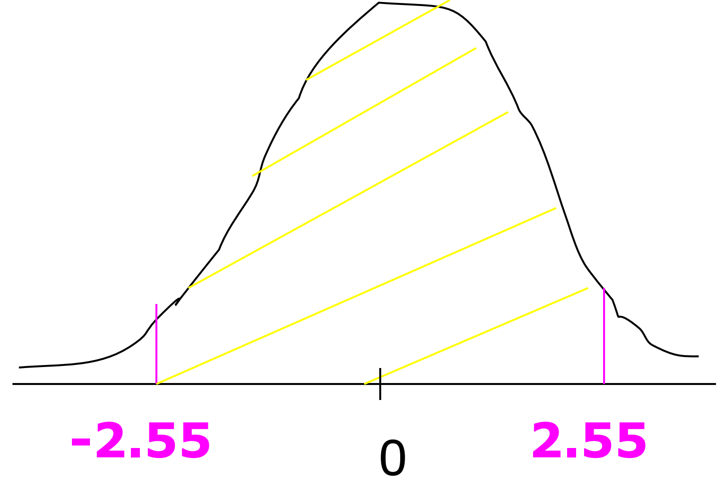
1) $P(Z < 2) = 0.9772$

is the area to the left to 2

and it equals 0.9772.

Example :

P(-2.55 < Z < 2.55) is the area between -2.55 and 2.55, Then it equals

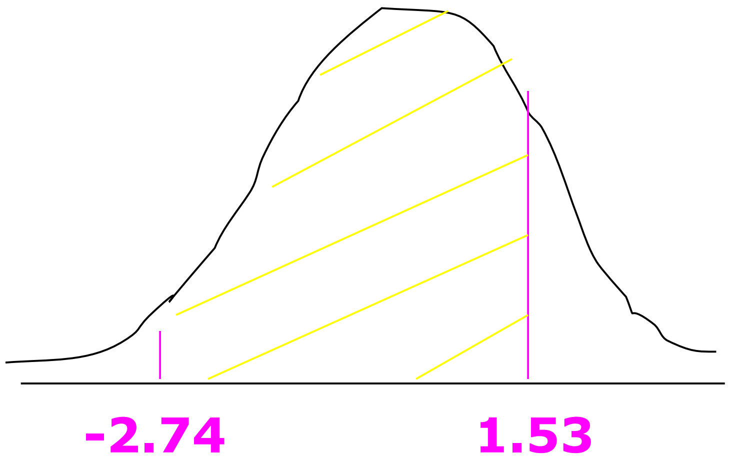


P(-2.55 < Z < 2.55) =0.9946 – 0.0054

= 0.9892.

Example :

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

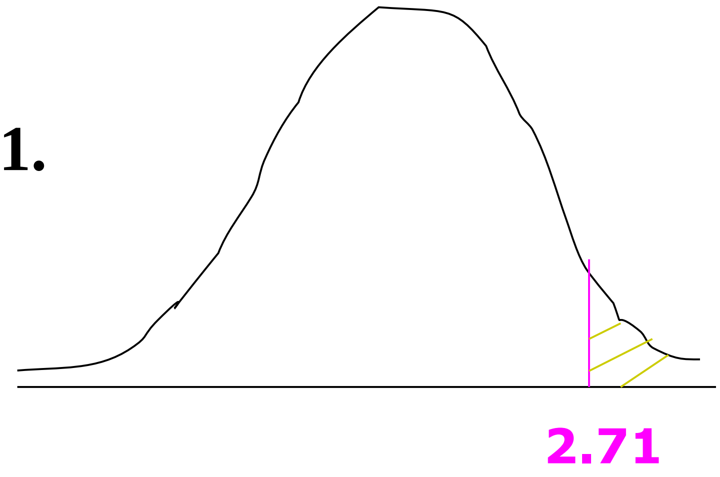
= 0.9339.

Example :

P(Z > 2.71) is the area to the right to 2.71.

So,

P(Z > 2.71) =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 $\mu = 3$, $\sigma = 2$. Find the value of standard normal Z, If X= 6?

Answer:

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.
- **Human stature** and **human intelligence** are frequently cited as examples of variables that are approximately normally distributed.
- We may answer simple probability questions about random variables **when we know, or are willing to assume, that they are, at least, approximately normally distributed**.

Example:

The ‘Uptime ’is a custom-made light weight battery-operated activity monitor t**hat 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.

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

$$P(X < 3) = P\left(\frac{X-\mu}{\sigma} < \frac{3-5.4}{1.3}\right) = P(Z < -1.85) = 0.0322$$

2-The probability that the child spend more than 5 hours in the upright position 24-hour period

$$P(X > 5) = P\left(\frac{X-\mu}{\sigma} > \frac{5-5.4}{1.3}\right) = P(Z > -0.31)$$

3-The probability that the child spend exactly 6.2 hours in the upright position 24-hour period

4-The probability that the child spend from 4.5 to 7.3 hours in the upright position 24-hour period

$$P(4.5 < X < 7.3) = P\left(\frac{4.5-5.4}{1.3} < \frac{X-\mu}{\sigma} < \frac{7.3-5.4}{1.3}\right)$$

$$= P(-0.69 < Z < 1.46) =$$

SAMPLING DISTRIBUTIONS

- We use sampling distributions **to answer probability questions about sample statistics**.
- **A sample statistic is a descriptive measure**, such as the mean, median, variance, or standard deviation, that is **computed from the data of a sample**.
- Sampling distributions make statistical inferences valid.

DEFINITION

- 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.

- 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.

- We usually are interested in knowing **three things** about a given sampling distribution: its **mean**, its **variance**, and its **functional form** (how it looks when graphed)

DISTRIBUTION OF THE SAMPLE MEAN

Example:

- Suppose we have a population of size $N = 5$ consisting of the ages of five children who are outpatients in a community mental health center. The ages are as follows: $x_1 = 6, x_2 = 8, x_3 = 10, x_4 = 12,$ and $x_5 = 14.$

- The mean, of this population is equal to $\sum x_i / N = 10$ and the variance is

$$\boxed{\sigma^2 = \frac{\sum (x_i - \mu)^2}{N} = \frac{40}{5} = 8}$$

- Let us draw **all possible samples of size** $n = 2$ from this population. These samples, along with their means, are shown in table on next slide.

All Possible samples of size n=2 from a population of size N=5. Samples above or below the principal diagonal result when sampling is without replacement. sample means are **in parentheses**

		Second Draw				
		6	8	10	12	14
First Draw	6	6, 6 (6)	6, 8 (7)	6, 10 (8)	6, 12 (9)	6, 14 (10)
	8	8, 6 (7)	8, 8 (8)	8, 10 (9)	8, 12 (10)	8, 14 (11)
	10	10, 6 (8)	10, 8 (9)	10, 10 (10)	10, 12 (11)	10, 14 (12)
	12	12, 6 (9)	12, 8 (10)	12, 10 (11)	12, 12 (12)	12, 14 (13)
	14	14, 6 (10)	14, 8 (11)	14, 10 (12)	14, 12 (13)	14, 14 (14)

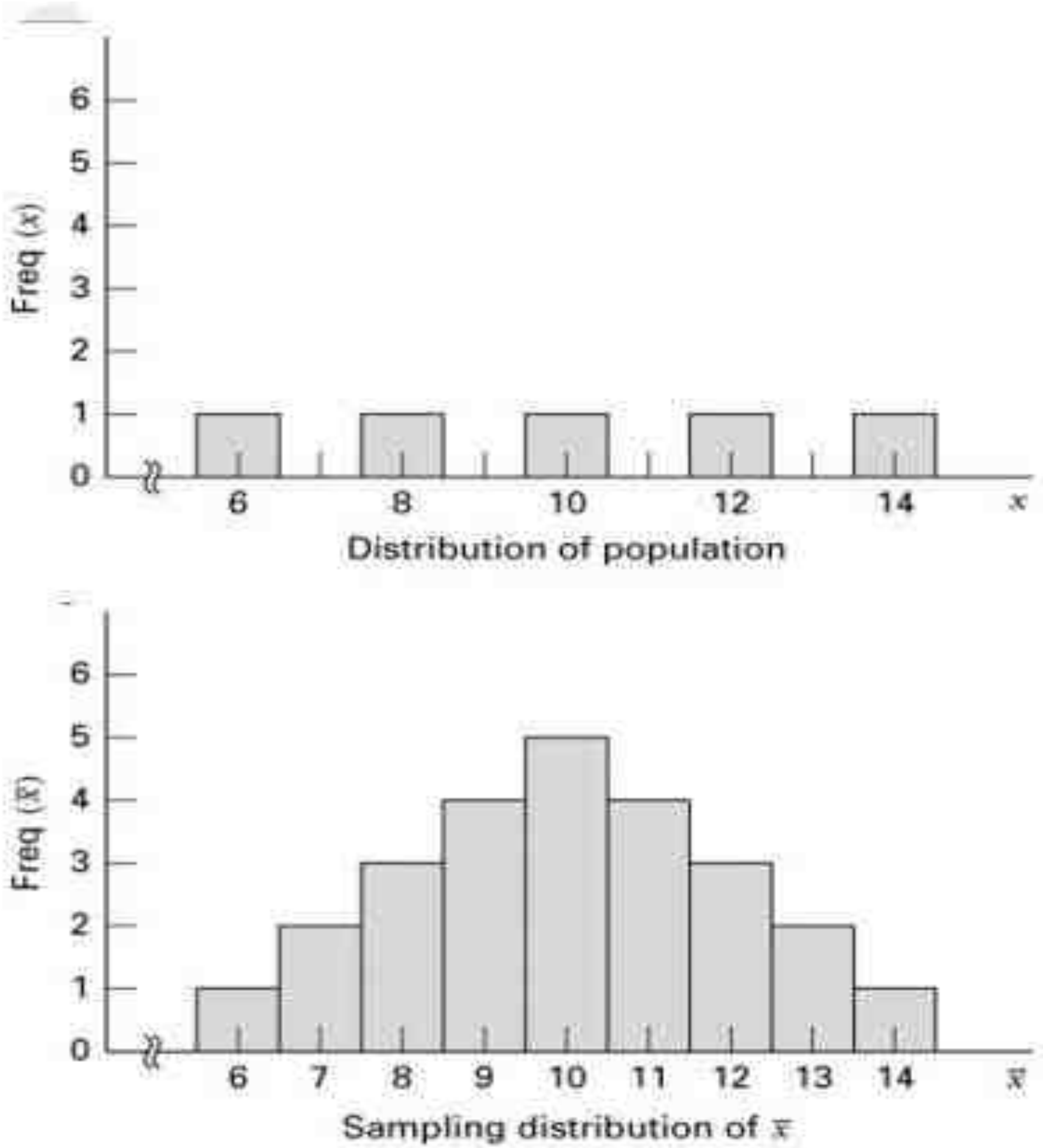
- In this example (previous slide), when **sampling is with replacement**, there are 25 possible samples. In general, when sampling is with replacement, the number of possible samples is equal to N^n.

Sampling distribution of $\bar{x}$ computed from samples in the table on the previous slide.

$\bar{x}$	Frequency	Relative Frequency
6	1	1/25
7	2	2/25
8	3	3/25
9	4	4/25
10	5	5/25
11	4	4/25
12	3	3/25
13	2	2/25
14	1	1/25
Total	25	25/25

6, 8, 10, 12, 14

- The **distribution of $\overline{x}$** plotted as a histogram, along with the **distribution of the population** are seen below:



Distribution of population and sampling distribution of $\overline{x}$.

We note the radical difference in appearance between the histogram of the population and the histogram of the sampling distribution of $\overline{x}$ Whereas the former is **uniformly distributed**, the latter **gradually rises to a peak and then drops** off with perfect symmetry.

- Now let us compute the mean, which we will call $\mu_{\bar{x}}$, of our sampling distribution. To do this we add the 25 sample means and divide by 25.

$$\mu_{\bar{x}} = \frac{\sum \bar{x}_i}{N^n} = \frac{6+7+7+8+\cdots+14}{25} = \frac{250}{25} = 10$$

- We note with interest that the mean of the sampling distribution of $\bar{x}$ has **the same value as the mean of the original population**.

Finally, we may compute the variance of $\bar{x}$ which we call as follows.

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{N^n}$$

$$= \frac{(6-10)^2 + (7-10)^2 + (7-10)^2 + \cdots + (14-10)^2}{25}$$

$$= \frac{100}{25} = 4$$

The variance of the sampling distribution is **equal to the population variance divided by the size of the sample** used to obtain the sampling distribution. That is, $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{8}{2} = 4$

- The square root of the variance of the sampling distribution, is called the **standard error** of the mean or, simply, the standard error.

$$\sqrt{\sigma_{\bar{x}}^2} = {}^{\sigma}/_{\sqrt{n}}$$

- When sampling is **from a normally distributed population**, the distribution of the sample mean will possess the following properties:

1. The distribution of will be **normal**.
2. The mean, $\mu_{\bar{x},}$ of the distribution of $\bar{x}$ will be **equal to the mean of the population** from which the samples were drawn.
3. The variance, of the distribution of $\bar{x}$ will be equal to **the variance of the population divided by the sample size**.

The Central Limit Theorem

Given a population of any nonnormal functional form with a mean μ and finite variance σ^2, the sampling distribution of $\bar{x}$ computed from samples of size n from this population, will have mean μ and variance ${}^{\sigma^2}/_{n}$ and will be approximately normally distributed **when the sample size is large**.

- A mathematical formulation of the central limit theorem is that the distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ approaches a normal distribution with mean 0 and variance 1 as $n \rightarrow \infty$.

- In the case of the sample mean, we are assured of at least an approximately normally distributed sampling distribution under **three conditions**:

(1) when sampling is from a **normally distributed population**;

(2) when sampling is from a nonnormally distributed population and our **sample is large**; and

(3) when sampling is from a population whose functional form is unknown to us as long as our **sample size is large**.

- One rule of thumb states that, in most practical situations, **a sample of size 30 is satisfactory**. In general, the approximation to normality of the sampling distribution of $\bar{x}$ **becomes better and better as the sample size increases**.

- The sample means that result when sampling is without replacement are those above the principal diagonal, which are the same as those below the principal diagonal (slide 20), **if we ignore the order ("8,6";"6,8" for example are the same) in which the observations were drawn**.
- We see that there are 10 possible samples. In general, when drawing samples of size n from a finite population of size N **without replacement**, and ignoring the order in which the sample values are drawn, **the number of possible samples is given by the combination of N things taken n at a time**.
- In our present example we have.

$$ {}_NC_n = \frac{N!}{n!(N-n)!} = \frac{5!}{2!3!} = \frac{5\cdot 4\cdot 3!}{2!3!} = 10 \text{ possible samples} $$

- The mean of the 10 sample means is $\mu_{\bar{x}} = \dfrac{\sum \bar{x}_i}{{}_NC_n} = \dfrac{7+8+9+\cdots+13}{10} = \dfrac{100}{10} = 10$

Once again the mean of the sampling distribution is equal to the population mean.

- The variance of this sampling distribution is found to be

$$\sigma_{\bar{x}}^2 = \frac{\sum(\bar{x}_i - \mu_{\bar{x}})^2}{{}_N C_n} = \frac{30}{10} = 3$$

- If we multiply the variance of the sampling distribution that would be obtained if sampling were with replacement, by the factor $(N-n)/(N-1)$, we obtain the value of the variance of the sampling distribution that results when sampling is without replacement. $$\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1} = \frac{8}{2} \cdot \frac{5-2}{4} = 3$$

- The factor $(N-n)/(N-1)$ is called **the finite population correction** and can be **ignored when the sample size is small** in comparison with the population size.

- Most practicing statisticians do not use the finite population correction **unless the sample is more than 5 percent of the size of the population**. That is, the finite population correction is usually ignored when $n/N \leq 0.05$

- The simplest application of our knowledge of the sampling distribution of the sample mean is in **computing the probability of obtaining a sample with a mean of some specified magnitude**.

Example

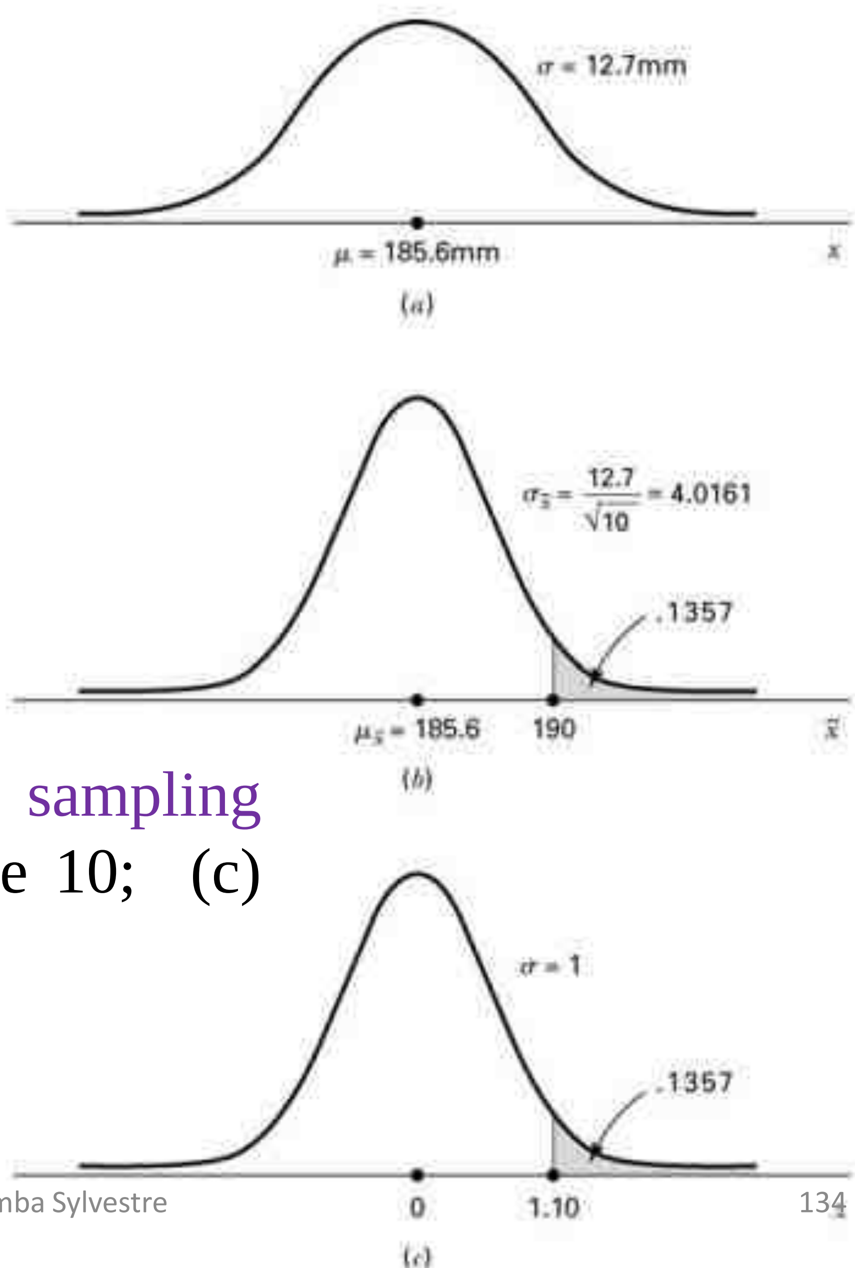
- Suppose it is known that in a certain large human population **cranial length** is approximately normally distributed with a mean of **185.6** mm and a standard deviation of **12.7** mm. What is the probability that a random sample of **size 10** from this population will have a mean **greater than 190?**

Solution:

- We know that the single sample under consideration is **one of all possible samples of size 10** that can be drawn from the population, so that the mean that it yields is **one of the** $\boxed{\bar{x}\text{'s}}$ **constituting the sampling distribution of** $\boxed{\bar{x}}$ **that, theoretically, could be derived from this population**.

- When we say that **the population is approximately normally distributed**, we assume that the sampling distribution of $\bar{x}$ will be, for all practical purposes, normally distributed. We also know that **the mean and standard deviation of the sampling distribution are equal to 185.6 and** $\sqrt{(12.7)^2/10} = 12.7/\sqrt{10} = 4.0161$ **respectively**.
- We assume that the population is large relative to the sample so that **the finite population correction can be ignored**.
- Our random variable now is $\bar{x}$, the mean of its distribution is $\mu_{\bar{x}}$, and its standard deviation is $\sqrt{\sigma_{\bar{x}}^2} = \sigma/\sqrt{n}$. By appropriately modifying the formula given previously (see slide 13), we arrive at the following formula for transforming the normal distribution of $\bar{x}$ to the standard normal distribution: $z = \dfrac{\bar{x} - \mu_{\bar{x}}}{\sigma/\sqrt{n}}$
- The probability that answers our question is represented by the area to the right of $\bar{x} = 190$ under the curve of the sampling distribution.

➢ This area is equal to the area to the right of $z = \frac{190 - 185.6}{4.0161} = \frac{4.4}{4.0161} = 1.10$



(a) **population** distribution; (b) sampling distribution of $\bar{x}$ for samples of size 10; (c) **standard normal** distribution.

- By consulting the standard normal table, we find that the area to the right of 1.10 is 0.1357; hence, we say that the probability is 0.1357 that a sample of size 10 will have a mean greater than 190.

Exercises

- If the mean and standard deviation of serum iron values for **healthy men** are 120 and 15 micrograms per 100 ml, respectively, what is the probability that a random sample of **50** normal men will yield a mean between 115 and 125 micr*ogr*ams per 100 ml?

DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

- An investigator may wish to know **something about the difference between two population means**.
- In one investigation, for example, a researcher may wish to know i … ns a
-] … ge a … wo]
- A … her (… on (
- … ns a … .

Working Table for Constructing the Distribution of the Difference Between Two Sample Means

Samples from Population 1	Samples from Population 2	Sample Means Population 1	Sample Means Population 2	All Possible Differences Between Means
n_{11}	n_{12}	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{11} - \bar{x}_{12}$
n_{21}	n_{22}	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{11} - \bar{x}_{22}$
n_{31}	n_{32}	$\bar{x}_{31}$	$\bar{x}_{32}$	$\bar{x}_{11} - \bar{x}_{32}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
$n_{N_1C_{n_1}1}$	$n_{N_2C_{n_2}2}$	$\bar{x}_{N_1C_{n_1}1}$	$\bar{x}_{N_2C_{n_2}2}$	$\bar{x}_{N_1C_{n_1}1} - \bar{x}_{N_2C_{n_2}2}$

- **A knowledge of the sampling distribution of the difference between two means is useful in investigations of this type.**

- The following example illustrates **the construction of and the characteristics** of the sampling distribution of the difference between sample means **when sampling is from two normally distributed populations**.
- Suppose we have two populations of individuals—one population (**population 1**) has **experienced some condition thought to be associated with mental retardation**, and the other population (**population 2) has not experienced the condition**. The distribution of intelligence scores in each of the two populations is believed to be approximately normally distributed with **a standard deviation of 20**. Suppose, further, that we take a sample of 15 individuals from each population and compute for each sample the mean intelligence score with the following results $\bar{x}_1 = 92$ and $\bar{x}_2 = 105$. If there is no difference between the two populations, with respect to their true mean intelligence scores, what is the probability of observing a difference this large or larger $(\bar{x}_1 - \bar{x}_2)$ between sample means?

- If we plotted the sample differences against their frequency of occurrence, we would obtain a normal distribution with a mean equal to $\mu_1 - \mu_2$, and a variance equal to $(\sigma_1^2/n_1) + (\sigma_2^2/n_2)$
- That is, the standard error of the difference between sample means would be equal to $\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$
- **We would have a normal distribution with a mean of 0 (if there is no difference between the two population means)** and a variance of $[(20)^2/15] + [(20)^2/15] = 53.3333.$

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- The area under the curve of $\bar{x}_1 - \bar{x}_2$ corresponding to the probability we seek is **the area to the left of** $\bar{x}_1 - \bar{x}_2 = 92 - 105 = -13.$
- The z value corresponding to -13, assuming that there is no difference between population means, is

$$z = \frac{-13 - 0}{\sqrt{\frac{(20)^2}{15} + \frac{(20)^2}{15}}} = \frac{-13}{\sqrt{53.3}} = \frac{-13}{7.3} = -1.78$$

- We find that the area under the standard normal curve to the left of -1.78 is equal to 0.0375. if there is no difference between population means, the probability of obtaining a difference between sample means **as large as or larger than** 13 is 0.0375.

Given two normally distributed populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, the sampling distribution of the difference, $\bar{x}_1 - \bar{x}_2$, between the means of independent samples of size n_1 and n_2 drawn from these populations is normally distributed with mean $\mu_1 - \mu_2$ and variance $\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}$.

- Suppose it has been established that for a certain type of client the average length of a home visit by a public health nurse is **45 minutes with a standard deviation of 15 minutes**, and that for a second type of client the average home visit is **30 minutes long with a standard deviation of 20 minutes**. If a nurse randomly visits 35 clients from the first and 40 from the second population, what is the probability that the average length of home visit will differ between the two groups **by 20 or more minutes?**

- No mention is made of the functional form of the two populations, so let us assume that this characteristic is unknown, or that the populations are not normally distributed. **Since the sample sizes are large (greater than 30) in both cases, we draw on the results of the central limit theorem to answer the question posed**. We know that the difference between sample means is at least approximately normally distributed with the following mean and variance:

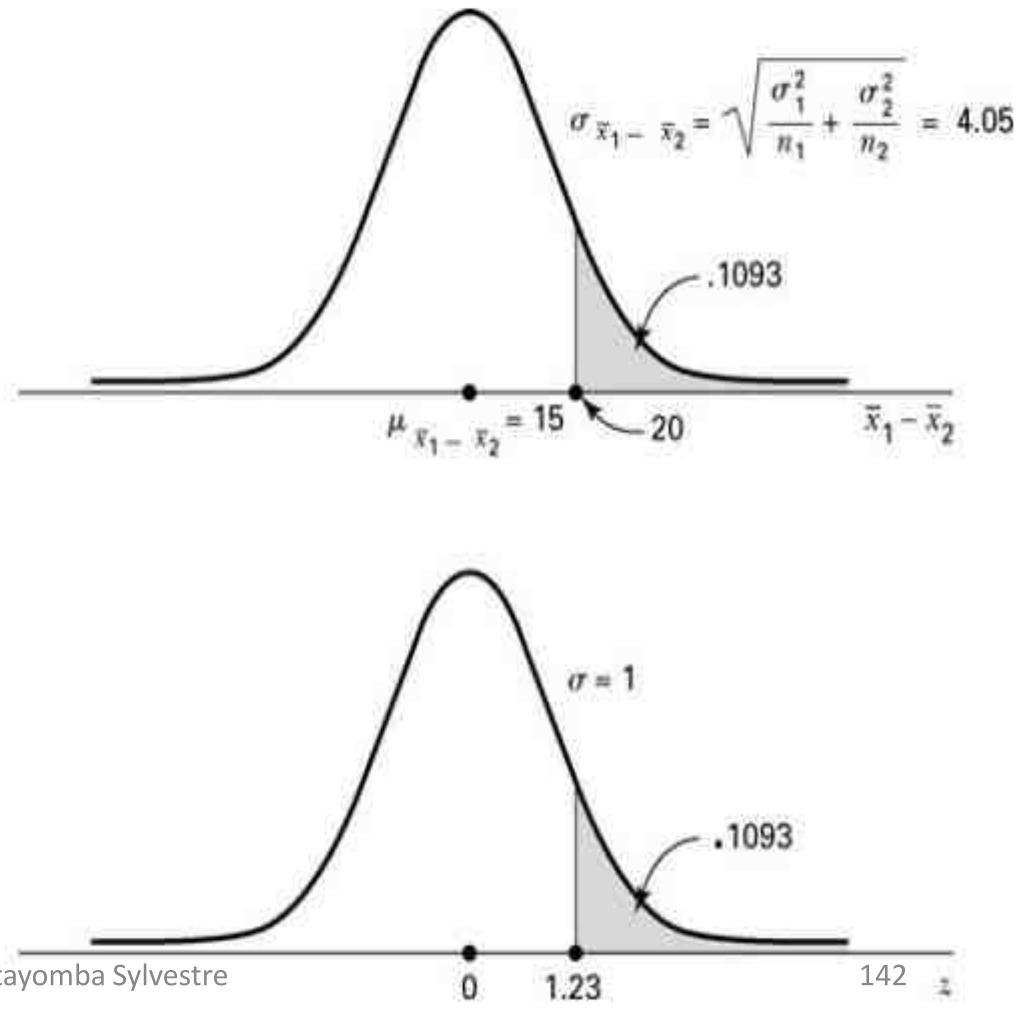
$$\mu_{\bar{x}_1-\bar{x}_2} = \mu_1 - \mu_2 = 45 - 30 = 15$$

$$\sigma^2_{\bar{x}_1-\bar{x}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(15)^2}{35} + \frac{(20)^2}{40} = 16.4286$$

The area under the curve of $\bar{x}_1 - \bar{x}_2$ that we seek is that area to the right of 20. The corresponding value of z in the standard normal is

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{20 - 15}{\sqrt{16.4286}} = \frac{5}{4.0532} = 1.23$$

- The area **to the right** of z=1.23 is 1-0.8907=0.1093
- We say, then, that the probability of the nurse's random visits resulting in a difference between the two means as great as or greater than 20 minutes is 0.1093.



DISTRIBUTION OF THE SAMPLE PROPORTION

- We are frequently interested, in the sampling distribution of a statistic, such as a sample proportion, that **results from counts or frequency data**.

Example:

- Results (A-3) from the 1999–2000 National Health and Nutrition Examination Survey (NHANES), show that **31 percent of U.S. adults ages 20–74 are obese** (obese as defined with body mass index greater than or equal to **30.0**). We designate this population proportion as $p = 0.31$. If we randomly select 150 individuals from this population, **what is the probability that the proportion in the sample who are obese will be as great as 0.40? (at least 0.40)**

Solution:

- We will designate the sample proportion by the symbol $\hat{p}$
- The variable obesity is a dichotomous variable, since an individual can be classified into one or the other of two mutually exclusive categories **obese** or **not obese**.

- When the **sample size is large**, the distribution of sample proportions is approximately normally distributed **by virtue of the central limit theorem**.
- The mean of the distribution, $\mu_{\hat{p}}$, that is, the average of all the possible sample proportions, will be equal to the true population proportion, p, and the variance of the distribution $\sigma^2_{\hat{p}}$, will be equal to $p(1-p)/n$ or pq/n.
- To answer probability questions about p, then, **we use the following formula:**

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

- The question that now arises is, **how large does the sample size have to be for the use of the normal approximation to be valid**? A widely used criterion is that both np and $\boxed{n(1-p)}$ must be greater than 5

Since both np and $n(1-p)$ are greater than $5(150 \times .31 = 46.5$ and $150 \times .69 = 103.5)$, we can say that, in this case, $\hat{p}$ is approximately normally distributed with a mean $\mu_{\hat{p}} = p = .31$ and $\sigma^2_{\hat{p}} = p(1-p)/n = (.31)(.69)/150 = .001426$.

- The probability we seek is the area under the curve of $\hat{p}$ that is **to the right of 0.40**. This area is equal to the area under the standard normal curve to the right of

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.40 - .31}{\sqrt{.001426}} = 2.38$$

- Using **Table D** we find that the area to the right of Z=2.38 is 1-0.9913=0.0087.
- **We may say, then, that the probability of observing $\hat{p} \geq 0.40$ in a random sample of size n=150 from a population in which p is 0.31 is 0.0087**. If we should, in fact, draw such a sample, most people would consider it a rare event.

CORRECTION FOR CONTINUITY

- The normal approximation **may be improved** by **the correction for continuity**, a device that makes an adjustment for the fact that **a discrete distribution is being approximated by a continuous distribution**. Suppose we let $x = n\hat{p}$ (see slide 43 part 2) the number in the sample with the characteristic of interest when the proportion is $\hat{p}$
- To apply the correction for continuity, we compute

$$z_c = \frac{\frac{x + .5}{n} - p}{\sqrt{pq/n}}, \qquad \text{for } x < np$$

or

$$z_c = \frac{\frac{x - .5}{n} - p}{\sqrt{pq/n}}, \qquad \text{for } x > np$$

The correction for continuity will not make a great deal of difference when n is large. In the above example $n\hat{p} = 150(.4) = 60$, and

$$z_c = \frac{\frac{60 - .5}{150} - .31}{\sqrt{(.31)(.69)/150}} = 2.30$$

and $P(\hat{p} \geq .40) = 1 - .9893 = .0107$, result not greatly different from that obtained without the correction for continuity.

Exercise

- Blanche Mikhail (A-4) studied the use of **prenatal care** among low-income African-American women. She found that **only 51 percent of these women had adequate prenatal care**. Let us assume that for a population of similar low-income African-American women, 51 percent had adequate prenatal care. If 200 women from this population are drawn at random, **what is the probability that less than 45 percent will have received adequate prenatal care?**

DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE PROPORTIONS

If independent random samples of size n_1 and n_2 are drawn from two populations of dichotomous variables where the proportions of observations with the characteristic of interest in the two populations are p_1 and p_2, respectively, the distribution of the difference between sample proportions, $\hat{p}_1 - \hat{p}_2$, is approximately normal with mean

$$\mu_{\hat{p}_1 - \hat{p}_2} = p_1 - p_2$$

and variance

$$\sigma^2_{\hat{p}_1 - \hat{p}_2} = \frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}$$

when n_1 and n_2 are large.

We consider n_1 and n_2 sufficiently large when $n_1 p_1, n_2 p_2, n_1(1 - p_1)$, and $n_2(1 - p_2)$ are all greater than 5.

- To answer probability questions about the difference between two sample proportions, then, we use the following formula:

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Exercise

1. The 1999 National Health Interview Survey, released in 2003 (A-7), reported that **28 percent of the subjects self-identifying as white said they had experienced lower back pain during the three months prior to the survey**. Among subjects of **Hispanic origin**, 21 percent reported lower back pain. Let us assume that 0.28 and 0.21 are the proportions for the respective races reporting lower back pain in the United States. What is the probability that independent random samples of size 100 drawn from each of the populations will yield a value of $\hat{p}_1 - \hat{p}_2$ as large as 0.10? (**at least 0.10**)

2. In the 1999 National Health Interview Survey (A-7), researchers found that among U.S. **adults ages 75 or older, 34 percent** had **lost all their natural teeth** and for U.S. Adults ages **65–74, 26 percent had lost all their natural teeth**. Assume that these proportions are the parameters for the United States **in those age groups**. If a random sample of 250 adults ages **75 or older** and an independent random sample of 200 adults ages 65–74 years old are drawn from these populations, **find the probability that the difference in percent of total natural teeth loss is less than 5 percent between the two populations.**

d.f.	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	3.078	6.3138	12.706	31.821	63.657
2	1.886	2.9200	4.3027	6.965	9.9248
3	1.638	2.3534	3.1825	4.541	5.8409
4	1.533	2.1318	2.7764	3.747	4.6041
5	1.476	2.0150	2.5706	3.365	4.0321
6	1.440	1.9432	2.4469	3.143	3.7074
7	1.415	1.8946	2.3646	2.998	3.4995
8	1.397	1.8595	2.3060	2.896	3.3554
9	1.383	1.8331	2.2622	2.821	3.2498
10	1.372	1.8125	2.2281	2.764	3.1693
11	1.363	1.7959	2.2010	2.718	3.1058
12	1.356	1.7823	2.1788	2.681	3.0545
13	1.350	1.7709	2.1604	2.650	3.0123
14	1.345	1.7613	2.1448	2.624	2.9768
15	1.341	1.7530	2.1315	2.602	2.9467
16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609
20	1.325	1.7247	2.0860	2.528	2.8453
21	1.323	1.7207	2.0796	2.518	2.8314
22	1.321	1.7171	2.0739	2.508	2.8188
23	1.319	1.7139	2.0687	2.500	2.8073
24	1.318	1.7109	2.0639	2.492	2.7969
25	1.316	1.7081	2.0595	2.485	2.7874
26	1.315	1.7056	2.0555	2.479	2.7787
27	1.314	1.7033	2.0518	2.473	2.7707
28	1.313	1.7011	2.0484	2.467	2.7633
29	1.311	1.6991	2.0452	2.462	2.7564
30	1.310	1.6973	2.0423	2.457	2.7500
35	1.3062	1.6896	2.0301	2.438	2.7239
40	1.3031	1.6839	2.0211	2.423	2.7045
45	1.3007	1.6794	2.0141	2.412	2.6896
50	1.2987	1.6759	2.0086	2.403	2.6778
60	1.2959	1.6707	2.0003	2.390	2.6603
70	1.2938	1.6669	1.9945	2.381	2.6480
80	1.2922	1.6641	1.9901	2.374	2.6388
90	1.2910	1.6620	1.9867	2.368	2.6316
100	1.2901	1.6602	1.9840	2.364	2.6260
120	1.2887	1.6577	1.9799	2.358	2.6175
140	1.2876	1.6558	1.9771	2.353	2.6114
160	1.2869	1.6545	1.9749	2.350	2.6070
180	1.2863	1.6534	1.9733	2.347	2.6035
200	1.2858	1.652[illegible]	[illegible]	2.345	2.6006
∞	1.282	1.645	1.96	2.326	2.576

$p(\chi^2_{20} \leq 31.410) = .95$

d.f.	$\chi^2_{.005}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.0000393	.000982	.00393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
4	.207	.484	.711	7.779	9.488	11.143	13.277	14.860
5	.412	.831	1.145	9.236	11.070	12.832	15.086	16.750
6	.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	18.549	21.026	23.336	26.217	28.300
13	3.565	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	11.688	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	33.196	36.415	39.364	42.980	45.558
25	10.520	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	36.741	40.113	43.194	46.963	49.645
28	12.461	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	16.791	18.493	40.256	43.773	46.979	50.892	53.672
35	17.192	20.569	22.465	46.059	49.802	53.203	57.342	60.275
40	20.707	24.433	26.509	51.805	55.758	59.342	63.691	66.766
45	24.311	28.366	30.612	57.505	61.656	65.410	69.957	73.166
50	27.991	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.535	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	107.565	113.145	118.136	124.116	128.299
100	67.328	74.222	77.929	118.498	124.342	129.561	135.807	140.169

Denominator Degrees of Freedom	Numerator Degrees of Freedom									
	10	12	15	20	24	30	40	60	120	∞
1	968.6	976.7	984.9	993.1	997.2	1001	1006	1010	1014	1018
2	39.40	39.41	39.43	39.45	39.46	39.46	39.47	39.48	39.49	39.50
3	14.42	14.34	14.25	14.17	14.12	14.08	14.04	13.99	13.95	13.90
4	8.84	8.75	8.66	8.56	8.51	8.46	8.41	8.36	8.31	8.26
5	6.62	6.52	6.43	6.33	6.28	6.23	6.18	6.12	6.07	6.02
6	5.46	5.37	5.27	5.17	5.12	5.07	5.01	4.96	4.90	4.85
7	4.76	4.67	4.57	4.47	4.42	4.36	4.31	4.25	4.20	4.14
8	4.30	4.20	4.10	4.00	3.95	3.89	3.84	3.78	3.73	3.67
9	3.96	3.87	3.77	3.67	3.61	3.56	3.51	3.45	3.39	3.33
10	3.72	3.62	3.52	3.42	3.37	3.31	3.26	3.20	3.14	3.08
11	3.53	3.43	3.33	3.23	3.17	3.12	3.06	3.00	2.94	2.88
12	3.37	3.28	3.18	3.07	3.02	2.96	2.91	2.85	2.79	2.72
13	3.25	3.15	3.05	2.95	2.89	2.84	2.78	2.72	2.66	2.60
14	3.15	3.05	2.95	2.84	2.79	2.73	2.67	2.61	2.55	2.49
15	3.06	2.96	2.86	2.76	2.70	2.64	2.59	2.52	2.46	2.40
16	2.99	2.89	2.79	2.68	2.63	2.57	2.51	2.45	2.38	2.32
17	2.92	2.82	2.72	2.62	2.56	2.50	2.44	2.38	2.32	2.25
18	2.87	2.77	2.67	2.56	2.50	2.44	2.38	2.32	2.26	2.19
19	2.82	2.72	2.62	2.51	2.45	2.39	2.33	2.27	2.20	2.13
20	2.77	2.68	2.57	2.46	2.41	2.35	2.29	2.22	2.16	2.09
21	2.73	2.64	2.53	2.42	2.37	2.31	2.25	2.18	2.11	2.04
22	2.70	2.60	2.50	2.39	2.33	2.27	2.21	2.14	2.08	2.00
23	2.67	2.57	2.47	2.36	2.30	2.24	2.18	2.11	2.04	1.97
24	2.64	2.54	2.44	2.33	2.27	2.21	2.15	2.08	2.01	1.94
25	2.61	2.51	2.41	2.30	2.24	2.18	2.12	2.05	1.98	1.91
26	2.59	2.49	2.39	2.28	2.22	2.16	2.09	2.03	1.95	1.88
27	2.57	2.47	2.36	2.25	2.19	2.13	2.07	2.00	1.93	1.85
28	2.55	2.45	2.34	2.23	2.17	2.11	2.05	1.98	1.91	1.83
29	2.53	2.43	2.32	2.21	2.15	2.09	2.03	1.96	1.89	1.81
30	2.51	2.41	2.31	2.20	2.14	2.07	2.01	1.94	1.87	1.79
40	2.39	2.29	2.18	2.07	2.01	1.94	1.88	1.80	1.72	1.64
60	2.27	2.17	2.06	1.94	1.88	1.82	1.74	1.67	1.58	1.48
120	2.16	2.05	1.94	1.82	1.76	1.69	1.61	1.53	1.43	1.31
∞	2.05	1.94	1.83	1.71	1.64	1.57	1.48	1.39	1.27	1.00

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
−3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.70
−3.60	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	−3.60
−3.50	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	−3.50
−3.40	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	−3.40
−3.30	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005	−3.30
−3.20	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007	−3.20
−3.10	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010	−3.10
−3.00	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013	−3.00
−2.90	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019	−2.90
−2.80	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026	−2.80
−2.70	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035	−2.70
−2.60	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047	−2.60
−2.50	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062	−2.50
−2.40	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082	−2.40
−2.30	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107	−2.30
−2.20	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139	−2.20
−2.10	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179	−2.10
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
−1.70	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446	−1.70
−1.60	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548	−1.60
−1.50	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668	−1.50
−1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808	−1.40
−1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968	−1.30
−1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151	−1.20
−1.10	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357	−1.10
−1.00	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587	−1.00
−0.90	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841	−0.90
−0.80	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119	−0.80
−0.70	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420	−0.70
−0.60	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743	−0.60
−0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085	−0.50
−0.40	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446	−0.40
−0.30	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821	−0.30
−0.20	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207	−0.20
−0.10	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602	−0.10
0.00	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000	0.00

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.20
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.30
0.40	.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
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	2.30
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	2.40
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	2.50
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	2.60
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	2.70
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	2.80
2.90	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	2.90
3.00	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990	3.00
3.10	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993	3.10
3.20	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995	3.20
3.30	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997	3.30
3.40	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998	3.40
3.50	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	3.50
3.60	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.60
3.70	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.70
3.80	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.80

ESTIMATION

- Estimation is **one of the two types of statistical inference**.
- Statistics, such as means and variances, can be **calculated from samples drawn from populations**.
- These statistics serve as **estimates of the corresponding population parameters**. We expect these estimates **to differ by some amount** from the parameters they estimate.
- Estimation procedures **take these differences into account**, thereby providing a foundation for statistical inference procedures.

DEFINITIONS

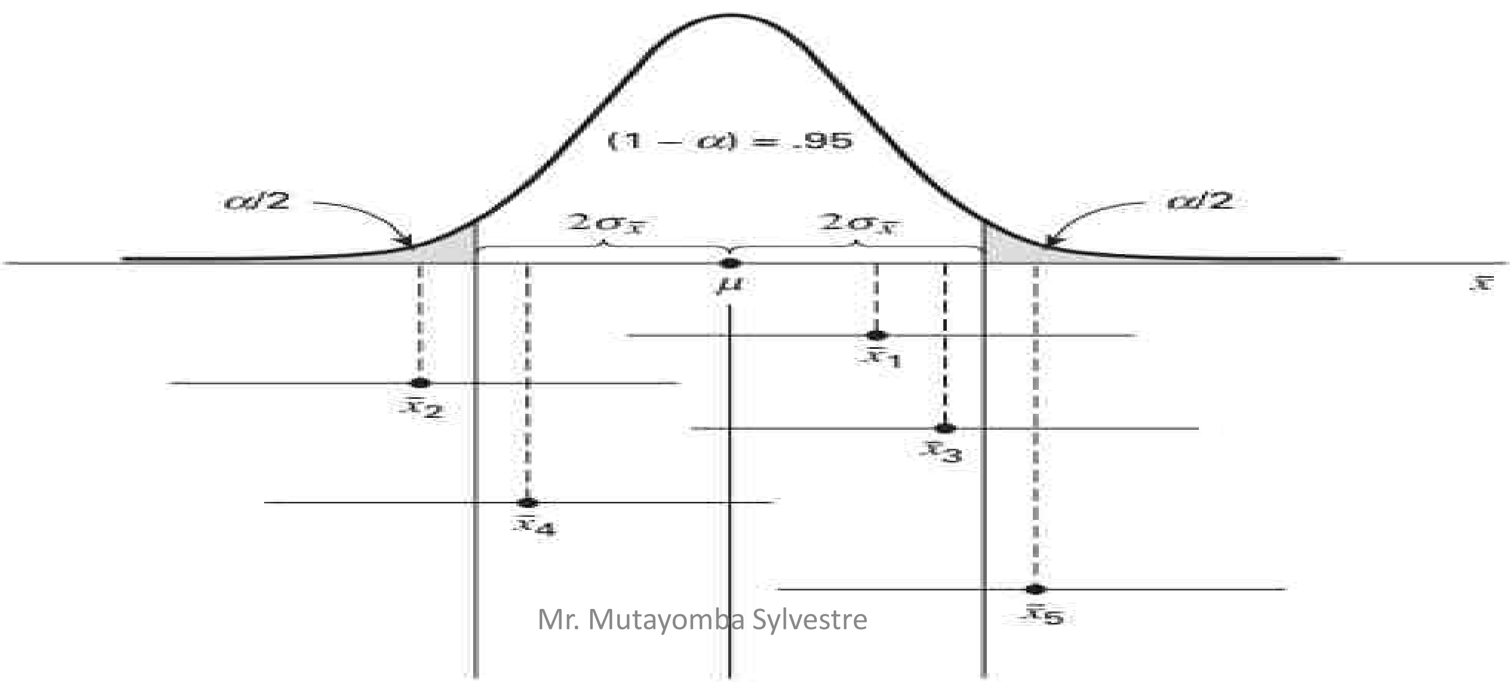
- ✓ **Statistical inference** is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.
- ✓ **A point estimate** is **a single numerical value** used to estimate the corresponding population parameter.

- **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**.
- An estimator, say, T, of the parameter is said to be an **unbiased** Estimator of θ if $E(T) = \theta$.
- For example, since **the mean of the sampling distribution of $\bar{x}$ is equal to μ** we know that $\bar{x}$ is an unbiased estimator of μ.
- **The sampled population** is the population from which one actually draws a sample.
- **The target population** is the population about which one wishes to make an inference.
- In many situations the sampled population and the target population are identical;

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 **$\bar{x}$, 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**.
- We could plot the sampling distribution if we only knew where to locate it on the $\bar{x}$ -axis.
- We know, for example, that **regardless of where the distribution of $\bar{x}$ is located**, **approximately 95 percent of the possible values** of $\bar{x}$ constituting the distribution are within two standard deviations of the mean.
- Since we do not know the value of μ **not a great deal is accomplished by the expression $\mu \pm 2\sigma_{\bar{x}}$**. **We do**, however, **have a point estimate of μ** which is $\bar{x}$.

- Suppose we constructed **intervals about every possible value of** $\bar{x}$ computed from all possible samples of size n from the population of interest.
- We would have **a large number of intervals of the form** $\bar{x} \pm 2\sigma_{\bar{x}}$ with **widths all** ***equal to the width of the interval about the unknown μ***.
- Approximately 95 percent of these intervals would have **centers falling within the interval** $\pm 2\sigma_{\bar{x}}$ **about μ**. Each of the intervals whose **centers fall within** $2\sigma_{\bar{x}}$ **of μ would contain μ**.



Interval Estimate Components

$$\text{estimator} \pm (\text{reliability coefficient}) \times (\text{standard error})$$

In particular, when sampling is from a normal distribution with known variance, an interval estimate for μ may be expressed as

$$\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$$

where $z_{(1-\alpha/2)}$ is the value of z to the left of which lies $1 - \alpha/2$ and to the right of which lies $\alpha/2$ of the area under its curve.

Probabilistic Interpretation

In repeated sampling, from a normally distributed population with a known standard deviation, $100(1 - \alpha)$ percent of all intervals of the form $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$ will in the long run include the population mean μ.

- 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)**

Practical Interpretation

When sampling is from a normally distributed population with known standard deviation, we are $100(1-\alpha)$ percent confident that the single computed interval, $\bar{x} \pm z_{(1-\alpha/2)}\sigma_{\bar{x}}$, contains the population mean μ.

In the example given here (**slide 7**) we might prefer, rather than 2, the more exact value of z, **1.96**, corresponding to a confidence coefficient of 0.95. Researchers may use any **confidence coefficient** they wish; the most frequently used values are 0.90, 0.95, and 0.99, which have associated reliability factors, respectively, of **1.645**, **1.96**, and **2.58**.

Precision

The quantity obtained by multiplying the reliability factor by the **standard error of the mean** is called **the *precision* of the estimate**. This quantity is also called **the *margin* of error**.

Exercise

1. A physical therapist wished to estimate, **with 99 percent confidence**, **the mean maximal strength** of a particular muscle in a certain group of individuals. He is willing to assume that strength scores are approximately normally distributed with **a variance of 144**. A sample of 15 subjects who participated in the experiment yielded **a mean of 84.3**
2. **Punctuality of patients in keeping appointments** is of interest to a research team. In a study of patient flow through the offices of general practitioners, it was found that **a sample of 35 patients** were **17.2 minutes** late for appointments, on the average. Previous research had shown the **standard deviation to be about 8 minutes**. The population distribution was felt to be non normal. What is the 90 percent **confidence interval** for the true mean amount of time late for appointments?

THE t DISTRIBUTION

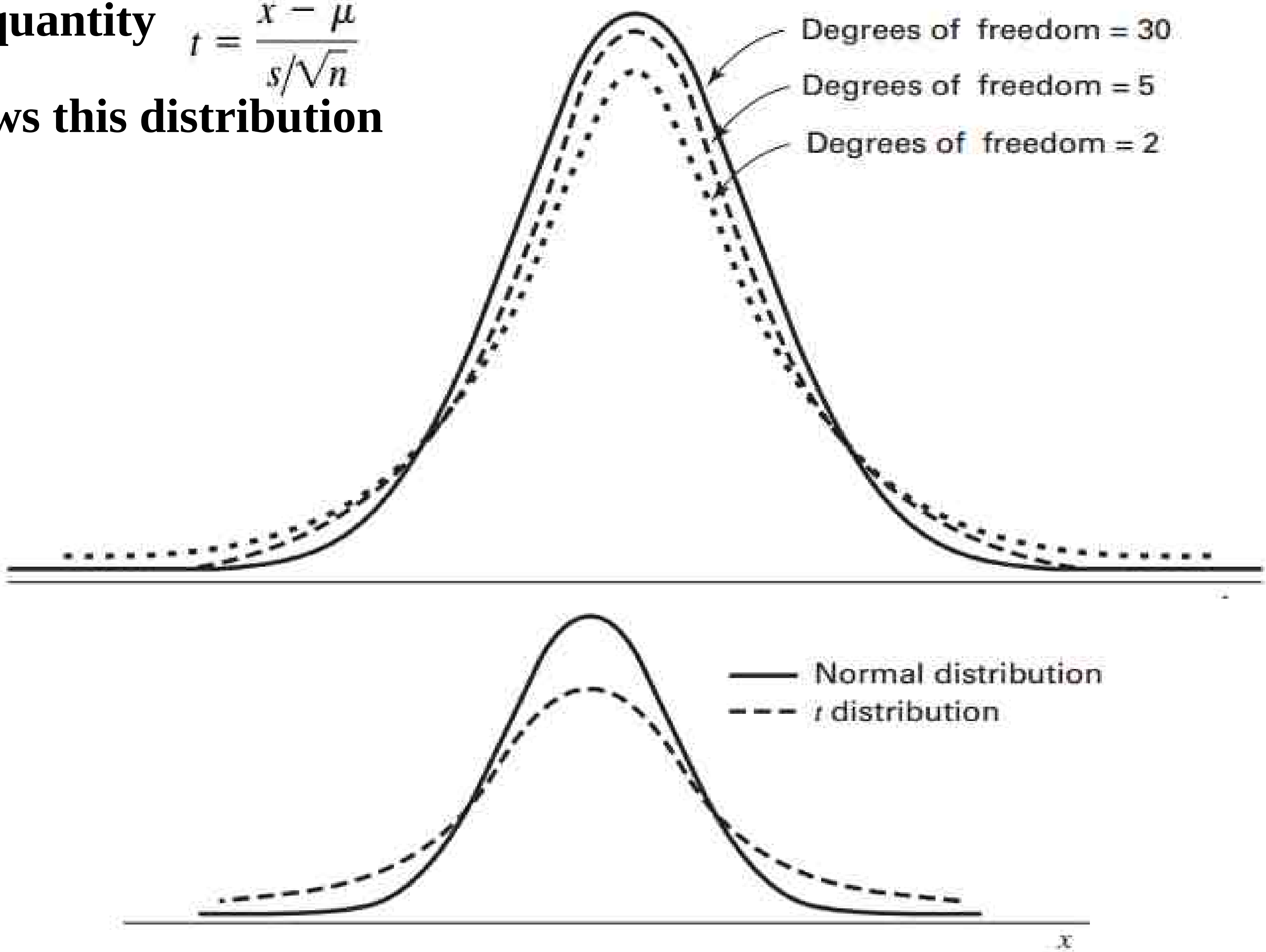
- It is the usual case that **the population variance, as well as the population mean, is unknown**. This condition presents a problem with respect to constructing confidence intervals.
- All is not lost, and the most logical solution to the problem is the **use the sample standard deviation to replace σ**.
- When the sample size is large, say, greater than 30, **our faith in s as an approximation of σ is usually substantial**, and we may be appropriately justified in using normal distribution theory to construct a confidence interval for the population mean.
- It is when we have small samples that **it becomes mandatory for us to find an alternative procedure** for constructing confidence intervals.
- An alternative, known as Student's t distribution, usually shortened to **t distribution**, is available to us.

Properties of the t Distribution The t distribution has the following properties.

1. It has a mean of 0.
2. It is symmetrical about the mean.
3. In general, it has a variance greater than 1, but the variance approaches 1 as the sample size becomes large. For $df > 2$, the variance of the t distribution is $df/(df - 2)$, where df is the degrees of freedom. Alternatively, since here $df = n - 1$ for $n > 3$, we may write the variance of the t distribution as $(n - 1)/(n - 3)$.
4. The variable t ranges from $-\infty$ to $+\infty$.
5. The t distribution is really a family of distributions, since there is a different distribution for each sample value of $n - 1$, the divisor used in computing s^2. We recall that $n - 1$ is referred to as degrees of freedom.
6. Compared to the normal distribution, the t distribution is less peaked in the center and has thicker tails.
7. The t distribution approaches the normal distribution as $n - 1$ approaches infinity.

The quantity $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

follows this distribution



Comparison of normal distribution and t distribution.

- The t distribution, like the standard normal, has been extensively **tabulated**.
- We must **take both the confidence coefficient and degrees of freedom into account** when using the table of the t distribution.

Confidence Intervals Using t

- When sampling is from a normal distribution **whose standard deviation, is unknown**, the percent confidence interval for the population mean, is given by

$$\bar{x} \pm t_{(1-\alpha/2)} \frac{s}{\sqrt{n}}$$

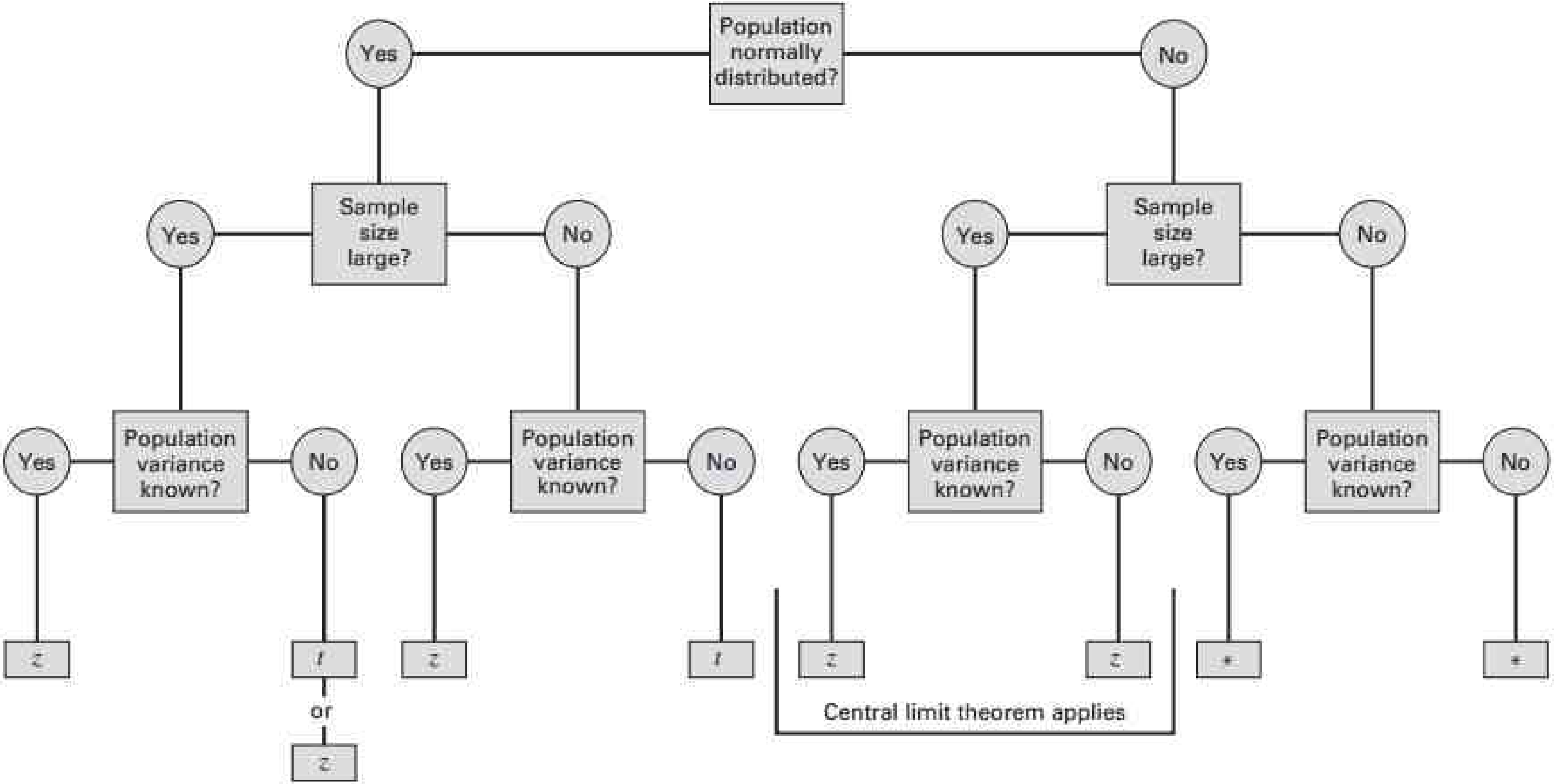
- The **strictly valid use** of the t distribution is that the sample **must be drawn from a normal distribution**.
- Experience has shown, however, that **moderate departures from this requirement can be tolerated**.
- As a consequence, the t distribution is used even when it is known that the parent population **deviates somewhat from normality**.

- Suppose a researcher , studied the effectiveness of **early weight bearing and ankle therapies** following acute repair of a ruptured Achilles tendon (the tendon that connects the heel bone to the calf muscles). One of the variables they measured following treatment was the muscle strength. In **19 subjects**, the mean of the strength was 250.8 with **standard deviation of 130.9**

we assume that *the sample* was taken from an approximately normally distributed population. Calculate 95% confident interval for the mean of the strength ?

Deciding Between z and t

- To make an appropriate choice we must consider **sample size**, whether the **sampled population** is normally distributed, and whether the **population variance is known**.



Flowchart for use in deciding between *z* and *t* when making inferences about population means.

CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION MEANS

- From each of the populations an independent random sample is drawn and, from the data of each, the sample means $\bar{x}_1$ **and** $\bar{x}_2$ respectively, are computed.
- When the population variances are known, the percent confidence interval for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- An examination of a confidence interval for the difference between population means provides **information that is helpful in deciding whether or not it is likely that the two population means are equal**. When the constructed interval **does not include zero**, we say that the interval provides **evidence that the two population means are not equal**.
- When the interval **includes zero**, we say that the population means ***may be equal***.

- The researcher team interested in the difference between **serum uric acid level** in a **patient with and without Down’s syndrome (**Down syndrome occurs when an individual has a full or partial extra copy of chromosome 21**)**. In a large hospital for the treatment of the mentally retarded, a sample of **12 individual with Down’s Syndrome** yielded a mean of $\bar{x} = 4.5$ 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 $\bar{x} = 3.4$ **.**

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 $\mu_1 - \mu_2$

Solution:

SAMPLING FROM NONNORMAL POPULATIONS

- If the sample sizes n_1 and n_2 are large. And the population variances are unknown, we use the sample variances to estimate them.

- E.g. Despite common knowledge of the adverse effects of doing so, **many women continue to smoke while pregnant**. Mayhew et al. (A-6) examined the effectiveness of a smoking cessation program for pregnant women. The mean number of cigarettes smoked daily at the close of the program by the 328 women who completed the program **was 4.3** with a standard deviation of 5.22. Among 64 women who did not complete the program, the mean number of cigarettes smoked per day at the close of the program **was 13** with a standard deviation of 8.97. We wish to construct **a 99 percent confidence interval for the difference between the means** of the populations from which the samples may be presumed to have been selected.

The t Distribution and the Difference Between Means

- When **population variances are unknown**, and we wish to estimate the difference between two population means with a confidence interval, **we can use the t distribution as a source of the reliability factor if certain assumptions are met**.
- We must know, or be willing to assume, that the **two sampled populations are normally distributed**.
- With regard to the population variances, we distinguish between two situations: (1) the situation in which the population variances are **equal**, and (2) the situation in which they are **not equal**. **Let us consider the situation where population variances are equal.**

Population Variances Equal

- The two sample variances may be considered as estimates of the same quantity, **the common variance**.
- This pooled estimate is obtained by computing **the weighted average** of the two sample variances. Each sample variance is **weighted by its degrees of freedom**.
- If the **sample sizes are equal**, this weighted average is the arithmetic mean of the two sample variances.
- The pooled estimate is given by the formula

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The standard error of the estimate, then, is given by

$$s_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

- And the 100(1-α) percent **confidence interval** for $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(1-\alpha/2)} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

❖ The purpose of **a study by Granholm et al.** (A-7) was to determine the effectiveness of an integrated outpatient dual-diagnosis treatment program for mentally ill subjects. **The authors were addressing the problem of substance abuse issues among people with severe mental disorders.** A retrospective chart review was carried out on 50 consecutive patient referrals to the Substance Abuse /Mental Illness program at the VA San Diego Healthcare System. One of the outcome variables examined was **the number of inpatient treatment days for psychiatric disorder during the year following the end of the program**. Among 18 subjects with schizophrenia, the mean number of treatment days was 4.7 with a standard deviation of 9.3. For 10 subjects with bipolar disorder, the mean number of psychiatric disorder treatment days was 8.8 with a standard deviation of 11.5. We wish to construct a 95 percent confidence interval for the difference between the means of the populations represented by these two samples.

CONFIDENCE INTERVAL FOR A POPULATION PROPORTION

- To estimate a population proportion we proceed in **the same manner** as when estimating a population mean.
- A sample is drawn from the population of interest, and the sample proportion, is computed.
- This sample proportion is **used as the point estimator of the population proportion**. A confidence interval is obtained by the general formula

$$\text{estimator} \pm (\text{reliability coefficient}) \times (\text{standard error of the estimator})$$

- When both **np** and **n(1-p)** are **greater than 5**, we may consider the sampling distribution of $\hat{p}$ to be quite close to the normal distribution.
- **When this condition (above) is met**, our reliability coefficient is some value of z from the standard normal distribution.

- The standard error, we have seen, is equal to $\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$.
- Since p, the parameter we are trying to estimate, is **unknown**, we must use $\hat{p}$ as an estimate. Thus, we estimate $\sigma_{\hat{p}}$ by $\sqrt{\hat{p}(1-\hat{p})/n}$,

, and our100(1-α) percent confidence interval for p is given by

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

- The Pew Internet and American Life Project (A-13) reported in 2003 that **18 percent of Internet users have used it to search for information regarding experimental treatments or medicines**. The sample consisted of **1220 adult Internet users**, and information was collected from telephone interviews. We wish to construct a 95 percent **confidence interval** for the proportion of Internet users in the sampled population who have searched for information on experimental treatments or medicines.
- What if a 98% confident interval for the above question is constructed

CONFIDENCE INTERVAL FOR THE DIFFERENCE BETWEEN TWO POPULATION PROPORTIONS

- We may want **to compare**, for example, men and women, two age groups, two socioeconomic groups, or two diagnostic groups with respect to **the proportion possessing some characteristic of interest**. An unbiased point estimator of the difference between two population proportions is provided by the difference between $\hat{p}_1 - \hat{p}_2$. proportions,
- **When n1 and n2 are large and the population proportions are not too close to 0 or 1**, the central limit theorem applies and normal distribution theory may be employed to obtain confidence intervals.
- The standard error of the estimate usually must be estimated by

$$\hat{\sigma}_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Because, as a rule, the population proportions are unknown.
- A 100(1-α) percent confidence interval for p1-p2 is given by:

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

- Connor et al. (A-17) investigated gender differences in proactive and reactive aggression in **a sample of 323 children and adolescents (68 females and 255 males)**. In the sample, **31 of the females and 53 of the males reported sexual abuse**. We wish to construct a 99 percent confidence interval for the difference between the proportions of sexual abuse in the two sampled populations.

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**.
- 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 (Reliability coefficient)x(Standard error of estimator)** since the total width of the interval is **twice** this amount.
- 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? That depends on the **size of σ**, the population standard deviation, the desired **degree of reliability, *z***, and the desired **interval width, 2*d***.
- Let us suppose we want an interval that extends **d** units on either side of the estimator. We can write **d = (reliability coefficient)x(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 $n = \frac{z^2\sigma^2}{d^2}$

- When sampling is **without replacement from a small finite population**, the finite population correction is required and

$$d = z\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

which, when solved for n, gives

$$n = \frac{Nz^2\sigma^2}{d^2(N-1) + z^2\sigma^2}$$

Estimating $\boldsymbol{\sigma^2}$

- The formulas for sample size require knowledge of $\boldsymbol{\sigma^2}$ but, as has been pointed out, **the population variance is, as a rule, unknown**. As a result, $\boldsymbol{\sigma^2}$ has to be estimated. The most frequently used sources of estimates for $\boldsymbol{\sigma^2}$ 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 $\boldsymbol{\sigma^2}$. **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**.

❖ 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**.

- 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 0.95 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, σ=20 and d=5. Let us assume that the population of interest is large so that the statistician may ignore the finite population correction. the value of n is found to be? (Find the solution)

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**. We make use of the fact that **one-half the desired interval, d, may be set equal to the product of the reliability coefficient and the standard error**.
- Assuming that random sampling and conditions warranting approximate normality of the distribution of $\hat{p}$ leads to the following formula for n when sampling is **with replacement**, when sampling is **from an infinite population**, or when the **sampled population is large enough** to make use of the finite population correction unnecessary,

$$n = \frac{z^2pq}{d^2}$$

- If the finite population correction cannot be disregarded, the proper formula for n is

$$n = \frac{Nz^2pq}{d^2(N-1)+z^2pq}$$

Estimating p

- Both formulas require **knowledge of p**, *the proportion in the population possessing the characteristic of interest*. Since this is **the parameter we are trying to estimate**, it, obviously, will be unknown. 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**.

- 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 **0.35**. A 95 percent confidence interval is desired with **d=0.05.** What size sample of families should be selected?

CONFIDENCE INTERVAL FOR THE VARIANCE OF A NORMALLY DISTRIBUTED POPULATION

- Let us see if the sample variance is an unbiased estimator of the population variance. To be unbiased, **the average value of the sample variance over all possible samples must be equal to the population variance.** That is, the expression $E(s^2) = \sigma^2$ must hold.
- All possible samples of size 2 from the population consisting of the values 6, 8, 10, 12, and 14 are found on slide 25 of part III notes.
- Two measures of dispersion for this population were computed as follows:

$$\sigma^2 = \frac{\Sigma(x_i - \mu)^2}{N} = 8 \quad \text{and} \quad S^2 = \frac{\Sigma(x_i - \mu)^2}{N - 1} = 10$$

- If we compute **the sample variance** $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ **for each of the possible samples** we obtain the sample variances shown in table on next slide.

		Second Draw				
		6	8	10	12	14
	6	0	2	8	18	32
	8	2	0	2	8	18
First Draw	10	8	2	0	2	8
	12	18	8	2	0	2
	14	32	18	8	2	0

- If sampling is with replacement, the expected value of s^2 is obtained by taking **the mean of all sample variances in table above.** $E(s^2) = \frac{\sum s_i^2}{N^n} = \frac{0 + 2 + \cdots + 2 + 0}{25} = \frac{200}{25} = 8$

- If we consider the case where sampling is without replacement, the expected value of S^2 is obtained by taking the mean of all variances above (or below) the principal diagonal. That is

$$E(s^2) = \frac{\sum s_i^2}{{}_N C_n} = \frac{2 + 8 + \cdots + 2}{10} = \frac{100}{10} = 10$$

In general

$E(s^2) = \sigma^2$ *when sampling is with replacement*

$E(s^2) = S^2$ *when sampling is without replacement*

- When N is large, **N-1 and N** will be approximately equal and, consequently, σ^2 and s^2 will be approximately equal. These results justify our use of $s^2 = \Sigma(x_i - \bar{x})^2/(n-1)$ when computing the sample variance.

Interval Estimation of a Population Variance

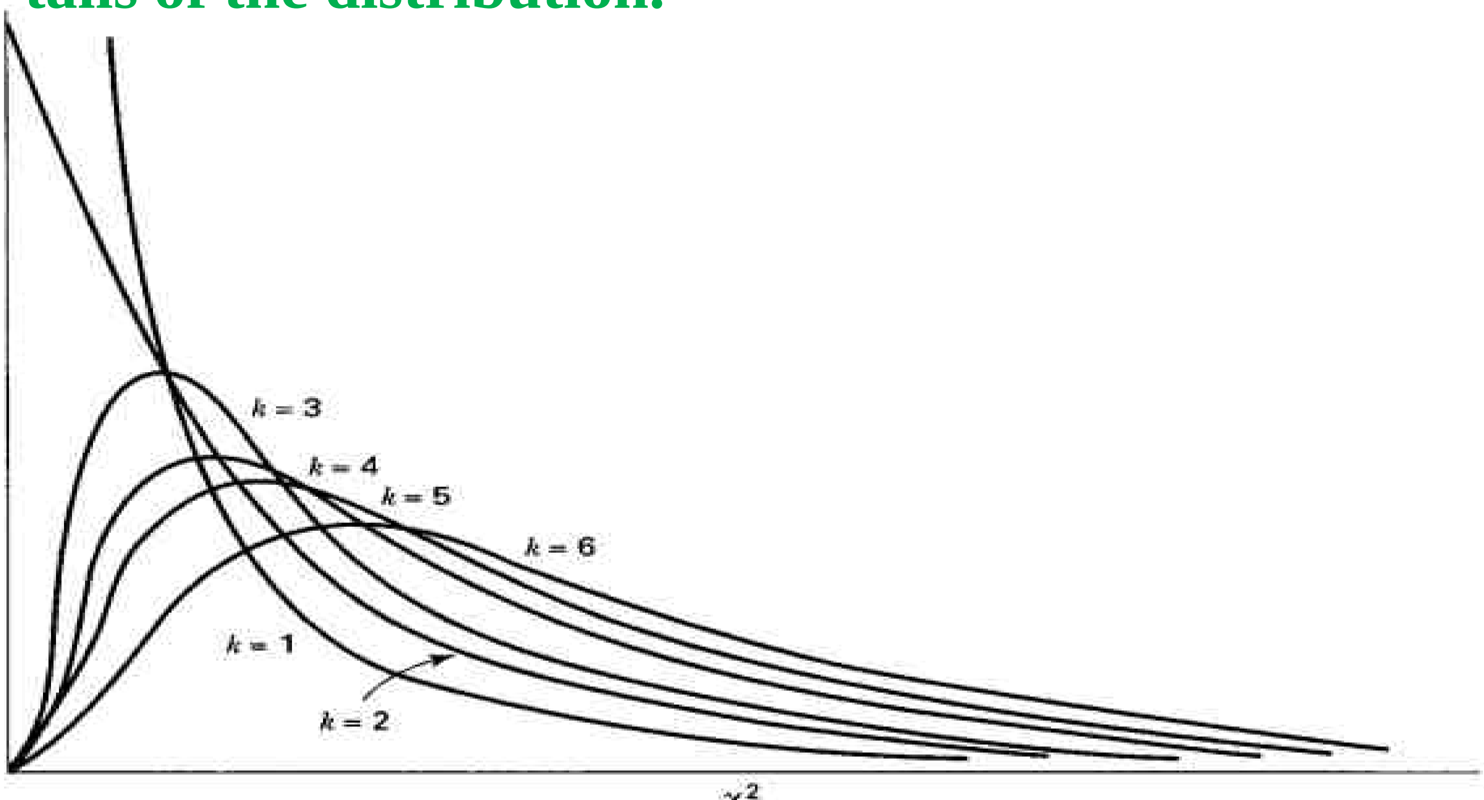
- **With a point estimate available**, it is logical to inquire about the construction of a confidence interval for a population variance.
- Whether we are successful in constructing a confidence interval for σ^2 will depend on our **ability to find an appropriate sampling distribution.**

The Chi-Square Distribution

- Confidence intervals for σ^2 are usually based on the sampling distribution of $\boxed{(n-1)s^2/\sigma^2}$
- If samples of size n are drawn from a normally distributed population, this quantity has a distribution known as the **chi-square (χ^2)** distribution with $n-1$ degrees of freedom.
- It is **useful in finding confidence intervals** for σ^2 when the assumption that the population is **normally distributed holds true**.
- **Figure aft next slide** shows **chi-square distributions for several values of degrees of freedom**.
- Percentiles of the chi-square distribution are given in Appendix Table F.
- The **column headings** give the values of (χ^2) to the left of which lies a proportion of the total area under the curve equal to the subscript of (χ^2). **The row** labels are the degrees of freedom.

d.f.	$\chi^2_{.005}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.0000393	.000982	.00393	2.706	3.841	5.024	6.635	7.879
2	.0100	.0506	.103	4.605	5.991	7.378	9.210	10.597
3	.0717	.216	.352	6.251	7.815	9.348	11.345	12.838
4	.207	.484	.711	7.779	9.488	11.143	13.277	14.860
5	.412	.831	1.145	9.236	11.070	12.832	15.086	16.750
6	.676	1.237	1.635	10.645	12.592	14.449	16.812	18.548
7	.989	1.690	2.167	12.017	14.067	16.013	18.475	20.278
8	1.344	2.180	2.733	13.362	15.507	17.535	20.090	21.955
9	1.735	2.700	3.325	14.684	16.919	19.023	21.666	23.589
10	2.156	3.247	3.940	15.987	18.307	20.483	23.209	25.188
11	2.603	3.816	4.575	17.275	19.675	21.920	24.725	26.757
12	3.074	4.404	5.226	18.549	21.026	23.336	26.217	28.300
13	3.565	5.009	5.892	19.812	22.362	24.736	27.688	29.819
14	4.075	5.629	6.571	21.064	23.685	26.119	29.141	31.319
15	4.601	6.262	7.261	22.307	24.996	27.488	30.578	32.801
16	5.142	6.908	7.962	23.542	26.296	28.845	32.000	34.267
17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401
22	8.643	10.982	12.338	30.813	33.924	36.781	40.289	42.796
23	9.260	11.688	13.091	32.007	35.172	38.076	41.638	44.181
24	9.886	12.401	13.848	33.196	36.415	39.364	42.980	45.558
25	10.520	13.120	14.611	34.382	37.652	40.646	44.314	46.928
26	11.160	13.844	15.379	35.563	38.885	41.923	45.642	48.290
27	11.808	14.573	16.151	36.741	40.113	43.194	46.963	49.645
28	12.461	15.308	16.928	37.916	41.337	44.461	48.278	50.993
29	13.121	16.047	17.708	39.087	42.557	45.722	49.588	52.336
30	13.787	16.791	18.493	40.256	43.773	46.979	50.892	53.672
35	17.192	20.569	22.465	46.059	49.802	53.203	57.342	60.275
40	20.707	24.433	26.509	51.805	55.758	59.342	63.691	66.766
45	24.311	28.366	30.612	57.505	61.656	65.410	69.957	73.166
50	27.991	32.357	34.764	63.167	67.505	71.420	76.154	79.490
60	35.535	40.482	43.188	74.397	79.082	83.298	88.379	91.952
70	43.275	48.758	51.739	85.527	90.531	95.023	100.425	104.215
80	51.172	57.153	60.391	96.578	101.879	106.629	112.329	116.321
90	59.196	65.647	69.126	107.565	113.145	118.136	124.116	128.299
100	67.328	74.222	77.929	118.498	124.342	129.561	135.807	140.169

To obtain a 100(1-α) percent confidence interval for σ^2 we first obtain the 100(1-α) percent confidence interval for $(n-1)s^2/\sigma^2$. To do this, we select the values of (χ^2) from Appendix Table F in such a way that $\alpha/2$ is to the left of the smaller value and $\alpha/2$ is to the right of the larger value. In other words, the two values of (χ^2) are **selected in such a way that α is divided equally between the two tails of the distribution.**



Chi-square distributions for several values of degrees of freedom k.

- We may designate these two values of χ^2 as $\chi^2_{\alpha/2}$ and $\chi^2_{1-(\frac{\alpha}{2})}$ respectively. The 100(1-α) percent confidence interval for $(n-1)\,{s^2}/{\sigma^2}$ then, is given by $\chi^2_{\alpha/2} < \frac{(n-1)s^2}{\sigma^2} < \chi^2_{1-(\alpha/2)}$ or $\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha/2}}$

which is the **percent confidence interval** for σ^2 . If we take the square root of each of its term we have the following percent confidence interval for σ, the population standard deviation:

$$\sqrt{\frac{(n-1)s^2}{\chi^2_{1-(\alpha/2)}}} < \sigma < \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}}}$$

- In a study of the effectiveness of a **gluten-free diet** in **first-degree relatives** of patients with type I diabetics, Hummel et al. (A-22) placed **seven subjects on a gluten-free diet for 12 months**. Prior to the diet, they took baseline measurements of several antibodies and autoantibodies, one of which was the diabetes related insulin auto antibody (IAA).

The IAA levels were measured by **radiobinding** assay. The seven subjects had IAA units of 9.7, 12.3, 11.2, 5.1, 24.8, 14.8, 17.7

- We wish **to estimate from the data in this sample the variance of the IAA units in the population from which the sample was drawn and construct a 95 percent confidence interval for this estimate**.

CONFIDENCE INTERVAL FOR THE RATIO OF THE VARIANCES OF TWO NORMALLY DISTRIBUTED POPULATIONS

- It is frequently of interest **to compare two variances**, and one way to do this is to form their ratio, **If two variances are equal, their ratio will be equal to 1.**
- The use of the ratio of two population variances for determining equality of variances has been formalized into a statistical test.
- The distribution of this test provides **test values for determining if the ratio exceeds the value 1 to a large enough extent that we may conclude that the variances are not equal**.
- If the confidence interval for the ratio of two population variances **includes 1**, we conclude that the **two population variances may, in fact, be equal**. Again, since this is a form of inference, we must rely on some sampling distribution, and this time the distribution of $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ is utilized **provided certain assumptions are met**.

- The assumptions are that s_1^2 and s_2^2 are computed from independent samples of size n_1 and n_2 respectively, drawn from two normally distributed populations. **We use S_1^2 to designate the larger of the two sample variances.**

The F Distribution

- If the assumptions are met, $(s_1^2/\sigma_1^2)/(s_2^2/\sigma_2^2)$ follows a distribution known as the F distribution.
- This distribution **depends on two-degrees-of freedom values**, one corresponding to the value of $n_1 - 1$ used in computing s_1^2 and the other corresponding to the value of $n_2 - 1$ used in computing s_2^2, **the numerator degrees of freedom and the denominator degrees of freedom.**
- Table G contains, for specified combinations of degrees of freedom and values of α, F values to the right of which lies α/2 of the area under the curve of F.

A Confidence Interval for σ_1^2/σ_2^2

- To find the 100(1-α) percent confidence interval fo σ_1^2/σ_2^2 we begin with the expression $$F_{\alpha/2} < \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} < F_{1-(\alpha/2)}$$

- where $F_{\alpha/2}$ and $F_{1-(\alpha/2)}$ are the values from the F table to the left and right of which, respectively, lies α/2 of the area under the curve. The middle term of this expression may be rewritten so that the entire expression is $$F_{\alpha/2} < \frac{s_1^2}{s_2^2} \cdot \frac{\sigma_2^2}{\sigma_1^2} < F_{1-(\alpha/$$ Which can be written as

$$\frac{s_1^2/s_2^2}{F_{1-(\alpha/2)}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2/s_2^2}{F_{\alpha/2}}$$

- Allen and Gross (A-25) examine **toe flexors strength** in subjects with plantar fasciitis (pain from heel spurs, or **general heel pain**), a common condition in patients with musculoskeletal problems. Inflammation of the plantar fascia is often costly to treat and **frustrating for both the patient and the clinician**. One of the baseline measurements was the **body mass index** (BMI). For the **16 women** in the study, the standard deviation for BMI **was 8.1** and for **four men** in the study, the standard deviation **was 5.9**. We wish to construct **a 95 percent confidence interval for the ratio of the variances of the two populations from which we presume these samples were drawn**.

- There exists a relationship that enables us **to compute the lower percentile values** from our limited table. The relationship is as follows:

$$F_{\alpha,df_1,df_2} = \frac{1}{F_{1-\alpha,df_2,df_1}}$$

- We proceed as follows: **Interchange the numerator and denominator degrees of freedom and locate the appropriate value of F**. For the problem at hand we locate 4.15, which is at the intersection of the column headed 3 and the row labeled 15. We now take the reciprocal of this value, 1/4.15=0.24096
- In summary, the lower confidence limit (LCL) and upper confidence limit (UCL) σ_1^2/σ_2^2 are as follows:

$$LCL = \frac{s_1^2}{s_2^2}\frac{1}{F_{(1-\alpha/2),df_1,df_2}}$$

$$UCL = \frac{s_1^2}{s_2^2}F_{1-(\alpha/2),df_2,df_1}$$

Critical Values of F_{max} for Hartley's Homogeneity of Variance Test

The upper value in each box is for $\alpha = 0.05$. The lower value is for $\alpha = 0.01$. The test assumes that there are equal sample sizes in each group (n). For unequal sample sizes, use the smaller of the df for the two variances being compared.

DF (n-1)	Number of treatments (k)										
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
2	39.0 199	87.5 448	142 729	202 1036	266 1362	333 1705	403 2063	475 2432	550 2813	626 3204	714 3605
3	15.4 47.5	27.8 85.0	39.2 120	50.7 151	62.0 184	72.9 216	83.5 249	93.9 281	104 310	114 337	124 361
4	9.6 23.2	15.5 37.0	20.6 49.0	25.2 59	29.5 69	33.6 79	37.5 89	41.1 97	44.6 106	48.0 113	51.4 120
5	7.2 14.9	10.8 22.0	13.7 28.0	16.3 33	18.7 38	20.8 42	22.9 46	24.7 50	26.5 54	28.2 57	29.9 60
6	5.82 11.1	8.38 15.5	10.4 19.1	12.1 22	13.7 25	15.0 27	16.3 30	17.5 32	18.6 34	19.7 36	20.7 37
7	0.99 8.89	6.94 12.1	8.44 14.5	9.70 16.5	10.8 18.4	11.8 20	12.7 22	13.5 23	14.3 24	15.1 26	15.8 27
8	4.43 7.50	6.00 9.90	7.18 11.7	8.12 13.2	9.03 14.5	9.78 15.8	10.5 16.9	11.1 17.9	11.7 18.9	12.2 19.8	12.7 21
9	4.03 6.54	5.34 8.50	6.31 9.9	7.11 11.1	7.80 12.1	8.41 13.1	8.95 13.9	9.45 14.7	9.91 15.3	10.3 16.0	10.7 16.6
10	3.72 5.85	4.85 7.40	5.67 8.6	6.34 9.6	6.92 10.4	7.42 11.1	7.87 11.8	8.28 12.4	8.66 12.9	9.01 13.4	9.34 13.9
12	3.28 4.91	4.16 6.1	4.75 6.9	5.30 7.6	5.72 8.2	6.09 8.7	6.42 9.1	6.72 9.5	7.00 9.9	7.25 10.2	7.43 10.6
15	2.86 4.07	3.54 4.9	4.01 5.5	4.37 6.0	4.68 6.4	4.95 6.7	5.19 7.1	5.40 7.3	5.59 7.5	5.77 7.8	5.95 8.0
20	2.46 3.32	2.95 3.8	3.29 4.3	3.54 4.6	3.76 4.9	3.94 5.1	4.10 5.3	4.24 5.5	4.37 5.6	4.49 5.8	4.59 5.9
30	2.07 2.63	2.40 3.0	2.61 3.3	2.78 3.4	2.91 3.6	3.02 3.7	3.12 3.8	3.21 3.9	3.29 4.0	3.36 4.1	3.39 4.2
60	1.67 1.96	1.85 2.2	1.96 2.3	2.04 2.4	2.11 2.4	2.17 2.5	2.22 2.5	2.26 2.6	2.30 2.6	2.33 2.7	2.36 2.7
∞	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00	1.00 1.00

z	−0.09	−0.08	−0.07	−0.06	−0.05	−0.04	−0.03	−0.02	−0.01	0.00	z
−3.80	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.80
−3.70	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	−3.70
−3.60	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0002	.0002	−3.60
−3.50	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	−3.50
−3.40	.0002	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	−3.40
−3.30	.0003	.0004	.0004	.0004	.0004	.0004	.0004	.0005	.0005	.0005	−3.30
−3.20	.0005	.0005	.0005	.0006	.0006	.0006	.0006	.0006	.0007	.0007	−3.20
−3.10	.0007	.0007	.0008	.0008	.0008	.0008	.0009	.0009	.0009	.0010	−3.10
−3.00	.0010	.0010	.0011	.0011	.0011	.0012	.0012	.0013	.0013	.0013	−3.00
−2.90	.0014	.0014	.0015	.0015	.0016	.0016	.0017	.0018	.0018	.0019	−2.90
−2.80	.0019	.0020	.0021	.0021	.0022	.0023	.0023	.0024	.0025	.0026	−2.80
−2.70	.0026	.0027	.0028	.0029	.0030	.0031	.0032	.0033	.0034	.0035	−2.70
−2.60	.0036	.0037	.0038	.0039	.0040	.0041	.0043	.0044	.0045	.0047	−2.60
−2.50	.0048	.0049	.0051	.0052	.0054	.0055	.0057	.0059	.0060	.0062	−2.50
−2.40	.0064	.0066	.0068	.0069	.0071	.0073	.0075	.0078	.0080	.0082	−2.40
−2.30	.0084	.0087	.0089	.0091	.0094	.0096	.0099	.0102	.0104	.0107	−2.30
−2.20	.0110	.0113	.0116	.0119	.0122	.0125	.0129	.0132	.0136	.0139	−2.20
−2.10	.0143	.0146	.0150	.0154	.0158	.0162	.0166	.0170	.0174	.0179	−2.10
−2.00	.0183	.0188	.0192	.0197	.0202	.0207	.0212	.0217	.0222	.0228	−2.00
−1.90	.0233	.0239	.0244	.0250	.0256	.0262	.0268	.0274	.0281	.0287	−1.90
−1.80	.0294	.0301	.0307	.0314	.0322	.0329	.0336	.0344	.0351	.0359	−1.80
−1.70	.0367	.0375	.0384	.0392	.0401	.0409	.0418	.0427	.0436	.0446	−1.70
−1.60	.0455	.0465	.0475	.0485	.0495	.0505	.0516	.0526	.0537	.0548	−1.60
−1.50	.0559	.0571	.0582	.0594	.0606	.0618	.0630	.0643	.0655	.0668	−1.50
−1.40	.0681	.0694	.0708	.0721	.0735	.0749	.0764	.0778	.0793	.0808	−1.40
−1.30	.0823	.0838	.0853	.0869	.0885	.0901	.0918	.0934	.0951	.0968	−1.30
−1.20	.0985	.1003	.1020	.1038	.1056	.1075	.1093	.1112	.1131	.1151	−1.20
−1.10	.1170	.1190	.1210	.1230	.1251	.1271	.1292	.1314	.1335	.1357	−1.10
−1.00	.1379	.1401	.1423	.1446	.1469	.1492	.1515	.1539	.1562	.1587	−1.00
−0.90	.1611	.1635	.1660	.1685	.1711	.1736	.1762	.1788	.1814	.1841	−0.90
−0.80	.1867	.1894	.1922	.1949	.1977	.2005	.2033	.2061	.2090	.2119	−0.80
−0.70	.2148	.2177	.2206	.2236	.2266	.2296	.2327	.2358	.2389	.2420	−0.70
−0.60	.2451	.2483	.2514	.2546	.2578	.2611	.2643	.2676	.2709	.2743	−0.60
−0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085	−0.50
−0.40	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446	−0.40
−0.30	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821	−0.30
−0.20	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207	−0.20
−0.10	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602	−0.10
0.00	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000	0.00

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359	0.00
0.10	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753	0.10
0.20	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141	0.20
0.30	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517	0.30
0.40	.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
2.30	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916	2.30
2.40	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936	2.40
2.50	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952	2.50
2.60	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964	2.60
2.70	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974	2.70
2.80	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981	2.80
2.90	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986	2.90
3.00	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990	3.00
3.10	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993	3.10
3.20	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995	3.20
3.30	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997	3.30
3.40	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998	3.40
3.50	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	3.50
3.60	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.60
3.70	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.70
3.80	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999	3.80

HYPOTHESIS TESTING

- **Interval estimation** and **hypothesis testing** are **based on similar concepts.** In fact, confidence intervals may be used to arrive at the same conclusions that are reached through the use of hypothesis tests.
- The purpose of hypothesis testing is to aid the clinician, researcher, or administrator in **reaching a conclusion concerning a population by examining a sample from that population**.

DEFINITION

- **A hypothesis** may be defined simply as **a statement about one or more populations**.
- ✓ A hospital administrator may hypothesize that **the average length of stay of patients admitted to the hospital is 5 days**; a public health nurse may hypothesize that **a particular educational program will result in improved communication between nurse and patient**; a physician may hypothesize that **a certain drug will be effective in 90 percent of the cases for which it is used**. By means of hypothesis testing one **determines whether or not such statements are *compatible with the available data***.

- **The research hypothesis** is the ***conjecture or supposition*** that motivates the research.
  - A public health nurse, for example, may have noted that **certain clients responded more readily to a particular type of health education program**.
  - A physician may **recall numerous instances in which certain combinations of therapeutic measures were more effective than any one of them alone**.
  - ***Research projects*** often result from the desire of such health practitioners ***to determine whether or not their theories or suspicions can be supported when subjected to the rigors of scientific investigation***. Research hypotheses lead directly to statistical hypotheses.
- **Statistical hypotheses** are hypotheses that are stated in such a way that they may be ***evaluated*** by appropriate statistical techniques.

Hypothesis Testing Steps

1. **Data**: The nature of the data that form the basis of the testing procedures must be understood, since this **determines the particular test to be employed.**
2. **Assumptions:** A general procedure is modified depending on the assumptions. These include assumptions about the normality of the population distribution, equality of variances, and independence of samples.
3. **Hypotheses**: There are **two statistical hypotheses** involved in hypothesis testing, **The null hypothesis** designated by the symbol H_0 and The **alternative hypothesis** designated by the symbol H_A

The null hypothesis is sometimes referred to as **a hypothesis of no difference**. The alternative hypothesis is a statement of **what we will believe is true if our sample data cause us to reject the null hypothesis**. Usually **the alternative hypothesis** and the **research hypothesis are the same**, and in fact the two terms are used interchangeably.

1. **Data**: The natu
procedures mus
particular test t
2. **Assumptions:** A
assumptions. Th
the population
independence of
3. **Hypotheses**: T
involved in hypo
the symbol H_0
symbol H_A
The null hypothesi
no difference. The
will believe is true
hypothesis. Usually
hypothesis are th
interchangeably.

- Indication of equality (either $=, \leq,$ or $\geq$) **must appear** in the null hypothesis
- ✓ Suppose, for example, that we want to answer the question: Can we conclude that a certain population mean is not 50? The **null** hypothesis is $H_0: \mu = 50$ and the **alternative** is $H_A: \mu \neq 50$
- ✓ Suppose we want to know if we can conclude that the population mean is **greater than 50**. Our hypotheses are $H_0: \mu \leq 50$ $H_A: \mu > 50$
- ✓ If we want to know if we can conclude that the population mean is less than 50, the hypotheses are $H_0: \mu \geq 50$ $H_A: \mu < 50$
- In summary
- **What you hope** or expect to be able to conclude as a result of the test usually should be placed in the alternative hypothesis.
- The null hypothesis **should contain a statement of equality**, either $=, \leq,$ or $\geq$.
- The null hypothesis **is the hypothesis that is tested**.
- The null and alternative hypotheses are **complementary.**

4. **Test statistic**. The test statistic is some statistic that may be **computed from the data of the sample**.

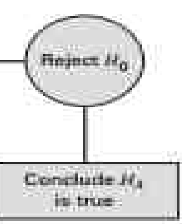
✓ As a rule, there are **many possible values that the test statistic may assume**, the particular value observed depending on the particular sample drawn. **The test statistic serves as a decision maker, since the decision to reject or not to reject the null hypothesis depends on the magnitude of the test statistic.**

$$\text{test statistic} = \frac{\text{relevant statistic} - \text{hypothesized parameter}}{\text{standard error of the relevant statistic}}$$ **EXAMPLE** $$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

5. **Distribution of test statistic**. It has been pointed out that **the key to statistical inference is the sampling distribution**.

6. **Decision rule**. All possible values that the test statistic can assume are points on the horizontal axis of the graph of the distribution of the test statistic and are divided into two groups.

✓ The **rejection region** and the **nonrejection region**.



testing procedure.

A

where μ_0 i
to the stati

with which

becomes nece
distribution of

for example, f
and the assum

SIGNIFICANCE LEVEL

- The decision as to **which values go into the rejection region and which ones go into the nonrejection region** is made on the basis of the desired **level of significance**, designated by α.
- The term level of significance reflects the fact that **hypothesis tests** are sometimes called **significance tests**, and a computed value of the test statistic that falls in the rejection region **is said to be significant**.
- The level of significance,α, specifies the area under the curve of the distribution of the test statistic that is **above the values on the horizontal axis constituting the rejection region**

*"The level of significance α is a probability and, in fact, is **the probability of rejecting a true null hypothesis**".*

- A small value of α is selected in order **to make the probability of rejecting a true null hypothesis small**.
- The more frequently encountered values of α are 0.01, 0.05, and 0.10.

TYPES OF ERRORS

- The error committed when a true null hypothesis is rejected is called the **type I error**.
- The type II error is the error committed **when a false null hypothesis is not rejected**. The probability of committing a type II error is **designated by β** .
- If the testing procedure leads to rejection of the null hypothesis, we can take **comfort from the fact that we made *α* small and, therefore**, the probability of committing a type I error was small.
- If we fail to reject the null hypothesis, we do not know the concurrent risk of committing a type II error, **since β** is usually unknown but, as has been pointed out, we do know that, in most practical situations, it **is larger than α** .

		Condition of Null Hypothesis	
		True	False
Possible Action	Fail to reject H_0	Correct action	Type II error
	Reject H_0	Type I error	Correct action

Conditions under which type I and type II errors may be committed.

7. **Calculation of test statistic**. From the data contained in the sample we compute a value of the test statistic and **compare it with the rejection and nonrejection regions that have already been specified**.

8. **Statistical decision**. The statistical decision consists of rejecting or of not rejecting the null hypothesis.

- ✓ It is rejected if the computed value of the test statistic **falls in the rejection region**, and it is not rejected if the computed value of the test statistic **falls in the nonrejection region**.

9. **Conclusion**. If $\boldsymbol{H_0}$ **is rejected**, we conclude that $\boldsymbol{H_A}$ **is true**. If $\boldsymbol{H_0}$ is not rejected, we conclude that $\boldsymbol{H_0}$ may be true.

10. P values. "***A p value is the probability that the computed value of a test statistic is at least as extreme as a specified value of the test statistic when the null hypothesis is true. Thus, the p value is the smallest value of α for which we can reject a null hypothesis*".**

- It is emphasized that when the null hypothesis is not rejected **one should not say that the null hypothesis is accepted**.

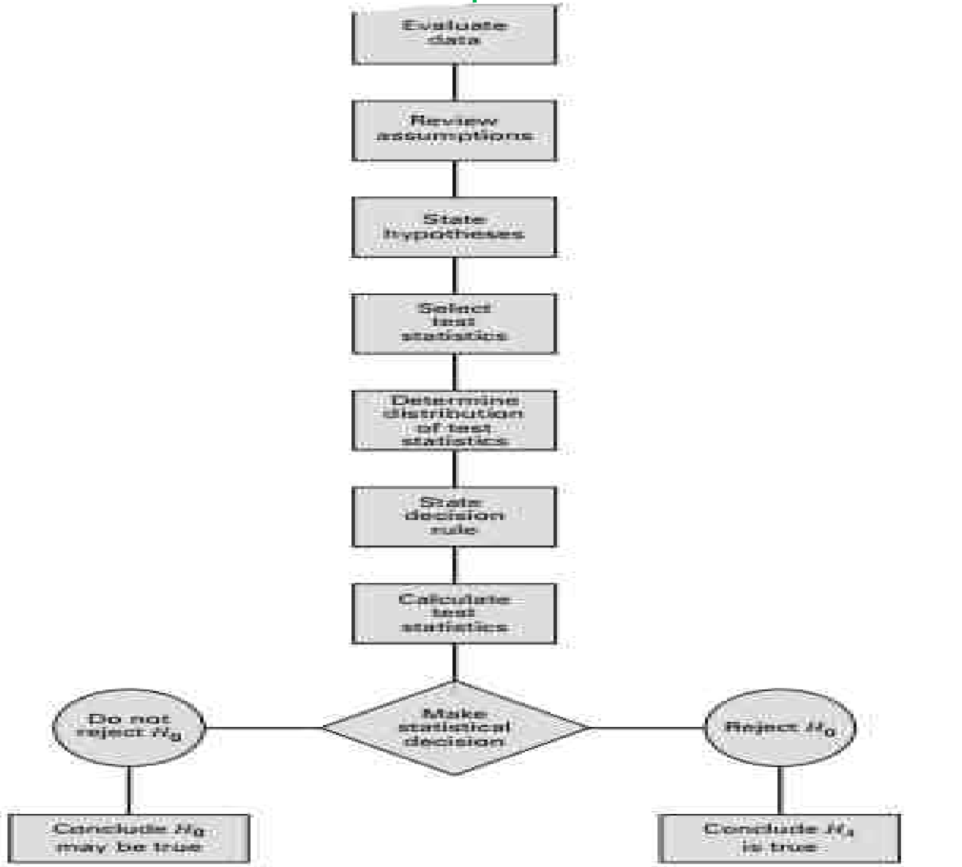
- We should say that the null hypothesis is "not rejected."

- One avoids using the word "accept" in this case because **we may have committed a type II error**. Since, frequently, the probability of committing a type II error can be quite high, ***we do not wish to commit ourselves to accepting the null hypothesis***.

Purpose of Hypothesis Testing

- The purpose of hypothesis testing is to assist administrators and clinicians in making decisions.
- If the *null hypothesis is rejected*, the administrative or clinical decision usually reflects this, in that the decision is **compatible with the alternative hypothesis**.
- The reverse is usually true if the null hypothesis is not rejected. The administrative or clinical decision, however, may take other forms, such as **a decision to gather more data**.

- The statistical decision **should not be interpreted as definitive** but should be considered along with all the other relevant information available to the experimenter.



Steps in the hypothesis testing procedure.

HYPOTHESIS TESTING: A SINGLE POPULATION MEAN

1.When sampling is from a normally distributed population of values with known variance

When sampling is from a normally distributed population and the population variance is known, the test statistic for testing $H_0: \mu = \mu_0$ is $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ which, **when H_0 is true, is distributed as the standard normal**

- ❖ Researchers are interested in the mean age of a certain population. Let us say that they are asking the following question: Can we conclude that the mean age of this population is different from 30 years?

1. Data: **a simple random sample** of a given number of individuals, 10 for example, is drawn from the population of interest and its mean is computed, suppose it is **27**
2. Assumptions: assume that the sample comes from a population whose ages are approximately normally distributed with a variance of **20**

Researchers are
of **10 individu**
Assuming that
20,can we conc
value is 0.0340

Flowchart for use in deciding between z and t when making inferences about population means.

3. Hypotheses. H_0: $\mu = 30$

H_A: $\mu \neq 30$

4. **Test statistic**: Since we are testing a hypothesis about a population mean, since we assume that the population is normally distributed, and since the population variance is known,

our test statistic is given by $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

5. **Distribution of test statistic**: Based on our knowledge of sampling distributions and the normal distribution, we know that the test statistic is normally distributed with a **mean of 0 and a variance of 1**, if H_0 is true. There are many possible values of the test statistic that the present situation can generate; one for every possible sample of size 10 that can be drawn from the population. **Since we draw only one sample, we have only one of these possible values on which to base a decision**.

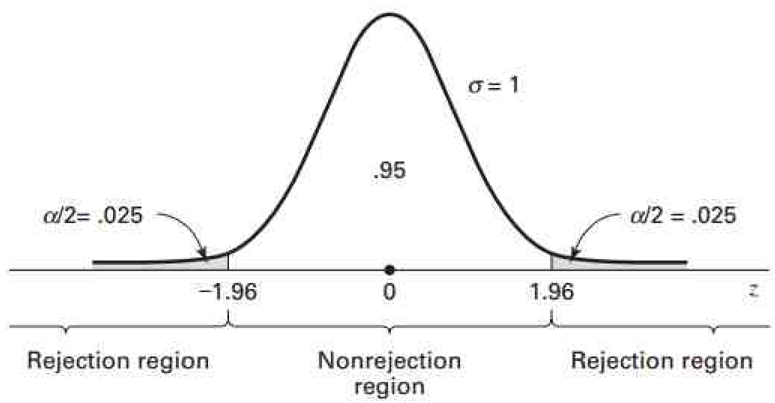
6. **Decision rule**: to reject H_0 if the computed value of the test statistic falls in the rejection region and to fail to reject H_0 if it falls in the nonrejection region.

We must now specify the rejection and nonrejection regions.

- If the null hypothesis is false, it may be so either because the population **mean is less than 30** or because the population **mean is greater than 30**.
- Therefore, either **sufficiently small** values or **sufficiently large** values of the test statistic will cause rejection of the null hypothesis.
- **These extreme values constitute the rejection region**.
- Let us say that we want **the probability of rejecting a true null hypothesis** to be $\alpha=0.05$.
- Since our rejection region is to consist of **two parts**, sufficiently small values and sufficiently large values of the test statistic, **part of α will have to be associated with the large values and part with the small values**.
- It seems reasonable that we should divide α equally and let **α/2**=0.025 be associated with small values and **α/2** be associated with large values

CRITICAL VALUE OF TEST STATISTIC

- The values of the test statistic that separate the rejection and nonrejection regions are called critical values of the test statistic, and the rejection region is sometimes referred to as the **critical region**.
- Our rejection region, then, consists of all values of the test statistic **equal to or greater than 1.96** and **less than or equal to -1.96**. The nonrejection region consists of **all values in between**.
- We may state the decision rule for this test as follows: reject H_0 if the computed value of the test statistic is either ≥ 1.96 or ≤ -1.96 . O ... w)



- The decision rule tells us to compute a value of the test statistic from the data of our sample and to reject H_0 if we get a value that is either equal to or greater than 1.96 or equal to or less than -1.96 **and to fail to reject H_0 if we get any other value**. The value of α and, hence, the decision rule should be **decided on before gathering the data**.
- This prevents our being accused of **allowing the sample results to influence our choice of α**. This condition of objectivity is **highly desirable and should be preserved in all tests**.

7. **Calculation of test statistic**: From our sample we compute

$$z= \frac{27-30}{\sqrt{20/10}}=\frac{-3}{1.4142}=-2.12$$

8. **Statistical decision:** Abiding by the decision rule, we are able to reject the null hypothesis **since -2.12 is in the rejection region**. We can say that **the computed value of the test statistic is significant at the 0.05 level**.

9. **Conclusion**: We conclude that μ is not equal to 30 and let our administrative or clinical actions be in accordance with this conclusion.

−0.50	.2776	.2810	.2843	.2877	.2912	.2946	.2981	.3015	.3050	.3085	−0.50
−0.40	.3121	.3156	.3192	.3228	.3264	.3300	.3336	.3372	.3409	.3446	−0.40
−0.30	.3483	.3520	.3557	.3594	.3632	.3669	.3707	.3745	.3783	.3821	−0.30
−0.20	.3859	.3897	.3936	.3974	.4013	.4052	.4090	.4129	.4168	.4207	−0.20
−0.10	.4247	.4286	.4325	.4364	.4404	.4443	.4483	.4522	.4562	.4602	−0.10
0.00	.4641	.4681	.4721	.4761	.4801	.4840	.4880	.4920	.4960	.5000	0.00

10. p values: Instead of saying that an observed value of the test statistic is **significant or is not significant**, most writers in the research literature prefer to report **the exact probability of getting a value as extreme as or more extreme than that observed if the null hypothesis is true**. In the present instance these writers would give the computed value of the test statistic along with the statement **p=0.0340**

- **The statement means that the probability of getting a value as extreme as 2.12 in either direction, when the null hypothesis is true, is .0340.**
- That is, when H_0 is true, the probability of obtaining **a value of z as large as or larger than 2.12 is .0170**, and the probability of observing **a value of z as small as or smaller than -2.12 is .0170**.
- The probability of one or the other of these events occurring, when H_0 is true, is equal to **the sum of the two individual probabilities**, and hence, in the present example, we say that P=0.0170+0.0170=0.0347

ch we can

- P value for a test may be defined also as **the smallest value of α for which the null hypothesis can be rejected**.
- We know that we could have chosen an α value as small as .0340 and still have rejected the null hypothesis.
- If we had chosen an α smaller than .0340, we would not have been able to reject the null hypothesis.

"*if the p value is less than or equal to α, we reject the null hypothesis; if the p value is greater than α, we do not reject the null hypothesis*".

Testing H_0 by I
that one can use conf
hypothesis testing pro
were able to reject H_0
tion region.

Let us see ho
$(1-\alpha)$ percent co

Since this interval c
are estimating and,
conclusion reached
If the hypothe
val, we would have
eral, *when testing c*
reject H_0 at the α l
within the 100(1 −
tained within the in

Flowchart for use in deciding between z and t when making inferences about population means.

ONE-SIDED HYPOTHESIS TESTS

- Whether a one-sided or a two-sided test is used **depends on the nature of the question being asked by the researcher**.
- **If both large and small values** will cause rejection of the null hypothesis, a two sided test is indicated. When **either sufficiently "small" values only or sufficiently "large" values only** will cause rejection of the null hypothesis, a one-sided test is indicated.
- Suppose, instead of asking if they could conclude that the mean is equal to 30, the researchers had asked: **Can we conclude that $\mu < 30$?** To this question we would reply that they can so conclude if they can reject the null hypothesis that $\mu \geq 30$. Find the solution

- If the researcher's question had been, "Can we conclude that the mean is greater than 30?," Find the solution

7.

8.

9.

10.

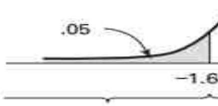


FIGURE 7.2.2 Re

If the researcher'
greater than 30?,
to a one-sided te
tribution of the t

Flowchart for use in deciding between z and t when making inferences about population means.

2.Sampling from a population that is not normally distributed

- If, as is frequently the case, the sample on which we base our hypothesis test about a population mean comes from a population that is not normally distributed, we may, **if our sample is large** (greater than or equal to 30), take advantage of the central limit theorem and use $z = (\bar{x} - \mu_0)/(\sigma/\sqrt{n})$ as the test statistic

- If the population standard deviation is not known, the usual practice is to **use the sample standard deviation as an estimate**.

- The test statistic for testing H_0 , then, is $z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ which, when H_0 is true, is distributed approximately as the standard normal distribution if n is large. The rationale for using s to replace σ is that the large sample, necessary for the central limit theorem to apply, will yield **a sample standard deviation that closely approximates σ.**

- 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 men is greater than 140. **Use $\alpha=0.05$**

Flowchart for use in deciding between z and t when making inferences about population means.

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 σ 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. $n \geq 30$).

3.Hypotheses:

- We have three cases
- ✓ Case I : H_0: $\mu 1 = \mu 2 \quad \rightarrow \quad \mu_1 - \mu_2 = 0$

  H_A: $\mu_1 \neq \mu_2 \quad \rightarrow \quad \mu_1 - \mu_2 \neq 0$

- ❖ e.g. We want to test that the mean for first population is **different** from second population mean.
- ✓ Case II : H_0: $\mu 1 \leq \mu 2 \quad \rightarrow \quad \mu_1 - \mu \leq 0$

  H_A: $\mu_1 > \mu_2 \quad \rightarrow \quad \mu_1 - \mu_2 > 0$

- ❖ e.g. We want to test that the mean for first population is **greater** than second population mean.
- ✓ Case III : H_0: $\mu 1 \geq \mu 2 \quad \rightarrow \quad \mu_1 - \mu_2 \geq 0$

  H_A: $\mu_1 < \mu_2 \quad \rightarrow \quad \mu_1 - \mu_2 < 0$

- ❖ e.g. We want to test that the mean for first population is **less** than second population mean.

or

z

Central limit theorem applies

Flowchart for use in deciding between z and t when making inferences about population means.

4.Test Statistic:

✓ **Case 1:** **Two populations are 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

ce not 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

$$\boxed{S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

between two pop
sampling is from
(2) when samplin
variances, and (3

Sampling fro
Variances Un
unknown, two possi
be unequal. We con
that they are equal.
described in Sectio

Population Va
assumed to be equal,
ances by means of t

When each of two i
distributed populati
test statistic for test

which, when H_0 is t

Flowchart for use in deciding between z and t when making inferences about population means.

- ✓ Case2: If population is not normally distributed and n_1, n_2 large ($n_1 \geq 0$, $n_2 \geq 0$) and population variances **is known**,

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

5.Decision Rule:

i) If H_A: $\mu_1 \neq \mu_2 \rightarrow \mu_1 - \mu_2 \neq 0$

- ❖ Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$

  (when use Z - test)

Or Reject H_0 if $T > t_{1-\alpha/2, (n_1+n_2-2)}$ or $T < -t_{1-\alpha/2, (n_1+n_2-2)}$ (when use T- test)

ii) H_A: $\mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$

Reject H_0 if $Z > Z_{1-\alpha}$ (when use Z - test)

Or Reject H_0 if $T > t_{1-\alpha, (n_1+n_2-2)}$ (when use T - test)

iii) If H_A: $\mu_1 < \mu_2 \rightarrow \mu_1 - \mu_2 < 0$

- Reject H_0 if $Z < -Z_{1-\alpha}$ (when use Z - test)

Or

- Reject H_0 if $T < -t_{1-\alpha,(n_1+n_2-2)}$ (when use T - test)

Note:

$Z_{1-\alpha/2}$, $Z_{1-\alpha}$, Z_α are tabulated values obtained from table D

$t_{1-\alpha/2}$, $t_{1-\alpha}$, t_α are tabulated values obtained from table E with (n_1+n_2-2) degree of freedom (df)

6. Conclusion: **reject or fail to reject H_0**

- Researchers **wish** to know if the data they 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 $\bar{X}_1 = 4.5mg/100$ and $\bar{X}_2 = 3.4mg/100$ **α=0.05.**

Solution:

1. **Data:** Variable is serum uric acid levels, $n_1=12$, $n_2=15$, $\sigma^2_1=1$, $\sigma^2_2=1.5$, $\alpha=0.05$.

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:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(4.5 - 3.4) - (0)}{\sqrt{\frac{1}{12} + \frac{1.5}{15}}} = 2.57$$

Solution: For a
reliab
be 1.

The 95 perce

We sa
$\mu_1 - \mu_2$, is
95 percent o
ference betw
Since t
ulation mean

5. Decision Rule:

Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$

$Z_{1-\alpha/2} = Z_{1-0.05/2} = Z_{0.975} = 1.96$ (from table D)

6-**Conclusion:** Reject H_0 since $2.57 > 1.96$

Or p-value =**0.0102**→ reject H_0 since if $p < \alpha$ → then reject H_0

- ❖ 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 wheelchair to incorporate a rigid seat surface to facilitate the specified experimental measurements. The data for measurements of the left **ischial tuberosity** for SCI and control C are shown below

C	131	115	124	131	122	117	88	114	150	169
SCI	60	150	130	180	163	130	121	119	130	148

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 is lower than the mean of left ischial tuberosity for SCI, Assume normal populations equal variances. α=0.05.

Solution:

1. Data:, $n_C=10$, $n_{SCI}=10$, $S_C=21.8$, $S_{SCI}=32.2$,α=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: $\mu_C \geq \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} \geq 0$

H_A: $\mu_C < \mu_{SCI} \rightarrow \mu_C - \mu_{SCI} < 0$

4.Test Statistic:

Where,

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(126.1 - 133.1) - 0}{\sqrt{756.04}\sqrt{\frac{1}{10} + \frac{1}{10}}} = -0.569$$

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{9(21.8)^2 + 9(32.3)^2}{10 + 10 - 2} = 756.04$$

16	1.337	1.7459	2.1199	2.583	2.9208
17	1.333	1.7396	2.1098	2.567	2.8982
18	1.330	1.7341	2.1009	2.552	2.8784
19	1.328	1.7291	2.0930	2.539	2.8609

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 $p > 0.10$ (**$\alpha = 0.05$**)

- The objective of a study by Sairam et al. (A-8) was to identify the role of various disease states and additional risk factors in the development of thrombosis. One focus of the study was **to determine if there were differing levels of the anticardiolipin antibody IgG in subjects with and without thrombosis**. Table below summarizes the researchers' findings:

Group	Mean IgG level	Sample Size	standard deviation
Thrombosis	59.01	53	44.89
No Thrombosis	46.61	54	34.85

- ✓ We wish to know if we may conclude, **on the basis of these results, that, in general, persons with thrombosis have, on the average, higher IgG levels than persons without thrombosis.**

$$\boldsymbol{\alpha = 0.01}$$

Solution:

1. Data:, $n_1=53$, $n_2=54$, $S_1= 44.89$, $S_2= 34.85$ $\alpha=0.01$.

2.Assumption: Two population are not normal, σ^2_1 , σ^2_2 are unknown and sample size large

3. Hypotheses: H_0: $\mu_1 \leq \mu_2 \rightarrow \mu_1 - \mu_2 \leq 0$

H_A: $\mu_1 > \mu_2 \rightarrow \mu_1 - \mu_2 > 0$

4.Test Statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{(59.01 - 46.61) - 0}{\sqrt{\frac{44.89^2}{53} + \frac{34.85^2}{54}}} = 1.59$$

5. Decision Rule:

Reject H_0 if $Z > Z_{1-\alpha}$

$Z_{1-\alpha} = Z_{0.99} = 2.33$ (from table D)

6-**Conclusion:** Fail to reject H_0 since 1.59 is in the non rejection region

Or Fail to reject H_0 since $p = 0.0559 > \alpha = 0.01$

Hypothesis Testing A single population proportion:

- Testing hypothesis about population proportion (P) is carried out in much **the same way as for mean** **when condition is necessary for using normal curve are met**

✓ We have the following steps:

1.Data: sample size (n), sample proportion($\hat{p}$) , P_0

$$\boxed{\hat{p} = \frac{no.of\ element\ in\ the\ sample\ with\ some\ characteristics}{Total\ number\ of\ element\ in\ the\ sample}}$$

2. Assumptions :normal distribution **,**

3.Hypotheses:

WE HAVE THREE CASES

- Case I : H_0: $p = P_0$
  H_A: $p \neq P_0$
- Case II : H_0:$p \leq P_0$
  H_A: $p > P_0$
- Case III : H_0: $p \geq P_0$
  H_A: $p < P_0$

4.Test Statistic:

$$\boxed{Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}}}$$

5.Decision Rule:

i) If H_A: $p \neq P_0$

Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$

ii) If H_A: $p > P_0$

Reject H_0 if $Z > Z_{1-\alpha}$

iii) If H_A: $p < P_0$

Reject H_0 if $Z < -Z_{1-\alpha}$

Note: $Z_{1-\alpha/2}$, $Z_{1-\alpha}$, Z_{α} are tabulated values obtained from table D

6. Conclusion: **reject or fail to reject H_0**

- Wagenknecht et al. (A-20) 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?** $\boldsymbol{\propto = 0.05}$

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

2. Assumptions : $\hat{p}$ is approximately normaly distributed

3.Hypotheses:

❖ H_0: p $\leq$ 0.063

H_A: p > 0.063

4.Test Statistic :

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.08 - 0.063}{\sqrt{\frac{0.063(0.937)}{301}}} = 1.21$$

5.Decision Rule: Reject H_0 if $Z > Z_{1-\alpha}$

Where $Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.645$

6. Conclusion: Fail to reject H_0 since $1.21 < 1.645$

P-value = 0.1131,

fail to reject H_0 $\rightarrow$ $P > \alpha$

Hypothesis Testing :The Difference between two population proportion:

- Testing hypothesis about two population proportion $(\hat{p}_{1,} \hat{p}_2)$ is

  carried out **in much the same way as for difference between two means when condition is necessary for using normal curve are met.**

- We have the following steps:

1.Data: Sample size ($n_{1,}$ n_2), sample proportions($\hat{P}_1, \hat{P}_2$),

Characteristic in two samples (x_1 , x_2), $\boxed{\bar{p} = \dfrac{x_1 + x_2}{n_1 + n_2}}$

2- Assumption : Two populations are independent .

3.Hypotheses:

We have three cases

- Case I : $H_0: p_1 = p_2 \rightarrow p_1 - p_2 = 0$
  $H_A: p_1 \neq p_2 \rightarrow p_1 - p_2 \neq 0$
- Case II : $H_0: p_1 \leq p_2 \rightarrow p_1 - p_2 \leq 0$
  $H_A: p_1 > p_2 \rightarrow p_1 - p_2 > 0$
- Case III : $H_0: p_1 \geq p_2 \rightarrow p_1 - p_2 \geq 0$
  $H_A: p_1 < p_2 \rightarrow p_1 - p_2 < 0$

4.Test Statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}}$$

Where H_0 is true ,is distributed approximately as the standard normal

5.Decision Rule:

i) If H_A: $P_1 \neq P_2$

- Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$

---

ii) If H_A: $P_1 > P_2$

Reject H_0 if $Z > Z_{1-\alpha}$

---

iii) If H_A: $P_1 < P_2$

Reject H_0 if $Z < -Z_{1-\alpha}$

Note: $Z_{1-\alpha/2}$, $Z_{1-\alpha}$, Z_α are tabulated values obtained from table D

6. Conclusion: reject or fail to reject H_0

Noonan is a genetic condition that **can affect the heart growth**, **blood clotting** and **mental and physical development**. Noonan examined the stature of men and women with Noonan. The study contained **29 Male** and 44 female adults. One of the cut-off values used to assess stature was the **third percentile of adult height**. Eleven of the males fell below the third percentile of adult male height ,while 24 of the female fell below the third percentile of female adult height. Does this study provide sufficient evidence for us to conclude that among subjects with Noonan, females are more likely than males to fall below the respective of adult height? Let α=0.05

$$\hat{p}_M = \frac{x_m}{n_M} = \frac{11}{29} = 0.379, \ \hat{p}_F = \frac{x_F}{n_F} = \frac{24}{44} = 0.545$$

Solution:

1.Data: $n_M = 29$, $n_F = 44$, $x_M = 11$, $x_F = 24$, $\alpha=0.05$

$$\bar{p} = \frac{x_M + x_F}{n_M + n_F} = \frac{11+24}{29+44} = 0.479$$

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

2- Assumption : Two populations are independent.

3.Hypotheses:

- Case II : $H_0: P_F \leq P_M \rightarrow P_F - P_M \leq 0$

  $H_A: P_F > P_M \rightarrow P_F - P_M > 0$

- 4.Test Statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.545 - 0.379) - 0}{\sqrt{\frac{(0.479)(0.521)}{44} + \frac{(0.479)(0.521)}{29}}} = 1.39$$

5.Decision Rule:

Reject H_0 if $Z > Z_{1-\alpha}$, Where $Z_{1-\alpha} = Z_{1-0.05} = Z_{0.95} = 1.645$

6. Conclusion: Fail to reject H_0

Since $Z = 1.39 < Z_{1-\alpha=} 1.645$

Or , If P-value = 0.0823 $\rightarrow$ fail to reject H_0 $\rightarrow P > \alpha$

ANALYSIS OF VARIANCE

- Let us imagine that we wish to compare the means of seven samples: **No less than 21 z** test are required to compare all possible pairs of means and there is a good chance that at least **one false conclusion** will be drawn if P=0.05.
- Analysis of variance (ANOVA) overcomes this by allowing **comparison to be made between any number of sample mean in a single test**.
- When it is used in this way **to compare the means of several samples**, statisticians speak of one way ANOVA.
- When the influence of two variables upon a sample mean is being analyzed, the technique involved is described as a **two-way ANOVA,** etc.

Explain

Compare the individual variances of the three samples below with the overall variance when all 15 observation n=15 are aggregated

Sample 1	Sample 2	Sample 3	Overall
8	9	3	
10	11	5	
12	13	7	
14	15	9	
16	17	11	
$\sum x$=60	$\sum x$=65	$\sum x$=35	$\sum x_T$=160
$\hat{x}$=12	$\hat{x}$=13	$\hat{x}$=7	$\bar{X}_T$ =10.667
S^2=10	S^2=10	S^2=10	$S^2{}_T$=16

Example

- Referring to table on previous slide, the increase in overall variance is due to **the difference between means of the samples**, in particular the **difference between mean of sample 3 and the two other means**.
- The samples give rise to **two sources of variability**.
- The variability **around each mean within a sample** (random scatter)
- The variability **between the samples** due to **differences between the means of the population from which the sample are drawn.**
- In other words: **Total Variability=variability within + variability between**.
- Analysis of variance may be defined as a technique whereby **the total variation present in a set of data is partitioned into two or more components.**

- Associated with each of these components is **a specific source of variation**, so that in the analysis **it is possible to ascertain the magnitude of the contributions of each of these sources to the total variation**.
- If samples are drawn from normally distributed populations with **equal means and variances**, **the within variance is the same as the between variance.**
- If a statistical test shows that this is not the case then the sample have been **drawn from populations with different means and/or variances**.
- **If it is assumed that variances are equal** (and this is an underlying assumption of ANOVA) then it is **concluded that the discrepancy is due to differences between means**.

- Thus H_0 = Samples are drown from normally distributed populations **with equal means and variances**.
- H_1 = Population **variances are assumed to be equal** and therefore **samples are drawn from populations with different means**

One way ANOVA

- A biologist wishes to know **if the mean masses of starlings sampled in four different roosts situations are different**.
- Cast the data into a table, labelling each sample 1-4 respectively. (see figure on next slide)
- Calculate, for each sample, **the mean**, **the standard deviation**, **variance**, $\sum \boldsymbol{x}$,$(\sum \boldsymbol{x})^2$ and $\sum \boldsymbol{x}^2$ **(see table on next slide)**
- **Check** if all four sample variances are similar to each other (test for the homogeneity of variances):
- ✓ If the largest and smallest variances of the samples are not significantly different from each other, then the other cannot be: **select the largest sample variances in the table and divide it by the smallest**.

Mass of starlings from four roost situation (g)

Situation1,sample 1	Situation2,sample 2	Situation3,sample 3	Situation4,sample 4	Total
78	78	79	77	
88	78	73	69	
87	83	79	75	
88	81	75	70	
83	78	77	74	
82	81	78	83	
81	81	80	80	
80	82	78	75	
80	76	83	76	
89	76	84	75	
n=10	n=10	n=10	n=10	n_T=40
$\bar{x} = 83.6$	$\bar{x}$=79.4	$\bar{x}$=78.6	$\bar{x}$=75.4	
S=4.03	S=2.50	S=3.31	S=4.14	
s^2=16,27	s^2=6.25	s^2=10.96	s^2=17.14	
$\sum x$=836	$\sum x$=794	$\sum x$=786	$\sum x$=754	$\sum x_T$=3170
$(\sum x)^2$=698896	$(\sum x)^2$=630436	$(\sum x)^2$=617796	$(\sum x)^2$=568516	
$\sum x^2$=70036	$\sum x^2$=63100	$\sum x^2$=61878	$\sum x^2$=57006	$\sum {x_T}^2$=252020

- Equate the results to F. $F_{max}=\frac{17.14}{6.25}$=2.74 (with 9 degree of freedom in each sample) which is **less than the critical value of 6.31 for number of samples** a=4 and df (n-1)=9 **Why is it called F_{max}?**

Calculate the correction term CT: $CT=\frac{(\sum x_T)^2}{n_T}$

- Calculate **the total sum of squares** of the aggregated samples:

$SS_T=\sum x_T{}^2$-CT

- Calculate **the between samples sum of squares**, $SS_{between}$, $\frac{(\sum x_1)^2}{n_1}+\frac{(\sum x_2)^2}{n_2}+\frac{(\sum x_3)^2}{n_3}+\frac{(\sum x_4)^2}{n_4}-\text{CT}$
- Calculate **the within samples sum of squares** ,SS_{within}, $(\sum x_1^2-\frac{(\sum x_1)^2}{n_1})+(\sum x_2^2-\frac{(\sum x_2)^2}{n_2})+(\sum x_3^2-\frac{(\sum x_3)^2}{n_3})+(\sum x_4^2-\frac{(\sum x_4)^2}{n_4})$ which is equal to **the sum of the individual SS_{Within}** for each sample.
- **Check** that the independently calculated SS_{within} and $SS_{between}$ add up to that of SS_T,$\sum x_T{}^2-CT$

- Determine the number of degree of freedom (df) for each of the calculated ss values. The rules for determining these are:
- ✓ df for SS_T=n_T-1
- ✓ df for $SS_{between}$ = a-1 (where a=number of samples)
- ✓ df for SS_{within}=n_T-a

- $S^2{}_{between}=\frac{SS_{between}}{df_{between}}$
- $S^2{}_{within}=\frac{SS_{within}}{df_{within}}$
- Compute F=$\frac{Between\ sample\ variance}{Within\ sample\ variance}=\frac{113.97}{12.66}$=9.002

25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12

- Enter the result in an ANOVA table:

Source of variation	SS	df	S^2	F
Between	341.9	3	113.97	9.002
Within	455.6	36	12.66	
Total	797.5	39		

- Consulting a table of the **one-tailed distribution** of F, we find that our calculated value of F at 3 and 36 degree of freedom exceed the critical value of 2.88 (interpolating between 30 and 40 df).
- We therefore reject the null hypothesis and conclude that the variation in the mean mass of the four starling samples is significantly different. we record the result as:

"***The difference in mean mass of the four samples, when n=10 in each case is statistically significant ($F_{3,36}$= 9.002,p <0.05)***

THE CHI-SQUARE DISTRIBUTION

The mathematical properties of the chi-square distribution

- The chi-square distribution **may be derived from normal distributions**. Suppose that from a normally distributed **random variable Y** with mean μ and variance σ^2 we randomly and independently select **samples of size n=1**.
- Each value selected may be transformed to the standard normal variable z by the familiar formula

$$z_i = \frac{y_i - \mu}{\sigma}$$

- **Each value of z may be squared to obtain z^2**.
- When we investigate the sampling distribution of z^2, we find that **it follows a chi-square distribution with 1 degree of freedom**. That is

$$\chi^2_{(1)} = \left(\frac{y - \mu}{\sigma}\right)^2 = z^2$$

- Now suppose that we randomly and independently select samples of size n=2 from the normally distributed population of Y values.
- Within each sample we may transform each value of y to the standard normal variable z and square as before.

- If the resulting values of z^2 **for each sample** are added, we may designate this su $\chi^2_{(2)} = \left(\frac{y_1 - \mu}{\sigma}\right)^2 + \left(\frac{y_2 - \mu}{\sigma}\right)^2 = z_1^2 + z_2^2$

- **The procedure may be repeated for any sample of size n**. **The sum of the resulting values** in each case will be distributed as chi-square with n degrees of freedom. In general, then $\chi^2_{(n)} = z_1^2 + z_2^2 + \cdots + z_n^2$

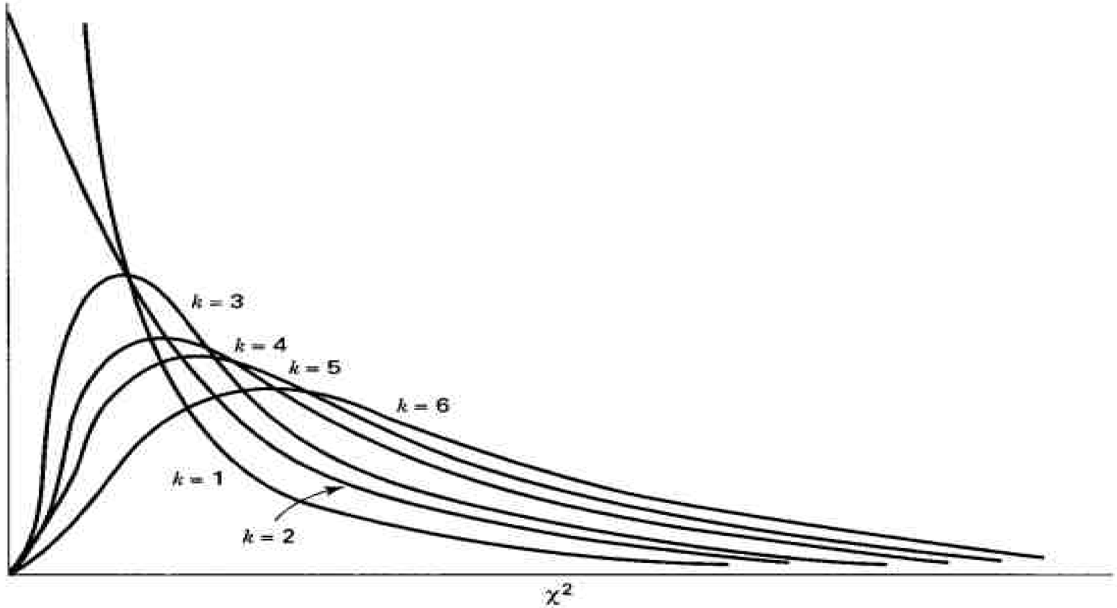
follows the chi-square distribution with n degrees of freedom.

- The mathematical form of the chi-square distribution is as follows:

$$f(u) = \frac{1}{\left(\frac{k}{2} - 1\right)!} \frac{1}{2^{k/2}} u^{(k/2)-1} e^{-(u/2)}, \quad u > 0$$

- Where e is the irrational number 2.71828 . . . And k is the number of degrees of freedom. **The variate u is usually designated by the Greek letter chi and, hence, the distribution is called the chi-square distribution**.
- The **mean** and **variance** of the chi-square distribution are **k** and **2k**, respectively.

- The **modal value** of the distribution is **k-2** for values of k greater than or equal to 2 and is zero for k=1
- Chi-square assumes **values between 0 and infinity**. It cannot take on negative values, since it is the sum of values that have been squared.
- A final characteristic of the chi-square distribution worth noting is that **the sum of two or more independent chi-square**



Chi-square distributions for several values of degrees of freedom *k*.

Types of Chi - Square Tests

- Tests of **goodness-of-fit**, tests of **independence**, and tests of **homogeneity**.
- **In a sense, all of the chi-square tests that we employ may be thought of as goodness-of-fit tests**, in that **they test the goodness-of-fit of observed frequencies to frequencies that one would expect if the data were generated under some particular theory or hypothesis**
- However, the phrase "goodness-of-fit" is reserved for use **in a more restricted sense**, **in comparison of a sample distribution to some theoretical distribution that it is assumed describes the population from which the sample came**.
- The chi-square distribution may be used as **a test of the agreement between observation and hypothesis** ***whenever the data are in the form of frequencies***.

Observed Versus Expected Frequencies

- There are two sets of frequencies with which we are concerned, **observed frequencies and expected frequencies**.
- The observed frequencies are **the number of subjects or objects in our sample that fall into the various categories of the variable of interest**.
- For example, if we have a sample of **100 hospital patients**, we may observe that **50 are married**, **30 are single**, **15 are widowed**, and **5 are divorced**.
- Expected frequencies are the number of subjects or objects in our sample that we would expect to observe **if some null hypothesis about the variable is true**.
- For example, our null hypothesis might be that **the four categories of marital status are equally represented in the population from which we drew our sample**.
- In that case we would expect our sample to contain 25 married, 25 single, 25 widowed, and 25 divorced patients.

The Chi-Square Test Statistic

- **When the null hypothesis is true**, χ^2 is distributed approximately as χ^2 with k-r degrees of freedom.
- In determining the degrees of freedom, **k is equal to the number of groups for which observed and expected frequencies are available**, and **r is the number of restrictions or constraints imposed on the given comparison**.

$$X^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

- O_i is the observed frequency for the i^{th} category of the variable of interest, and E_i is the expected frequency (given that H_0 is true) for the i^{th} category.
- The quantity χ^2 **is a measure of the extent to which, in a given situation, pairs of observed and expected frequencies agree**.
- The nature of χ^2 is such that when there is **close agreement** between observed and expected frequencies **it is small**, and when the agreement is poor **it is large**.

- When there is disagreement between a pair of observed and expected frequencies, **the difference may be either positive or negative**, depending on which of the two frequencies is the larger.
- χ^2 is **a summary statistic that reflects the extent of the overall agreement between observed and expected frequencies.**

The Decision Rule

- The quantity $\Sigma[(O_i - E_i)^2/E_i]$ will be small if the observed and expected frequencies are close together and will be large if the differences are large.
- **The computed value of χ^2 is compared with the tabulated value of χ^2 with k-r degrees of freedom**.
- The decision rule, then, is: **Reject H_0 if χ^2 is greater than or equal to the tabulated χ^2 for the chosen value of α**

TESTS OF GOODNESS-OF-FIT

- A goodness-of-fit test is appropriate when one wishes to decide **if an observed distribution of frequencies is incompatible with some preconceived or hypothesized distribution**.

- Cranor and Christensen (A-1) conducted a study to assess short-term clinical, economic, and humanistic outcomes of pharmaceutical care services for patients with diabetes in community pharmacies. For 47 of the subjects in the study, cholesterol levels are summarize

Cholesterol Level (mg/dl)	Number of Subjects
100.0–124.9	1
125.0–149.9	3
150.0–174.9	8
175.0–199.9	18
200.0–224.9	6
225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

We wish to know whether these data provide sufficient evidence to indicate that the sample did not come from a normally distributed population. Let $\alpha = 0.05$

Steps in the hypothesis testing procedure.

Solution

1.Data. See table on previous slide

2.Assumptions. We assume that the sample available for **analysis is a simple random sample**.

3. Hypotheses.

H_0: In the population from which the sample was drawn, cholesterol levels are normally distributed.

H_A: The sampled population is not normally distributed.

4. Test statistic. The test statistic is

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

5. **Distribution of test statistic**. If H_0 is true, **the test statistic is distributed approximately as chi-square with k-r degrees of freedom**.

The values of k

6. Decision rule. We will reject H_0 if the computed value of χ^2 is equal to or greater than the critical value of chi-square.

Steps in the hypothesis testing procedure.

7.Calculation of test statistic. Since the **mean and variance** of the hypothesized distribution are not specified, **the sample data must be used to estimate them**. These parameters, or their estimates, will be needed to compute **the frequency that would be expected in each class interval when the null hypothesis is true**. **The mean and standard deviation computed from the grouped data of table on slide 53 are**

$$\bar{x} = 198.67$$
$$s = 41.31$$

- We must **obtain for each class interval the frequency of occurrence of values that we would expect when the null hypothesis is true**, that is, **if the sample were, in fact, drawn from a normally distributed population of values**.
- To do this, **we first determine the expected relative frequency of occurrence of values for each class interval** and then multiply these expected relative frequencies by the total number of values to obtain the expected number of values for each interval.

225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

- The first step consists of obtaining **z values corresponding to the lower limit of each class interval**. The area between two successive z values will give the expected relative frequency of occurrence of values for the corresponding class interval.
- For example, to obtain the expected relative frequency of occurrence of values in the interval 100.0 to 124.9 we proceed as follows:

$$\text{The } z \text{ value corresponding to } X = 100.0 \text{ is } z = \frac{100.0 - 198.67}{41.31} = -2.39$$

$$\text{The } z \text{ value corresponding to } X = 125.0 \text{ is } z = \frac{125.0 - 198.67}{41.31} = -1.78$$

- The area to the left of -2.39 is .0084, and the area to the left of -1.78 is .0375. The area between -1.78 and -2.39 is equal to 0.0375-0.0084=**0.0291** which is **equal to the expected relative frequency of occurrence of cholesterol levels within the interval 100.0 to 124.9.**
- This tells us that if the null hypothesis is true, that is, **if the cholesterol levels are normally distributed, we should expect 2.91 percent of the values in our sample to be between 100.0 and 124.9.**

225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

- When we multiply our total sample size, 47, by .0291 we find the expected frequency for the interval to be 1.4. Similar calculations will give the expected frequencies for the other intervals as shown in table below:

Class Interval	$z = (x_i - \bar{x})/s$ At Lower Limit of Interval	Expected Relative Frequency	Expected Frequency
< 100		.0084	.4 } 1.8
100.0–124.9	−2.39	.0291	1.4
125.0–149.9	−1.78	.0815	3.8
150.0–174.9	−1.18	.1653	7.8
175.0–199.9	−.57	.2277	10.7
200.0–224.9	.03	.2269	10.7
225.0–249.9	.64	.1536	7.2
250.0–274.9	1.24	.0753	3.5
275.0–299.9	1.85	.0251	1.2 } 1.5
300.0 and greater	2.45	.0071	.3

$\bar{x} = 198.67$

$s = 41.31$

.01	0.00	z
01	.0001	−3.80
01	.0001	−3.70
02	.0002	−3.60
02	.0002	−3.50
03	.0003	−3.40
05	.0005	−3.30
07	.0007	−3.20
09	.0010	−3.10
13	.0013	−3.00
18	.0019	−2.90
25	.0026	−2.80
34	.0035	−2.70
45	.0047	−2.60
60	.0062	−2.50
80	.0082	−2.40
04	.0107	−2.30
36	.0139	−2.20
74	.0179	−2.10
22	.0228	−2.00
81	.0287	−1.90
51	.0359	−1.80
36	.0446	−1.70
37	.0548	−1.60
55	.0668	−1.50
93	.0808	−1.40
51	.0968	−1.30
31	.1151	−1.20
35	.1357	−1.10
62	.1587	−1.00
14	.1841	−0.90
90	.2119	−0.80
89	.2420	−0.70
09	.2743	−0.60
50	.3085	−0.50
09	.3446	−0.40
83	.3821	−0.30
68	.4207	−0.20
62	.4602	−0.10
60	.5000	0.00

z	0.00	0.
0.00	.5000	.50
0.10	.5398	.54
0.20	.5793	.58
0.30	.6179	.62
0.40	.6554	.65
0.50	.6915	.69
0.60	.7257	.72
0.70	.7580	.76
0.80	.7881	.79
0.90	.8159	.81
1.00	.8413	.84
1.10	.8643	.86
1.20	.8849	.88
1.30	.9032	.90
1.40	.9192	.92
1.50	.9332	.93
1.60	.9452	.94
1.70	.9554	.95
1.80	.9641	.96
1.90	.9713	.97
2.00	.9772	.97
2.10	.9821	.98
2.20	.9861	.98
2.30	.9893	.98
2.40	.9918	.99
2.50	.9938	.99
2.60	.9953	.99
2.70	.9965	.99
2.80	.9974	.99
2.90	.9981	.99
3.00	.9987	.99
3.10	.9990	.99
3.20	.9993	.99
3.30	.9995	.99
3.40	.9997	.99
3.50	.9998	.99
3.60	.9998	.99
3.70	.9999	.99
3.80	.9999	.99

Comparing Observed and Expected Frequencies

- We are now interested in examining the magnitudes of the discrepancies between the observed frequencies and the expected frequencies, since we note that **the two sets of frequencies *do not***

Class Interval	$z = (x_i - \bar{x})/s$ At Lower Limit of Interval	Expected Relative Frequency	Expected Frequency
< 100		.0084	.4 } 1.8
100.0–124.9	−2.39	.0291	1.4
125.0–149.9	−1.78	.0815	3.8
150.0–174.9	−1.18	.1653	7.8
175.0–199.9	−.57	.2277	10.7
200.0–224.9	.03	.2269	10.7
225.0–249.9	.64	.1536	7.2
250.0–274.9	1.24	.0753	3.5
275.0–299.9	1.85	.0251	1.2 } 1.5
300.0 and greater	2.45	.0071	.3

Cholesterol Level (mg/dl)	Number of Subjects
100.0–124.9	1
125.0–149.9	3
150.0–174.9	8
175.0–199.9	18
200.0–224.9	6
225.0–249.9	4
250.0–274.9	4
275.0–299.9	3

perfectly. We know that even if our sample were drawn from a normal distribution of values, sampling variability alone would make it highly unlikely that the observed and expected frequencies would agree

- We wonder, then, if the discrepancies between the observed and expected frequencies are small enough that we feel it **reasonable that they could have occurred by chance alone**, when the null hypothesis is true.
- If they are of this magnitude, **we will be unwilling to reject the null hypothesis that the sample came from a normally distributed population**.
- If the discrepancies are so large that it does not seem reasonable that they could have occurred by chance alone when the null hypothesis is true, we will want to reject the null hypothesis. **The criterion against which we judge whether the discrepancies are "large" or "small" is provided by the chi-square distribution**.
- The observed and expected frequencies along with each value of are shown in table on next slide
- The first entry in the last column, for example, is computed from $(1 - 1.8)^2/1.8 = 0.356$. The other values of are computed in a similar manner.

$$(O_i - E_i)^2/E_i$$

The appropriate degrees of freedom are 8 (the number of groups or class intervals)-3 (for the three restrictions: making $\sum E_i$ - $\sum O_i$ and estimating μ and σ from the sample data)=5

Class Interval	Observed Frequency (O_i)	Expected Frequency (E_i)		$(O_i - E_i)^2/E_i$
< 100	0	.4	1.8	.356
100.0–124.9	1	1.4		
125.0–149.9	3	3.8		.168
150.0–174.9	8	7.8		.005
175.0–199.9	18	10.7		4.980
200.0–224.9	6	10.7		2.064
225.0–249.9	4	7.2		1.422
250.0–274.9	4	3.5		.071
275.0–299.9	3	1.2	1.5	1.500
300.0 and greater	0	.3		
Total	47	47		10.566

$$X^2 = \sum[(O_i - E_i)^2/E_i] = 10.566.$$

$\alpha = 0.05$

8. Statistical decision. When we compare χ^2=10.566 with values of χ^2 in appendix table F, we see that **it is less than $\chi^2_{0.95}$** so that, at the .05 level of significance, **we cannot reject the null hypothesis that the sample came from a normally distributed population**.

9. **Conclusion.** We conclude that **in the sampled population, cholesterol levels may follow a normal distribution**.

10.p value. Since 11.070> 10.566 > 9.236, $0.05 < p < 0.10$. In other words, the probability of obtaining a value of χ^2 as large as 10.566, when the null hypothesis is true, is **between 0.05 and 0.10**. Thus we conclude that such an event is not sufficiently rare to reject the null hypothesis that the data come from a normal distribution.

Sometim
that had the r
hypothesis in
ple and our de

Alternativ
square to test
esized distribu
13, was especi

TESTS OF INDEPENDENCE

- We say that two criteria of classification are independent **if the distribution of one criterion is the same no matter what the distribution of the other criterion.**
- For example, if **socioeconomic status** and area of **residence of the inhabitants of a certain** city are independent, we would expect to find the **same proportion of families in the low, medium, and high socioeconomic groups** in all areas of the city.

The Contingency Table

- The classification, **according to two criteria**, of a set of entities, say, people, can be shown by a table in which the **r rows** represent the various levels of one criterion of classification and **the c columns** represent the various levels of the second criterion.
- We will be interested in **testing the null hypothesis that in the population the two criteria of classification are independent**.
- If the hypothesis is rejected, we will conclude that the two criteria of classification **are not independent**.

- A sample of size n is drawn from the population of entities, and **the frequency of occurrence of entities in the sample corresponding to the cells formed by the intersections of the rows and columns** along with the marginal totals is displayed in a table like the one below.

Two-Way Classification of a Sample of Entities

Second Criterion of Classification Level	First Criterion of Classification Level					
	1	2	3	...	c	Total
1	n_{11}	n_{12}	n_{13}	...	n_{1c}	$n_{1.}$
2	n_{21}	n_{22}	n_{23}	...	n_{2c}	$n_{2.}$
3	n_{31}	n_{32}	n_{33}	...	n_{3c}	$n_{3.}$
⋮	⋮	⋮	⋮		⋮	⋮
r	n_{r1}	n_{r2}	n_{r3}	...	n_{rc}	$n_{r.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	...	$n_{.c}$	n

CALCULATING THE EXPECTED FREQUENCIES

- The expected frequency, under the null hypothesis that the two criteria of classification are independent, is **calculated for each cell.**
- If two events are independent, **the probability of their joint occurrence is equal to the product of their individual probabilities**.
- Under the assumption of independence, for example, we compute **the probability** that one of the n subjects represented in table on **previous slide** will be counted in Row 1 **and** Column 1 of the table (that is, in Cell 11).
- In the notation of the table, the desired calculation is $\left(\frac{n_{1.}}{n}\right)\left(\frac{n.1}{n}\right)$
- To obtain the expected **frequency** for Cell 11, **we multiply this probability by the total number of subjects**, n. That is, the expected frequency for Cell 11 is given by $\left(\frac{n_{1.}}{n}\right)\left(\frac{n.1}{n}\right)(n) = \frac{(n1.)(n.1)}{n}$

Two-Way Classification of a Sample of Entities

- The **expected frequencies and observed frequencies are compared**.
- If the discrepancy is **sufficiently small**, the null hypothesis is tenable.
- If the discrepancy is **sufficiently large**, the null hypothesis is rejected, and we conclude that the two criteria of classification are not independent.
- It will be helpful to think of the cells as being numbered **from 1 to k**, where **1 refers to Cell 11** and **k refers to Cell rc**.
- It can be shown that χ^2 as defined in this manner is distributed approximately as χ^2 with $(r-1)(c-1)$ degrees of freedom when the null hypothesis is true.
- If the computed value of χ^2 is **equal to or larger than the tabulated value** of χ^2 for some α, the null hypothesis is **rejected at the α level of significance.**

- In 1992, the U.S. Public Health Service and the Centers for Disease Control and Prevention **recommended that all women of childbearing age consume 400 mg of folic acid daily to reduce the risk of having a pregnancy that is affected by a neural tube defect such as spina bifida or anencephaly**. In a study by Stepanuk et al. (A-3), **636** pregnant women **called a teratology information service** about their use of folic acid supplementation. The researchers wished to determine **if preconceptional use of folic acid and race are independent**. The data appear in table below.

	Preconceptional Use of Folic Acid		
	Yes	No	Total
White	260	299	559
Black	15	41	56
Other	7	14	21
Total	282	354	636

Solution

1. Data:

2. Assumptions. We assume that the sample available for analysis is equivalent to a simple random sample drawn from the population of interest.

3.Hypotheses:

H_0:Race and preconceptional use of folic acid are independent.

H_A:The two variables are not independent.

Let α =0.05

4.Test statistic:

$$X^2 = \sum_{i=1}^{k} \left[\frac{(O_i - E_i)^2}{E_i} \right]$$

5.Distribution of test statistic. When H_0 is true, χ^2 is distributed approximately as χ^2 with $(r-1)(c-1) = (3-1)(2-1) = 2$ degrees of freedom.

17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401

6.Decision rule. Reject H_0 if the computed value of χ^2 is equal to or greater than 5.991.

7.Calculation of test statistics: The expected frequency for the first cell is $\frac{559x282}{636}$=247.86.

- The other expected frequencies are calculated in a similar manner.
- From the observed and expected frequencies we may compute.

$$X^2 = \sum\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

$$= \frac{(260 - 247.86)^2}{247.86} + \frac{(299 - 311.14)^2}{311.14} + \cdots + \frac{(14 - 11.69)^2}{11.69}$$

$$= .59461 + .47368 + \cdots + .45647 = 9.08960$$

8. **Statistical decision.** We reject H_0 since $9.08960 > 5.991$.
9. **Conclusion.** We conclude that H_0 is false, and that there is a relationship between race and preconceptional use of folic acid.
10. ***p* value.** Since $7.378 < 9.08960 < 9.210$, $.01 < p < .025$.

- In the case of a 2×2 contingency table, however, χ^2 may be calculated by the following **sho**

$$X^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$$

- Where a, b, c, and d are the observed cell frequencies as shown in table below.
- When we apply the $(r - 1)(c - 1)$ rule for finding degrees of freedom to a $2x2$ table, the result is **1 degree of freedom**.

Second Criterion of Classification	**First Criterion of Classification**		
	1	**2**	**Total**
1	*a*	*b*	*a* + *b*
2	*c*	*d*	*c* + *d*
Total	*a* + *c*	*b* + *d*	*n*

17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401

lass of week end

- According to Silver and Aiello (A-4), falls are of major concern among polio survivors. Researchers wanted to determine **the impact of a fall on lifestyle changes**. Table below shows the results of a study of 233 polio survivors on **whether fear of falling resulted in lifestyle changes**.

	Made Lifestyle Changes Because of Fear of Falling		
	Yes	No	Total
Fallers	131	52	183
Nonfallers	14	36	50
Total	145	88	233

Solution:

1. **Data.** gency
2. **Assum** dom s
3. **Hypot** H_0: Fa ind H_1: Th Let α
4. **Test st**
5. **Distri** mately degree
6. **Decisi** greater
7. **Calcul** X^2
8. **Statist**
9. **Conclusion.** W between experi falling.
10. **p value.** Since

TESTS OF HOMOGENEITY

- Either row or column totals may be under the control of the investigator; that is, the investigator may **specify that independent samples be drawn from each of several populations**.
- In this case, **one set of marginal totals is said to be fixed**, while the other set, **corresponding to the criterion of classification** applied to the samples, is **random**.
- The test of independence is concerned with the question: **Are the two criteria of classification independent?**
- The homogeneity test is concerned with the question: **Are the samples drawn from populations that are homogeneous with respect to some criterion of classification**?
- In the latter case the null hypothesis states that the samples are **drawn from the same population**.
- Despite these differences in concept and sampling procedure, the two tests are mathematically identical.

Calculating Expected Frequencies

- **Either the row categories or the column** categories may represent the different populations from which the samples are drawn. If, for example, **three populations** are sampled, they may be designated as populations 1, 2, and 3, in which case these labels may serve as **either row or column headings**. If the variable of interest has three categories, say, A, B, and C, these labels may serve as **headings for rows or columns**, **whichever is not used for the populations**.
- The contingency table for this situation, with columns used to represent the populations, is shown as table below:

- **If the populations are indeed homogeneous, or, equivalently, if the samples are all drawn from the same population, with respect to the categories A, B, and C, our best estimate of the proportion in the combined population who belong to category A is $n_{A.}/n$. We interpret this probability as applying to each of the populations individually.**

A Contingency Table for Data for a Chi-Square Test of Homogeneity

	Population			
Variable Category	**1**	**2**	**3**	**Total**
A	n_{A1}	n_{A2}	n_{A3}	$n_{A.}$
B	n_{B1}	n_{B2}	n_{B3}	$n_{B.}$
C	n_{C1}	n_{C2}	n_{C3}	$n_{C.}$
Total	$n_{.1}$	$n_{.2}$	$n_{.3}$	n

Chi-Square
Variable Ca
A
B
C
Total

- For example, **under the null hypothesis, $n_{A.}/n$ is our best estimate of the probability that a subject picked at random from the combined population will belong to category A**.
- We would expect, then, to find $n_{.1}(n_{A.}/n)$ of those in the sample from population 1 to belong to category A, $n_{.2}(n_{A.}/n)$ of those in the sample from population 2 to belong to category A, and $n_{.3}(n_{A.}/n)$ of those in the sample from population 3 to belong to category A.
- These calculations **yield the expected frequencies** for the first row of table on previous slide.
- Similar reasoning and calculations yield the expected frequencies for the other two rows.

"*The shortcut procedure of multiplying appropriate marginal totals and dividing by the grand total yields the expected frequencies for the cells*".

- From the data in table on previous slide we compute the following test statistic:

$$X^2 = \sum_{i=1}^{k}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

- For exampl
of the prob
combined
- We would e
population
sample fro
those in the
- These calcu
table on pre
- Similar rea
the other tw
"The shortcu
and dividing
cells".
- From the da
statistic:

17	5.697	7.564	8.672	24.769	27.587	30.191	33.409	35.718
18	6.265	8.231	9.390	25.989	28.869	31.526	34.805	37.156
19	6.844	8.907	10.117	27.204	30.144	32.852	36.191	38.582
20	7.434	9.591	10.851	28.412	31.410	34.170	37.566	39.997
21	8.034	10.283	11.591	29.615	32.671	35.479	38.932	41.401

p_2, the sam-
we pool the

of 1.96.

roach, our contin-

- Narcolepsy is a disease involving **disturbances of the sleep–wake cycle**. Members of the German Migraine and Headache Society (A-8) studied the relationship between migraine headaches in 96 subjects diagnosed with narcolepsy and 96 healthy controls. The results are shown in table below. *We wish to know if we may conclude, on the basis of these data, that the narcolepsy population and healthy populations represented by the samples are not homogeneous with respect to migraine frequency*.

	Reported Migraine Headaches		
	Yes	No	Total
Narcoleptic subjects	21	75	96
Healthy controls	19	77	96
Total	40	152	192

critical value of
oth methods. This
z^2. We note that

Solution:

1. Da
2. As
ea
3. Hy
H_0
H_A
Le
4. Te
5. Dis
ma
6. De
gre
7. Ca
12.

RELATIVE RISK and ODDS RATIO

- Another important class of scientific investigation that is widely used is the **observational study**.
- "***An observational study is a scientific investigation in which*** *neither the subjects under study nor any of the variables of interest are manipulated in any way*".
- It may be defined simply as **an investigation that is not an experiment**.
- The simplest form of observational study is one in which there are **only two variables of interest**.
- One of the variables is called the **risk factor**, or **independent variable**, and the other variable is referred to as the **outcome**, or **dependent variable.**
- ***"The term risk factor is used to designate a variable that is thought to be related to some outcome variable. The risk factor may be a suspected cause of some specific state of the outcome variable".***

Types of Observational Studies

- There are two basic types of observational studies, **prospective studies** and **retrospective studies**.
- **"*A prospective*** *study is an observational study in which two random samples of subjects are selected. One sample consists of subjects who possess the risk factor, and the other sample consists of subjects who do not possess the risk factor.*
- *The subjects are followed into the future (that is, they are followed prospectively), and* ***a record is kept on the number of subjects in each sample who, at some point in time, are classifiable into each of the categories of the outcome variable***".
- The data resulting from a prospective study involving **two dichotomous variables** can be displayed in **a contingency table** that usually provides information regarding the number of subjects with and without the risk factor and the number who did and did not succumb to the disease of interest as well as the frequencies for each combination of categories of the two variables.

to Disease St

Risk Factor
Present
Absent
Total

“***A retrospective study is the reverse of a prospective study***. *The samples are selected* ***from those falling into the categories of the outcome variable***. *The investigator then looks back (that is, takes a retrospective look) at the subjects and determines which ones have (or had) and which ones do not have (or did not have) the risk factor*”.

- In general, the prospective study is more expensive to conduct than the retrospective study.
- The prospective study, however, more closely resembles an experiment.

Relative Risk

- The data resulting from a prospective study in which the **dependent variable** and the **risk factor** are both dichotomous may be displayed in a contingency table such as table below

Classification of a Sample of Subjects with Respect to Disease Status and Risk Factor

	Disease Status		
Risk Factor	**Present**	**Absent**	**Total at Risk**
Present	a	b	$a + b$
Absent	c	d	$c + d$
Total	$a + c$	$b + d$	n

- **The risk** of the development of the disease among the subjects with the risk factor is $\boldsymbol{a/(a+b)}$.
- **The risk** of the development of the disease among the subjects without the risk factor is $\boldsymbol{c/(c+d)}$. We define relative risk as follows.

“***Relative risk*** *is the ratio of the risk of developing a disease among subjects with the risk factor to the risk of developing the disease among subjects without the risk factor*”.

- We represent the relative risk from a prospective study symbolically as $\boldsymbol{\widehat{RR} = \frac{a/(a+b)}{c/(c+d)}}$ $P(D\backslash R)/P(D\backslash \bar{R})$
- Where a, b, c, and d are as defined in table on previous slide, and $\widehat{RR}$ indicates that the relative risk is computed from a sample to be used as an estimate of the relative risk, RR, for the population from which the sample was drawn.
- We may construct a confidence interval for RR

$$100(1-\alpha)\%\text{CI} = \widehat{RR}^{1 \pm (z_\alpha/\sqrt{X^2})}$$

- Where z_α is **the two-sided z value** corresponding to the chosen confidence coefficient and χ^2 is computed by

$$X^2 = \frac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

Interpretation of RR

- RR may range anywhere **between zero and infinity**.
- **A value of 1** indicates that there is no association between the status of the risk factor and the status of the dependent variable.
- In most cases the two possible states of the dependent variable are **disease present** and **disease absent**.
- We interpret **an RR of 1** to mean that the risk of acquiring the disease is the **same for those subjects with the risk factor and those without the risk factor**.
- **A value of RR greater than 1** indicates that the risk of acquiring the disease is greater among subjects with the risk factor than among subjects without the risk factor.
- **An RR value that is less than 1** indicates less risk of acquiring the disease **among subjects with the risk factor** than among subjects without the risk factor.
- For example, a relative risk of 2 is taken to mean that those subjects with the risk factor are twice as likely to acquire the disease as compared to subjects without the risk factor.

❖ In a prospective study of pregnant women, Magann et al. (A-16) collected extensive information on **exercise level** of low-risk pregnant working women. A group of **217** women did no voluntary or mandatory exercise during the pregnancy, while a group of **238** women exercised extensively. **One outcome variable of interest was experiencing preterm labor**. The results are summarized in table below. **We wish to estimate the relative risk of preterm labor when pregnant women exercise extensively.**

Subjects with and without the Risk Factor Who Became Cases of Preterm Labor

Risk Factor	Cases of Preterm Labor	Noncases of Preterm Labor	Total
Extreme exercising	22	216	238
Not exercising	18	199	217
Total	40	415	455

Solution:

- By $\widehat{RR} = \dfrac{a/(a+b)}{c/(c+d)}$ we compute $\widehat{RR} = \dfrac{22/238}{18/217} = \dfrac{.0924}{.0829} = 1.1$
  - These data indicate that the risk of experiencing preterm labor when a woman exercises heavily is **1.1 times** as great as it is among women who do not exercise at all.
  - We compute the 95 percent confidence interval for RR as follows: By Equation $X^2 = \dfrac{n(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$, we compute from the data in table on previous slide $X^2 = \dfrac{455[(22)(199)-(216)(18)]^2}{(40)(415)(238)(217)} = .1274$
  - By Equation $100(1-\alpha)\%CI = \widehat{RR}^{1 \pm (z_{\alpha}/\sqrt{X^2})}$,the lower and upper confidence limits are, respectively, $1.1^{1-1.96/\sqrt{0.1274}}$=0.65 and $1.1^{1+1.96/\sqrt{0.1274}}$=1.86. Since the interval includes 1, we conclude, at the 0.05 level of significance, that **the population risk may be 1**. In other words, we conclude that, in the population, **there may not be an increased risk of experiencing preterm labor when a pregnant woman exercises extensively**.

Preterm Labor

Risk Factor
Extreme exercisi
Not exercising
Total

By Equation 12
tively, $1.1^{1-1.96/\sqrt{}}$
includes 1, we c
risk may be 1. In
not be an increa
woman exercises
The data w
12.7.1. The relat
of the output, al
errors, these valu

ODDS RATIO

- When the data to be analyzed come from a retrospective study, **relative risk is not a meaningful measure** for comparing two groups.
- A retrospective study is based on a sample of subjects **with the disease** (cases) and a separate sample of subjects **without the disease** (controls or non cases).
- Given the results of a retrospective study involving two samples of subjects, cases, and controls, we may display the data in a $2x2$ table such as **table on next slide**, in which subjects are dichotomized with respect to the presence and absence of the risk factor.
- Note that the column headings in **table on next slide** differ from those in **table on slide 82** *to emphasize the fact that the data are from a retrospective study and that the subjects were selected because they were either cases or controls.*

According t They Are C
Risk Factor
Present
Absent
Total

Subjects of a Retrospective Study Classified According to Status Relativeto a Risk Factor and Whether They Are Cases or Controls

	Sample		
Risk Factor	**Cases**	**Controls**	**Total**
Present	a	b	$a + b$
Absent	c	d	$c + d$
Total	$a + c$	$b + d$	n

- When the data from a retrospective study are displayed as in **table above**, the ratio $a/(a+b)$, for example, is **not an estimate of the risk of disease for subjects with the risk factor**.
- The appropriate measure for comparing cases and controls in a retrospective study is **the odds ratio**

"*The odds for success* ***are the ratio of the probability of success to the probability of failure***".

- We use this definition of odds to define two odds that we can calculate from data displayed as in **table on previous slide**:

1. The odds of being a case (having the disease) to being a control (not having the disease) among subjects with the risk factor is $[a/(a+b)]/[b/(a+b)] = a/b$.
2. The odds of being a case (having the disease) to being a control (not having the disease) among subjects without the risk factor is $[c/(c+d)]/[d/(c+d)] = c/d$.

- We now define the odds ratio that we may compute from the data of a retrospective study.
- The symbol $\widehat{OR}$ indicates that the measure is computed from sample data and used as an estimate of the population odds ratio, OR.

According t
They Are C

Risk Factor
Present
Absent
Total

- The estimate of the population odds ratio is $\widehat{OR} = \frac{a/b}{c/d} = \frac{ad}{bc}$. where a, b, c, and d are as defined in **table on slide 88**
- **We may construct a confidence interval for OR by the following method:** $100(1-\alpha)\%\text{CI} = \widehat{OR}^{1 \pm (z_\alpha/\sqrt{X^2})}$
- Where z_α is **the two-sided z** value corresponding to the chosen confidence coefficient and χ^2 is computed by $X^2 = \frac{n(ad - bc)^2}{(a+c)(b+d)(a+b)(c+d)}$

According t
They Are C

Risk Factor
Present
Absent
Total

Interpretation of the Odds Ratio

- **In the case of a rare disease**, the population odds ratio provides **a good approximation to the population relative risk**.
- Consequently, the sample odds ratio, being an estimate of the population odds ratio, provides **an indirect estimate of the population relative risk in the case of a rare disease**.

- ➢ The odds ratio can assume values **between zero and ∞**.
- ➢ A value of **1** indicates **no association between the risk factor and disease status**.
- ➢ A value less than 1 indicates **reduced odds** of the disease among subjects with the risk factor.
- ➢ A value greater than 1 indicates **increased odds** of having the disease among subjects in whom the risk factor is present.
- ❖ Toschke et al. (A-17) collected data on obesity status of children ages 5–6 years and the smoking status of the mother during the pregnancy. **Table on next slide** shows 3970 subjects classified as cases or noncases of obesity and also classified according to smoking status of the mother during pregnancy (the risk factor). ***We wish to compare the odds of obesity at ages 5–6 among those whose mother smoked throughout the pregnancy with the odds of obesity at age 5–6 among those whose mother did not smoke during pregnancy***.

Status and

Smoking St During Preg
Smoked thr
Never smok
Total

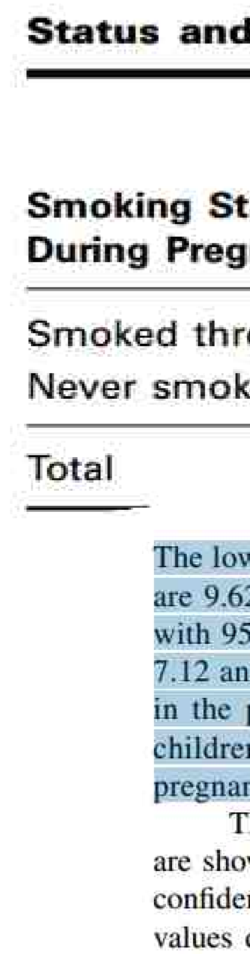
Subjects Classified According to Obesity Status and Mother's Smoking Status during Pregnancy

	Obesity Status		
Smoking Status During Pregnancy	Cases	Noncases	Total
Smoked throughout	64	342	406
Never smoked	68	3496	3564
Total	132	3838	3970

Solution

- We compute $\widehat{OR} = \frac{(64)(3496)}{(342)(68)} = 9.62.$
- We see that obese children (cases) **are 9.62 times as likely as nonobese children** (noncases) to have had a mother who smoked throughout the pregnancy.

- We compute the 95 percent confidence interval for OR as follows. By $X^2 = \frac{n(ad - bc)^2}{(a + c)(b + d)(a + b)(c + d)}$ we compute from the data in Table **on previous slide** $X^2 = \frac{3970[(64)(3496) - (342)(68)]^2}{(132)(3838)(406)(3564)} = 217.6831$
- The lower and upper confidence limits for the population OR, respectively, are $9.62^{1-1.96/\sqrt{217.6831}} = 7.12$ and $9.62^{1+1.96/\sqrt{217.6831}} = 13.00.$
- We conclude with 95 percent confidence that the population OR is somewhere between 7.12 and 13.00. Because the interval **does not include 1**, we conclude that, in the population, obese children (cases) are **more likely than nonobese children (noncases) to have had a mother who smoked throughout the pregnancy**.



SIMPLE LINEAR REGRESSION AND CORRELATION

- It is frequently desirable to learn something about **the relationship between two numeric variables**.
- We may, for example, be interested in studying the relationship between **blood pressure and age**, **height and weight**, **the concentration of an injected drug and heart rate**, the consumption level of some nutrient and weight gain, the intensity of a stimulus and reaction time, or total family income and medical care expenditures.
- **The nature and strength** of the relationships between variables such as these may be examined by **regression** and **correlation analysis**, two statistical techniques that, although **related**, serve **different purposes.**

https://stats.libretexts.org/Bookshelves/Introductory_Statistics/Book%3A_Introductory_Statistics_(Shafer_and_Zhang)/10%3A_Correlation_and_Regression/10.E%3A_Correlation_and_Regression_(Exercises)

Regression

- Regression analysis is helpful in assessing specific **forms of the relationship** between variables, and the ultimate objective when this method of analysis is employed usually is **to predict or estimate the value of one variable corresponding to a given value of another variable**

CORRELATION

Is concerned with measuring **the strength of the relationship between variables**. When we compute measures of correlation from a set of data, we are interested in **the degree of the correlation between variables.**

THE REGRESSION MODEL

- It is important, therefore, that the researchers understand **the nature** of the population in which they are interested to be able either **to construct a mathematical model for its representation or to determine if it reasonably fits some** ***established model***.
- Researchers, should be able **to distinguish between the occasion when their chosen models and the data are sufficiently compatible** for them to proceed and the case where **their chosen model must be abandoned**.

ASSUMPTIONS UNDERLYING SIMPLE LINEAR REGRESSION

- **Two** variables, usually labeled X and Y, are of interest.

The letter X is usually used to designate a variable referred to as the **independent** variable, since frequently it is **controlled by the investigator:** values of X may be selected by the investigator and, corresponding to each preselected value of X, **one or more values of another variable, labeled Y, are obtained**.

- The variable, Y, accordingly, is called the **dependent variable**, and we speak of **the regression of Y on X**.
- The following are the **assumptions underlying the simple linear regression model**.

1. Values of the independent variable X are said to be "**fixed**."

   X is referred to by some writers as **a nonrandom variable** and by others as **a mathematical variable (classical regression model):** Regression analysis also can be carried out on data in which X is a random variable.

2. The variable X is **measured without error**. Since no measuring procedure is perfect, this means that **the magnitude of the measurement error in X is negligible.**

3. **For each value of X there is a subpopulation of Y values**. For the usual inferential procedures of estimation and hypothesis testing to be valid, **these subpopulations must be normally distributed**

Inkoko zitera ar
bangahe, amag

4. **The variances** of the subpopulations of Y **are all equal** and denoted by σ^2.

5. **The means** of the subpopulations of Y all **lie on the same straight line**. This is known as the assumption of linearity. This assumption may be expressed symbolically as $\mu_{y|x} = \beta_0 + \beta_1 x$

- ✓ where $\mu_{y/x}$ is **the mean of the subpopulation of Y values for a particular value of X**, and β_0 and β_1 are called **the population regression coefficients**.

6. The Y values are **statistically independent**. In other words, in drawing the sample, it is assumed that the values of Y chosen at one value of X **in no way** depend on the values of Y chosen at another value of X.

- ➢ These assumptions may be summarized by means of the following equation, which is called **the regression model**:

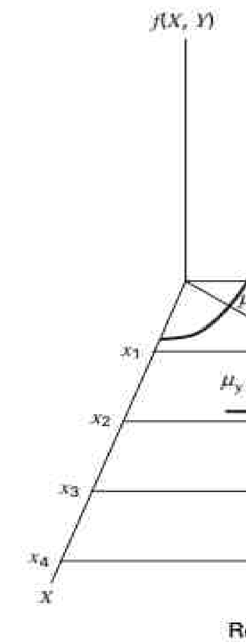
$$y = \beta_0 + \beta_1 x + \epsilon$$

Where **y is a typical value** from one of the subpopulations of Y and ϵ is called **the error term**

$$\begin{aligned} \epsilon &= y - (\beta_0 + \beta_1 x) \\ &= y - \mu_{y|x} \end{aligned}$$

ϵ shows the **amount by which y deviates from the mean of the subpopulation of Y values from which it is drawn**.

As a consequence of the assumption that the subpopulations of Y values are normally distributed with equal variances, **the ϵ's for each subpopulation are normally distributed with a variance equal to the common variance of the subpopulations of Y values**.



THE SAMPLE REGRESSION EQUATION

- The researcher's interest is **the population regression equation**—the equation that describes the true relationship between the dependent variable Y and the independent variable X.
- The variable designated by Y is sometimes called the **response variable** and X is sometimes called the **predictor variable**.
- In an effort to reach a decision regarding **the likely form** of this relationship, **the researcher draws a sample from the population of interest and using the resulting data**, computes ***a sample regression equation that forms the basis for reaching conclusions regarding the unknown population regression equation.***

STEPS IN REGRESSION ANALYSIS

1. Determine **whether or not the assumptions underlying a linear relationship are met** in the data available for analysis.
2. Obtain the **equation for the line** that best fits the sample data.
3. **Evaluate the equation** to obtain some idea of the strength of the relationship and the usefulness of the equation for predicting and estimating.
4. If the data appear to conform satisfactorily to the linear model, use the equation obtained from the sample data to predict and to estimate.

- When we use **the regression equation to predict**, we will be predicting **the value Y is likely to have** when X has a given value.
- When we use **the equation to estimate**, we will be estimating **the mean of the subpopulation of Y values assumed to exist** at a given value of X.
- When the equation is used to predict and to estimate Y, **only the corresponding values of X will be known**.

Steps in Regr
regarding the nature
assume initially that
lowing steps.

32	88.10	89.31	69	105.00	97.13	106	93.30	62.20
33	90.80	78.94	70	107.00	166.00	107	101.80	133.00
34	89.40	83.55	71	107.00	87.99	108	107.90	208.00
35	102.00	127.00	72	101.00	154.00	109	108.50	208.00
36	94.50	121.00	73	97.00	100.00			
37	91.00	107.00	74	100.00	123.00			

❖ Després et al. (A-1) point out **that the topography of adipose tissue (AT) is associated with metabolic complications considered as risk factors for cardiovascular disease**. *"It is important, they state, to measure the amount of intra abdominal AT as part of the evaluation of the cardiovascular-disease risk of an individual"*. Computed tomography (CT), the only available technique that precisely and reliably measures the amount of deep abdominal AT, however, **is costly and requires irradiation of the subject**. In addition, the technique is not available to many physicians. Després and his colleagues conducted a study to develop equations to predict the amount of deep abdominal AT from simple anthropometric measurements. Their subjects were men between the ages of 18 and 42 years who were free from metabolic disease that would require treatment. Among the measurements taken on each subject were **deep abdominal AT** obtained by CT and waist circumference as shown on next slide.

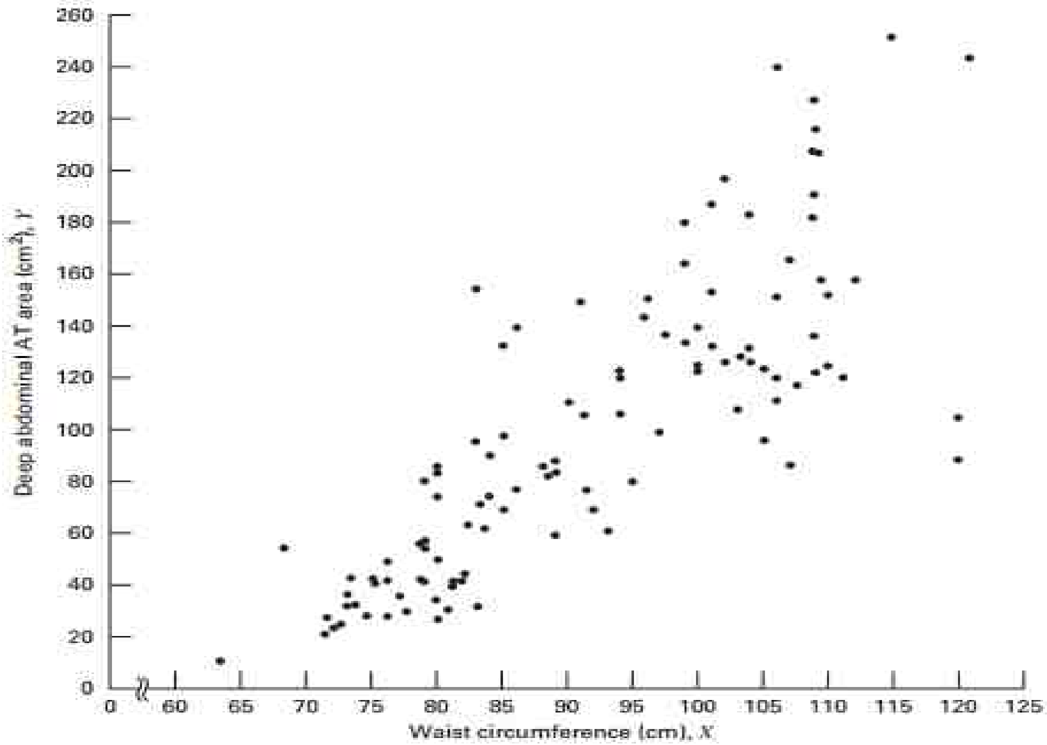
TABLE 9.3.1 Waist Circumference (cm), *X*, and Deep Abdominal AT, *Y*, of 109 Men

	Subject	*X*	*Y*	Subject	*X*	*Y*	Subject	*X*	*Y*
	1	74.75	25.72	38	103.00	129.00	75	108.00	217.00
	2	72.60	25.89	39	80.00	74.02	76	100.00	140.00
	3	81.80	42.60	40	79.00	55.48	77	103.00	109.00
	4	83.95	42.80	41	83.50	73.13	78	104.00	127.00
	5	74.65	29.84	42	76.00	50.50	79	106.00	112.00
3	6	71.85	21.68	43	80.50	50.88	80	109.00	192.00
	7	80.90	29.08	44	86.50	140.00	81	103.50	132.00
	8	83.40	32.98	45	83.00	96.54	82	110.00	126.00
1	9	63.50	11.44	46	107.10	118.00	83	110.00	153.00
	10	73.20	32.22	47	94.30	107.00	84	112.00	158.00
	11	71.90	28.32	48	94.50	123.00	85	108.50	183.00
	12	75.00	43.86	49	79.70	65.92	86	104.00	184.00
	13	73.10	38.21	50	79.30	81.29	87	111.00	121.00
	14	79.00	42.48	51	89.80	111.00	88	108.50	159.00
	15	77.00	30.96	52	83.80	90.73	89	121.00	245.00
2	16	68.85	55.78	53	85.20	133.00	90	109.00	137.00
	17	75.95	43.78	54	75.50	41.90	91	97.50	165.00
	18	74.15	33.41	55	78.40	41.71	92	105.50	152.00
	19	73.80	43.35	56	78.60	58.16	93	98.00	181.00
	20	75.90	29.31	57	87.80	88.85	94	94.50	80.95
	21	76.85	36.60	58	86.30	155.00	95	97.00	137.00
	22	80.90	40.25	59	85.50	70.77	96	105.00	125.00
	23	79.90	35.43	60	83.70	75.08	97	106.00	241.00
	24	89.20	60.09	61	77.60	57.05	98	99.00	134.00
	25	82.00	45.84	62	84.90	99.73	99	91.00	150.00
	26	92.00	70.40	63	79.80	27.96	100	102.50	198.00
	27	86.60	83.45	64	108.30	123.00	101	106.00	151.00
	28	80.50	84.30	65	119.60	90.41	102	109.10	229.00
	29	86.00	78.89	66	119.90	106.00	103	115.00	253.00
	30	82.50	64.75	67	96.50	144.00	104	101.00	188.00
	31	83.50	72.56	68	105.50	121.00	105	100.10	124.00
	32	88.10	89.31	69	105.00	97.13	106	93.30	62.20
	33	90.80	78.94	70	107.00	166.00	107	101.80	133.00
	34	89.40	83.55	71	107.00	87.99	108	107.90	208.00
	35	102.00	127.00	72	101.00	154.00	109	108.50	208.00
	36	94.50	121.00	73	97.00	100.00			
	37	91.00	107.00	74	100.00	123.00			

- A question of interest is ***how well one can predict and estimate deep abdominal AT from knowledge of the waist circumference***. **This question is typical of those that can be answered by means of regression analysis**.
- Since deep abdominal AT is the variable about which we wish to make predictions and estimations, it is **the dependent variable**. The variable waist measurement, knowledge of which will be used to make the predictions and estimations, **is the independent variable**.

The Scatter Diagram

- A **first step** that is usually useful in studying the relationship between two variables is to prepare **a scatter diagram** of the data such as is shown on next slide.
- The points **seem to be scattered around an invisible straight line**.
- It looks as if it would be simple to draw, freehand, through the data points the line that describes the relationship between X and Y. It is highly unlikely, however, that the lines drawn by any two people would be exactly the same.



Scatter diagram of data shown on slide 103

- In other words, for every person drawing such a line by eye, or freehand, we would expect **a slightly different line**.
- The question then arises as to **which line best describes the relationship between the two variables**.

THE LEAST-SQUARES LINE

- The method usually employed for obtaining the desired line is known as **the method of least squares**, and the resulting line is called the least-squares line.
- We recall from algebra that **the general equation for a straight** line may be written as $y = a + bx$ where y is a value on the vertical axis, x is a value on the horizontal axis, a is the point where the line crosses the vertical axis, and b shows the amount by which y changes for each unit change in x.We refer to a as the **y-intercept** and b as the **slope of the line**.

Obtaining the Least-Square Line

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_1 - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

Where x_i and y_i are the corresponding values of each data point (X, Y), $\bar{x}$ and $\bar{y}$ are the means of the X and Y sample data values, respectively, and $\widehat{B}_0$ and $\hat{B}_1$ and are **the estimates** of the intercept B_0 and slope B_1 , respectively, of the population regression line

$\hat{y}$ = -216+ 3.46x

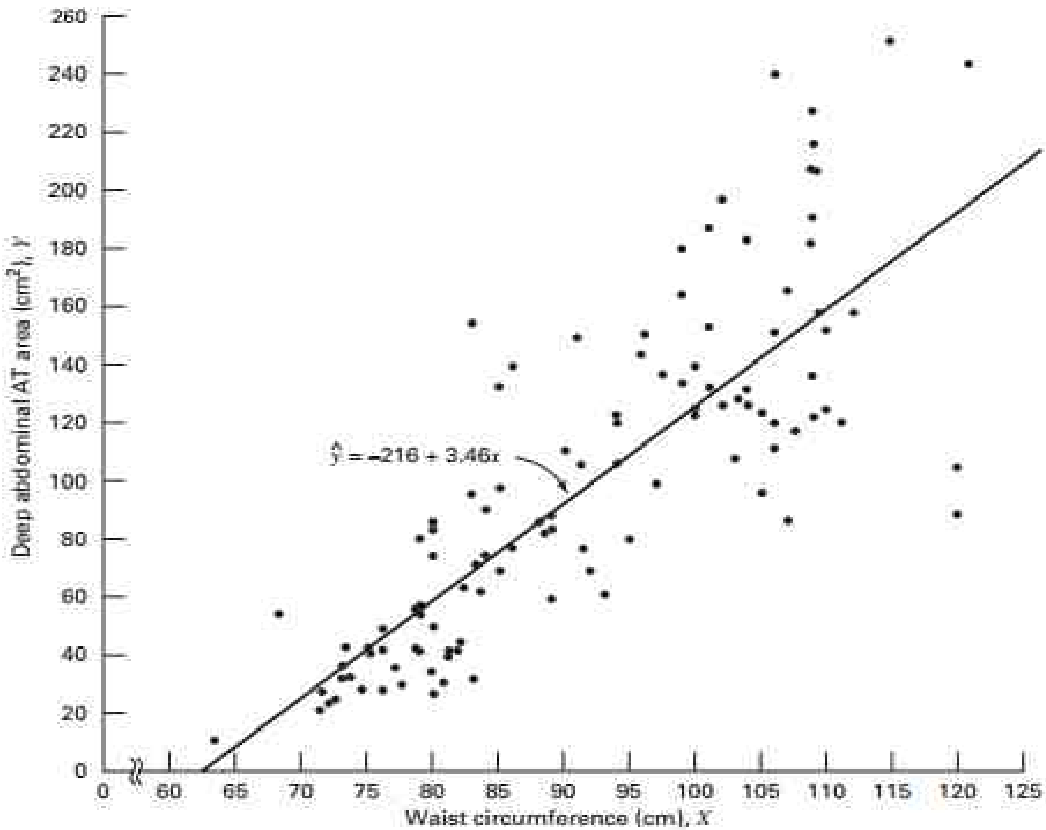
- This equation tells us that since $\boldsymbol{B_0}$ **is negative**, the line crosses the Y-axis below the origin, and that since $\boldsymbol{B_1}$ **the slope, is positive**, the line extends from the lower left-hand corner of the graph to the upper right-hand corner.
- We see further that for **each unit increase in x, y increases by an amount equal to 3.46.**
- The symbol y denotes **a value of y computed from the equation, rather than an observed value of** Y.
- By substituting **two convenient values of X into Equation**, we may obtain the necessary coordinates for drawing the line (see next slide)

$$\hat{y} = -216 + 3.46(70) = 26.2$$

If we let $X = 110$ we obtain

$$\hat{y} = -216 + 3.46(110) = 164$$

The line, along with the original data, is shown in Figure 9.3.3.



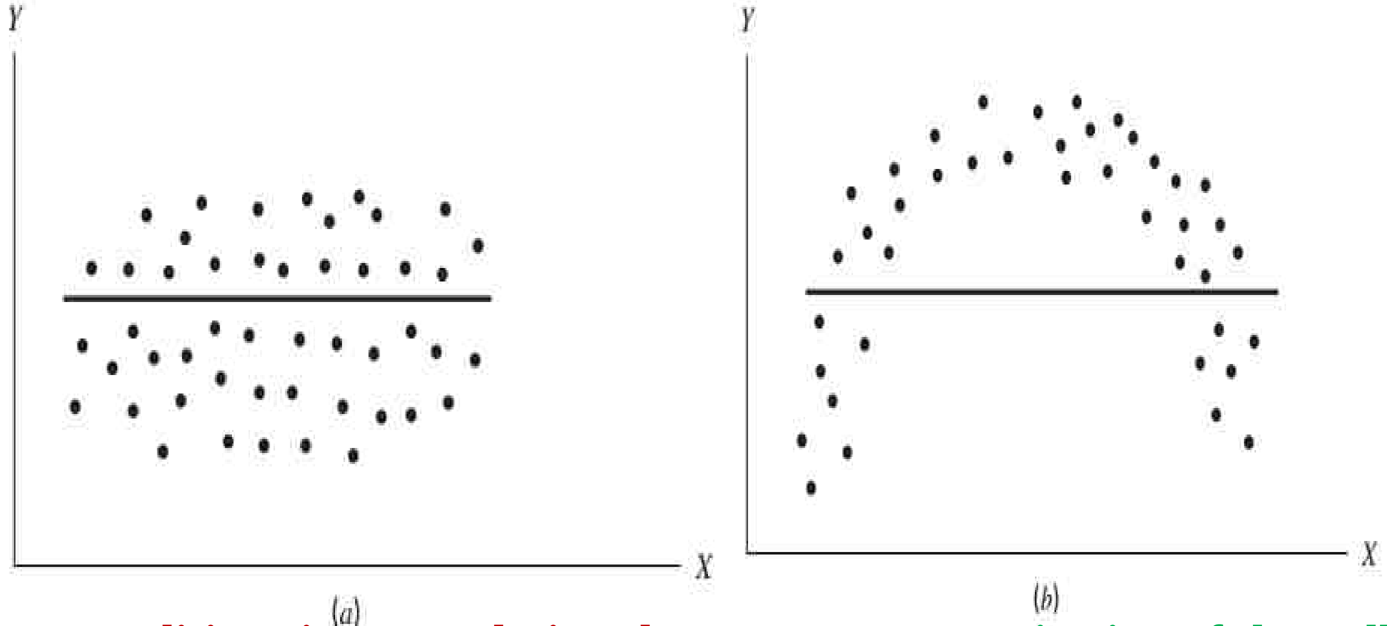
- ➢ The line that we have drawn through the points is **best in this sense**:
- ❖ The sum of the **squared vertical deviations** of the observed data points (y_i) **from the least-squares** line is **smaller** than the sum of the squared vertical deviations of the data points from any other line.
- ✓ In other words, if we square the vertical distance from each observed point (y_i) to the least-squares line and add these squared values for all points, the resulting total will be smaller than the similarly computed total for any other line that can be drawn through the points.
- ✓ **For this reason the line we have drawn is called the least-squares line**.

EVALUATING THE REGRESSION EQUATION

- Once the regression equation has been obtained it must be evaluated **to determine whether it *adequately* describes the relationship between the two variables** and *whether it can be used effectively for prediction and estimation purposes*.

When $H_0 : B_1$=0 Is Not Rejected.

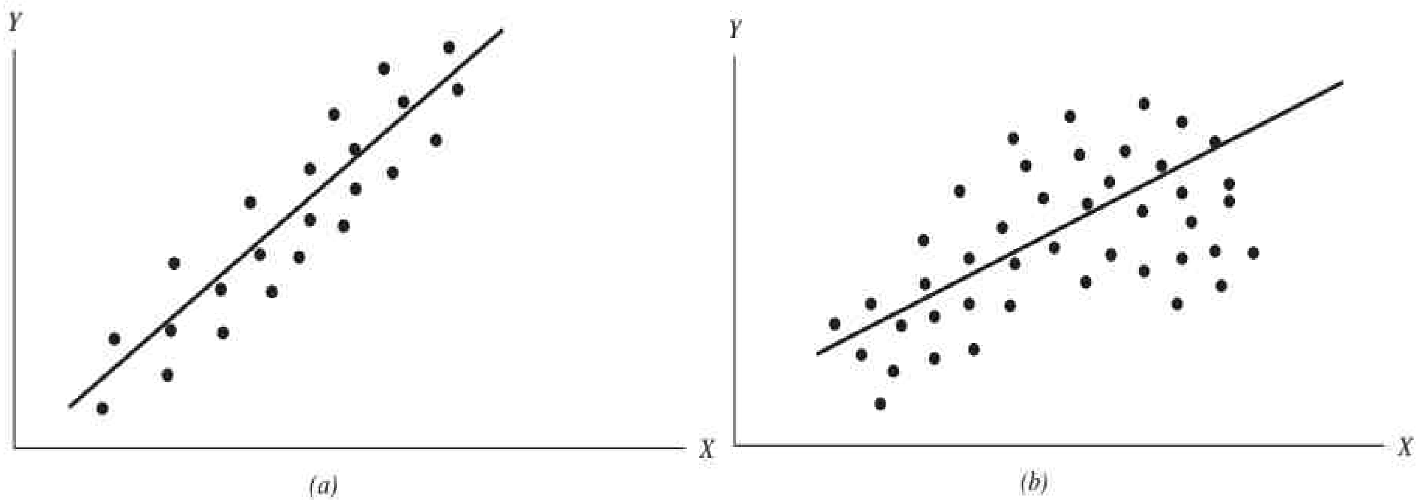
- If in the population the relationship between X and Y is linear, B_1 , the slope of the line that describes this relationship, will be either **positive**, **negative**, or **zero**.
- Following a test in which **the null hypothesis that equals zero is not rejected**, **we may conclude** (assuming that we have not made a type II error by accepting a false null hypothesis) either **(1)** that although the relationship between X and Y **may be linear** it is ***not strong enough* for X to be of much value** in predicting and estimating Y, or **(2)** that the relationship between X and Y is **not linear**; that is, some ***curvilinear*** model provides a better fit to the data.

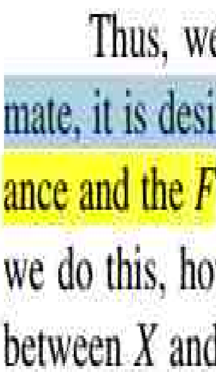


Conditions in a population that may prevent rejection of the null hypothesis that (a) The relationship between X and Y is linear, but B_1 is so close to zero that sample data are not likely to yield equations that are useful for predicting Y when X is given. (b) The relationship between X and Y is not linear; **a curvilinear model provides a better fit to the data**; sample data are not likely to yield equations that are useful for predicting Y when X is given.

WHEN H_0: β_1= 0 IS REJECTED

- Rejection of the null hypothesis that β_1=0 may be attributed to one of the following conditions in the population: **(1)** the relationship **is linear and of sufficient strength to justify the use of sample regression equations to predict and estimate Y for given values of X**; and **(2)** there is a good fit of the data to a linear model, but some curvilinear model might provide an even better fit

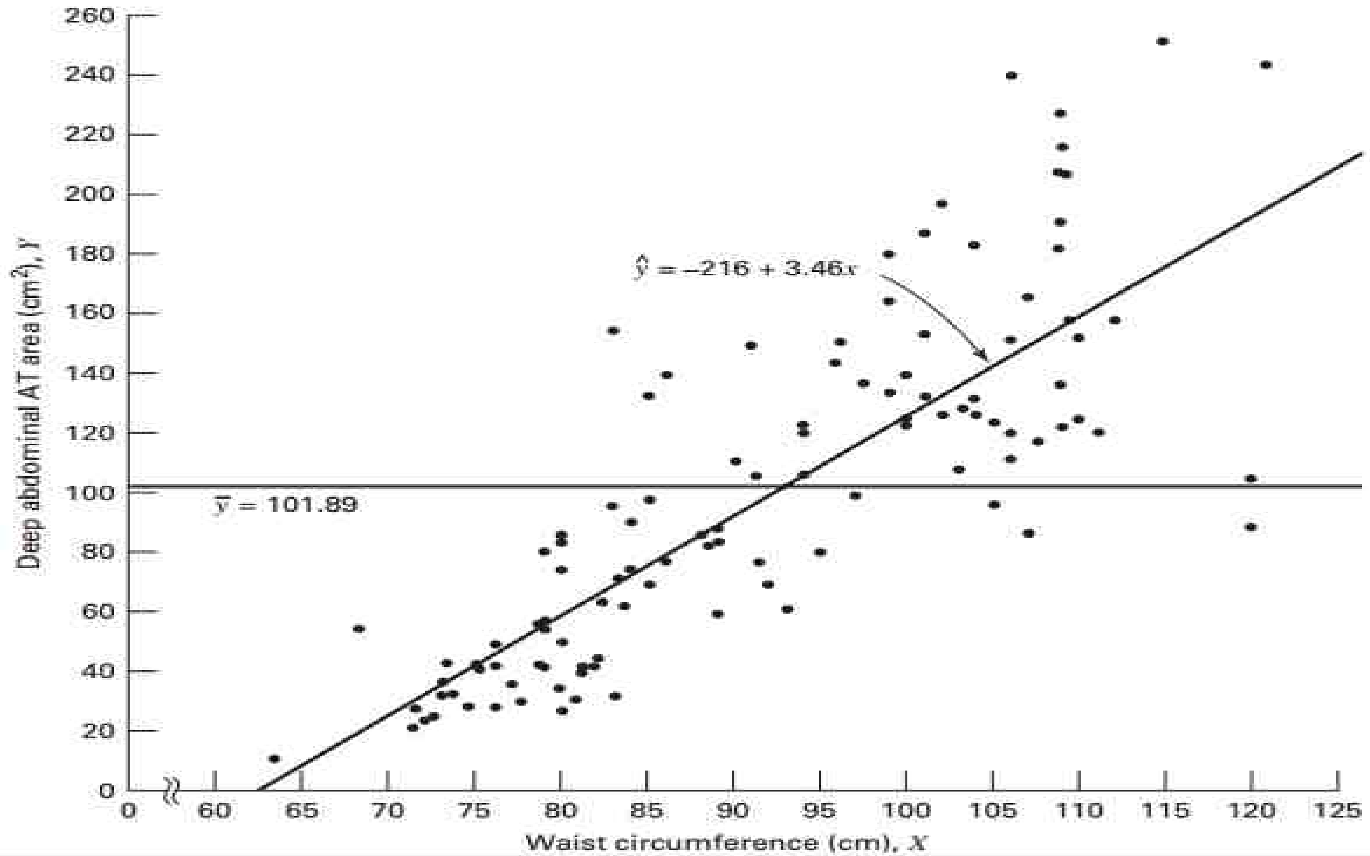



Thus, we
mate, it is desi
ance and the F
we do this, ho
between X and

Before using a sample regression equation to predict and estimate, it is desirable to test H_0: β_1= 0

THE COEFFICIENT OF DETERMINATION

One way ***to evaluate the strength*** of the regression equation is **to compare the scatter of the points about the regression line with the scatter about the mean of the sample values of Y**.



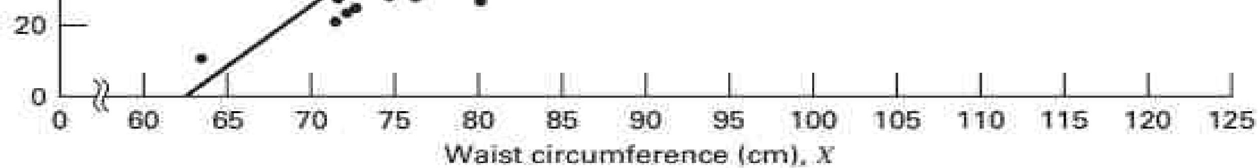
- It appears rather obvious from figure on previous slide that the scatter of the points about the regression line is **much less** than the scatter about the $\bar{y}$ line but **the situation may not be always this clear-cut**, so that an ***objective* measure of some sort would be much more desirable**.
- Such an objective measure, is called **the coefficient of determination**.
- Let us first justify the use of **the coefficient of determination** by examining the logic behind its computation.

The Total Deviation:

- The measurement of the vertical distance of any observed value from the $\bar{y}$ line is called the total deviation, $(y_i - \bar{y})$.

The Explained Deviation

- The measurement of the vertical distance from the regression line to the line $\bar{y}$, $(\hat{y}_i - \bar{y})$ is called the **explained deviation**, since it shows by **how much the total deviation is reduced when the regression line is fitted to the points**.



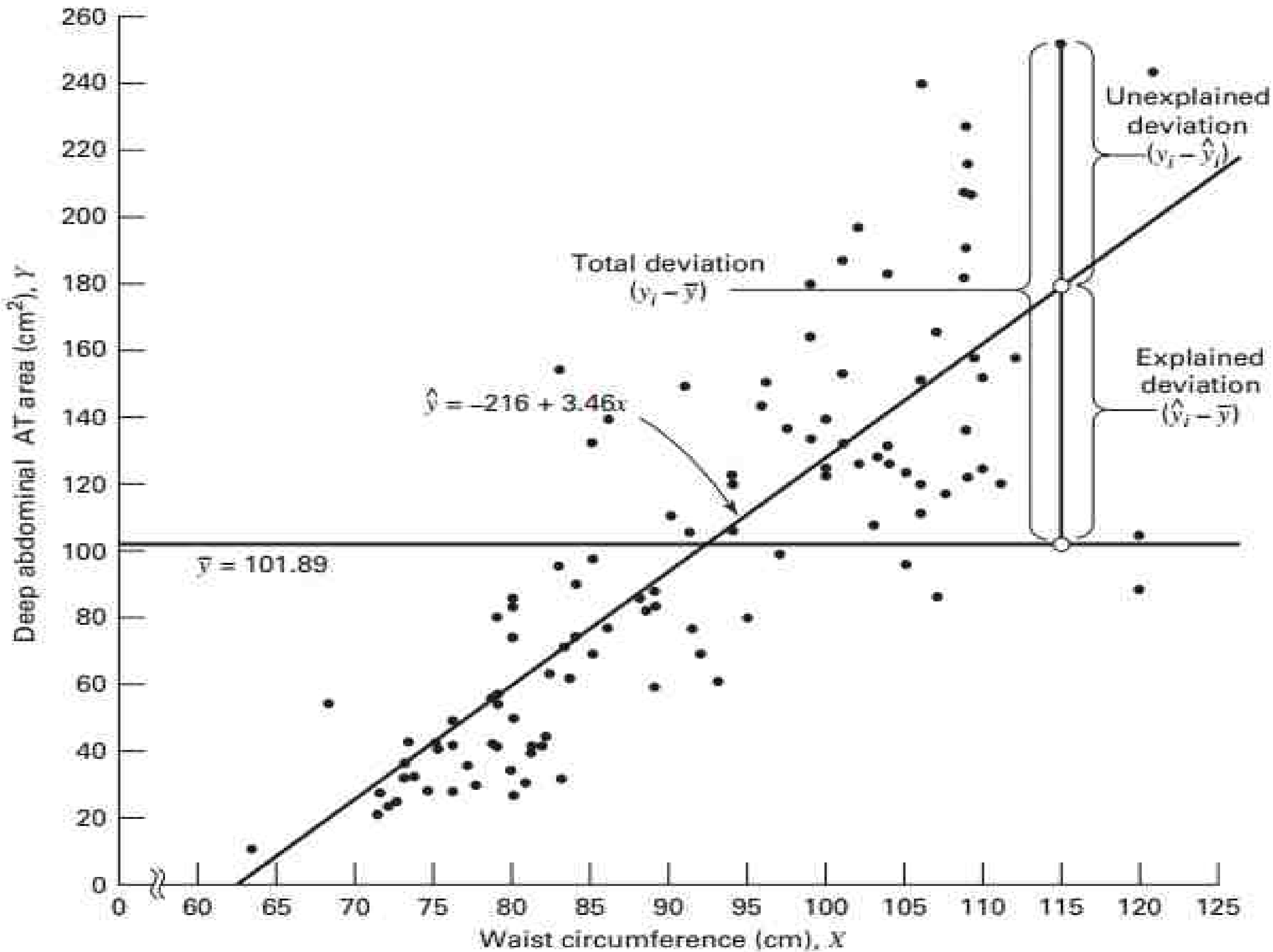
Unexplained deviation

- The measurement of the vertical distance of the observed point from the regression line, **$(y_i - \hat{y}_i)$,** is called the unexplained deviation, since it represents the portion of the total deviation not "explained" or accounted for by the introduction of the regression line.
- The difference between the observed value of Y and the predicted value of Y, $(y_i - \hat{y}_i)$, is also referred to as a **residual**.
- **The total deviation for a particular y_i is equal to the sum of the explained and unexplained deviation**

$$\underbrace{(y_i - \bar{y})}_{\text{total deviation}} = \underbrace{(\hat{y}_i - \bar{y})}_{\text{explained deviation}} + \underbrace{(y_i - \hat{y}_i)}_{\text{unexplained deviation}}$$

- If we measure these deviations for each value of y_i and $\hat{y}_i$, square each deviation, and add up the squared deviations, we have

$$\underbrace{\Sigma (y_i - \bar{y})^2}_{\text{total sum of squares}} = \underbrace{\Sigma (\hat{y}_i - \bar{y})^2}_{\text{explained sum of squares}} + \underbrace{\Sigma (y_i - \hat{y}_i)^2}_{\text{unexplained sum of squares}}$$



Unexplained deviation $(y_i - \hat{y}_i)$
Total deviation $(y_i - \bar{y})$
Explained deviation $(\hat{y}_i - \bar{y})$
$\hat{y} = -216 + 3.46x$
$\bar{y} = 101.89$
Deep abdominal AT area (cm²), Y
Waist circumference (cm), X

TOTAL SUM OF SQUARES

- The total sum of squares (SST), is **a measure of the dispersion of the observed values of Y about their mean $\bar{y}$** which is a measure of the total variation in the observed values of Y, **the numerator of the familiar formula for the sample variance**.

Explained Sum of Squares

- Measures the amount of the total variability in the observed values of Y that is **accounted for by the linear relationship between the observed values of X and Y**. This quantity is referred to also as **the sum of squares due to linear regression (SSR)**

Unexplained Sum of Squares

- Is a measure of **the dispersion of the observed Y values about the regression line** and is sometimes called the **error sum of squares**, or the residual sum of squares (SSE).

"It is this quantity that is minimized when the least-squares line is obtained"

- We may express the relationship among the three sums of squares values as $SST = SSR + SSE$

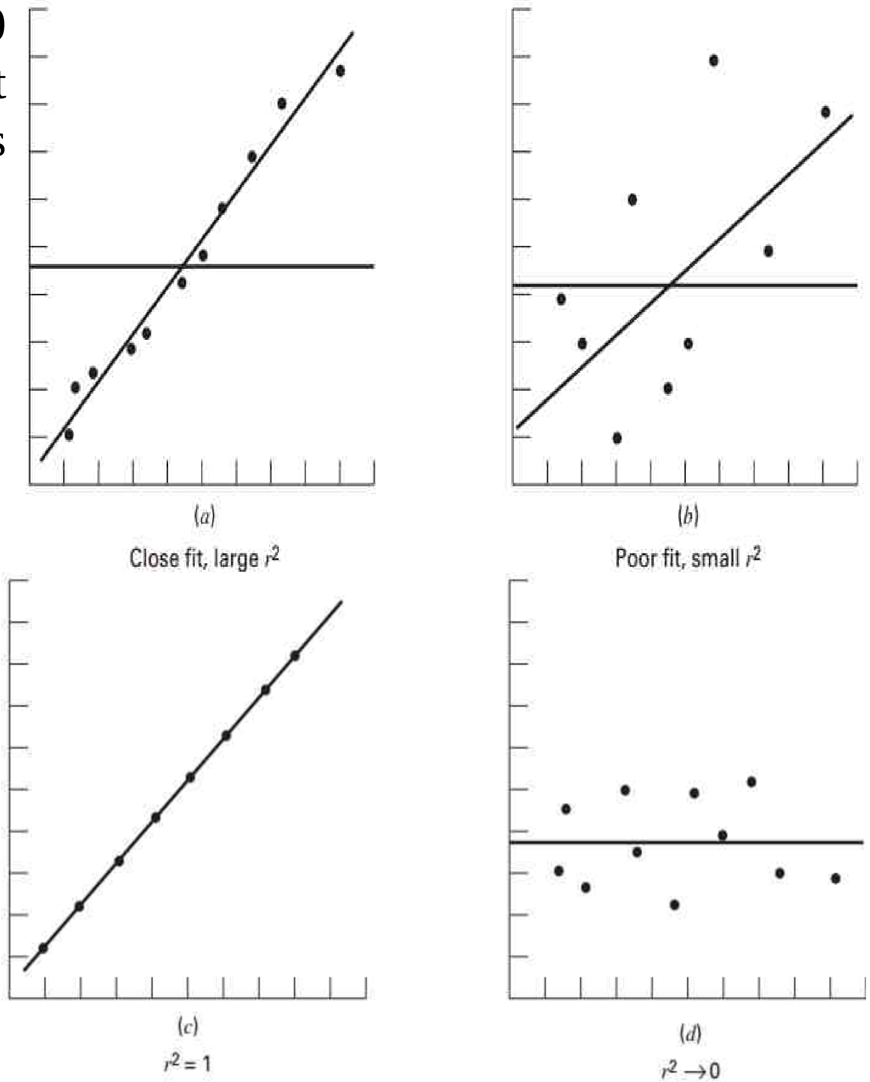
CALCULATING r^2

- It is intuitively appealing to speculate that if a regression equation does a good job of describing the relationship between two variables, **the explained or regression sum of squares should constitute a large proportion of the total sum of squares.**
- It would be of interest, then, to determine the magnitude of this proportion by **computing the ratio of the explained sum of squares to the total sum of squares**.
- This is *exactly what is done in evaluating a regression equation based on sample data*, and the result is called the **sample coefficient of determination,** r^2

$$r^2 = \frac{\Sigma(\hat{y}_i - \bar{y})^2}{\Sigma(y_i - \bar{y})^2} = \frac{SSR}{SST}$$

- When the quantities $(y_i - \hat{y}_i)$, **the vertical distances of the observed values of Y from the equations**, are small, the unexplained sum of squares is small.
- This leads to a large explained sum of squares that leads, in turn, to **a large value of r^2** (see figures on the following slide).
- **When r^2 is large**, then, the regression has accounted for a large proportion of the total variability in the observed values of Y, and ***we look with favor on the regression equation***.
- On the other hand, **a small r^2** which indicates a failure of the regression to account for a large proportion of the total variation in the observed values of Y, **tends to cast doubt on the usefulness of the regression equation for predicting and estimating purposes**.
- **We do not, however, pass final judgment on the equation until it has been subjected to an objective statistical test**.

e data is 0
99 percent
n the is
n.



a) The observations all lie close to the regression line, and we would **expect r^2 to be large**

b) A case in which the y_i are widely scattered about the regression line, and there we **suspect that r^2 is small**

c) When r^2 =1 all the observations **fall on the regression line**.

d)The result r^2 =0 is obtained when the regression line and the line drawn through $\bar{y}$ coincide. none of the variation in the y_i is explained by the regression. For the d case,r^2 **is close to zero**.

TESTING H_0: B_1=0 WITH THE F STATISTIC

- Referring to the table on slide 104 (9.3.1) We wish to know if we can conclude that, in the population from which our sample was drawn, X and Y are linearly related.

ANOVA Table for Simple Linear Regression

Source of Variation	SS	d.f.	MS	V.R.
Linear regression	SSR	1	MSR = SSR/1	MSR/MSE
Residual	SSE	$n-2$	MSE = SSE/($n-2$)	
Total	SST	$n-1$		

- The degrees of freedom associated with the sum of squares due to regression is equal to the number of constants in the regression equation minus 1
- In the simple linear case we have two estimates, β_0 and β_1 ; hence the degrees of freedom for regression are 2-1=1

Solution: The

1. D
9
2. A
it
3. H

4. **Test statistic**
that follows.
From th
of freedom th
In gene
squares due to
sion equation
β_0 and β_1; h
5. **Distribution**
of no linear r
tions underly
regression mo
with 1 and n
6. **Decision rul**
greater than t
7. **Calculation**
value of F is
8. **Statistical de**
of F (obtaine
null hypothes
9. **Conclusion.**
the data.
10. p **value.** For

28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

- The ratio obtained by dividing the **regression mean square** by the **residual mean square** is distributed as F with 1 and $n-2$ degrees of freedom.
- **Calculation of test statistic**. The computed value of F is 217.28.
- **Statistical decision**. Since 217.28 is greater than 3.94, the critical value of F (obtained by interpolation) for 1 and 107 degrees of freedom, **the null hypothesis is rejected**.
- **Conclusion**. We conclude that the linear model provides a good fit to the data.
- P value. For this test, since we have $217.28 > 8.25$ *we have* $p <$ 0.0

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	237549	237549	217.28	0.000
Error	107	116982	1093		
Total	108	354531			

Estimating the Population Coefficient of Determination

- The sample coefficient of determination provides a point estimate of ρ^2, the population coefficient of determination.
- The population coefficient of determination, ρ^2 has the same function relative to the population as r^2 has to the sample.
- It shows **what proportion of the total population** variation in Y is explained by the regression of Y on X.
- When the number of degrees of freedom is small, r^2 is positively biased.
- That is, r^2 tends to be large. An unbiased estimator of ρ^2 is provided by

$$\tilde{r}^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2/(n-2)}{\sum(y_i - \bar{y})^2/(n-1)}$$

$$-\frac{\sum(y_i - \hat{y}_i)^2}{-\bar{y})^2} = 1$$

❖ The numerator of the fraction is the unexplained mean square and the denominator is the total mean square.

Observe that the
and the denomi
variance table. I

This quantity is
that this value i

We see that the
large, this factor

THANKS

VITAL STATISTICS
INTRODUCTION

- The **physician** arrives at **a diagnosis and treatment plan** for an ***individual*** patient by means of a **case history**, a **physical examination**, and various **laboratory tests**.
- The ***community*** may be thought of as **a living complex organism** for which the **public health team** is the physician.
- To carry out this role satisfactorily the public health team must **also** make use of **appropriate tools and techniques** for **evaluating the health status of the community**.
- The idea of community-based, public health medicine is most often studied using the **concepts and tools of epidemiology**.
- Epidemiologists study **the mechanisms** ***by which diseases and other health-related conditions arise and how they are distributed among populations***.

- While **physicians** diagnose and treat individual patients who have a medical condition, **epidemiologists and public health professionals** are **additionally** interested in studying those members of a population who are well, and how those who have an illness differ from those who are free of the illness.
- To that end, the use of **vital statistics and epidemiological tools** are employed in determining the prevalence of a given condition **at a point in time** and incidence at which new conditions arise in the population.

SOME DEFINITIONS

1.Rate: Although there are some exceptions, the term rate usually is reserved to refer to those calculations that involve the **frequency of the occurrence of some event**. A rate is expressed in the form $\left(\frac{a}{a+b}\right)k$ where

- a = the frequency with which an event has occurred during some specified period of time
- $a + b$= the number of persons **exposed** to the risk of the event during the same period of time.

- k= some number such as 10, 100, 1000, 10,000, or 100,000.
- Notice that **the numerator of a rate is a component part of the denominator.**
- The purpose of the multiplier, k, called the base, is **to avoid results involving the *very small numbers* that may arise in the calculation of rates and to facilitate comprehension of the rate**.
- The value chosen for k will **depend on the magnitudes of the numerator and denominator**

2. Ratio: A ratio is a fraction of the form $(\frac{c}{d})\ k$

- Where **both c and d** refer to the **frequency of occurrence of some event or item.**
- In the case of a ratio, as opposed to a rate, **the numerator is not a component part of the denominator**.
- We can speak, for example, of **the person–doctor ratio** or the **person–hospital-bed** ratio of a certain geographic area.
- The values of k most frequently used in ratios are **1 and 100.**

- All of the girls in Ms. Smith’s class have either brown hair or blonde hair. There are 15 girls in the class and 5 of them have blonde hair. What is the ratio of **blonde-haired girls** to **brown-haired girls?**

DEATH RATES AND RATIOS

- Death rates express the **relative frequency of the occurrence of death within some specified interval of time in a specific population**. The denominator of a death rate is referred to as the **population at risk**. The numerator represents only those deaths that occurred in the population specified by the denominator.

1.Annual crude death rate: The annual crude death rate is defined as

$$\frac{\text{total number of deaths during year (January 1 to December 31)}}{\text{total population as of July 1}} \cdot k$$

The most widely used rate for measuring the overall health of a community.

- where the value of k is usually chosen as **1000**.

Variables that enter into the picture include **age**, **race**, **sex**, and **socioeconomic status**.

When two populations must be compared on the basis of death rates, **adjustments may be made to reconcile the population differences with respect to these variables**.

The same precautions should be exercised when comparing the annual death rates for the same community for 2 different years.

2. Annual specific death rates: It is usually more meaningful and enlightening to observe the death rates of small, well-defined subgroups of the total population.

Rates of this type are called specific death rates and are defined as

$$\frac{\text{total number of deaths in a specific subgroup during a year}}{\text{total population in the specific subgroup as of July 1}} \cdot k$$

Where k is usually equal to **1000**.

- Subgroups for which specific death rates may be computed include those groups that may be distinguished on the basis of **sex**, **race**, and **age**.
- Specific rates **may be computed for two or more characteristics simultaneously.**
- For example, we may compute the death rate for **white males**, thus obtaining **a RACE-SEX specific rate**.
- **Cause-specific death rates** may also be computed by including in the numerator only those deaths due to a particular cause of death, say, cancer, heart disease, or accidents.

- Because of the small fraction that results, the base, k, for a cause-specific rate is usually 100,000 or 1,000,000.

3.Adjusted or standardized death rates

- As already pointed out, the usefulness of the crude death rate is restricted by the fact that **it does not reflect the composition of the population with respect to certain characteristics by which it is influenced.**
- By means of specific death rates, **various segments of the population may be investigated individually**. If, however, we attempt to obtain **an overall impression of the health of a population** by looking at individual specific death rates, we are soon overwhelmed by their great number.
- For adjustment calculations, a population of 1,000,000, reflecting the composition of the standard population and called the **standard million, is usually used**.
- In the following example, **the direct method of adjustment to obtain an age-adjusted death rate is illustrated**.

Sources: [a] *Georgia Vital and Morbidity Statistics 2000,* Georgia Division of Public Health, Atlanta (A-1).

[b] *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).

[c] Total does not reflect actual sum because of rounding to the nearest person.

- **The 2000 crude death rate for Georgia was 7.8 deaths per 1000 population**.
- Let us obtain an age-adjusted death rate for Georgia by *using the **2000 United States census as the standard population***.
- In other words, we want a death rate that could have been expected in Georgia if the age composition of the Georgia population had been the same as that of the United States in 2000.
- Solution: The data necessary for the calculations are shown in Table on next slide.
- The procedure for calculating an **age-adjusted death rate** by the direct method consists of the **following steps (after next slide)**:

Calculations of Age-Adjusted Death Rate for Georgia, 2000, by Direct Method

1 Age (Years)	2 Population[a]	3 Deaths[a]	4 Age-Specific Death Rates (per 100,000)	5 Standard Population Based on U.S. Population 2000[b]	6 Number of Expected Deaths in Standard Population
0–4	595,150	1,299	218.3	68,139	149
5–9	615,584	101	16.4	73,020	12
10–14	607,759	136	22.4	72,944	16
15–19	596,277	447	75.0	71,849	54
20–44	3,244,960	5,185	159.8	369,567	591
45–64	1,741,448	13,092	751.8	220,141	1,655
65 and over	785,275	43,397	5526.3	124,339	6,871
Total	8,186,453	63,657		1,000,000[c]	9,348

Sources: [a] *Georgia Vital and Morbidity Statistics 2000,* Georgia Division of Public Health, Atlanta (A-1).

[b] *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).

[c] Total does not reflect actual sum because of rounding to the nearest person.

Sources: [a] *Georgia Vital and Morbidity Statistics 2000,* Georgia Division of Public Health, Atlanta (A-1).
[b] *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).
[c] Total does not reflect actual sum because of rounding to the nearest person.

1. The population of interest is listed (Column 2) according to age group (Column 1).
2. The deaths in the population of interest are listed (Column 3) by age group.
3. The **age-specific death rates** (Column 4) for each age group are calculated by dividing Column 3 by Column 2 and multiplying by 100,000 ”k”.
4. The standard population (Column 5) is listed by age group

- ➢ The standard population is obtained as follows:
- ✓ The 2000 U.S. population by age group is shown in Table on next slide.
- ✓ The total for each age group is **divided by the grand total and multiply by 1,000,000**.
- ✓ For example, to obtain the standard population count for the 0– 4 age group, we divide 19,175,798 by 281,421,906 and multiply the result by 1,000,000. That is, 1,000,000(19175798/281421906)=68,139 Similar calculations yield he standard population counts for the other age groups as shown in table on slide (134)

Population of the United States, 2000

Age (Years)	Population
0–4	19,175,798
5–9	20,549,505
10–14	20,528,072
15–19	20,219,890
20–44	104,004,252
45–64	61,952,636
65 and over	34,991,753
Total	281,421,906

Source: *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).

Sources: [a] *Georgia Vital and Morbidity Statistics 2000*, Georgia Division of Public Health, Atlanta (A-1).
[b] *Profile of General Demographic Characteristics: 2000*, U.S. Census Bureau DP-1 (A-2).
[c] Total does not reflect actual sum because of rounding to the nearest person.

5. **The expected number of deaths in the standard population for each group** (Column 6) is computed by **multiplying Column 4 by Column 5 and dividing by 100,000.**

"The entries in column 6 are the ***deaths that would be expected in the standard population if the persons in this population had been exposed to the same risk of death experienced by the population being adjusted".***

6. The entries in Column 6 are summed to obtain the total number of expected deaths in the standard population.

- The age-adjusted death rate is computed in the same manner as a crude death rate. That is, the age-adjusted death rate is equal to

$$\frac{\text{total number of expected deaths}}{\text{total standard population}} \cdot 1000$$

Sources: [a] *Georgia Vital and Morbidity Statistics 2000,* Georgia Division of Public Health, Atlanta (A-1).
[b] *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).
[c] Total does not reflect actual sum because of rounding to the nearest person.

- In the present example we have an age-adjusted death rate of $\frac{9348}{1000000}.1000 = 9.3$
- We see, then, that by adjusting the 2000 population of Georgia to the age distribution of the standard population, we obtain an adjusted death rate that is **1.5 per 1000 greater than the crude death rate (9.3 − 7.8)**
- This increase in the death rate following adjustment reflects the fact that **in 2000 the population of Georgia was slightly younger than the population of the United States as a whole**.
- For example, only 9.6 percent of the Georgia population was 65 years of age or older, whereas 12.4 percent of the U.S. population was in that age group.

4. Maternal mortality rate: This rate is defined as

$$\frac{\text{deaths from all puerperal causes during a year}}{\text{total live births during the year}} \cdot k$$

Where k is taken as **1000 or 100,000**. ***The preferred denominator for this rate is the number of women who were pregnant during the year***. But it is impossible to determine.

Some limitations of the maternal mortality rate include the following:

a. **Fetal deaths** are not included in the denominator. This results in an *inflated rate*, since **a mother can *die from a puerperal cause without producing a live birth***.

b. A maternal death can be counted **only once**, **although twins or larger multiple births may have occurred**. Such cases cause the denominator to be too large and, hence, *there is a too small rate*.

c. **Under-registration of live births**, which result in a too small denominator, *causes the rate to be too large*.

d. A maternal death may occur in a year later than the year in which the birth occurred. Although there are exceptions, in most cases **the transfer of maternal deaths will balance out in a given year**.

5. **Infant mortality rate**: This rate is defined as

$$\frac{\text{number of deaths under 1 year of age during a year}}{\text{total number of live births during the year}} \cdot k$$

- Where k is generally taken as **1000**. **Use and interpretation of this rate must be made in light of its limitations**, which are **similar to those that characterize the maternal mortality rate**.
- Many of the infants who die in a given calendar year **were born during the previous year**; and, similarly, many children born in a given calendar year **will die during the following year**. **In populations with a stable birth rate this does not pose a serious problem**.

In periods of rapid change, however, some adjustment should be made. One way to make an adjustment is to allocate the infant deaths to the calendar year in which the infants were born before computing the rate.

6. **Neonatal mortality rate**:

In **an effort to better understand the nature of infant deaths**, rates for ages **less than a year** are frequently computed. Of these, the one most frequently computed is the neonatal mortality rate, which is defined as

$$\frac{\text{number of deaths under 28 days of age during a year}}{\text{total number of live births during the year}} \cdot k$$

where k=1000

7. Fetal death rate: This rate is defined as

$$\frac{\text{total number of fetal deaths during a year}}{\text{total deliveries during the year}} \cdot k$$

Where k is usually taken to be **1000**. A fetal death is defined as a **product of conception that shows no sign of life after complete birth**. There are several problems associated with the use and interpretation of this rate. There is variation among reporting areas with respect to the duration of gestation.

- Some areas report **all fetal deaths** regardless of length of gestation, while others have **a minimum gestation** period that must be reached before reporting is required.
- Another objection to the fetal death rate is that **it does not take into account the extent to which a community is trying to reproduce**. The ratio to be considered next has been proposed to overcome this objection.

8. Fetal death ratio: This ratio is defined as

$$\frac{\text{total number of fetal deaths during a year}}{\text{total number of live births during the year}} \cdot k$$

Where k is taken as 100 or 1000.

- Some authorities have suggested that the number of fetal deaths as well as live births be included in the denominator in an attempt to include all pregnancies in the computation of the ratio. The objection to this suggestion rests on the incompleteness of fetal death reporting.

9. Perinatal mortality rate

- Since **fetal deaths occurring late in pregnancy and neonatal deaths** frequently have the same underlying causes, it has been suggested that **the two be combined to obtain what is known as the perinatal mortality rate**. This rate is computed as

Where k=**1000** $$\frac{(\text{number of fetal deaths of 28 weeks or more}) + (\text{infant deaths under 7 days})}{(\text{number of fetal deaths of 28 weeks or more}) + (\text{number of live births})} \cdot k$$

10. Cause-of-death ratio. This ratio is defined as

$$\frac{\text{number of deaths due to a specific disease during a year}}{\text{total number of deaths due to all causes during the year}} \cdot k$$

where this index is used **to measure the relative importance of a given cause of death**. It should be used with caution in comparing one community with another.

- A higher cause-of-death ratio in one community than that in another may be because the first community has a **low mortality from other causes**.

11. Proportional mortality ratio

- This index has been suggested as a single measure for **comparing the overall health conditions of different communities**. It is defined as.

$$\frac{\text{number of deaths in a particular subgroup}}{\text{total number of deaths}} \cdot k$$

Where k=**100**. The specified class is usually an age group such as 50 years and over, or a cause of death category, such as accidents.

MEASURES OF FERTILITY

- The **term fertility** as used by American demographers **refers to the actual bearing of children** as opposed to the capacity to bear children, for which phenomenon the term fecundity is used.
- Knowledge of the "rate" of childbearing in a community is important to the health worker in **planning services and facilities for mothers, infants, and children**.
- The following are the six basic measures of fertility.

1. **Crude birth rate**. This rate is the most widely used of the fertility measures. It is obtained from

$$\frac{\text{total number of live births during a year}}{\text{total population as of July 1}} \cdot k$$

where k=1000. For an illustration of the computation of this and the other five rates, see table on next slide.

Sources: [a] *Georgia Vital and Morbidity Statistics 2000,* Georgia Division of Public Health, Atlanta (A-1).
[b] *Profile of General Demographic Characteristics: 2000,* U.S. Census Bureau DP-1 (A-2).
[c] Total does not reflect actual sum because of rounding to the nearest person.

Illustration of Procedures for Computing Six Basic Measures of Fertility, for Georgia, 2000

1 Age of Woman (Years)	2 Number of Women in Population[a]	3 Number of Births to Women of Specified Age[b]	4 Age-Specific Birth Rate per 1000 Women	5 U.S. Population for Year 2000[c]	6 Standard Population Based on U.S. Population 2000	7 Expected Births	8 Cumulative Fertility Rate
10–14	296,114	396	1.3	20,528,072	112,524	146	6.7
15–19	286,463	17,915	62.5	20,219,890	110,835	6,927	319.4
20–24	285,733	36,512	127.8	18,964,001	103,951	13,285	958.3
25–29	316,000	35,206	111.4	19,381,336	106,239	11,835	1,515.4
30–34	326,709	27,168	83.2	20,512,388	112,438	9,355	1,931.1
35–39	350,943	12,685	36.1	22,706,664	124,466	4,493	2,111.9
40–54	887,104	2,404[d]	2.7	60,119,815	329,546	890	2,152.5
Total	2,749,066	132,286		182,432,166	1,000,000	46,931	

Computation of six basic rates:

(1) Crude birth rate = total births divided by total population
= (132,286/8,186,453)(1000) = 16.2.

(2) General fertility rates = (132,286/2,749,066)(1000) = 48.1.

(3) Age-specific fertility rates = entries in Column 3 divided by entries in Column 2 multiplied by 1000 for each group. Results appear in Column 4.

(4) Expected births = entries in Column 4 multiplied by entries in Column 6 divided by 1000 for each group. Results appear in Column 7.

(5) Total fertility rate = the sum of each age-specific rate multiplied by the age interval width = 1.3(5) + 62.5(5) + (127.8)(5) + 111.4(5) + 83.2(5) + 36.1(5) + 2.7(15) = 2,152.5.

(6) Cumulative fertility rate = age-specific birth rate multiplied by age interval width cumulated by age. See Column 8.

(7) Standardized general fertility rate = (46,943)/(1,000,000)(1000) = 46.9.

(6) Cumulative fertility rate = age-specific birth rate multiplied by age interval width cumulated by age. See Column 8.

(7) Standardized general fertility rate = (46,943)/(1,000,000)(1000) = 46.9.

2. General fertility rate. This rate is defined as $\frac{\text{number of live births during a year}}{\text{total number of women of childbearing age}} \cdot k$

Where k=1000 and the childbearing age is usually defined as ages **15 through 44** or ages 15 through 49.

- The attractive feature of this rate, when compared to the crude birth rate, is the fact that the **denominator approximates the number of persons actually exposed to the risk of bearing a child**.

3. Age-specific fertility rate. Since the rate of childbearing is not uniform throughout the childbearing ages, a rate that permits the analysis of **fertility rates for shorter maternal age** intervals is desirable. T[illegible]ific fertility rate, which is defined as $\frac{\text{number of births to women of a certain age in a year}}{\text{total number of women of the specified age}} \cdot k$

(6) Cumulative fertility rate = age-specific birth rate multiplied by age interval width cumulated by age. See Column 8.
(7) Standardized general fertility rate = (46,943)/(1,000,000)(1000) = 46.9.

- ✓ Where Age-specific rates may be computed **for single years of age or any age interval**.
- ✓ Rates for **5-year age groups** are the ones most frequently computed.
- ✓ Specific fertility rates may be computed also for other population subgroups such as those defined by **race**, **socioeconomic status**, and **various demographic characteristics**.

4. **Total fertility rate**. If the age-specific fertility rates for all ages are added and multiplied by the interval into which the ages were grouped, the result is called the total fertility rate.
   - ➢ The resulting figure is **an estimate of the number of children** a cohort of 1000 women would have if, during their reproductive years, **they reproduced at the rates represented by the age-specific fertility rates from which the total fertility rate is computed**.
5. Cumulative fertility rate. The cumulative fertility rate is computed in **the same manner as the total fertility rate except that the adding process can terminate at the end of any desired age group**.

(6) Cumulative fertility rate = age-specific birth rate multiplied by age interval width cumulated by age. See Column 8.

(7) Standardized general fertility rate = (46,943)/(1,000,000)(1000) = 46.9.

- The numbers in Column 8 of table on slide 146 are the cumulative fertility rates through the ages indicated in Column 1.
- **The final entry** in the cumulative fertility rate column **is the total fertility rate**.

6. Standardized fertility rate.

Just as the crude death rate may be standardized or adjusted, so may we standardize the general fertility rate. The procedure is identical to that already discussed for adjusting the crude death rate. The necessary computations for computing the age-standardized fertility rate are shown in table on slide 146.

MEASURES OF MORBIDITY

- Another area that concerns the **health worker who is analyzing the health of a community is morbidity**.
- The word "morbidity" refers to **the community's status with respect to disease**.
- Data for the study of the morbidity of a community are **not, as a rule, as readily available and complete as are the data on births and deaths** because of incompleteness of reporting and differences among states with regard to laws requiring the reporting of diseases.
- The two rates most frequently used in the study of diseases in a community are **the incidence rate** and the **prevalence rate**.

1.Incidence rate. This rate is defined as

$$\frac{\text{total number of new cases of a specific disease during a year}}{\text{total population as of July 1}} \cdot k$$

Where the value of k depends on the magnitude of the numerator.

- ✓ A base of 1000 is used when convenient, but 100 can be used for the more common diseases, and 10,000 or 100,000 can be used for those less common or rare.
- ✓ This rate, which **measures the degree to which new cases are occurring in the community**, is useful in helping determine the need for initiation of preventive measures.
- ✓ It is a meaningful measure for both chronic and acute diseases.

2. Prevalence rate. Although prevalence rate is really a ratio,

$$\frac{\text{total number of cases, new or old, existing at a point in time}}{\text{total population at that point in time}} \cdot k$$

where the value of k is selected by the same criteria as for the incidence rate. This rate is especially useful in the study of chronic diseases, but it may also be computed for acute diseases.

3. Case-fatality ratio. This ratio is useful in determining how well the treatment program for a certain disease is succeeding. It is defined as $\frac{\text{total number of deaths due to a disease}}{\text{total number of cases due to the disease}} \cdot k$

where k=**100**. The period of time covered is arbitrary, depending on the nature of the disease, and it may cover several years for an endemic disease.

Note that this ratio can be interpreted as **the probability of dying following contraction of the disease in question and, as such, reveals the seriousness of the disease**.

4. Immaturity ratio. This ratio is defined as

Where k=100 $\frac{\text{number of live births under 2500 grams during a year}}{\text{total number of live births during the year}} \cdot k$

5. Secondary attack rate. This rate measures **the occurrence of a contagious disease** among susceptible persons who have been exposed to a primary case and is defined as

$$\frac{\text{number of additional cases among contacts of a primary case within the maximum incubation period}}{\text{total number of susceptible contacts}} \cdot k$$

Where k=100. This rate is used to measure the spread of infection and is usually applied to closed groups such as a household or classroom, where it can reasonably be assumed that all members were, indeed, contacts.

THANKS