- **First Preliminary Topics**

CHAPTER 1: Overview

A. Definition of Statistics

Statistics – is the study of how to collect, organize, analyze, and interpret numerical information from data.

It is the branch of mathematics used to summarize quantities of data and help investigators draw sound conclusions. People are exposed to statistics every day from weather predictions, newspaper ads, election results, surveys, to report cards, to name a few.

Statistics, thus attempts to infer the properties of a large collection of data from inspection of a sample of the collection thereby allowing educated guesses to be made with a minimum of expense.

B. Applications of Statistics

1. Business Statistics
2. Educational Statistics
3. Psychological Statistics
4. Medical Statistics
5. Statistics for Historians

C. Methods of Statistics

- **Statistical methods**

  Methods of collecting, summarizing, analyzing, and interpreting variable numerical data. statistical methods are widely used in the life sciences, in economics, and in agricultural science, they also have an important role in the

physical sciences in the study of measurement errors, of random phenosuch as radioactivity or meteorological events, and in obtaining approximate results where deterministic solutions are hard to apply.

- **Data collection** involves deciding what to observe in order to obtain information relevant to the questions whose answers are required, and then making the observations.

- **Data summarization** is the calculation of appropriate statistics and the display of such information in the form of tables, graphs, or charts.

- **Statistical analysis** relates observed statistical data to theoretical models, such as probability distributions or models used in regression analysis.

D. Methods of Collecting Data

The choice of method is influenced by the data collection strategy, the type of variable, the accuracy required, the collection point and the skill of the enumerator.

The main data collection methods are:

Registration: registers and licenses are particularly valuable for complete enumeration, but are limited to variables that change slowly, such as numbers of fishing vessels and their characteristics.

Questionnaires: forms which are completed and returned by respondents. An inexpensive method that is useful where literacy rates are high and respondents are co-operative.

Interviews: forms which are completed through an interview with the respondent. More expensive than questionnaires, but they are better for more complex questions, low literacy or less co-operation.

Direct observations: making direct measurements is the most accurate method for many variables, such as catch, but is often expensive.

Reporting: the main alternative to making direct measurements is to require fishers and others to report their activities. Reporting requires literacy and co-operation, but can be backed up by a legal requirement and direct measurements.

E. Divisions of Statistics

1. **Descriptive Statistics** – Involves methods organizing, picturing, and summarizing information from samples or populations.

It is the branch of statistics that presents techniques for summarizing and describing sets of measurements. The following are samples of descriptive statistics: pie charts, line charts, bar charts or numerical tables.

2. **Inferential Statistics** – Involves methods of using information from a sample to draw conclusions regarding the population.

It is the branch of statistics that presents techniques for making inferences about the characteristics of the population, based on the information contained in a sample drawn from the population.

F. Populations and Samples

- **Data Collections (Population)**

A population is an entire set of individuals or objects, which may be finite or infinite.

1. Personal Interview Surveys
2. Telephone Surveys
3. Mailed Questionnaire Surveys
4. Other methods include surveying records and direct observation of situations.

- **The Sampling Methods**

A sample must also be large enough in order for its data to reflect the population. A sample that is too small may bias population estimates. When larger samples are used, data collected from idiosyncratic individuals have less influence than when smaller samples are used.

Sampling Techniques	
Random Sampling	Subjects are selected by random numbers.
Systematic Sampling	Subjects are selected by using every kth number after the first subject is randomly selected from 1 to k.

Stratified Sampling	Subjects are selected by dividing up the population into groups (strata), and subject are randomly selected within groups.
Cluster Sampling	Subjects ae selected by using an intact group that is representative of the population.

CHAPTER 2: Sample Size and Summations Notations

A. The Sloven's Formula

You can use Sloven's formula to figure out what sample size you need to take, which is written as n = N / (1 + Ne2) where n = Number of samples, N = Total population and e = Error tolerance
Sample question: Use **Sloven's formula** to find out what sample of a population of 1,000 people you need to take for a survey on their soda preferences.

Step 1: Figure out what you want your **confidence level** to be. For example, you might want a confidence level of 95 percent (which will give you a margin error of 0.05), or you might need better accuracy at the 98 percent confidence level (which produces a margin of error of 0.02).

Step 2. **Plug your data into the formula.** In this example, we'll use a 95 percent confidence level with a population size of 1,000.
n = N / (1 + N e2) =1,000 / (1 + 1000 * 0.05^2) = 285.714286

Step 3: ***Round your answer to a whole number*** *(because you can't sample a fraction of a person or thing) 285.714286 = 286*

1.7.1 Sloven's Formula in Determining the Sample Size

Let N be the population size and the margin of error e denotes the allowed probability of committing an error in selecting a small representative of the population.

The sample size n can be obtained by the formula

$$n = \frac{N}{1 + Ne^2}$$

B. The Lynch Formula

Lynch et al 1972, and cited by Ardoles, 1992, suggested the formula below to determine the sample size:

$$n = \frac{NZ^2 \times p(1-p)}{Nd^2 + Z^2 p(1-p)}$$

Where: n= Sample Size
N= Population

Z= the value of the normal variables (1.96) for a reliability level of 0.95

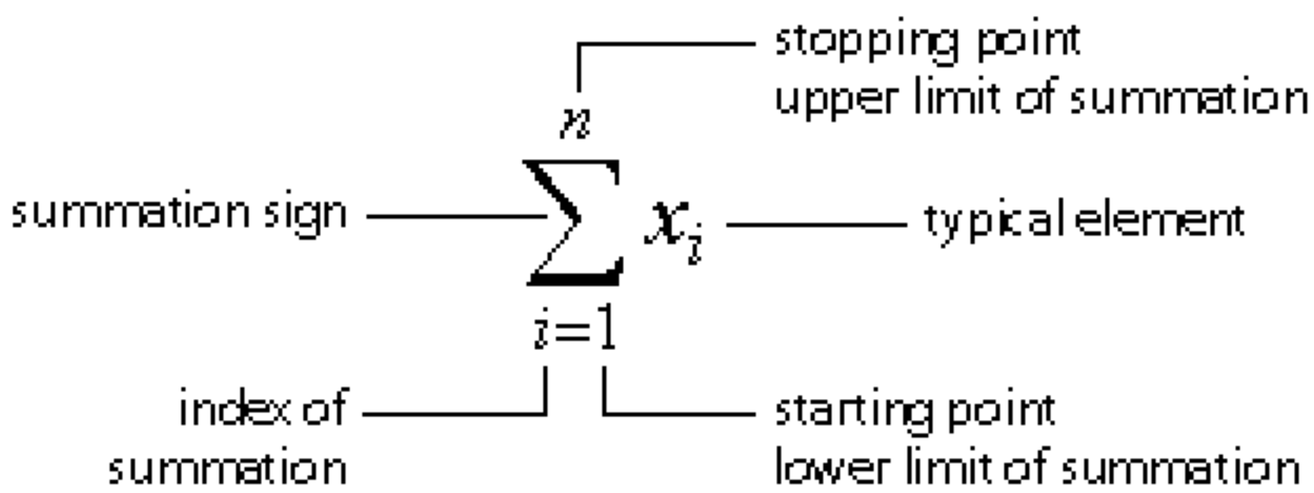
p= the largest possible proportion (0.50)

C. Summation Notations

- **Summation notation**

The summation sign. This appears as the symbol, S, which is the Greek upper case letter, S. The summation sign, S, instructs us to sum the elements of a sequence. A typical element of the sequence which is being summed appears to the right of the summation sign.

The variable of summation, i.e. the variable which is being summed. The variable of summation is represented by an index which is placed beneath the summation sign. The index is often represented by i. (Other common possibilities for representation of the index are j and t.) The index appears as the expression i = 1. The index assumes values starting with the value on the right hand side of the equation and ending with the value above the summation sign. The starting point for the summation or the lower limit of the summation. The stopping point for the summation or the upper limit of summation



Some typical examples of summation

$\sum_{i=1}^{n} x_i$ This expression means sum the values of x, starting at x_1 and ending with x_n.

- $\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \cdots + x_n$

$\sum_{i=1}^{10} x_i$ This expression means sum the values of x, starting at x_1 and ending with x_{10}.

- $\sum_{i=1}^{10} x_i = x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$

$\sum_{i=3}^{10} x_i$ This expression means sum the values of x, starting at x_3 and ending with x_{10}.

- $\sum_{i=3}^{10} x_i = x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10}$

CHAPTER 3: Data collection in the Health Care Profession

A. Patient's Data Collection

Data collection is defined as the ongoing systematic collection, analysis, and interpretation of health data necessary for designing, implementing, and evaluating public health prevention programs. To develop effective prevention strategies, countries need to improve their information. In particular, countries need to know about the numbers and types of injuries that occur and about the circumstances in which those injuries occur. Such information will indicate how serious the injury problem is, and where prevention measures are most urgently needed.

B. Uses of Data

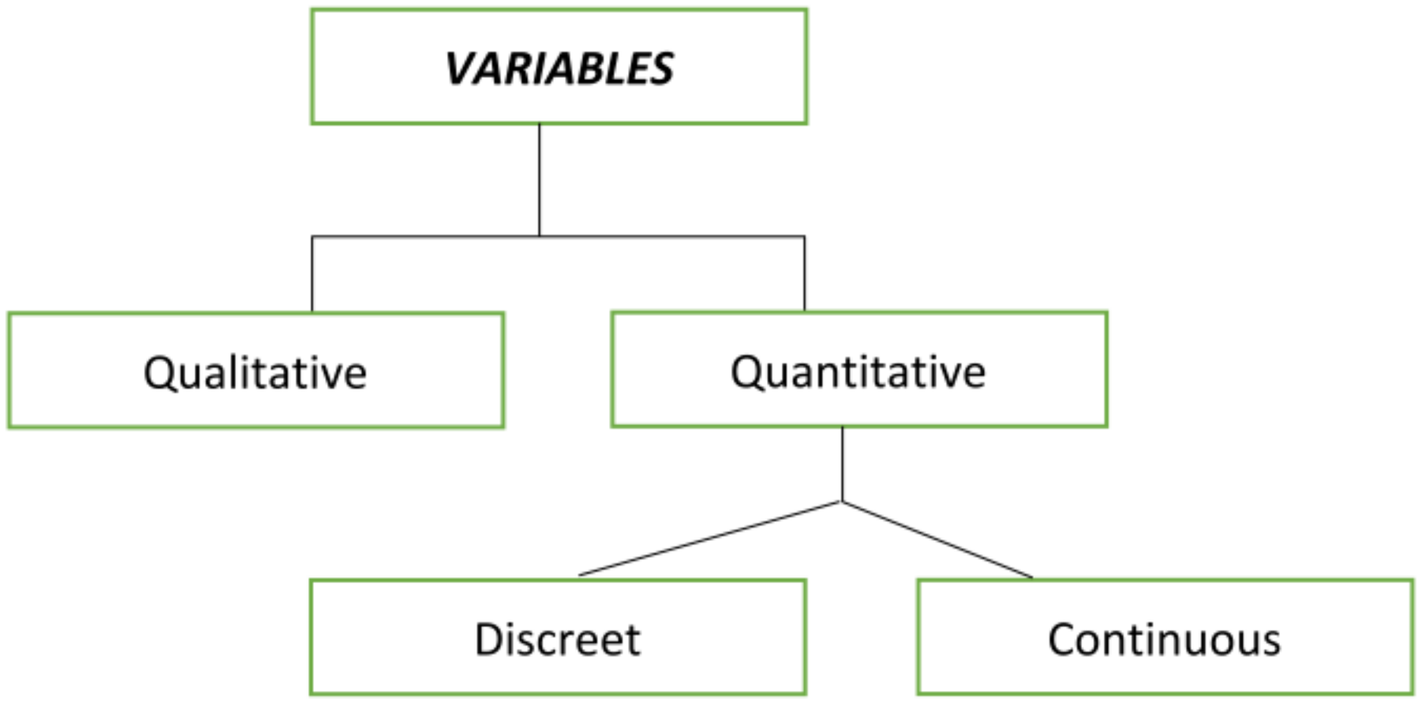
Quantitative research guides health care decision makers with statistics--numerical data collected from measurements or observation that describe the characteristics of specific population samples. Descriptive statistics summarize the utility, efficacy and costs of medical goods and services. Increasingly, health care organizations employ statistical analysis to measure their performance outcomes. Hospitals and other large provider service organizations implement data-driven, continuous quality improvement programs to maximize efficiency. Government health and human service agencies gauge the overall health and well-being of populations with statistical information.

C. Total Quality Management (TQM)

Total Quality Management (TQM) describes a management approach to long–term success through customer satisfaction. In a TQM effort, all members of an organization participate in improving processes, products, services, and the culture in which they work.

CHAPTER 4: Presentation and Organization of Data

A. Variables



Variables - The word variable is often used in the study of statistics, so it is important to understand its meaning. A variable is a characteristic that may assume more than one set of values to which a numerical measure can be assigned.

Height, age, amount of income, province or country of birth, grades obtained at school and type of housing are all examples of variables. Variables may be classified into various categories, some of which are outlined in this section.

Qualitative Variable vs. Quantitative Variable

Qualitative Variable – describes and individual by placing the individual into a category or group.

- Variables that can be placed into distinct categories, according to some characteristics or attribute.

Examples: gender, nationality

Quantitative Variable – has a value or numerical measurement for which operations such as addition or averaging make sense.

- Variables are numerical and can be ordered or ranked.

Examples: height, weight, age, and income

Types of Quantitative Variables

- **Discrete Variable -** assume values that cab be counted.

Example: number of siblings in the family, number of students in the class, number of calls received by a call center agent.

- **Continuous Variable –** can assume an infinite number of values between any two specific values. They are both obtained by measuring.

Example: temperature, weights, heights

B. Levels of Measurement

Levels of Measurement helps you decide how to interpret the data from that variable. When you know that a measure is nominal (like the one just described), then you know that the numerical values are just short codes for the longer names. Second, knowing the level of measurement helps you decide what statistical analysis is appropriate on the values that were assigned.

There are typically four levels of measurement that are defined:

- **Nominal measurement** the numerical values just "name" the attribute uniquely. No ordering of the cases is implied. Can be used as tags or labels, where the size of the number is arbitrary.
- **Ordinal measurement** the attributes can be rank-ordered. Here, distances between attributes do not have any meaning.
- **Interval measurement** the distance between attributes *does* have meaning.
- **Ratio measurement** there is always an absolute zero that is meaningful. This means that you can construct a meaningful fraction with a ratio variable.

Examples of Measurement Scales

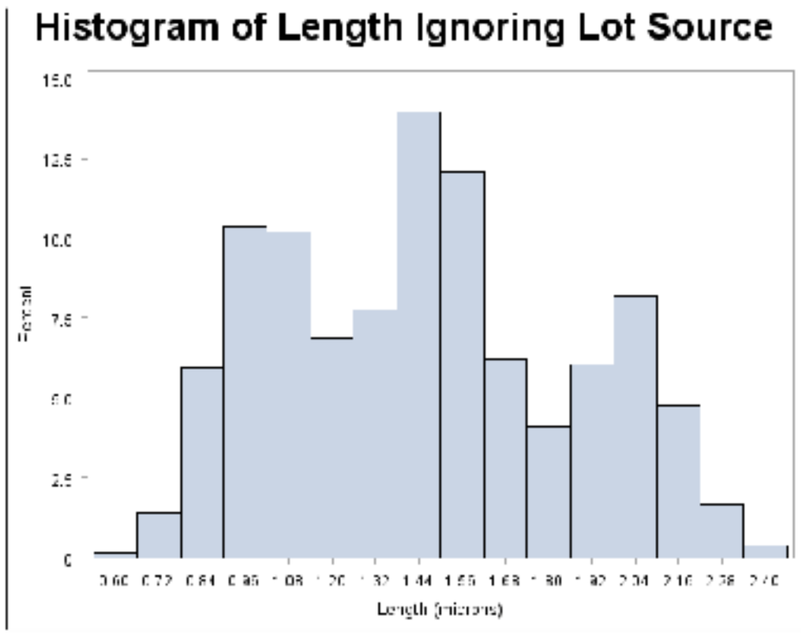
Nominal	Ordinal	Interval	Ratio
• Zip code • Gender • Eye color • Political affiliation • Religious affiliation • College course • Nationality	• Grade (A, B, C, D) • Judging (1st, 2nd, 3rd, etc) • Rating scale (poor, good, excellent) • Class rankings • Military ranks	• IQ • Temperature • SAT score	• Height • Weight • Time • Salary • Age

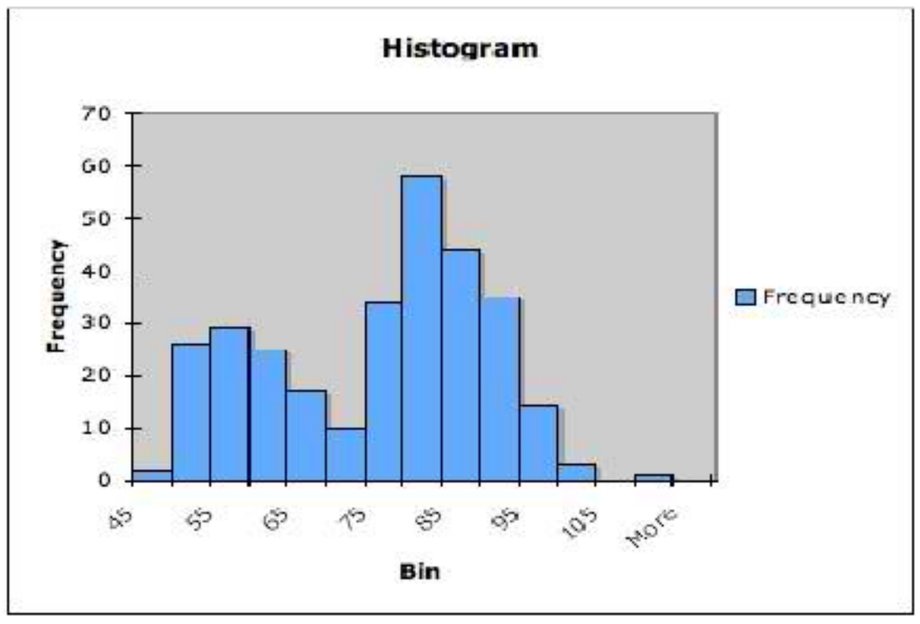
C. Graphical and Tubular Data Presentations

A. Graph of a Categorical Data

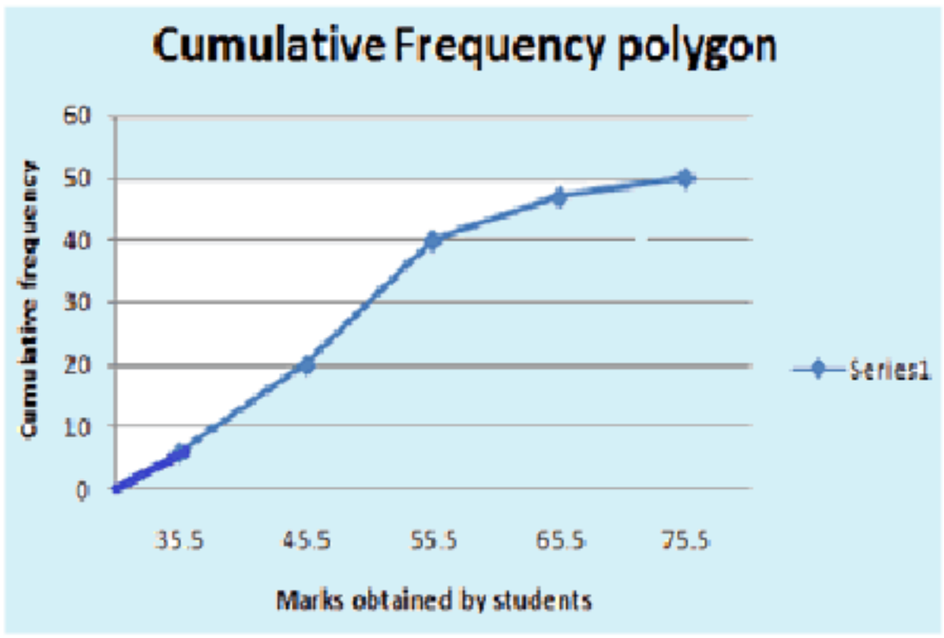
Categorical Data – a type of data that is classified according to criterion and can be a mixture of numerical or non-numerical observations.

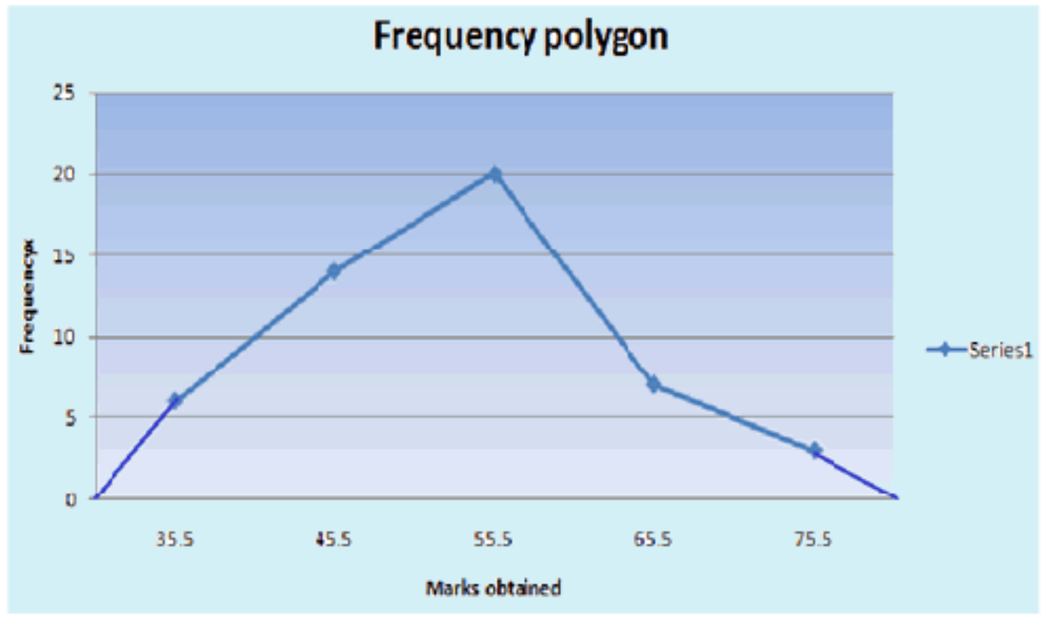
1. **Histogram –** is used to represent measurements of observations that are grouped.





2. **Frequency Polygon –** is used to present class frequencies plotted at the class mark and successive points are connected by means of straight lines.

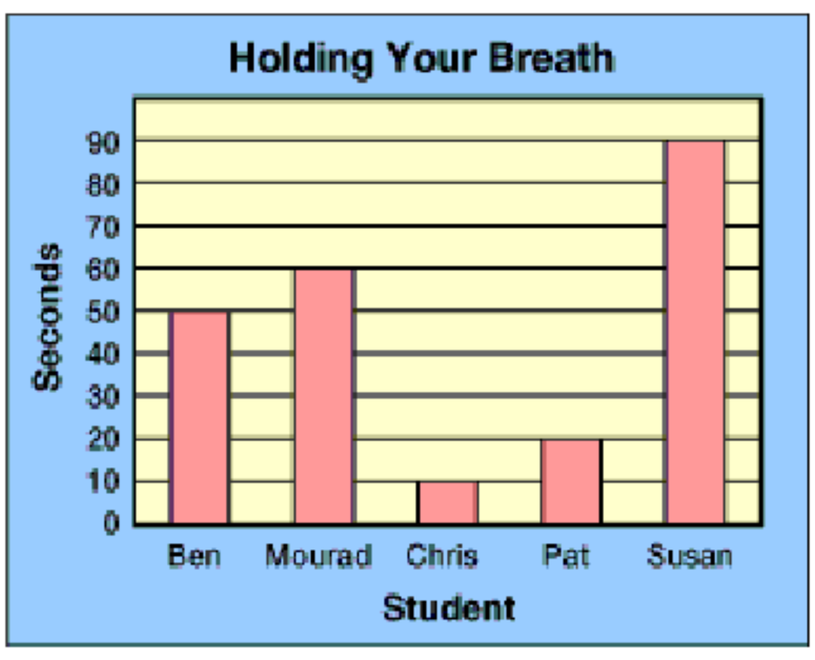


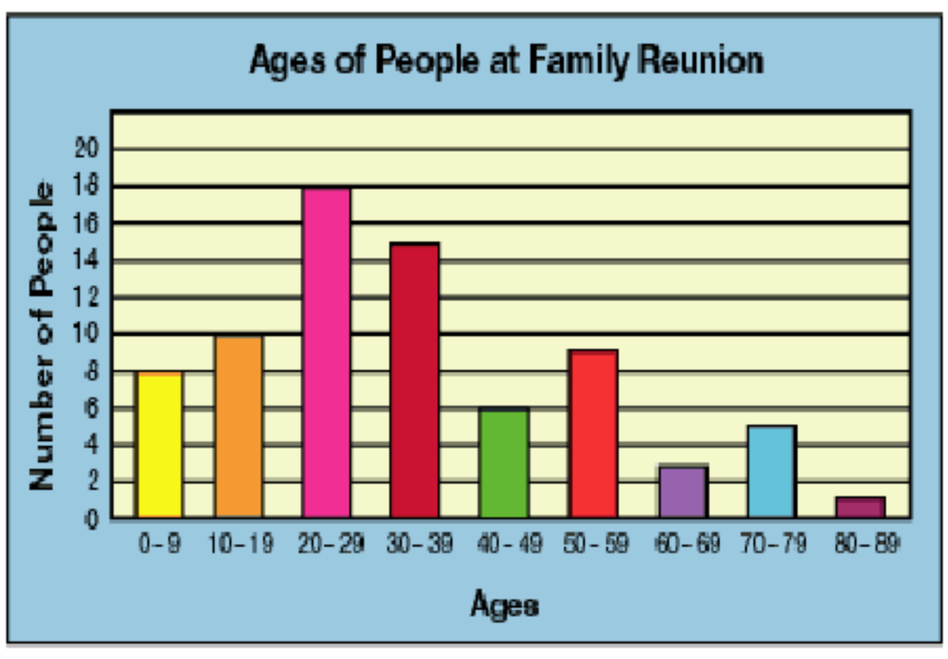


B. Graph of a Quantitative Data

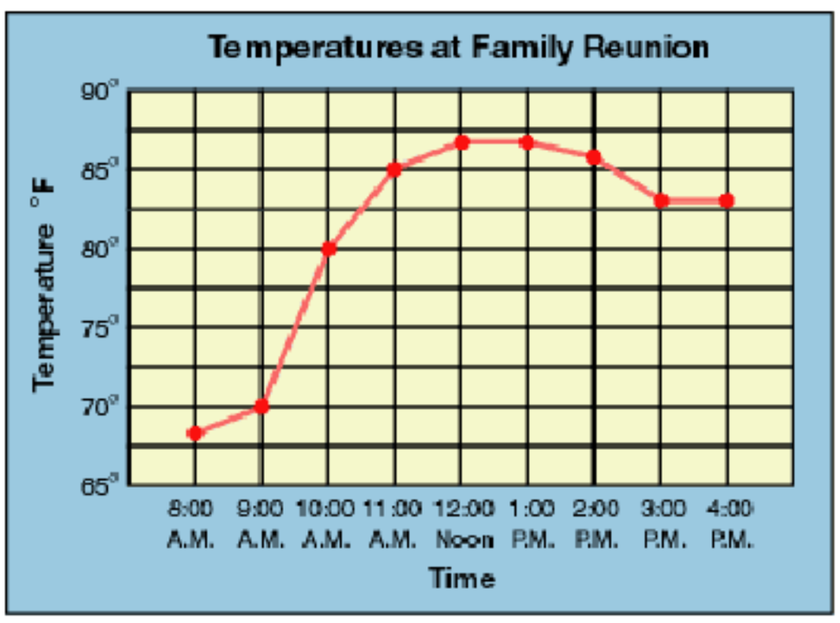
Quantitative Data – is a type of data where values of x and y are both numerical.

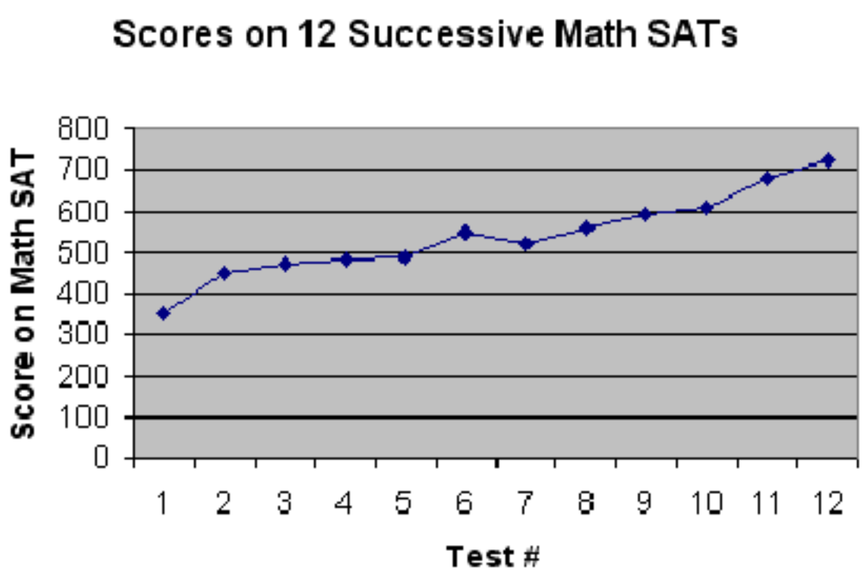
1. **Bar Graph –** consists of rectangle bars whose heights are the frequencies of the different categories.



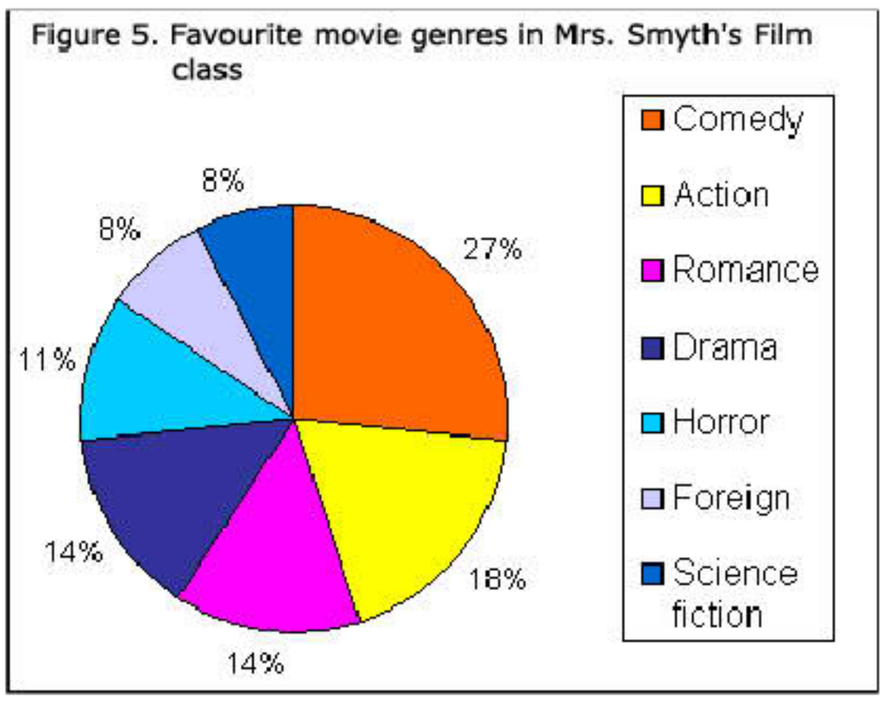


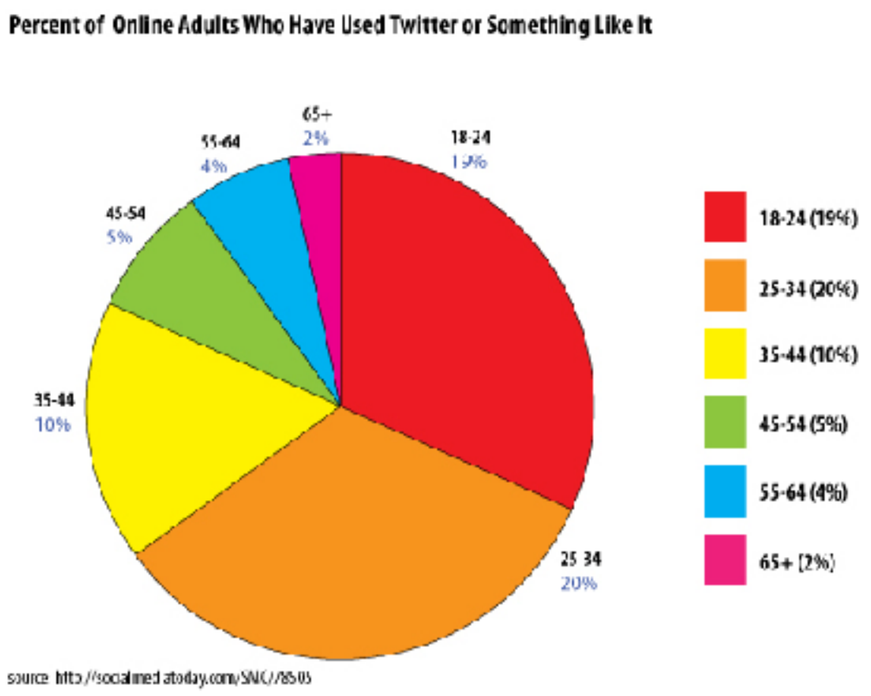
2. **Line Graph –** the frequencies are plotted and connected with lines.





3. **Pie Graph –** a circle who is divided into portions that represent the relative frequencies.





D. Other Approaches in Presenting Data

A. **Stem and Leaf –** are a method for showing the frequency with which certain classes of values occur. You could make a frequency distribution table or a histogram for the values, or you can use a stem-and-leaf plot and let the numbers themselves to show pretty much the same information.

For instance, suppose you have the following list of values: 12, 13, 21, 27, 33, 34, 35, 37, 40, 40, 41. You could make a frequency distribution table showing how many tens, twenties, thirties, and forties you have:

Frequency Class	Frequency
10 - 19	2
20 - 29	2
30 - 39	4
40 - 49	3

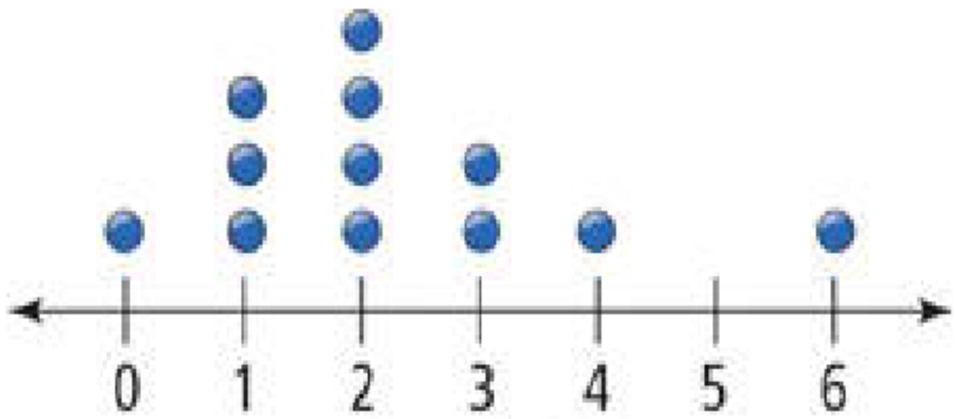
On the other hand, you could make a stem-and-leaf plot for the same data:

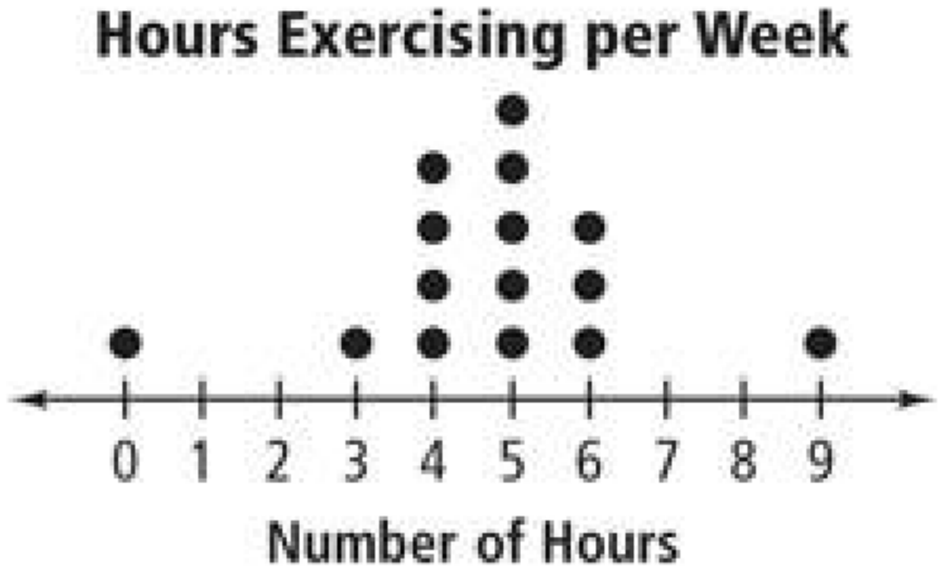
stem	leaf
1	2 3
2	1 7
3	3 4 5 7
4	0 0 1

The "stem" is the left-hand column which contains the tens digits. The "leaves" are the lists in the right-hand column, showing all the ones digits for each of the tens, twenties, thirties, and forties. As you can see, the original values can still be determined; you can tell, from that bottom leaf, that the three values in the forties were 40, 40, and 41.

Note that the horizontal leaves in the stem-and-leaf plot correspond to the vertical bars in the histogram, and the leaves have lengths that equal the numbers in the frequency table.

B. **Dot Plots -** is a statistical chart consisting of data points plotted on a fairly simple scale, typically using filled in circles. There are two common, yet very different, versions of the dot chart.





Box Whiskers – The "box" in the box-and-whisker plot contains, and thereby highlights, the middle half of these data points.

To create a box-and-whisker plot, you start by ordering your data (putting the values in numerical order), if they aren't ordered already. Then you find the median of your data. The median divides the data into two halves. To divide the data into quarters, you then find the medians of these two halves. Note: If you have an even number of values, so the first median was the average of the two middle values, then you include the middle values in your sub-median computations. If you have an odd number of values, so the first median was an actual data point, then you do not include that value in your sub-median computations. That is, to find the sub-medians, you're only looking at the values that haven't yet been used.

You have three points: the first middle point (the median), and the middle points of the two halves (what I call the "sub-medians"). These three points divide the entire data set into quarters, called "quartiles". The top point of each quartile has a name, being a "Q" followed by the number of the quarter. So the top point of the first quarter of the data points is "Q_1", and so forth. Note that Q_1 is also the middle number for the first half of the list, Q_2 is also the middle number for the whole list, Q_3 is the middle number for the second half of the list, and Q_4 is the largest value in the list.
Once you have these three points, Q_1, Q_2, and Q_3, you have all you need in order to draw a simple box-and-whisker plot. Here's an example of how it works.

- Draw a box-and-whisker plot for the following data set:

  4.3, 5.1, 3.9, 4.5, 4.4, 4.9, 5.0, 4.7, 4.1, 4.6, 4.4, 4.3, 4.8, 4.4, 4.2, 4.5, 4.4

  My first step is to order the set. This gives me:

  3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The first number I need is the median of the entire set. Since there are seventeen values in this list, I need the ninth value:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4, 4.4, 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The median is $Q_2 = 4.4$.

The next two numbers I need are the medians of the two halves. Since I used the "4.4" in the middle of the list, I can't re-use it, so my two remaining data sets are:

3.9, 4.1, 4.2, 4.3, 4.3, 4.4, 4.4, 4.4 and 4.5, 4.5, 4.6, 4.7, 4.8, 4.9, 5.0, 5.1

The first half has eight values, so the median is the average of the middle two:

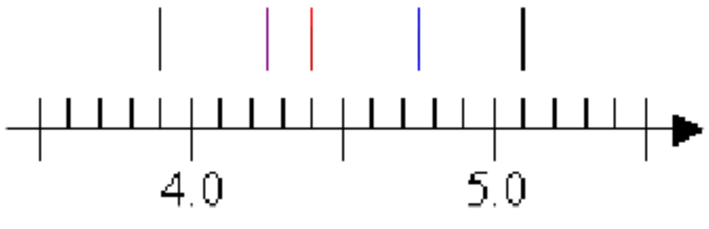
$$Q_1 = (4.3 + 4.3)/2 = 4.3$$

The median of the second half is:

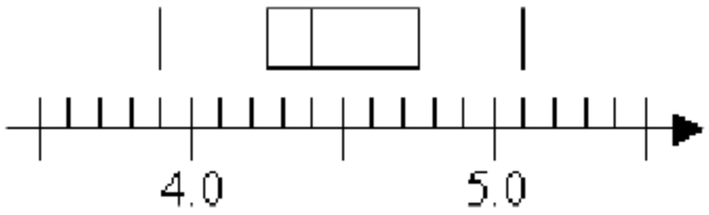
$$Q_3 = (4.7 + 4.8)/2 = 4.75$$

Since my list values have one decimal place and range from 3.9 to 5.1, I won't use a scale of, say, zero to ten, marked off by ones. Instead, I'll draw a number line from 3.5 to 5.5, and mark off by tenths.

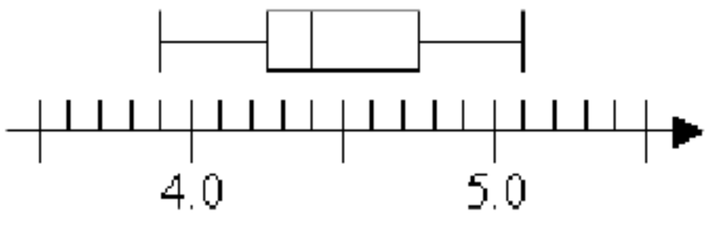
Now I'll mark off the minimum and maximum values, and Q_1, Q_2, and Q_3:



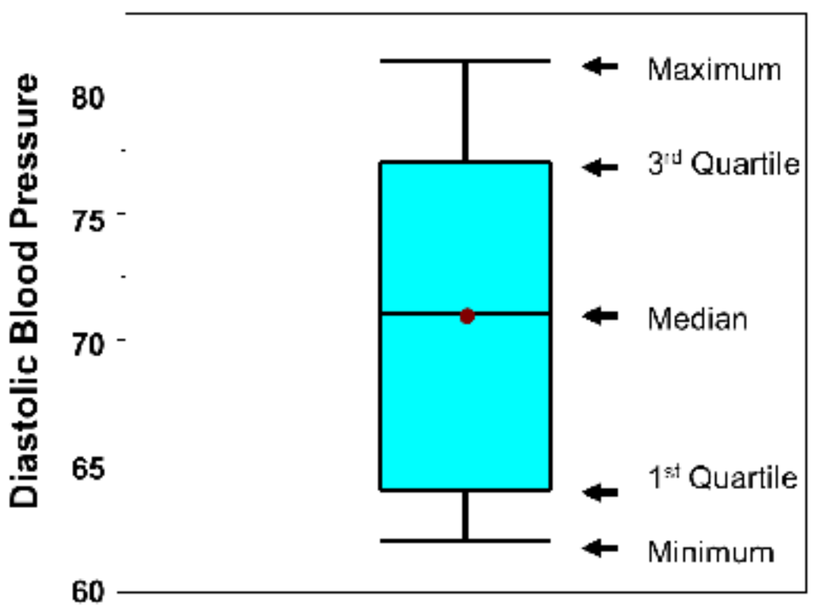
The "box" part of the plot goes from Q_1 to Q_3:

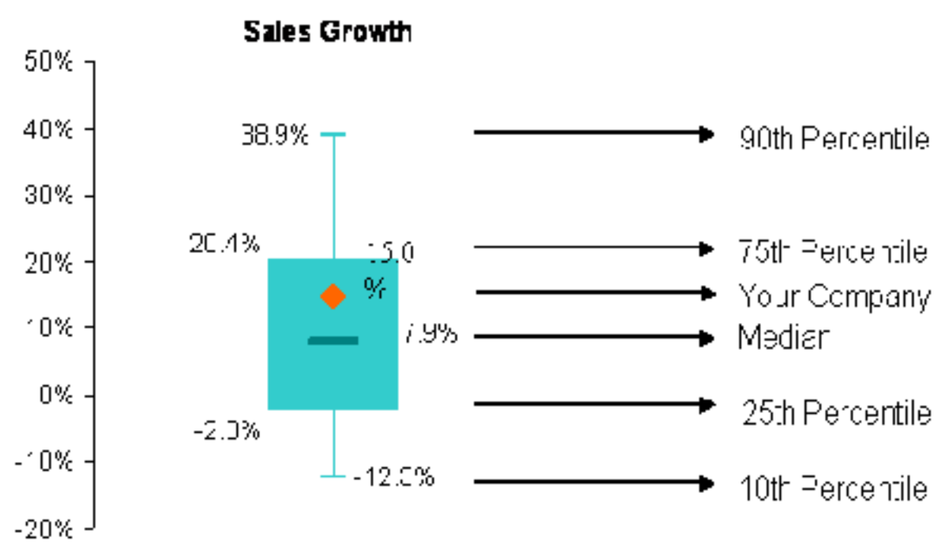


And then the "whiskers" are drawn to the endpoints:



By the way, box-and-whisker plots don't have to be drawn horizontally as I did above; they can be vertical, too.





- The Second Preliminary Topics

Chapter 5: Describing Data with numerical measures

The Frequency Distribution Table

Frequency distributions can show either the actual number of observations falling in each range or the percentage of observations. In the latter instance, the distribution is called a relative frequency distribution.

Frequency distribution tables can be used for both categorical and numeric variables. Continuous variables should only be used with class intervals, which will be explained shortly.

Constructing a frequency distribution table
A survey was taken on Maple Avenue. In each of 20 homes, people were asked how many cars were registered to their households. The results were recorded as follows:
1, 2, 1, 0, 3, 4, 0, 1, 1, 1, 2, 2, 3, 2, 3, 2, 1, 4, 0, 0

Use the following steps to present this data in a frequency distribution table.

- Divide the results (**x**) into intervals, and then count the number of results in each interval. In this case, the intervals would be the number of households with no car (0), one car (1), two cars (2) and so forth.
- Make a table with separate columns for the interval numbers (the number of cars per household), the tallied results, and the frequency of results in each interval. Label these columns Number of cars, Tally and Frequency.
- Read the list of data from left to right and place a tally mark in the appropriate row. For example, the first result is a 1, so place a tally mark in the row beside where 1 appears in the interval column (Number of cars). The next result is a 2, so place a tally mark in the row beside the 2, and so on. When you reach your fifth tally mark, draw a tally line through the preceding four marks to make your final frequency calculations easier to read.
- Add up the number of tally marks in each row and record them in the final column entitled Frequency.

Your frequency distribution table for this exercise should look like this:

Number of cars (x)	Tally	Frequency (f)
0	𝍸	4
1	𝍸 |	6
2	𝍸	5
3	|||	3
4	||	2

Table 1. Frequency table for the number of cars registered in each household

By looking at this frequency distribution table quickly, we can see that out of 20 households surveyed, 4 households had no cars, 6 households had 1 car, etc.

A. Types of Data

1. **Ungrouped Data –** is a type of data not presented in charts or tables. A typical data wherein all values of the observation is presented as how it is granted.

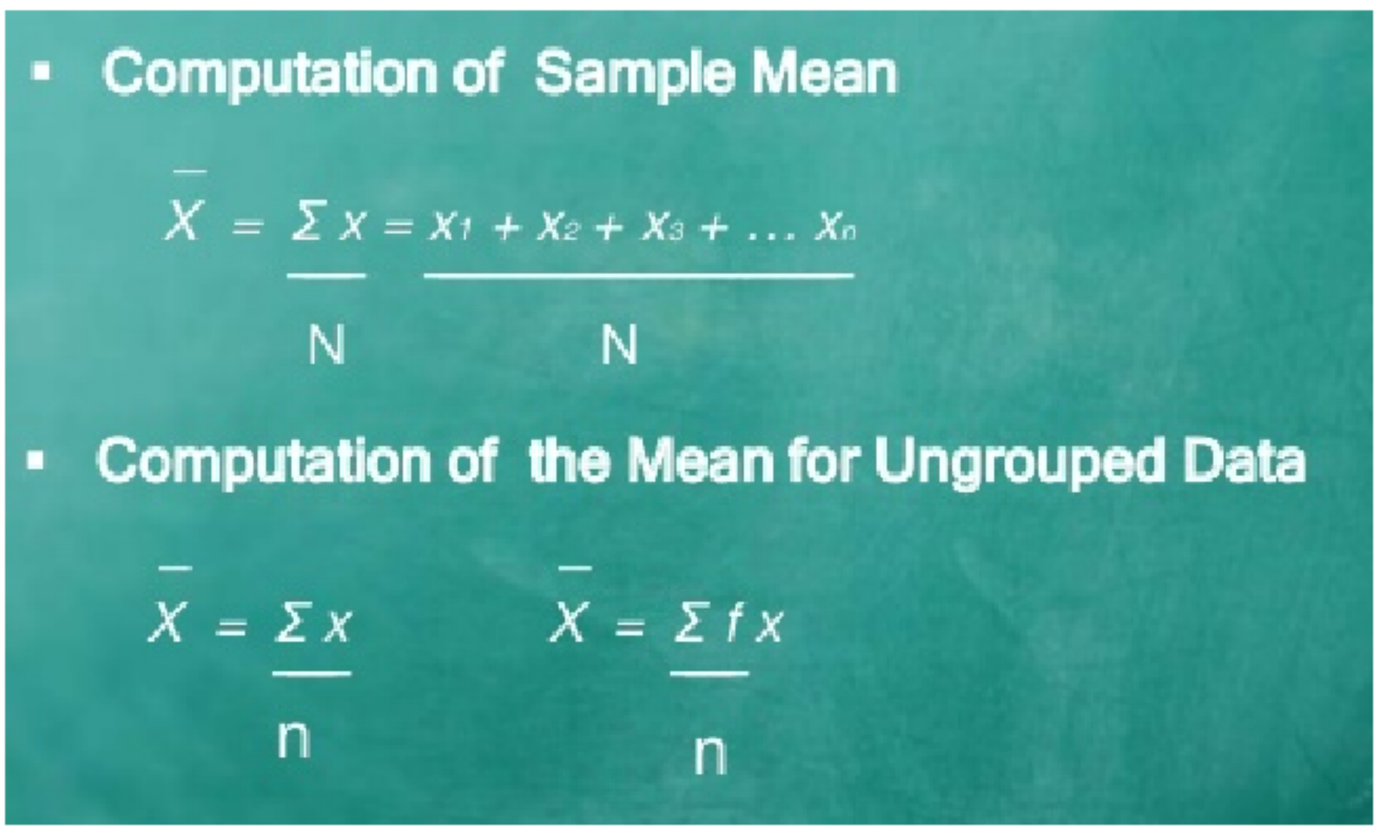
2. **Grouped Data –** is a type of data presented in tables or charts, making it more distinguishable from an ungrouped data.

B. Measures of Central Locations

Central Location – is defined as the tendency of the observations to coverage or meet at point at the center of a frequency distribution.

- **Mean -** Also known as the average. The mean is found by adding up all of the given data and dividing by the number of data entries.
- **Median -** is the middle number. First you arrange the numbers in order from lowest to highest, then you find the middle number by crossing off the numbers until you reach the middle.
- **Mode** - this is the number that occurs most often.

Mean for Ungrouped Data:



Median for Ungrouped Data:

1. Arrange the scores (from lowest to highest or highest to lowest).
2. Determine the middle most score in a distribution if n is an odd number and get the average of the two middle most scores if n is an even number.

a. Odd= middle value
b. Even= $\frac{n + 1}{2}$ $\longleftrightarrow$ $\frac{m1 + m2}{2}$

Mode for Ungrouped Data:

1. arrange in a descending or ascending order
2. Identify the most frequent observation

The mode or the modal score is a score or scores that occurred most in the distribution

It is classified as unimodal, bimodal, trimodal or mulitimodal.

- **Unimodal** is a distribution of scores that consists of only one mode.
- **Bimodal** is a distribution of scores that consists of two modes.
- **Trimodal** is a distribution of scores that consists of three modes or multimodal is a distribution of scores that consists of more than two modes.

Mean for Grouped Data:

Grouped data are the data or scores that are arranged in a frequency distribution.

Frequency is the number of observations falling in a category.

Frequency distribution is the arrangement of scores according to category of classes including the frequency.

$$\bar{X} = \frac{\Sigma f x_m}{n}$$

Where $\bar{X}$ = mean value

x_m = midpoint of each class or category

f = frequency in each class or category

$\Sigma f x_m$ = summation of the product of $f x_m$

CrackADAT

Median for Grouped Data:

1. Complete the table for cf<.
2. Get n/2 of the scores in the distribution so that you can identify MC.
3. Determine LB, cfp, fm, and c.i.
4. Solve the median using the formula.

Formula:

$$\tilde{x} = L_B + \frac{\frac{n}{2} - cfp}{fm} \times c.i$$

$\tilde{X}$ = median value

MC = median class is a category containing the $\frac{n}{2}$

L_B = lower boundary of the median class (MC)

cfp = cumulative frequency before the median class if the scores are arranged from lowest to highest value

fm = frequency of the median class

c.i = size of the class interval

Mode for Grouped Data:

In solving the mode value in grouped data, use the formula:

$$\hat{X} = L_B + \frac{d_1}{d_1 + d_2} \times c.i$$

- **LB** = lower boundary of the modal class Modal Class
- **(MC**) = is a category containing the highest frequency
- **d1** = difference between the frequency of the modal class and the frequency above it, when the scores are arranged from lowest to highest.
- **d2** = difference between the frequency of the modal class and the frequency below it, when the scores are arranged from lowest to highest.
- **c.i** = size of the class interval

CrackADAT

C. Measures of Variability and Dispersion

Ungrouped Data

1. **Range -** The difference between largest and smallest data point. Highly affected by outliers. Best for symmetric data with no outliers.
2. **Variance -** Measures average squared deviation of data points from their mean. Highly affected by outliers. Best for symmetric data. Problem is units are squared.

- If measuring variance of population, denoted by σ^2 ("**sigma-squared**").
- If measuring variance of sample, denoted by s^2 ("**s-squared**").

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

3. **Standard Deviation** - is square root of sample variance, and so is denoted by **s**. Units are the original units. Greek letter sigma **σ** or **s**) is a measure that is used to quantify the amount of variation or dispersion of a set of data value.

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \quad \sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Grouped Data

1. **Range** – is obtained by getting the difference of the highest value of the upper boundary and the lowest value of the lower boundary.
2. **Variance** – considers the position of each observation relative to the mean of the set, or sometimes termed to as the mean deviator.

   a. **Population variance:**

$$Variance(population) = \sigma^2 = \sum_{i=1}^{k} \frac{f_i(x_i-\mu)^2}{n}$$

b. **Sample variance:**

$$Variance(sample) = \sigma^2 = \sum_{i=1}^{k} \frac{f_i(x_i - \mu)^2}{n-1}$$

3. **Standard Deviation** – is square root of sample variance, and so is denoted by **s**. Units are the original units. Greek letter sigma **σ** or ***s***) is a measure that is used to quantify the amount of variation or dispersion of a set of data value.

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \quad \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$$

D. Other measures of location and variability

1. **Quantiles for Grouped**
   a. Quartiles
   b. Deciles
   c. Percetiles

2. **Quantiles for Ungrouped**
   a. Quartiles
   b. Deciles
   c. Percetiles

Quantiles - division of items in the distribution into equal parts.

Quartiles - division of the distribution into 4 equal parts

Deciles - division of the distribution into 10 equal parts.

Percentiles - division of the distribution into 100 equal parts

Steps in computing Quantiles for ungrouped data.

1. Arrange the given data in an array. (increasing or decreasing)
2. Compute for the location of the desired quartile using the following formula:
3. The value obtained in no. 2 is the nth number of the data from the array.
4. Inn case of values not exact, apply fundamental rules in rounding off numbers.

$$First\ Quartile\ Q_1 = P_{25} \qquad First\ Decile\ D_1 = P_{10}$$

$$Second\ Quartile\ Q_2 = P_{50} \qquad Second\ Decile\ D_2 = P_{20}$$

$$Third\ Quartile\ Q_3 = P_{75} \qquad Fifth\ Decile\ D_5 = P_{50}\ and\ so\ on$$

$$Second\ Quartile = Fifth\ Decile = 50th\ Percentile = Median$$

$$Q_2 = D_5 = P_{50} = Median$$

Steps in computing Quantiles for grouped data.

1. Prepare a cumulative frequency distribution table (including CB's).
2. Solve the location of the desired quartile using the formula:

$$K_N = \frac{n(N+1)}{X}$$

3. **Locate the nth items location in the distribution table according to the cumulative frequency (CF), which will then be termed as location of the quantile class.**
4. **Compute for the actual quantile value using the formula:**

$$Q_k = L_i + \frac{\frac{k \cdot N}{4} - F_{i-1}}{f_i} \cdot a_i \qquad k = 1,2,3$$

- $\mathbf{L_i}$ is the lower limit of the quartile class.
- **N** is the sum of the absolute frequency.
- $\mathbf{F_{i-1}}$ is the absolute frequency immediately below the quartile class.
- $\mathbf{a_i}$ is the width of the class containing the quartile.
  The quartiles are independent of the widths of the classes.

5. Apply rules in rounding off numbers.

Other Measures of Variablity or Dispresion

1. **Interquartile Range (Qr)** – an alternate measure of variability whose main components are the difference between Q3 and Q1 which actually measures the length of the interval that contains the middle 50% of the data.

   $$Qr = Q3 - Q1$$

2. **Quartile Deviation (QD)** – is the arithmetic average of the third quartile and the first quartile.

   $$QD = \frac{Q3 - Q1}{2}$$

3. **Mean Absolute Deviation** – is the summation of the absolute deviations of each values of x from the computed Mean divided by the total sample size.
   a. Ungrouped Data
   b. Grouped Data

$$MD = \frac{\sum |X - \bar{X}|}{n} \quad \text{for ungrouped data for mean}$$

$$MD = \frac{\sum f|X - \bar{X}|}{\sum f} \quad \text{for grouped data for mean}$$

$$MD = \frac{\sum |X - \tilde{X}|}{n} \quad \text{for ungrouped data for median}$$

$$MD = \frac{\sum f|X - \tilde{X}|}{\sum f} \quad \text{for grouped data for median}$$

Chapter 6: Graphical Representation of Frequency Distribution

1. **Histogram** – is a form of a frequency distribution. This is a set of vertical bars having their bases on the horizontal axes and whose height represents the frequency. There are two ways in which we can construct a histogram. That is either (1) f vs. x or (2) f vs. class boundary. Each technique should represent or show the same figure or trend when dealing with the same data.

2. **Frequency Polygon** – is useful way of presenting a frequency distribution by plotting f vs. class mark. Then the coordinates or points are simply connected by a straight line, forming a figure that resembles a polygon.

3. **Ogive** – is graphical presentation of a frequency distribution. The main difference of which with respect to a histogram and a frequency polygon is that instead of the frequency, is graphed vs. the class mark.

Chapter 7: Permutations and Combinations

A. Counting Sample Points and Three Diagrams

- **Sample Space** – is the list of all possible outcomes in an experiment or an event.
- **Sample Point** – is an individual outcome in the sample space.

Example: Tossing a coin twice

1st Toss (Sample Space)	2nd Toss (Sample Space)	Sample Points
H	H	HH
	T	HT
T	H	TH
	T	TT

B. Permutations

Permutations of a set of objects is an arrangement of the said objects based on a given order and positions of the one objects is importantly related to another.

Example: The permutation of event A,B,C.
Answer: ABC,ACB,BAC,BCA,CAB, and CBA

Rule 1: Permutation of n different objects taken altogether

$$nPn = n(n-1)(n-2)... = n!$$

Example

In how many ways can a number of permutations of 6 distinct objects taken altogether at a time in a se of {1,2,3,4,5,6}.

Solution

$6P6 = 6\,(6-1)(6-2)(6-3)(6-4)(6-5)$
$= 6 \times 5 \times 4 \times 3 \times 2 \times 1$
$= 720$ times

Rule 2: Permutation of n different objects taken r at a time

$$nPr = n(n-1)(n-2)...(n-r+1), \text{ where } r = 0,1,2...n$$
$$= \frac{n!}{(n-r)!}; \text{ where } r < n$$

Example 1

In how many ways can a number of 3 digits be formed from the digits 1,2,3,4,5,6 without repetitions.

Solution

$n = 6; r = 3$

$${}_6P_3 = \frac{n!}{(n-r)!} = \frac{6!}{(6-3)!} = 120 \text{ ways}$$

CrackADAT

Rule 3: Permutation of Different Kinds

$$P = \frac{n_T!}{n_1!n_2!n_3!...n_k!}$$

where $n_1 = first\ kind$
n_2= second kind
n_3= third kind
$n_T = total$
n_k= th kind

Example 2

How many 10-letter code-words can be formed from the word ASSIGNMENT?

Solution

$n_1 = A = 1; n_2 = S = 2; n_3 = I = 1; n_4 = G = 1;$
$n_5 = N = 2; n_6 = E = 1; n_7 = T = 1; n_8 = M = 1; n_T = 11;$

$P = \frac{11!}{1!2!1!1!2!1!1!1!} = 9\,979\,200$ ways

Rule 4: Circular Permutations

$$P = (n-1)!$$

Example 1

In how many ways can 9 people sit at a round table?

Solution

$n = 9$

$P = (n-1)! = 8! = 40\,320$ ways

C. Combinations

Combinations – Relative position of each element is not emphasized and only the existence of element is considered.

Rule 1: Combination of n objects taken r at a time

$$_nC_r = \frac{_nP_r}{r!} \quad \text{when } 0 < r < n$$

or

$$_nC_r = \frac{n!}{(n-r)!\,r!} \quad \text{when r = 0,1,2,3,...n}$$

Example 2

In how many ways can 6 generic drugs of paracetamol be selected out of 10?

Solution

$n = 10; r = 6$

$_nC_r = \frac{n!}{(n-r)!\,r!} = \frac{10!}{(10-6)!6!} = 210\ ways$

Chapter 8: Probability

Probability

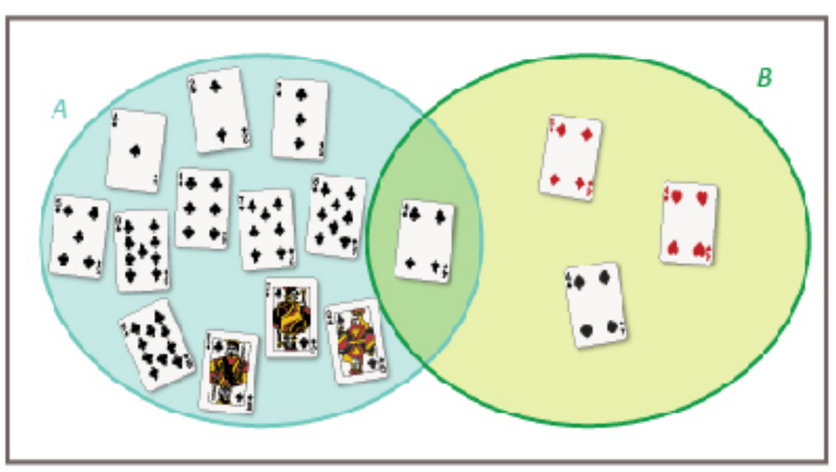
Empirical Probability – is based on consideration of the theoretical number of ways in which it is possible for an event (E) to occur.

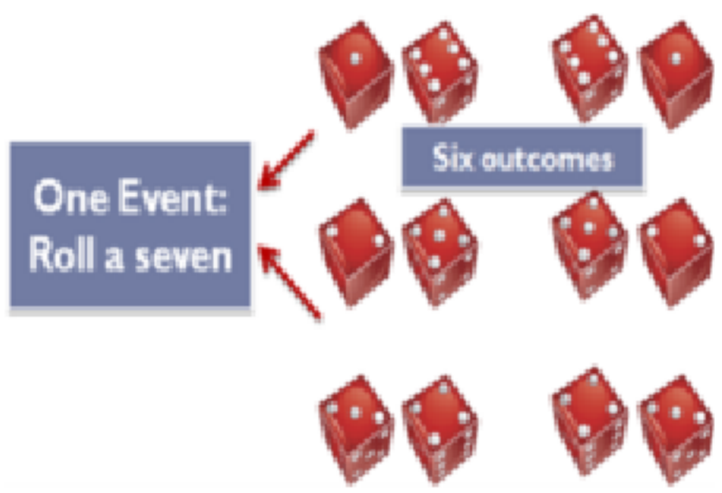
Subjectivity Probability is based on knowledge intuition, or simple guess.

$$P(E) = \frac{\text{\# of favorable outcomes}}{\text{Sample space}} = \frac{n(E)}{n(S)}$$

A. Set of Events

1. A ∪ B is the collection of elements present in A or B.
2. A ∩ B is the collection of elements present in A and B.





Example:

Given: A= {1,2,3,4,5} ; B= {1,3,5,7}

Example 1: A coin is tossed twice, what is the probability of at least 1 head occurs?

Solution:

n(S) = {hh,ht,th,tt} = 4
N(E) = probability that at least 1 head occurs = {hh,ht,th} = 3

$$P(E) = \frac{n(E)}{N(S)} = \frac{3}{4} = 0.75$$

Example 2: If a pair of dice is tossed, what is the probability of getting the same side or a sum of 7?

Solution:

n(S) = 6 x 6 =36
Event A = same side = (1,1)(2,2)(3,3)(4,4)(5,5)(6,6) = 6
Event B = sum of 7 = (4,3) (3,4)(5,2)(2,5)(6,1)(1,6) = 6

$P = \frac{6}{36} + \frac{6}{36} = \frac{1}{3} = 0.33$

Example 3: What is the probability of drawing an ace or a heart from a deck of card in a single draw?

Solution: If two events A and B are mutually exclusive events, then P = P (A) + P(B)

n(S) = 52
Event A = ACE = $\frac{4}{52} = \frac{1}{13}$
Event B = heart = $\frac{13}{52} = \frac{1}{13}$

$P = \frac{1}{13} + \frac{1}{4} = 0.3269 = 0.33$

B. Mutually Exclusive

Two events are mutually exclusive if not more than one of them can happen at the same in any trial.

Example: In a deck of cards, drawing of an ace and the drawing of a jack in the same draw of a single card.

C. Independent Events

Two events are independent if one or both of the events can happen at the same time without interference with other event in atrial.

Example: The drawing of a 6 or a 3 in a roll of a dice.

D. Mutuality Exclusive Event with Common Sample Points

Events A and B are mutually exclusive if A ∩ B contains no sample points - that is if A and B have no sample points in common. For mutually exclusive events,

$$P(A \cap B) = 0$$

Probability of Union of Two Mutually Exclusive Event

If two events A and B mutually exclusive, the probability of the union of A and B equals the sum of the probability of A and the probability of B; that is,

$$P(A \cup B) = P(A)+P(B).$$

Chapter 9: Screening Tests

- A **FALSE POSITIVE** results when test indicates a positive status when the true status is negative.
- A **FALSE NEGATIVE** results when a test indicates a negative status when the true status is positive.

A. Sensitivity and Specificity

Two columns indicate the actual condition of the subjects, diseased or non-diseased. The rows indicate the results of the test, positive or negative.

		Truth		
		Disease (number)	*Non Disease (number)*	*Total (number)*
Test Result	*Positive (number)*	***A*** *(True Positive)*	***B*** *(False Positive)*	$T_{Test\ Positive}$
	Negative (number)	***C*** *(False Negative)*	***D*** *(True Negative)*	$T_{Test\ Negative}$
		$T_{Disease}$	$T_{Non\ Disease}$	***Total***

Cell A contains true positives, subjects with the disease and positive test results. Cell D subjects do not have the disease and the test agrees.

A good test will have minimal numbers in cells B and C. Cell B identifies individuals without disease but for whom the test indicates 'disease'. These are false positives. Cell C has the false negatives.

If these results are from a population-based study, prevalence can be calculated as follows:

- **Prevalence of Disease= $T_{disease}$/ Total × 100**
  The population used for the study influences the prevalence calculation.

  Sensitivity is the probability that a test will indicate 'disease' among those with the disease:

- **Sensitivity: A/(A+C) × 100**
  Specificity is the fraction of those without disease who will have a negative test result:

- **Specificity: D/(D+B) × 100**
  Sensitivity and specificity are *characteristics of the test*. The population does not affect the results.

B. Positive and Negative Predictive Value

A clinician and a patient have a different question: what is the chance that a person with a positive test truly has the disease? If the subject is in the first row in the table above, what is the probability of being in cell A as compared to cell B? A clinician calculates across the row as follows:

- **Positive Predictive Value: A/(A+B) × 100**
- **Negative Predictive Value: D/(D+C) × 100**

Positive and negative predictive values are influenced by the prevalence of disease in the population that is being tested. If we test in a high prevalence setting, it is more likely that persons who test positive truly have disease than if the test is performed in a population with low prevalence..

Test Result \ Truth	Disease (number)	Non Disease (number)	Total (number)
Positive (number)	10 **A** (True Positive)	40 **B** (False Positive)	50 $T_{Test\ Positive}$
Negative (number)	5 **C** (False Negative)	45 **D** (True Negative)	50 $T_{Test\ Negative}$
	15 $T_{Disease}$	85 $T_{Non\ Disease}$	100 Total

t's see how this works out with some numbers...

Hypothetical Example 1 - Screening Test A

100 people are tested for disease. 15 people have the disease; 85 people are not diseased. So, prevalence is 15%:

- Prevalence of Disease:
  $T_{disease}$/ Total × 100,
  15/100 × 100 = 15%
  Sensitivity is two-thirds, so the test is able to detect two-thirds of the people with disease. The test misses one-third of the people who have disease.

- Sensitivity:
  A/(A + C) × 100
  10/15 × 100 = 67%
  The test has 53% specificity. In other words, 45 persons out of 85 persons with negative results are truly negative and 40 individuals test positive for a disease which they do not have.

- Specificity:
  D/(D + B) × 100
  45/85 × 100 = 53%
  The sensitivity and specificity are characteristics of this test. For a clinician, however, the important fact is among the people who test positive, only 20% actually have the disease.

- Positive Predictive Value:
  A/(A + B) × 100
  10/50 × 100 = 20%
  For those that test negative, 90% do not have the disease.

- Negative Predictive Value:
  D/(D + C) × 100
  45/50 × 100 = 90%
  Now, let's change the prevalence.

C. Drug Efficacy

Efficacy is the capacity for beneficial change of a given intervention, most commonly used in the practice of medicine and pharmacology.

In pharmacology, efficacy (E_{max}) is the maximum response achievable from a drug. Intrinsic activity is a relative term that describes a drug's efficacy relative to a drug with the highest observed efficacy.
Effectiveness refers to the ability of a drug to produce a beneficial effect.

Chapter 10: Sampling Distributions

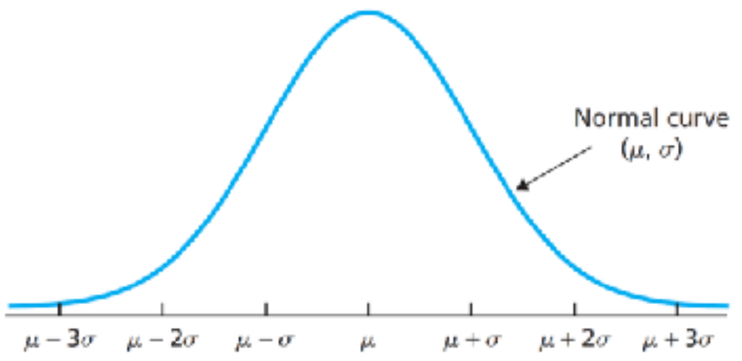
The **sampling distribution** of a statistic is the distribution of that statistic, considered as a random variable, when derived from a random sample of size n. It may be considered as the distribution of the statistic for all possible samples from the same population of a given size.

A. Normal Distribution

Normal distributions - are extremely important because they occur so often in real applications and they play such an important role in methods of inferential statistics.

If a continuous random variable has a distribution with a graph that is symmetric and bell- shaped, as in the Figure on the right, and it can be described by the function

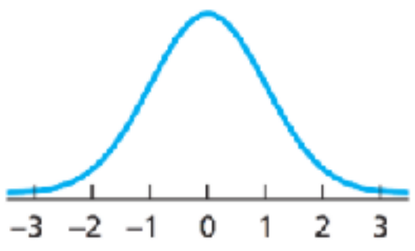
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$



Standardizing a Normally Distributed Variable

How do we find areas under a normal curve? Conceptually, we need a table of areas for each normal curve. This, of course, is impossible because there are infinitely many different normal curves — one for each choice of μ and σ. The way out of this diffi- culty is standardizing, which transforms every normal distribution into one particular normal distribution, the standard normal distribution.

A normally distributed variable having mean 0 and standard deviation 1 is said to have the standard normal distribution. Its associated normal curve is called the standard normal curve, which is shown in the Figure below.



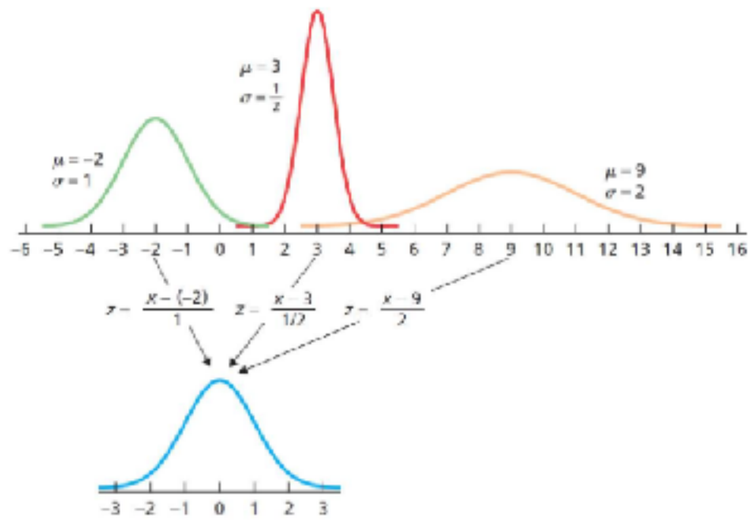
Standardized Normally Distributed Variable

The standardized version of a normally distributed variable x,

$$z = \frac{x - \mu}{\sigma},$$

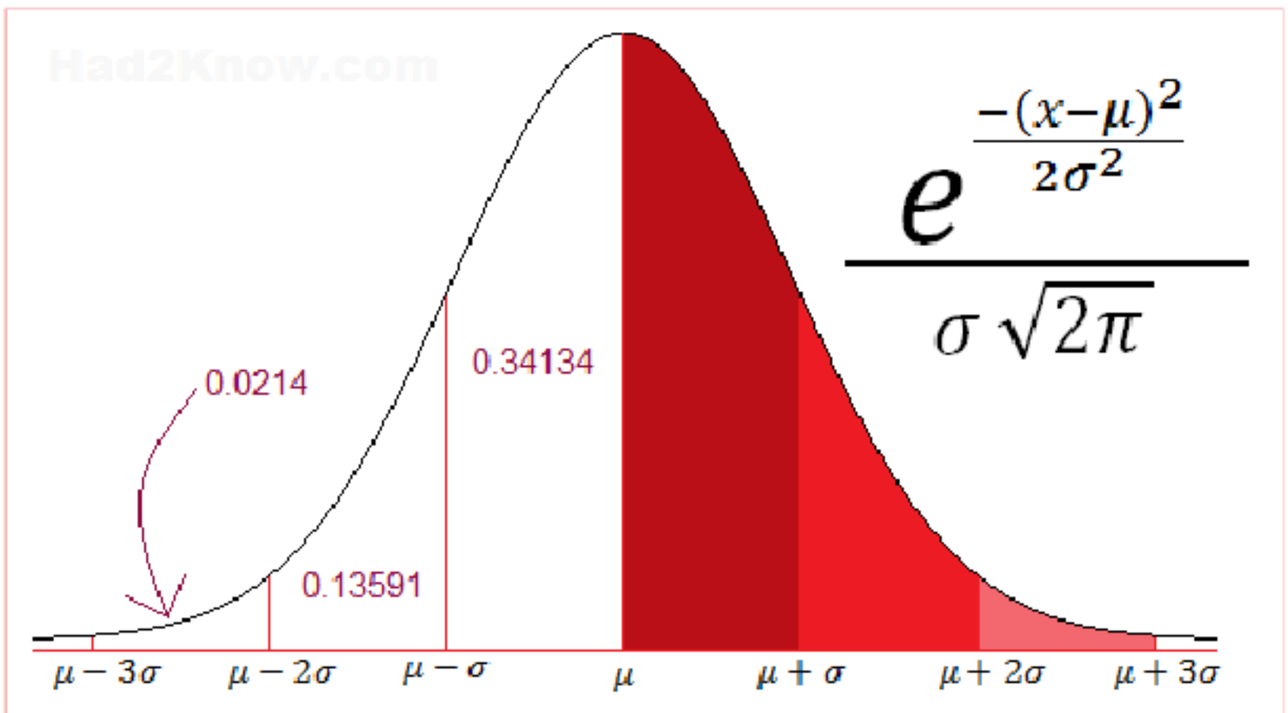
has the standard normal distribution.

We can interpret the above statement in several ways. Theoretically, it says that standardizing converts all normal distributions to the standard normal distribution, as depicted in the Figure below.



Basic Properties of the Standard Normal Curve:

Property 1: The total area under the standard normal curve is 1.
Property 2: The standard normal curve extends indefinitely in both directions, approaching, but never touching, the horizontal axis as it does so.
Property 3: The standard normal curve is symmetric about 0.
Property 4: Almost all the area under the standard normal curve lies between -3 and 3.



B. Binomial Distribution

Binomial Distribution is frequently used to model the number of successes in a sample of size n drawn with replacement from a population of size N. If the sampling is carried out without replacement, the draws are not independent and so the resulting distribution is a hypergeometric distribution, not a binomial one. However, for N much larger than n, the binomial distribution is a good approximation, and widely used.

The Binomial and the Normal Distributions Compared
For large *n* (say n > 20) and *p* not too near 0 or 1 (say 0.05 < p < 0.95) the distribution approximately follows the Normal distribution.

CrackADAT

This can be used to find binomial probabilities.
If X ~ binomial (n,p) where n > 20 and 0.05 < p < 0.95 then approximately X has the Normal distribution with mean E(X) = np

$$\sigma = \sqrt{var(X)} = \sqrt{npq} \quad \text{so} \quad Z = \frac{X - np}{\sqrt{npq}} \quad \text{is approximately N(0,1).}$$

C. Gaussian distribution

Distribution	Functional Form	Mean	Standard Deviation
Gaussian	$f_g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-a)^2}{2\sigma^2}}$	a	σ

If the number of events is very large, then the Gaussian distribution function may be used to describe physical events. The Gaussian distribution is a continuous function which approximates the exact binomial distribution of events.
The Gaussian distribution shown is normalized so that the sum over all values of x gives a probability of 1. The nature of the Gaussian gives a probability of 0.683 of being within one standard deviation of the mean. The mean value is a=np where n is the number of events and p the probability of any integer value of x (this expression carries over from the binomial distribution). The standard deviation expression used is also that of the binomial distribution.

- **Finals Topics**

CHAPTER 11: Standard Scores and Normal Distributions

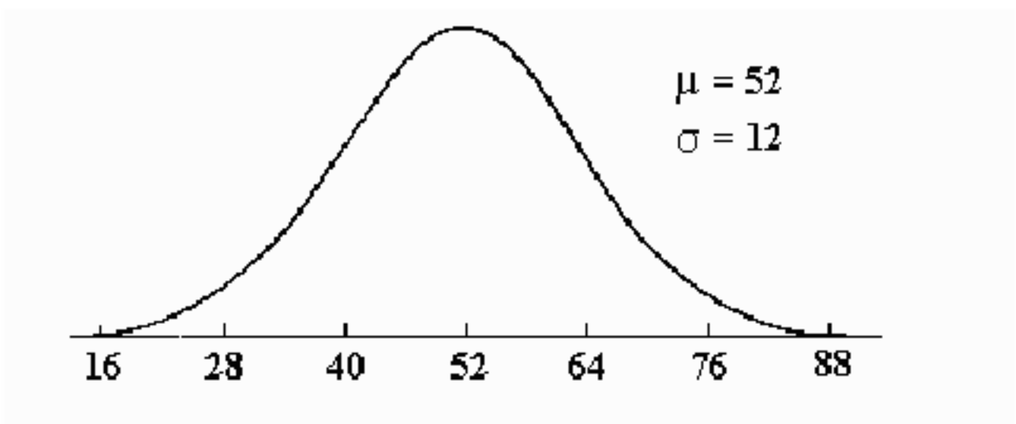
A. The Normal Curve

The **normal curve** is one of a number of possible models of probability distributions. Because it is widely used and an important theoretical tool, it is given special status as a separate chapter. The normal curve is called a family of distributions. Each member of the family is determined by setting the parameters (m and d) of the model to a particular value (number). Because the m parameter can take on any value, positive or negative, and the s parameter can take on any positive value, the family of normal curves is quite large, consisting of an infinite number of members. This makes the normal curve a general-purpose model, able to describe a large number of naturally occurring phenomena, from test scores to the size of the stars.

The normal curve is not a single curve, rather it is an infinite number of possible curves, all described by the same algebraic expression:

$$p(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \quad e^{\frac{-(X-\mu)^2}{2\sigma^2}}$$

The standard procedure for drawing a normal curve is to draw a bell-shaped curve and an X-axis. A tick is placed on the X-axis in corresponding to the highest point (middle) of the curve. Three ticks are then placed to both the right and left of the middle point. These ticks are equally spaced and include all but a very small portion under the curve. The middle tick is labeled with the value of m ; sequential ticks to the right are labeled by adding the value of d . Ticks to the left are labeled by subtracting the value of d from m for the three values. For example, if m =52 and d =12, then the middle value would be labeled with 52, points to the right would have the values of 64 (52 + 12), 76, and 88, and points to the left would have the values 40, 28, and 16. An example is presented below:



B. Standard Scores

(Z – SCORES)

- A technique used to transform original random variables obtained from sampling or original scores to units of standard deviation.

$$z = \frac{x - \mu}{s}$$

Where x = is any value in the distribution
μ = mean of the distribution
s = sample or population standard deviation

Examples:

1. A post operational requirement on renal adgenesis cases is a serum creatinine test for patients 6 months after the surgery. Calculate the standard score of the patient, aged 6, whose serum creatinine level is at 4.4 mg/dl against the normal level of 5.0 mg/dl (minimum), and a standard deviation of 0.5 mg/dl.

Solution:

$$z = \frac{x - \mu}{s} = \frac{4.4 - 5.0}{0.5} = -1.2$$

2. A patient checks her systolic blood pressure daily at home and finds her average systolic BP for 1 week to be 90mmHg. Assuming that her systolic BP to be normally distributed with standard deviation of 5 mmHg, what is her standard score if her systolic BP on a Tuesday is 105 mmHg?

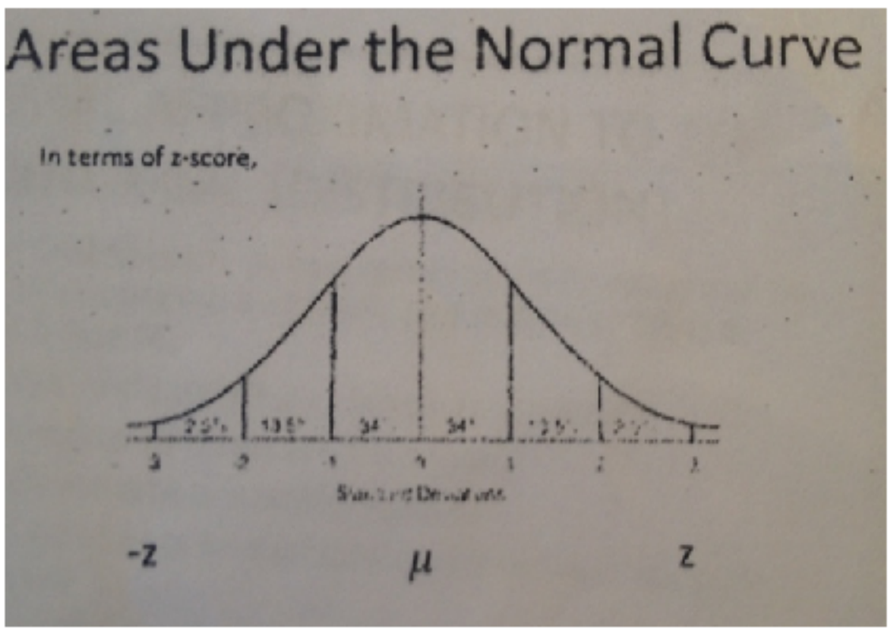
Solution:

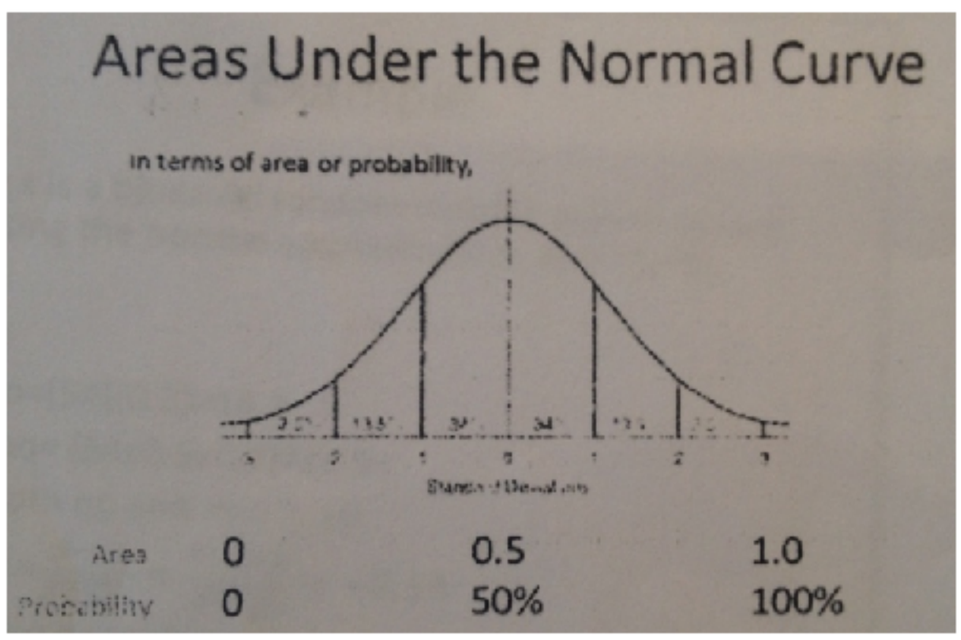
$$z = \frac{x - \mu}{s} = \frac{105 - 90}{5} = 3.0$$

A negative value of z indicates that its position in the normal curve is before the mean.

A positive value of z indicates that its position in the normal curve is after the mean.

C. Areas Under the Normal Curve





Hints in finding the area and probability

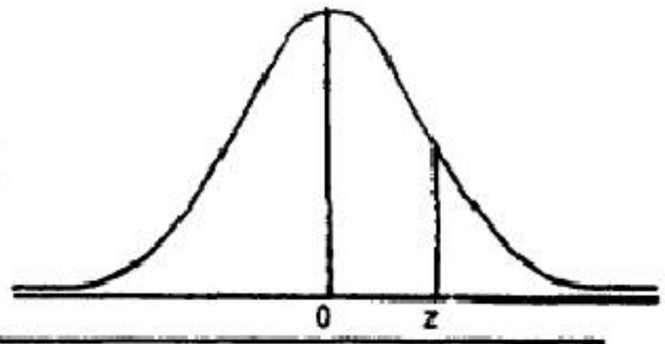
- If z is positive, it means that the data is located after the mean.
- If z is negative, it means that the data is located before the mean.
- If the required probability or area is less than the given random variable or data:

Area (At) = Area from the table (A)
Probability (P%) = At x 100

- If the required probability or area is greater than the given random variable or data:

Area (At) = 1- Area from the table (A)
Probability (P%) = At x 100

TABLE FOR AREAS UNDER THE STANDARD NORMAL CURVE



z	0	1	2	3	4	5	6	7	8	9
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0754
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2258	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2518	.2549
0.7	.2580	.2612	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2996	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.8	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.9	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000	.5000

D. Skewness and Kurtosis

Skewness

The coefficient of Skewness is a measure for the degree of symmetry in the variable distribution.

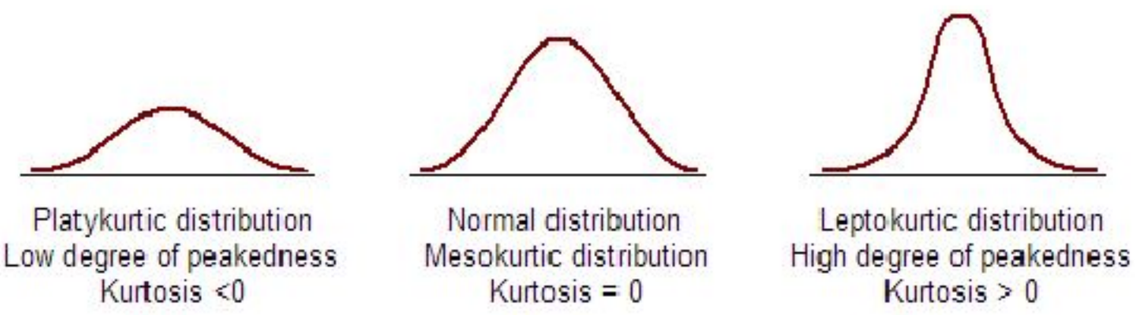
Negatively skewed distribution
or Skewed to the left
Skewness <0

Normal distribution
Symmetrical
Skewness = 0

Positively skewed distribution
or Skewed to the right
Skewness > 0

Kurtosis

The coefficient of Kurtosis is a measure for the degree of peakedness/flatness in the variable distribution.



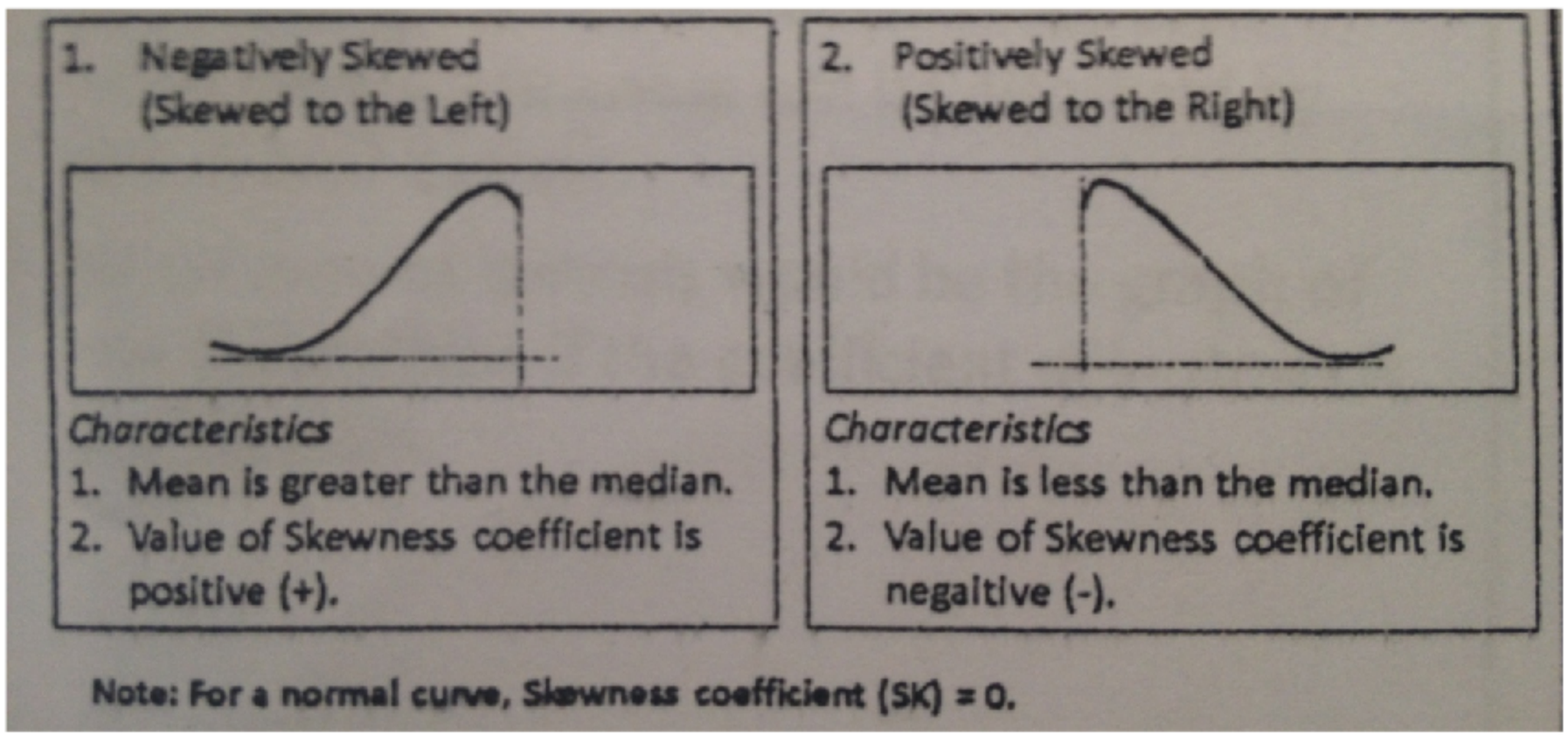
Platykurtic distribution
Low degree of peakedness
Kurtosis <0

Normal distribution
Mesokurtic distribution
Kurtosis = 0

Leptokurtic distribution
High degree of peakedness
Kurtosis > 0

Skewness is the degree of asymmetry, or departure from symmetry of distribution.

Types:



PERSONIAN COEFFICINET OF SKEWNESS (SK)

1. $SK_1 = \frac{\mu - Mo}{s}$

where:

μ = mean
Mo = mode
s = standard deviation

2. $SK_2 = \frac{3(\mu - Md)}{s}$

where:

μ = mean
Md = median
s = standard deviation

SKEWNESS FOR QUANTILES

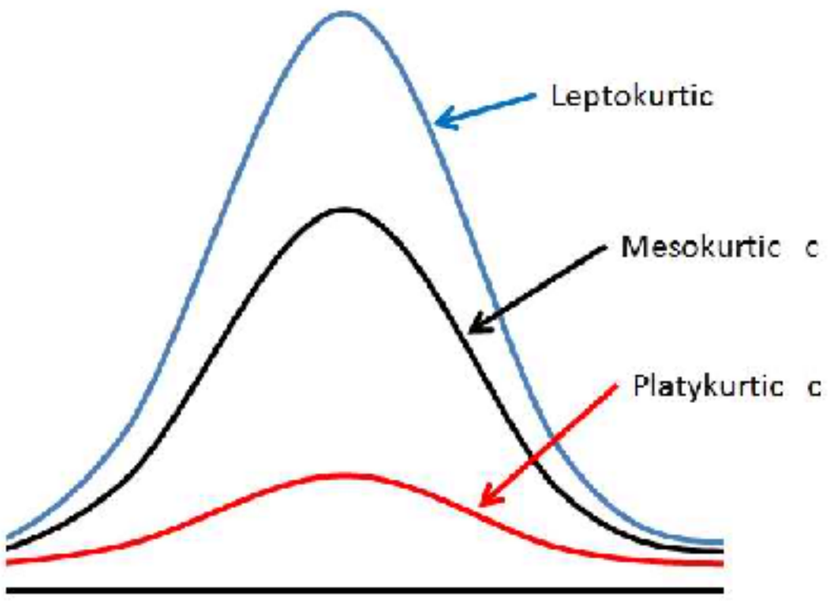
1. Quartile Coefficient of Skewness (S_{QC})

$$S_{QC} = \frac{Q_3 - 2Q_2 + Q_1}{Q_3 - Q_1}$$

2. 10 – 90 Percentile Coefficient of Skewness (S_{PC})

$$S_{PC} = \frac{P_{90} - 2P_{50} + P_{10}}{P_{90} - P_{10}}$$

Kurtosis is the degree of peakedness of a distribution, usually taken relative to a normal distribution.



A. Leptokurtic

- High peak
- Values are connected at the center of the curve with narrow intervals
- K is higher than 3

B. Mesokurtic

- Moderate peakness
- Values are moderately distributed about the center of the curve
- K is equal to 3

C. Platykurtic

- Flat-topped peak
- Values are distributed over a wide range of intervals
- K is lower than 3

E. Normal Approximations to the Binomial

Steps to working a normal approximation to the binomial distribution:

- Identify success, the probability of success, the number of trials, and the desired number of successes. Since this is a binomial problem, these are the same things which were identified when working a binomial problem.
- Convert the discrete x to a continuous x. Some people would argue that step 3 should be done before this step, but go ahead and convert the x before you forget about it and miss the problem.
- Find the smaller of np or nq. If the smaller one is at least five, then the larger must also be, so the approximation will be considered good. When you find np, you're actually finding the mean, mu, so denote it as such.
- Find the standard deviation, sigma = sqrt (npq). It might be easier to find the variance and just stick the square root in the final calculation - that way you don't have to work with all of the decimal places.
- Compute the z-score using the standard formula for an individual score (not the one for a sample mean).
- Calculate the probability desired.

F. Theories governing sampling distribution

- **The Central Theorem**

  Under general conditions, sums and means of random samples of measurements drawn from a population tend to have an approximately normal distribution.

- **Tchebysheff's Theorem**

  Given a number k greater than or equal to 1 and a set of n measurements, at least $[1 - (\frac{1}{k^2})]$ of the measurement will lie within k standard deviations of their mean.

- **Empirical rule**

  It is applicable to mound-shape distributions.

  - The interval m ± s contains approximately 68% of the measurements.
  - The interval m ± 2s contains approximately 95% of the measurements.
  - The interval m ± 3s contains approximately 99.7% of the measurements.

Chapter 12: Statistical Interface

A. Definition and Scope

Statistical inference is the process of deducing properties of an underlying distribution by analysis of data. Inferential statistical analysis infers properties about a population: this includes testing hypotheses and deriving estimates. The population is assumed to be larger than the observed data set; in other words, the observed data is assumed to be sampled from a larger population.

Hypothesis - is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it. Scientists generally base scientific hypotheses on previous observations that cannot satisfactorily be explained with the available scientific theories.

B. Types of Hypothesis

- **Null hypothesis**. The null hypothesis, denoted by H_0, is usually the hypothesis that sample observations result purely from chance.
- **Alternative hypothesis**. The alternative hypothesis, denoted by H_1 or H_a, is the hypothesis that sample observations are influenced by some non-random cause.

C. Types of Error

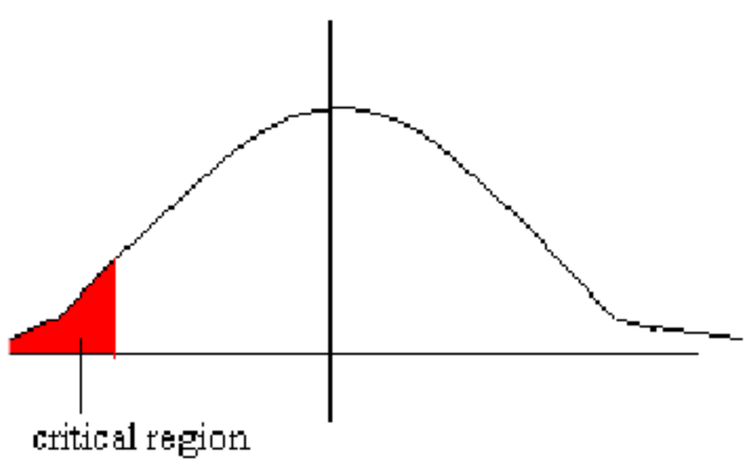
- **Type I error**. A Type I error occurs when the researcher rejects a null hypothesis when it is true. The probability of committing a Type I error is called the **significance level**. This probability is also called **alpha**, and is often denoted by a.
- **Type II error**. A Type II error occurs when the researcher fails to reject a null hypothesis that is false. The probability of committing a Type II error is called **Beta**, and is often denoted by β. The probability of *not* committing a Type II error is called the **Power** of the test.

D. Critical Regions and tailed - test

One and Two Tailed Tests

Suppose we have a null hypothesis H0 and an alternative hypothesis H1. We consider the distribution given by the null hypothesis and perform a test to determine whether or not the null hypothesis should be rejected in favour of the alternative hypothesis.
There are two different types of tests that can be performed. A one-tailed test looks for an increase or decrease in the parameter whereas a two-tailed test looks for any change in the parameter (which can be any change- increase or decrease).
We can perform the test at any level (usually 1%, 5% or 10%). For example, performing the test at a 5% level means that there is a 5% chance of wrongly rejecting H0.
If we perform the test at the 5% level and decide to reject the null hypothesis, we say "there is significant evidence at the 5% level to suggest the hypothesis is false".

One-Tailed Test
We choose a critical region. In a one-tailed test, the critical region will have just one part (the red area below). If our sample value lies in this region, we reject the null hypothesis in favour of the alternative.
Suppose we are looking for a definite decrease. Then the critical region will be to the

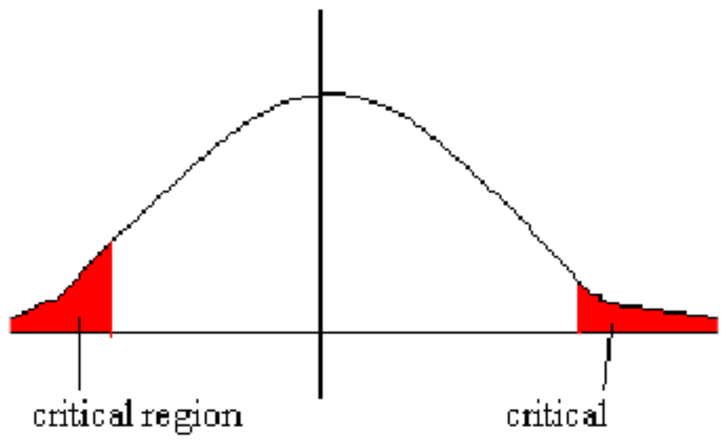


left. Note, however, that in the one-tailed test the value of the parameter can be as high as you like.

Example
Suppose we are given that X has a Poisson distribution and we want to carry out a hypothesis test on the mean, l, based upon a sample observation of 3.
Suppose the hypotheses are:
H0: l = 9
H1: l < 9
We want to test if it is "reasonable" for the observed value of 3 to have come from a Poisson distribution with parameter 9. So what is the probability that a value as low as 3 has come from a Po(9)?
P(X < 3) = 0.0212 (this has come from a Poisson table)
The probability is less than 0.05, so there is less than a 5% chance that the value has come from a Poisson(3) distribution. We therefore reject the null hypothesis in favour of the alternative at the 5% level.

However, the probability is greater than 0.01, so we would not reject the null hypothesis in favour of the alternative at the 1% level.

Two-Tailed Test
In a two-tailed test, we are looking for either an increase or a decrease. So, for example, H0 might be that the mean is equal to 9 (as before). This time, however, H1 would be that the mean is not equal to 9. In this case, therefore, the critical region has two parts:



Example
Lets test the parameter p of a Binomial distribution at the 10% level.
Suppose a coin is tossed 10 times and we get 7 heads. We want to test whether or not the coin is fair. If the coin is fair, p = 0.5 . Put this as the null hypothesis:
H0: p = 0.5
H1: p =(doesn' equal) 0.5
Now, because the test is 2-tailed, the critical region has two parts. Half of the critical region is to the right and half is to the left. So the critical region contains both the top 5% of the distribution and the bottom 5% of the distribution (since we are testing at the 10% level).
If H0 is true, X ~ Bin(10, 0.5).
If the null hypothesis is true, what is the probability that X is 7 or above?
P(X > 7) = 1 - P(X < 7) = 1 - P(X < 6) = 1 - 0.8281 = 0.1719
Is this in the critical region? No- because the probability that X is at least 7 is not less than 0.05 (5%), which is what we need it to be.

So there is not significant evidence at the 10% level to reject the null hypothesis.

E. Test on Means

A. **Z – Test** is any statistical test for which the distribution of the test statistic under the null hypothesis can be approximated by a normal distribution. Because of the central limit theorem, many test statistics are approximately normally distributed for large samples. For each significance level, the Z-test has a single critical value (for example, 1.96 for 5% two tailed) which makes it more convenient than the Student's *t*-test which has separate critical values for each sample size. Therefore, many statistical tests can be conveniently

performed as approximate Z-tests if the sample size is large or the population variance known. If the population variance is unknown (and therefore has to be estimated from the sample itself) and the sample size is not large ($n < 30$), the Student's *t*-test may be more appropriate.

B. **T- Test** is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution if the null hypothesis is supported. It can be used to determine if two sets of data are significantly different from each other, and is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistic (under certain conditions) follows a Student's *t* distribution.

F. Test on Relationships

A. Pearson Relationships

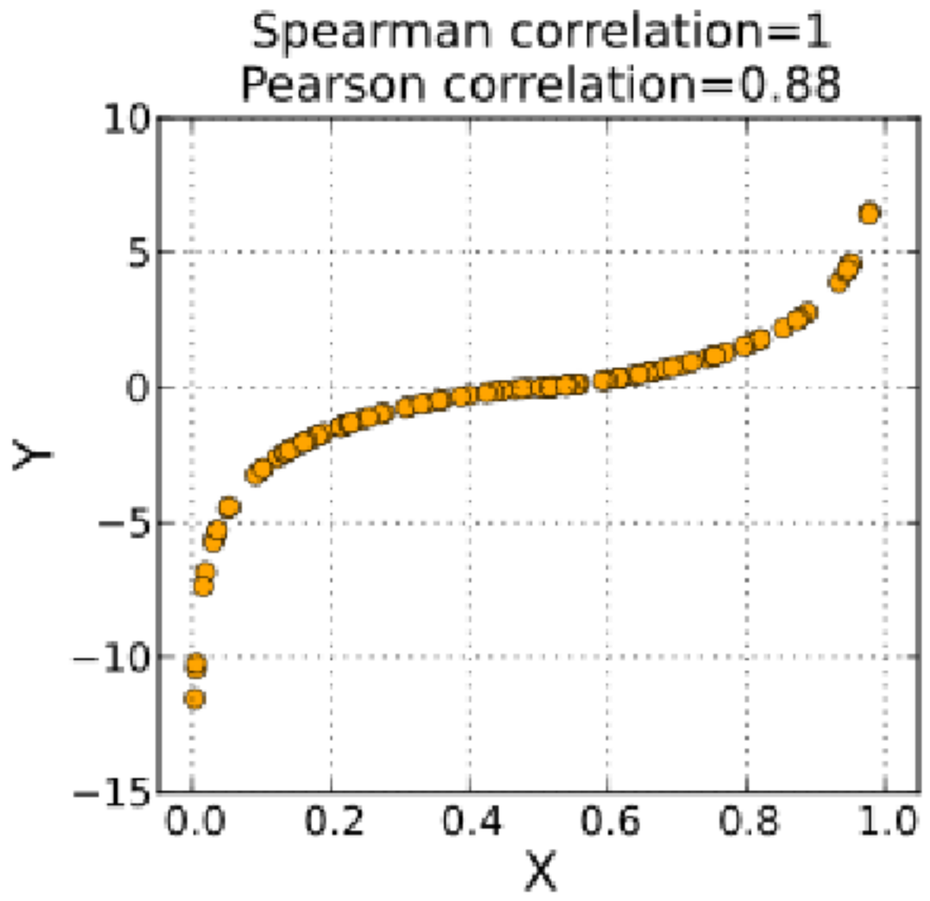
The Pearson product-moment correlation coefficient is a measure of the strength of the linear relationship between two variables. It is referred to as Pearson's correlation or simply as the correlation coefficient. If the relationship between the variables is not linear, then the correlation coefficient does not adequately represent the strength of the relationship between the variables. he symbol for Pearson's correlation is "ρ" when it is measured in the population and "r" when it is measured in a sample. Because we will be dealing almost exclusively with samples, we will use r to represent Pearson's correlation unless otherwise noted

B. Spearman-Rho rank correlation

Spearman's rank correlation coefficient or **Spearman's rho**, named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s, is a nonparametric measure of statistical dependence between two variables. It assesses how well the relationship between two variables can be described using a monotonic function. If there are no repeated data values, a perfect Spearman correlation of +1 or −1 occurs when each of the variables is a perfect monotone function of the other.

Spearman's coefficient, like any correlation calculation, is appropriate for both continuous and discrete variables, including

ordinal variables. Spearman's ρ and Kendall's τ can be formulated as special cases of a more general correlation coefficient.



A Spearman correlation of 1 results when the two variables being compared are monotonically related, even if their relationship is not linear. This means that all data-points with greater x-values than that of a given data-point will have greater y-values as well. In contrast, this does not give a perfect Pearson correlation.

C. Linear Regression

Linear regression is the most basic and commonly used predictive analysis. Regression estimates are used to describe data and to explain the relationship between one dependent variable and one or more independent variables.
At the center of the regression analysis is the task of fitting a single line through a scatter plot. The simplest form with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent, c = constant, b = regression coefficients, and x = independent variable.

Sometimes the dependent variable is also called a criterion variable, endogenous variable, prognostic variable, or regressand. The independent variables are also called exogenous variables, predictor variables or regressors.

Linear regression equation

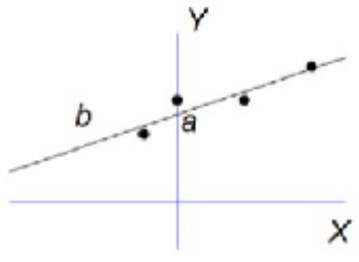
(without error)

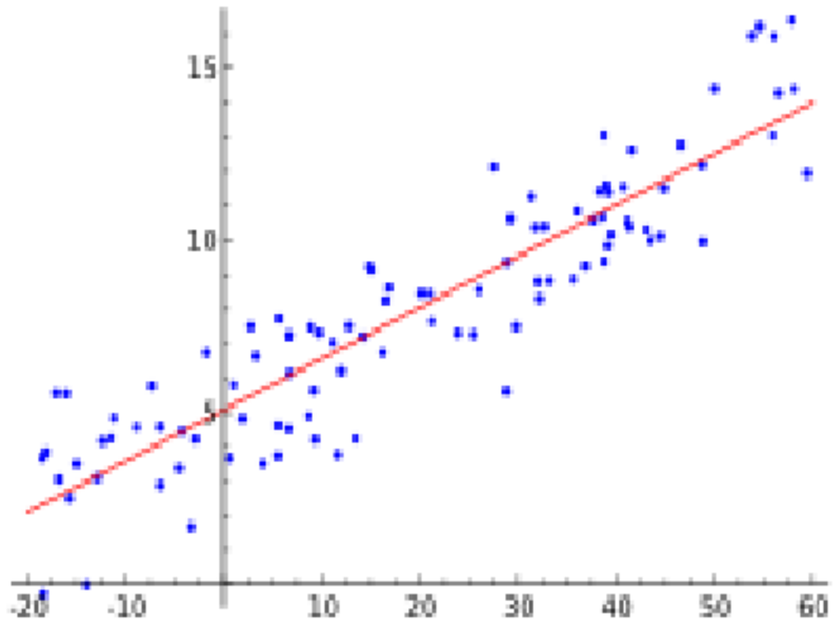
$$\hat{Y} = bX + a$$

predicted values of *Y*

slope = rate of increase/decrease of Y hat for each unit increase in *X*

Y-intercept = level of *Y* when X is *0*.





G. Simple Analysis of Variance (ANOVA)

ANOVA – is a technique in inferential statistics to test whether 2 or more samples (group) are significantly different from one another.

F-test:

$$F = \frac{between\ column\ sum\ of\ squares}{Within-column\ sum\ of\ squares}$$

Simple of One-way ANOVA (Steps):

1. **State the null and alternative hypothesis**

2. **Set the desired level of significance**

3. **Compute the sum of squares by following the formula:**

Total Sum of Squares:

$$TSS = \sum X^2 - \frac{(\sum x)2}{N}$$

Between-column Sum of Squares:

$$SS_b = \frac{1}{r} \sum (\sum X_{ij})^2 - \frac{(\sum x)\,2}{N}$$

Within-column of Sum of Squares:

$$SSw = TSS - SS_b$$

4. Compute for the degrees of freedom using any formula

Total degrees of formula: r = rows

$df_1 = rk - 1 = N - 1$ k = columns

Bet – column df:

$df_b = k - 1$

Within column:

$dfw = df_t - df_b$

5. Compute the mean sum of squares

$MSS_b = SS_b / df_b$

MSSW = SSw / dfw

6. Determine the computed value of F using the formula

F =

7. Determine the tabular value from appendix C

8. Compare the computed F with that if the tabular. Then state the conclusion arrived at:

If Fc<Ft, the null hypothesis is accepted

If Fc>Ft, the null hypothesis is rejected

H. Chi-Square Test

Chi-squared distribution (also chi-square or χ^2-distribution) with k degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. It is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, e.g., in hypothesis testing or in construction of confidence intervals. When it is being distinguished from the more general non central chi-squared distribution, this distribution is sometimes called the central chi-squared distribution.

You would use chi-squared if you were investigating difference between what you actually observed (by collecting primary or secondary data) and what you might normally expect to find.

Here is an example from a piece of human geography research into urban development relating to the 2012 Olympics site. First, look at the chi-squared formula:

$$X^2 = \sum \frac{(o - e)^2}{e}$$

o = the observed frequencies
e = the expected frequencies
Σ = the 'sum of'

Chapter 13: Non-Parametric Tests

Nonparametric statistics are statistics not based on parameterized families of probability distributions. They include both descriptive and inferential statistics. The typical parameters are the mean, variance, etc. Unlike parametric statistics, nonparametric statistics make no assumptions about the probability distributions of the variables being assessed. The difference between **parametric model** and **non-parametric model** is that the former has a fixed number of parameters, while the latter grows the number of parameters with the amount of training data. Note that the *non*-parametric model is not *none*-parametric: parameters are determined by the training data, not the model.

Chapter 14: Introductory Healthcare Statistics

A. Abbreviations used in a Healthcare Facility

Census Taking – is the process of counting patients.

Census Taking Time (CTT) – the time census taking is done.

Daily Inpatient Census (DIPC) – the number of patients present at the CTT each day plus any inpatients who are admitted and discharged after the CTT the previous day.

Inpatient Service Day (IPSD) – a unit measure denoting the services received by one inpatient during one 24 – hour period.

B. Census Formulas

$$\text{Average DIPC} = \frac{\text{Total IPSD for a period}}{\text{Total number of days in the period}}$$

$$\text{A \& C Average DIPC} = \frac{\text{Total IPSD (excluding NB) for a period}}{\text{Total number of days in the period}}$$

$$\text{NB Average DIPC} = \frac{\text{NB IPSD for a period}}{\text{Total number of days in a period}}$$

$$\text{Clinical Unit Ave. DIPC} = \frac{\text{Total IPSD for the Clinical unit for a period}}{\text{Total number of days in a period}}$$

TIPS in Solving Census Problems:

1. A & Ds are not included in an inpatient census but are included in the computation of DIPC.
2. In the computation of Average DIPC, there are separate computations for A & C and NB.
3. Always remember that NB is not included in A & C.
4. Even the following isents not included in the census:

   a. Fetal death

   b. DOA and

   c. OP deaths

Sample Computation:

1. Ward: PEDIA

Date	Census	Statistic	
May 31	Midnight	45	
June 1	admitted	10	
	Discharged		2
	TRF – in	1	
	TRF – out		0
	A & D		2

Census for June 1: (45 + 10 – 2 +1) = 54

DIPC: (45 + 10 -2 + 1 - 2) = 56

IPSD: 56

*DIPC and IPSD are usually numerically equal.

2. (From Basic Allied Health Sciences, 2nd ed.)
Applegate hospital records for the following:

May 31: Census:		141	10	
June 1:	Admissions	8	3	
	Discharged live	2	2	
	Deaths	1	0	
	Fetal deaths			2
	DOAs			2
	A & D	2	0	

Calculate:

a.) Census for June 1: 141 + 8 -2 -1 = 146

b.) IP census for June 1: 141 + 8 – 2 -1 = 146

c.) DIPC for June 1: 141 + 8 – 2 – 1 + 2 = 148

d.) IPSD for June 1: 148

C. Rate Formulas

OCCUPANCY FORMULA

Terms:

- **Bassinet** – are beds or isolettes in the between nursery.
- **Beds** – is a hospital facility that provides patient a permanent place to stay while in the hospital.
- **Inpatient Bed Count/ Bed Complement** – is the number of available hospital inpatient beds, both occupied and vacant at any given day.
- **Newborn Bassinet Count** – is the number of available hospital newborn bassinets, both occupied and vacant at any given day.

a.) % Daily IP/NB Bed/Bassinet Occupancy

$$= \frac{\text{Daily IPSD}}{\text{IP Bed/ Bassinet count for that day}} \times 100$$

b.) % IP/ND Bed/ Bassinet Occupancy for a period

$$= \frac{\text{Total IPSD for a period}}{\text{(Total IP/NB Bed/ Bassinet count X Number of days in the period)}} \times 100$$

c.) % Clinical Unit Occupancy for a Period

$$= \frac{\text{Total IPSD in a clinical unit for a period}}{\text{(Total Bed Count for that unit X number of days in a period)}} \times 100$$

d.) % Occupancy Percentage for a period

$$= \frac{\text{Total IPSD for the period}}{\text{(Total Bed Count for that unit X number of days in a period)}} \times 100$$

Examples:

	Bed Count	Bassinet Count
July 1	201	27
IPSD	187	21

Calculate:

a.) A&C Occupancy Rate

b.) NB Occupancy Rate

Solution:

a.) A&C Occ. Rate = $\frac{187}{210} \times 100 = 89\%$

b.) NB Occ. Rate = $\frac{21}{27} \times 100 = 78\%$

OTHER RATE FORMULA

1. **Mortality Death Rate –** refers to the occurrence of death or fatality at a giver, period with respect to the number of live births.

$$\textbf{Infant Mortality rate} = \frac{\textbf{infant death (neonatal + post natal)}}{\textbf{Number of live births}} \times \textbf{100}$$

Example: If 250,456 live births were reported in Pasay City where 2,321 are infant death, then the infant mortality rate is…

$$= \frac{2{,}321}{250\ ,\ 456} \times 1000 = 9.27\%$$

2. **Fetal death rate** – is defined as the number of fetal deaths over the number of live births plus the number of fetal deaths, quotient multiplied to 1000. This can be death of the fetus due to abortion or still birth.

Fetal death rate is computed as:

$$= \frac{4{,}834}{(501{,}320 + 4{,}834)} \times 1000 = 9.55$$

3. **Morbidity rate** - Is the rate associated with the disease. This is also called infection rates or nosocomial infection (infections originated from the hospital)
4. **Prevalence or Prevalence rate** – is defined as the number of existing cases of the disease in particular population in a given period. This can be multiplied to a factor cipher by 1000, 100,000, 1,000 which serves as a reference.

Prevalence rate formula:

$$= \frac{\text{known cases of disease (for period)}}{\text{population (for period)}} \times \text{factor}$$

Example. In the united states, a total of 400,000 people are diagnosed with AIDS during the year 2000. In june the same year, the US statistics reported that there are around 280 million americans living in the US. What is the prevalence of AIDS in the US

Incidence Rate =

$$\frac{400{,}000}{280{,}000{,}000} \times 100{,}000$$

= 142.86 per 100,000 population/ year.

5. **Incidence or incidence rate** In morbidity, incidence refers to the frequency or extent of the occurrence of the disease.

Incidence rate formula:

$$= \frac{\text{newly reported cases (in a period)}}{\text{population at the mid-period}} \times \text{factor}$$

Example : the population of the bay city in july 2001was 1.2 million with 5,546 New cases of dengue fever. Determine the incidence rate of dengue in bay city.

$$\text{Incidence} = \frac{5{,}546}{1{,}200{,}000} \times 1000$$

$$= 4.62$$

6. Nosocomial Rate or hospital infection rate

Nosocomial rate formula:

$$= \frac{\text{total number of infections.}}{\text{total number of discharges}} \times 100$$

(including death)