# SUPPLEMENTARY NOTES IN ELEMENTARY STATISTICS

JOHN CARLO P. DAQUIS

STATISTICS 101

# Supplementary Notes In Elementary Statistics

## Statistics 101

John Carlo P. Daquis

These notes serve as an accompaniment to the Stat 101 Course Syllabus.

# *Course Outline*

**1. Introduction**
- 1.1. Nature of Statistics
- 1.2. Population & Sample

**2. Collection & Presentation of Data**
- 2.1. Preliminaries
- 2.2. Methods of Data Collecting
- 2.3. Sampling
- 2.4. Tabular & Graphical Presentation
- 2.5. The Frequency Distribution
- 2.6. The Stem-and-Leaf Display

**3. Measures of Central Tendency & Location**
- 3.1. Notations and Symbols
- 3.2. The Arithmetic Mean
- 3.3. The Median
- 3.4. The Mode
- 3.5. Measures of Location

**4. Measure of Dispersion and Skewness**
- 4.1. Measures of Absolute Dispersion
- 4.2. Measures of Relative Dispersion
- 4.3. Measures of Skewness
- 4.4. The Boxplot

**5. Probability**
- 5.1. Random Experiments, Sample Spaces, Events
- 5.2. Properties of Probabilities

**6. Random Variables**
- 6.1. Concept of a Random Variable
- 6.2. Discrete & Continuous Random Variables
- 6.3. Expected Values
- 6.4. The Normal Distribution
- 6.5. Other Common Distributions

**7. Sampling Distribution**

**8. Estimation**

- 8.1. Basic Concepts
- 8.2. Estimating the Mean
- 8.3. Estimating the Difference of Two Population Means
- 8.4. Estimating Proportions
- 8.5. Estimating the Difference Between Two Proportions
- 8.6. Sample Size Determination

**9. Tests of Hypothesis**

- 9.1. Basic Concepts
- 9.2. Testing a Hypothesis on the Population Mean
- 9.3. Testing the Difference between Two Population Means
- 9.4. Testing a Hypothesis on Proportions
- 9.5. Testing the Difference between Two Proportions
- 9.6 Test of Independence

**10. Regression and Correlation**

- 10.1. Correlation Coefficient
- 10.2. Testing the Correlation Coefficient
- 10.3. Simple Linear Regression

Reference *Introduction to Statistics*
*(3rd Edition)*
Ronald E. Walpole

*Any introductory college statistics book*

*“Whenever I look through my notes on probability distributions, my eyes will look for her.*

*For I am always fascinated by the normal distribution –*

*the elegance she has formed from irrationalities,*

*the majestic curve that flows from a point in eternity towards the other unfathomable infinity,*

*the power emanating from her that challenges the impossible.*

*She has given me the power that love could only give:*

*Now I can defy probabilities.”*

JCDMJE

# 1

# Introduction

- *History of Statistics*
- *Statistics Today*
- *Nature of Statistics*
- *Population & Sample*

*"There are three kinds of lies: lies, damned lies, and statistics"*

*-Benjamin Disraeli*

## HISTORY OF STATISTICS

- First used in *statisticum collegium* (Modern Latin) meaning "lecture course on state affairs".
- The "first" statisticians are *statistas* (Italian) which means "one skilled in statescraft" or the politicians.
- In 1770, it was formally defined as a science dealing with data about the condition of a state or a community (political arithmetic). The German word *statistik* was coined by political scientist Gottfried Aschenwall (1719-72) in his paper *"Vorbereitung zur Staatswissenschaft"*.
- The mathematical foundation of Statistics, the theory of probability grew independently. Probability theory had its roots from gambling. The study of modern probability can be traced back from the correspondences between two famous mathematicians: Pierre de Fermat and Blaise Pascal (1654)

## STATISTICS TODAY

Statistics is used in virtually any field and seen in everyday life. Whenever there is data and numbers, there is statistics. There is always statistics in the news: survey results, speeches made by politicians, stocks and Peso-Dollar exchange rates to name a few. We see and hear numbers almost always in billboards, TV and radio advertisements. In your bachelor's course, you will certainly pass through some kinds of data. Below are still some examples of different fields wherein statistics is used.

### Quality Management in Manufacturing - $6\sigma$ (Sigma) Process

Statistically defined as "six standard deviations away from the mean", $6\sigma$ sigma consists of methods, particularly in statistical quality control which seeks to achieve a near-perfect standard in manufacturing (near perfect means 3.4 defective parts per million opportunities)

### Insurance and Finance Industries – Actuarial Science

Actuarial science is the discipline that applies mathematical and statistical methods to assess risk in the insurance and finance industries through a number of interrelating subjects, including probability and statistics, finance, and economics.

### Tourism Research – Forecasting Tourism Flow

A tourism researcher can use a regression and forecasting model in order to predict the travel destination of an individual based on the individual's profile (age, sex, monthly income, education, nationality) and his geographic origin relative to his destination.

## DEFINITION OF STATISTICS (CPAI)

In defining the word statistics, one may wish to remember CPAI: Statistics is the science dealing with the COLLECTION, PRESENTATION, ANALYSIS and INTERPRETATION of data. The science of statistics is divided into three fields: Descriptive, Mathematical and Inferential Statistics.

Aside from defining statistics as a science, it can also be defined as any numerical data. For example, the win-loss record of UAAP basketball teams is statistics. Exit poll results, Philippine population growth rate and the number of times your Multiply profile is viewed are all examples of statistics.

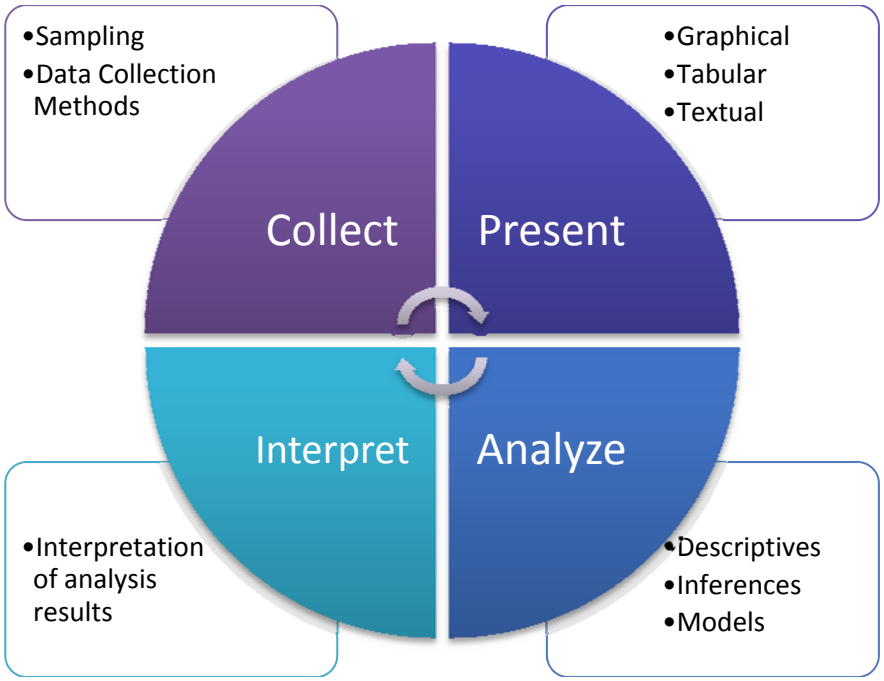



Figure 1-1 The Statistical Cycle

**DESCRIPTIVE STATISTICS**

Descriptive statistics are statistical methods which are used to **describe a collection of data at hand** gathered in various ways. This area of statistics is very essential to understand the behavior of the data being studied in order to make intelligent decisions. It forms the basis of virtually every quantitative (and sometimes qualitative) data analysis

There are various ways in describing data at hand:

- Graphical Presentations (charts, stem-and-leaf display, box-and-whisker plot)
- Tabular Descriptions (frequency distribution tables)
- Summary Statistics or Measures (central tendency, location, dispersion, skewness, kurtosis)

**INFERENTIAL STATISTICS**

Inferential statistics refers to methods which uses data at hand to make generalizations or conclusions from the given information. In other words, this area of statistics is used to draw inferences from the sample to population.

There are two main methods used in inferential statistics:

- **Estimation**
  - **Point Estimation**
  - **Confidence Interval Estimation**
- **Hypothesis Testing**

### DESCRIPTIVE VERSUS INFERENTIAL STATISTICS

| | Descriptive | Inferential |
|---|---|---|
| **Purpose** | Collection, Description and Analysis of data | Making Predictions/Inferences on a larger set of data |
| **Main Concern** | Describe the sample | Infer about the population |
| **Conclusion** | Applies only to data at hand (sample) | Applies to the whole population |
| **Keywords** | historical data, records | predict, estimate |

Table 1-1 Comparing Descriptive and Inferential Statistics

### MATHEMATICAL STATISTICS (PROBABILITY THEORY)

Mathematical statistics refers to the study of statistics from a mathematical standpoint. It often provides a theoretical foundation and a rigorous background on methods used in applied statistics. For example, generalizations or conclusions based on the sample data will most likely contain a certain degree of uncertainty. Hence, one must have a good grasp on probability theory.

### POPULATION VERSUS SAMPLE

The population is simply the entire group of individuals from which information is derived. While the sample is a subset of the population that we actually examine in order to gather information. A numerical characteristic of the population is called a parameter, and the numerical characteristic of the sample is a statistic. Most of the time, since information is based only from the sample, only the statistics can be (computationally) determined. Unknown parameter values are estimated or tested via inferential statistics.

| | Population | Sample |
|---|---|---|
| Nature | All elements considered under the study | Actual information is obtained |
| Size | Superset (greater than or equal) of the sample | Subset of the population |
| Numerical Characteristic | Parameter (may not be known) | Statistic (known) |
| Method of Determination | Estimated, predicted | Computed |
| Process of Obtaining Info | Census | Survey |

Table 1-2 Comparing Population and Sample

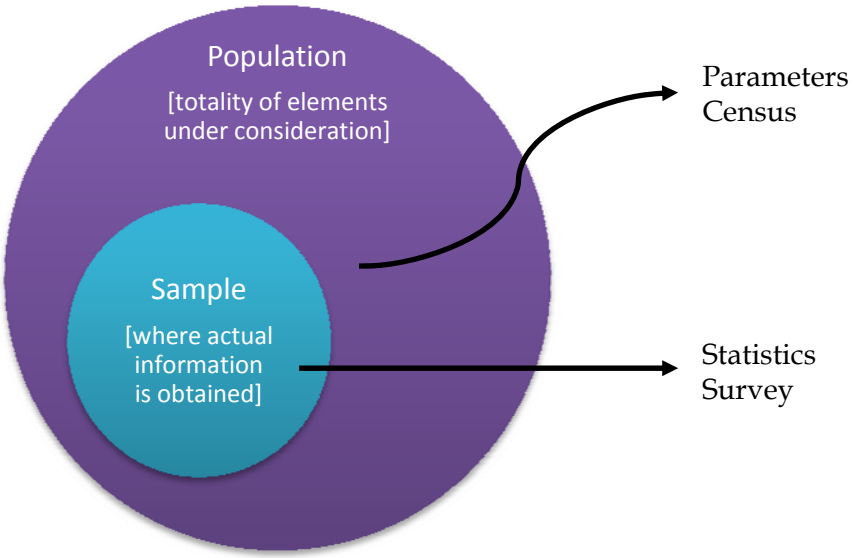



Figure 1-2 The Sample and Population

- In order to avoid exaggerated inferences, vague conclusions, or the sample being irrelevant relative to the study, one MUST clearly define the population in such a way that every element in the population will be relevant to the study regardless if that element is sampled or not.

- One may think: Why do we have to get a sample from the population when the population is always available? Sampling is required, recommended, in some cases even inevitable because of at least two reasons: studying the whole population is very costly in terms of time and resource and sometimes it is infeasible to study the whole population.

- Given the fact that we are only dealing with the sample, how can we be sure that the inferences about the population are close to the true value or scenario of the population? The whole statistics process is quantitatively and rigorously done in such a way error is minimized or measured. Statistics has its foundations in the theory of probability.

# 2

# *Collection and Presentation of Data*

- *Preliminaries*
- *Levels of Measurement*
- *Methods of Data Collection*
- *Sampling*
- *Tabular & Graphical Representation*
- *The Frequency Distribution Table*

*"God cannot be reduced to a sample for analysis"*

*-Kenneth Lee Pike*

## THE RESEARCH PROCESS

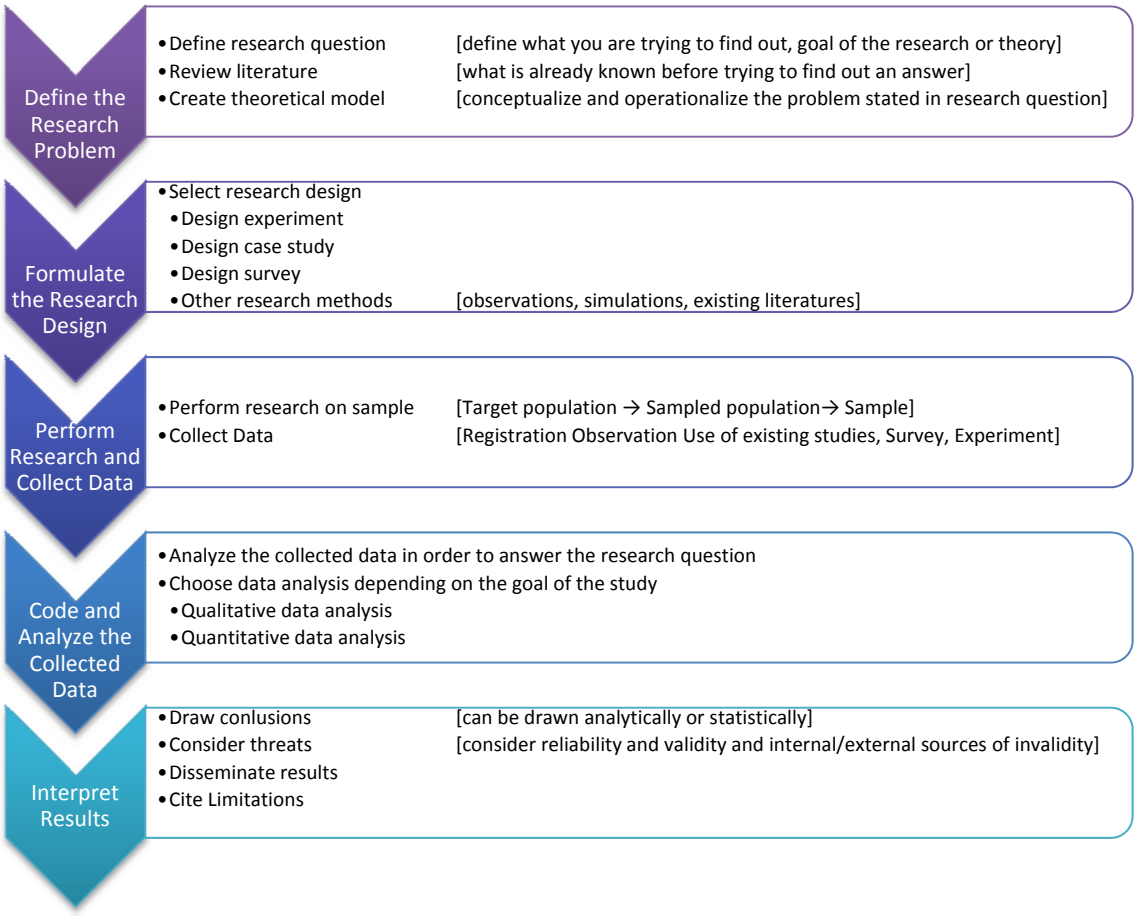



Figure 2-1 The Research Process

## VARIABLES

Variables are properties or characteristics of some event, object, or person that can take on different values or amounts. In general, it is an attribute of a physical or an abstract system which can assume different values which are determined by the measurement process.

## QUALITATIVE VERSUS QUANTITATIVE VARIABLES

Qualitative variables are sometimes called categorical variables. These are variables that express a qualitative attribute which do not imply numerical ordering. For example, the variable "religion" is a qualitative variable. And values of religion do not imply any ordering of religion.

There are some variables which are numeric in form but do not imply numerical ordering. These variables, like telephone number and student number are still classified as qualitative variables.

On the other hand, quantitative variables are variables measured on an ordinal, interval or ratio scale. These variables take on numerical values representing an amount or quantity.

## DISCRETE VERSUS CONTINUOUS VARIABLES

If a variable can take on any value between two specified values, it is called a continuous variable; otherwise, it is called a discrete variable. Some examples will clarify the difference between discrete and continuous variables.

Suppose the PNP mandates that all police officers must weigh between 150 and 200 pounds. The weight of a police officer would be an example of a continuous variable; since a policeman's weight could take on any value between 150 and 200 pounds.

Suppose we flip a coin and count the number of heads. The number of heads could be any integer value between 0 and plus infinity. However, it could not be any number between 0 and plus infinity. We could not, for example, get 2.5 heads. Therefore, the number of heads must be a discrete variable.

Here's a quick-check on how to determine whether a variable is discrete or continuous:

1. Can the values be counted? (Yes – Discrete, No – Continuous)
2. If discrete, will finish counting? (Yes – Finite, No – Countably Infinite)

- Regardless how many the values are (one, thousands, billions, quadrillions) as long as there is an END in enumerating them, the variable which can assume these values is discrete – finite.
- An example of a countably infinite set is the set of natural numbers (1, 2, 3, 4, ….). If we flip a coin and count the number of heads, the number of heads can be equal to any integer between zero to plus infinity.
- An example of a continuous variable is weight. When given two values of weight, one can ALWAYS find a number between the two values

## LEVELS OF MEASUREMENT (NOIR)

Determining the level of measurement of a particular variable is important because it partly determines the arithmetic and statistical procedures you can carry out on them. Remember NOIR: Weakest to strongest – Nominal (categorical), Ordinal (ordered categories), Interval (distance between two numbers is known) and Ratio (has "true zero" point).

### Nominal Level

In the nominal level, numbers, words or letters are used to classify data. Let us say you want to classify right-handed boxers (orthodox) from the left handed ones (southpaws). The term orthodox and southpaw here are used to classify boxers according to their stance. Another example is blood groups where the letters A, B, O and AB represent the different classes.

### Ordinal Level

In the ordinal level, the values given to measurements can be ordered. Furthermore, the numbers in an ordinal scale represents and the ordering of measurements but the difference or ratios between any two measurements along the scale cannot be determined or will not be the same.

As for the nominal scale, one can use textual labels instead of numbers to represent the categories. For example, a scale for the measurement of customer satisfaction on a certain product might look like this: | Not satisfied | Fairly satisfied | Satisfied | Very satisfied |.

There are many everyday examples of measurements assigned to ordinal scales: year level (Grade I – Grade VII), floor number in a certain shopping mall, accession number of books in the library.

**Interval Level**

On an interval scale, measurements are not only classified and ordered therefore having the properties of the two previous scales, but the distances between each interval on the scale are equal right along the scale from the low end to the high end. Two points next to each other on the scale, no matter whether they are high or low, are separated by the same distance.

For example, when you measure temperature in centigrade the distance between 96 and 98 degrees Celsius, for example, is the same as between 100 and 102 degrees Celsius. Remember though is that for interval scales, a measurement of 100 degrees C does not mean that the temperature is 10 times hotter than 10 degrees C even though the value given on the scale IS 10 times as large. That's because there is no absolute zero: the zero is arbitrary. On the centigrade scale, the zero value is taken as the point at which water freezes and the $100^{o}$C value when water begins to boil and between these extreme values the scale is divided into a hundred equal divisions.

**Ratio Level**

Measurements expressed on a ratio scale can have an actual zero. Apart from this difference, ratio scales have the same properties as interval scales. The divisions between the points on the scale have the same distance between them and numbers on the scale are ranked according to size.

There are many examples of ratio scale measurements, length, weight, temperature on the Kelvin scale, speed and counted values like numbers of people, exam marks – a score of zero really does mean no marks!! Returning to the Kelvin scale of temperatures, at the temperature of 0 K the lowest temperature possible, it is so cold that all molecules stop moving.

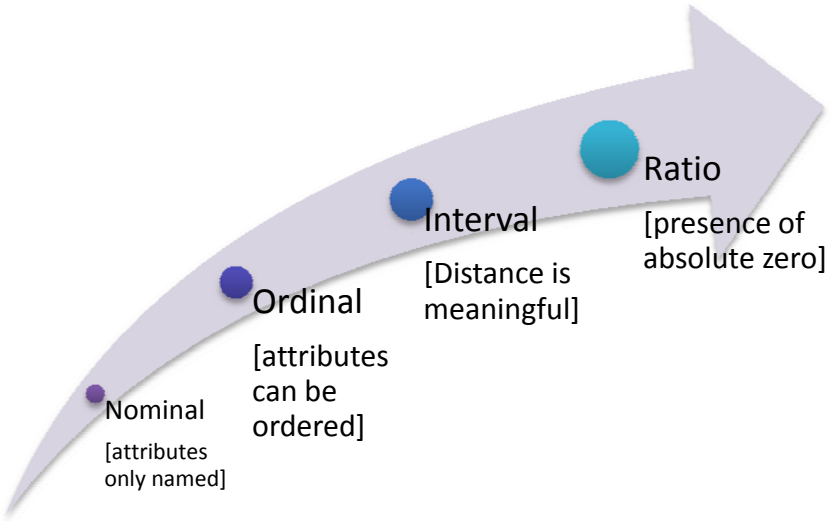



Figure 2-2 The 4 Levels of Measurement

| | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Frequency Distribution | Yes | Yes | Yes | Yes |
| Median and Percentiles | No | Yes | Yes | Yes |
| Sum or Difference | No | No | Yes | Yes |
| Mean, Standard Deviation | No | No | Yes | Yes |
| Ratio, Coefficient of variation | No | No | No | Yes |

Table 2-1 Levels of Measurement of Different Statistics

| | Nominal | Ordinal | Interval | Ratio |
|---|---|---|---|---|
| Sex | x | | | |
| Hair Color | x | | | |
| Pulse | | | | x |
| Temp Celsius | | | X | |
| Team Number | x | | | |
| Shoe Size | | x | | |

Table 2-2 Levels of Some Variables

| | **Nominal** | **Ordinal** | **Interval** | **Ratio** |
|---|---|---|---|---|
| **Sex** | | | | |
| **Height** | | | | |
| **Weight** | | | | |
| **Exercise** | | | | |
| **Year of birth** | | | | |
| **Friends** | | | | |
| **Health** | | | | |
| **Color** | | | | |
| **Home Address** | | | | |

**DATA COLLECTION METHODS (ROUSE)**

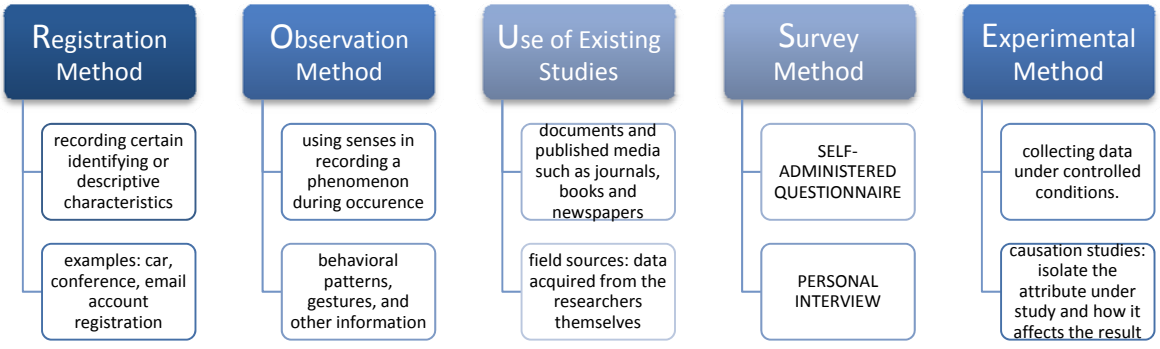



Figure 2-3 Data Collection Methods

### USING THE TABLE OF RANDOM NUMBERS

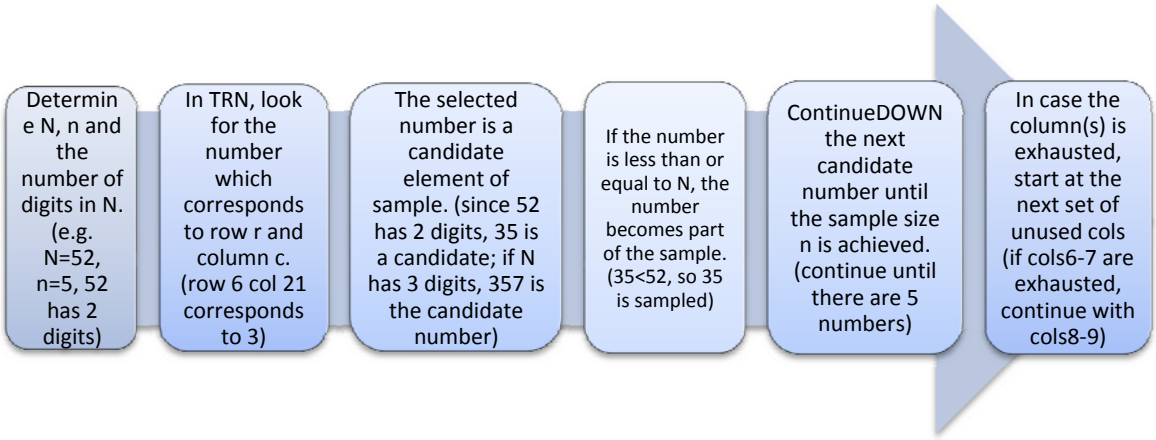


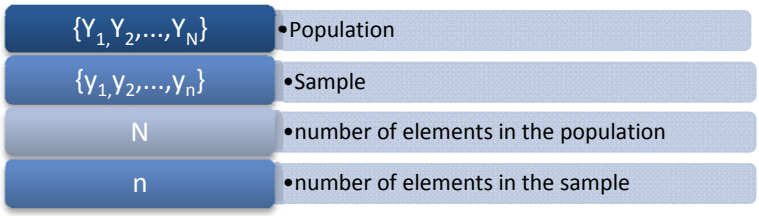

Figure 2-4 Steps in Using the Table of Random Numbers

### SAMPLING

In both qualitative and quantitative research, studies are conducted on a certain population (census). In many cases, only a portion of the total population is selected for study (survey). The goal of the researcher in quantitative studies is to select a sample that replicates the total population. In qualitative research, generalizing to the total population is not a major concern.

Sampling involves the selection of a sample from the population. The goal is to have the sample resemble as much as possible the population.

In sampling, it is important to know such notations:




### PROBABILITY SAMPLING

Probability sampling are methods or sampling plans that gives every element a (i) known and (ii) nonzero probability of being included in the sample. If at least one of the two conditions is not satisfied, then the method is called non-probability sampling. In more technical terms, below is the definition of probability sampling:

A sampling method is considered to be probability sampling if the following conditions are satisfied: Given elements of the population Yi's i = 1 – N, S as the sample and p as the *inclusion probability*, the probability of an element in the population be included in the sample is known and nonzero. In other words, the inclusion probability denoted by

$$P(Y_i \in S; i = 1 - N) = p \text{ is}$$

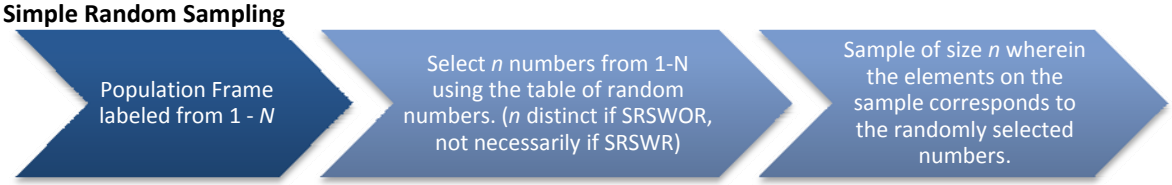
(i) **known** (p can be computed) and
(ii) **nonzero** (p>0).

## PROBABILITY SAMPLING METHODS

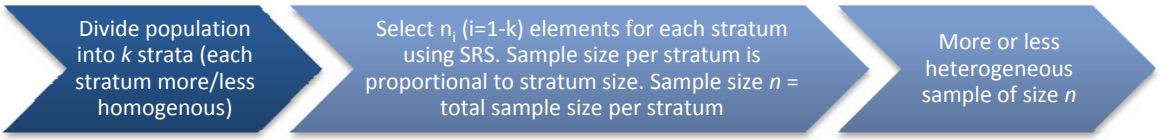
There are 6 probability sampling methods mentioned in this course, 5 of which will be discussed.

1. Simple Random Sampling (with and without replacement)
2. Stratified Random Sampling
3. Systematic Sampling
4. Cluster Sampling
5. Multistage Sampling
6. Sequential Sampling

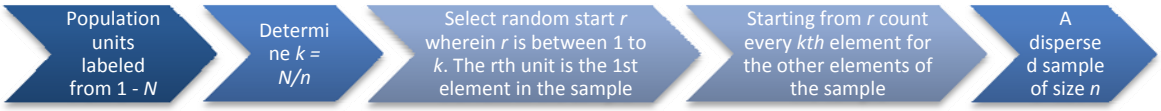
**Simple Random Sampling**

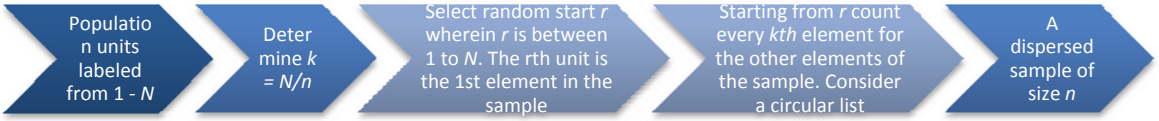



**Stratified Random Sampling**




**Systematic Sampling**




**Cluster Sampling**

### COMPARISON OF PROBABILITY SAMPLING METHODS

| | SRSWR | SRSWOR | STRATIFIED | SYSTEMATIC | CLUSTER | MULTISTAGE |
|---|---|---|---|---|---|---|
| Manner of sampling | Random, replace sampled element | Random, no replacement | SRS for every stratum | From a random start r, sample every kth element | Divide pop'n into clusters, get a sample of clusters | Combination of sampling methods |
| Subpop'n/ sample formed | Heterogeneous /Homogenous sample equally likely | Heterogeneous / Homogenous sample equally likely | Homogenous Stratum --> Heterogeneou s Sample | Dispersed. Sample hetero if there's no periodic behavior | Sample of clusters similar with respect to the population | Depends on the sampling technique per stage |
| Notations | N = pop'n size | N = pop'n size | N = pop'n size | N = pop'n size | N = pop'n size in clusters | |
| | n = sample size | n = sample size | n = sample size | n = sample size | n = sample size in clusters | |
| | | | k = # of strata | k = N/n (sampling interval) | M = # of elements per cluster (if equal # of elements per cluster) | |
| | | | $N_k$ = kth stratum/subp op'n size | r = random start (between 1-k in method 1) | nM = # of elements in the sample (equal # of elts per cluster) | |
| | | | $n_k$ = sample size from the kth stratum | r = random start (between 1-N in method 2) | | |

Table 2-3 A Comparison of Probability Sampling Methods

### NON-PROBABILITY SAMPLING METHODS

There are 4 non-probability sampling methods discussed in this course:

- purposive sampling
- quota sampling
- convenience sampling
- judgment sampling

These non-probability methods are much easier to implement than the probability methods. However, probability sampling is preferred since most statistical methods, like estimation apply objectively to the results of probability sampling.

### THE NEED FOR NON-PROBABILITY SAMPLING

- population has very few elements / the sample itself is difficult to obtain
- a particular study requires highly-sensitive or confidential data
- the number of elements in the sample is more important than the sample itself
- need to extract information impromptu
- budget and resources are limited

*"Words and pictures can work together to communicate more powerfully than either alone."*
*-William Albert Allard*

## DATA PRESENTATION

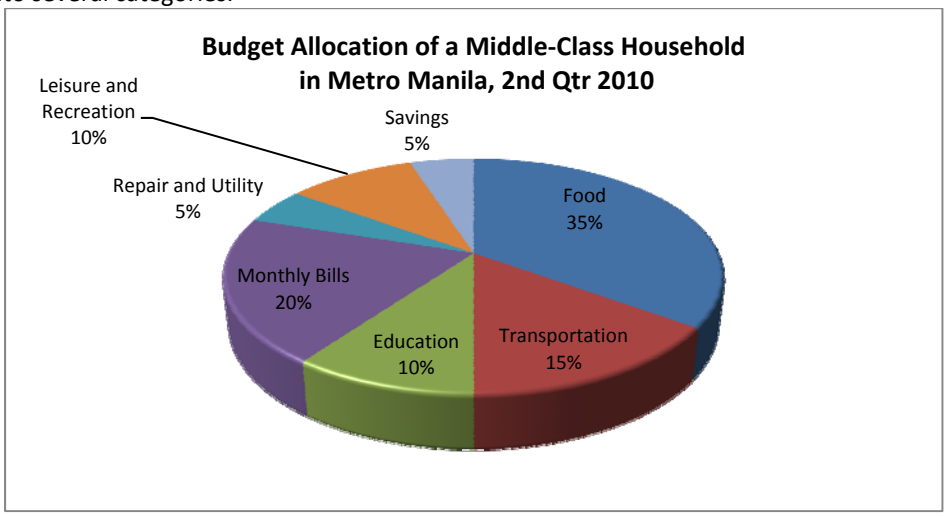
Data can be powerfully and effectively presented to the audience with the use of graphs. Graphs, charts or diagrams make it easier to understand data especially in large quantities and its relationships with other data and/or variables.

It is important to note that many graphs are applicable to one set of data, at which the author is left to choose which graph to use. These charts present the same data graphically but convey different messages or views. Hence, it is important to know certain rules of thumb on choosing the most effective graph type. All examples in the succeeding section are made for illustration purposes only and do not convey actual figures.

## TYPES OF GRAPHS

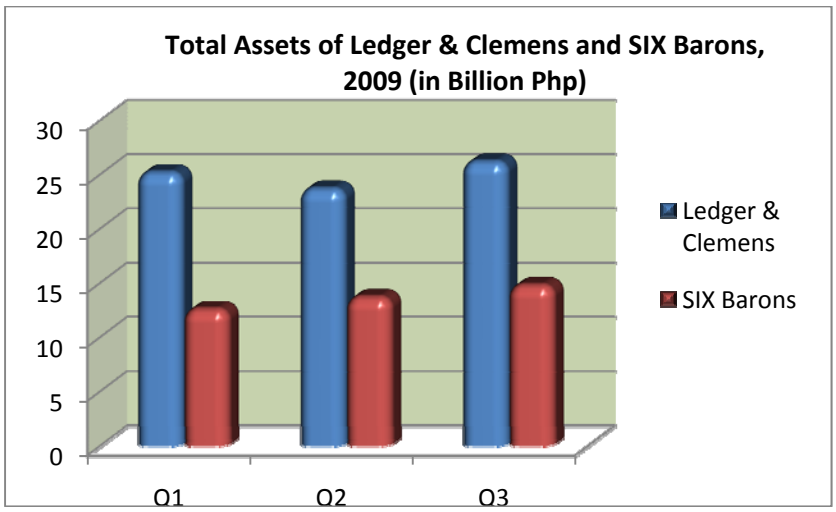
### Pie Charts

Use the pie chart when considering displaying one set of data as part of a whole, or when a main item is divided into several categories.



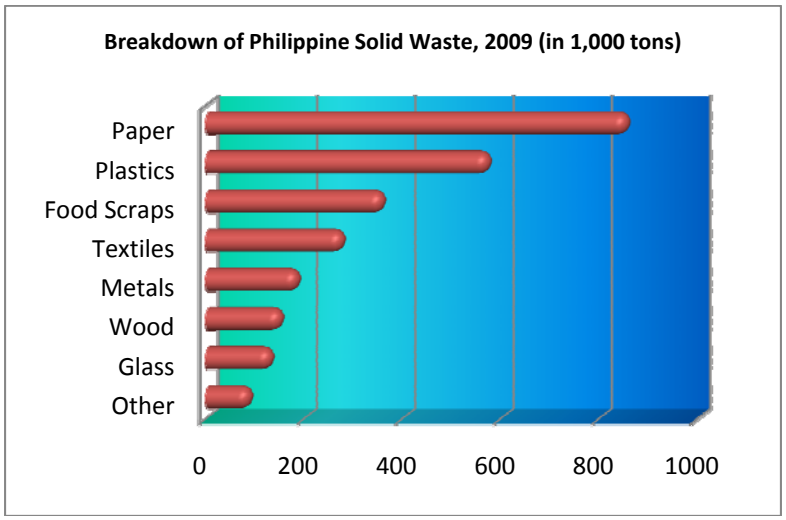
**Vertical Column Charts**

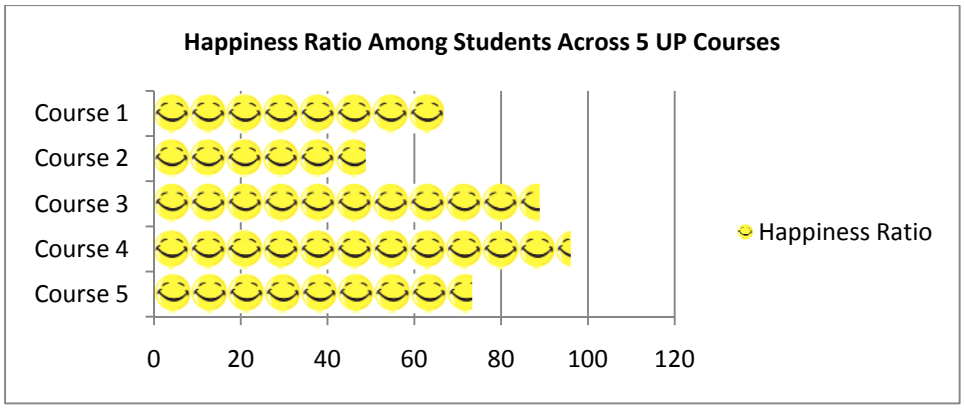
Vertical Column Charts are useful in showing data changes over a period of time or for comparing among several items.




**Horizontal Bar Charts**

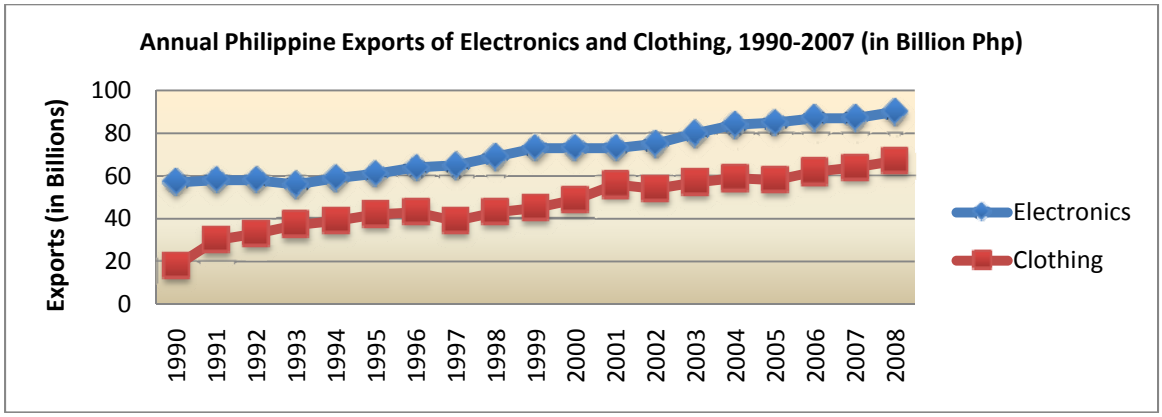
Horizontal bar charts are used to illustrate comparisons among individual items. Sometimes, pictures are used to substitute the bars. Such graphs are called pictographs.





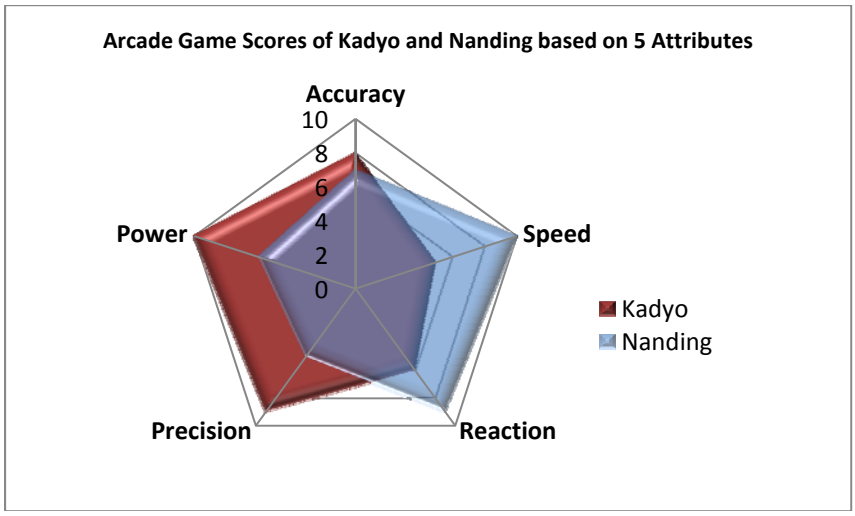

**Line Charts**

Like vertical column charts, line graphs are used to display data across a period of time. The emphasis of line charts are not the actual values, but the trend or movement of the data across time.



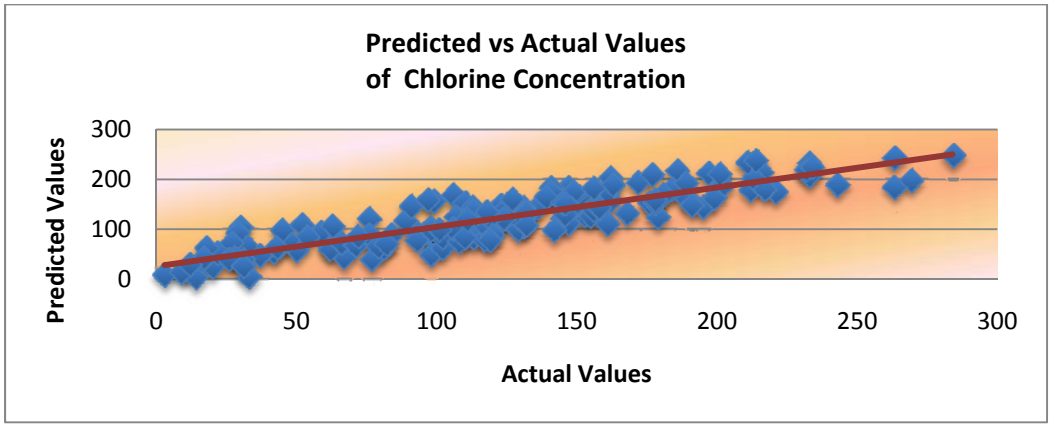

**Radar Graphs**

Radar graphs are used to compare two or more data based on several attributes.



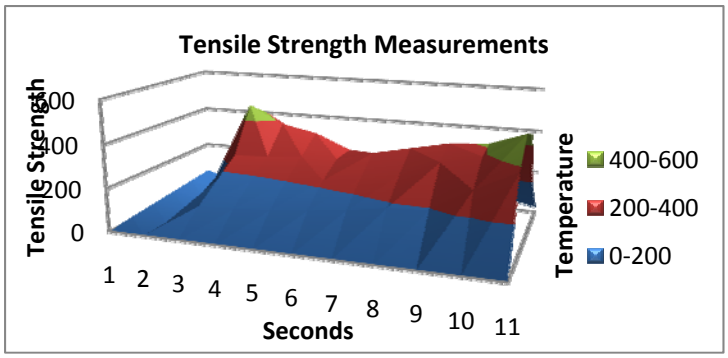
**Scatter Plots**

Like Scatter plots are used mainly in determining the relationship between two variables.



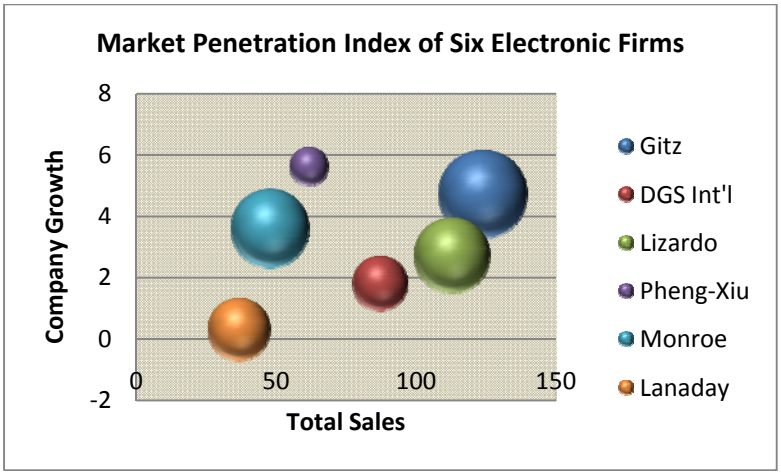

**Surface Charts**

Surface charts are used to display the effects of two variables, X, Y in determining the value of variable Z.




**Bubble Charts**

Bubble Charts are used to display three values in one graph: the X values, the Y values and the radius of the bubble.



*"It is not the number of books you read, nor the variety of sermons you hear, nor the amount of religious conversation in which you mix, but it is the frequency and earnestness with which you meditate on these things until the truth in them becomes you"*

-Frederick W. Robertson

## GETTING STARTED WITH THE FREQUENCY DISTRIBUTION TABLE

A frequency distribution table is defined as an organized tabulation of the number of individual observations located in each category called a class interval.

The Frequency Distribution Table (FDT) is a way to show frequencies in an organized way. It is part of the textual presentation of data. In the FDT, the number of times an observation has occurred or the frequency of the observation is being tallied. In most cases wherein there are many possible outcomes, instead of counting the frequencies of individual observations, they are grouped into class intervals. Then the "frequency" of a class interval is the number of observations that fall within that particular interval.

### PARTS OF THE FREQUENCY DISTRIBUTION TABLE

The FDT is composed of 9 columns. Refer to the table below:

| CLASS INTERVALS | | FREQ | LCB | UCB | CM | RF | RFP | <CF | >CF |
|---|---|---|---|---|---|---|---|---|---|
| 16 | 21 | 10 | 15.5 | 21.5 | 18.5 | 0.13 | 13 | 10 | 80 |
| 22 | 27 | 14 | 21.5 | 27.5 | 24.5 | 0.18 | 18 | 24 | 70 |
| 28 | 33 | 8 | 27.5 | 33.5 | 30.5 | 0.10 | 10 | 32 | 56 |
| 34 | 39 | 10 | 33.5 | 39.5 | 36.5 | 0.13 | 13 | 42 | 48 |
| 40 | 45 | 9 | 39.5 | 45.5 | 42.5 | 0.11 | 11 | 51 | 38 |
| 46 | 51 | 3 | 45.5 | 51.5 | 48.5 | 0.04 | 4 | 54 | 29 |
| 52 | 57 | 9 | 51.5 | 57.5 | 54.5 | 0.11 | 11 | 63 | 26 |
| 58 | 63 | 2 | 57.5 | 63.5 | 60.5 | 0.03 | 3 | 65 | 17 |
| 64 | 69 | 9 | 63.5 | 69.5 | 66.5 | 0.11 | 11 | 74 | 15 |
| 70 | 75 | 6 | 69.5 | 75.5 | 72.5 | 0.08 | 8 | 80 | 6 |
| n | = | 80 | | | | | | | |

Table 2-4. The Frequency Distribution Table

***Class Intervals:*** These are pairs of numbers, called *class limits* which define a class. The lower number in the interval is called the *lower class limit* and the upper number in the interval is called the *upper class limit*.

***Class Frequency (FREQ):*** These are the number of observations that fall within the class interval. It is a good practice to include the sum of observations at the last row in this column.

***Class Boundaries (LCB & UCB):*** These numbers are the true class limits. The *upper class boundary* is halfway between the upper class limit of the same class and the lower class limit of the next class while the *lower class boundary* is a number halfway between the lower class limit of the same class and the upper class limit of the previous class.

***Class Marks (CM):*** Numbers halfway between class limits or class boundaries of the same class.

***Relative Frequency (RF):*** The RF is computed as the frequency of the class divided by the total number of observations.

***Relative Frequency Percentage (RFP):*** It is equal to RF x 100%.

***Less than Cumulative Frequency (<CF):*** It is the number of observations less than the UCB of the same class.

***Greater than Cumulative Frequency (>CF):*** The number of observations greater than the LCB of the same class.

### CONSTRUCTING THE FDT

Creating the FDT can be summarized in three steps. Of course, the second and third steps require a number of processes:


**1. Arrange the Raw Data into an Array**

Let us first start with a raw data (80 observations):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 45 | 70 | 45 | 48 | 37 | 68 | 35 | 74 |
| 25 | 75 | 37 | 71 | 69 | 21 | 53 | 31 |
| 37 | 22 | 65 | 31 | 62 | 16 | 28 | 23 |
| 27 | 36 | 27 | 35 | 45 | 68 | 41 | 30 |
| 19 | 57 | 70 | 39 | 40 | 53 | 57 | 57 |
| 26 | 42 | 41 | 21 | 55 | 54 | 20 | 24 |
| 38 | 25 | 21 | 60 | 64 | 45 | 23 | 70 |
| 69 | 33 | 65 | 20 | 31 | 54 | 40 | 26 |
| 31 | 35 | 47 | 25 | 48 | 55 | 33 | 35 |
| 25 | 66 | 20 | 25 | 67 | 16 | 27 | 20 |

The raw data below is then ordered (preferably ascending) to form what we call an *array*.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 16 | 22 | 26 | 33 | 38 | 45 | 57 | 68 |
| 16 | 23 | 27 | 33 | 39 | 47 | 57 | 68 |
| 19 | 23 | 27 | 35 | 40 | 48 | 57 | 69 |
| 20 | 24 | 27 | 35 | 40 | 48 | 60 | 69 |
| 20 | 25 | 28 | 35 | 41 | 53 | 62 | 70 |
| 20 | 25 | 30 | 35 | 41 | 53 | 64 | 70 |
| 20 | 25 | 31 | 36 | 42 | 54 | 65 | 70 |
| 21 | 25 | 31 | 37 | 45 | 54 | 65 | 71 |
| 21 | 25 | 31 | 37 | 45 | 55 | 66 | 74 |
| 21 | 26 | 31 | 37 | 45 | 55 | 67 | 75 |

**2. Determine the Number of Classes**

Let K be the number of classes and n = number of observations. Approximating K can be done by using the Sturges' Formula:

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

So in our example:

$$K = 1 + 3.322 \log(80)$$
$$= 7.322065$$
$$\approx 8 \text{ (round-up)}$$

Alternatively, one can also use the $2^K$ rule which chooses the smallest K at which $2^K > n$ is satisfied:

K = 6: $2^6 = 64 > 80$? No.

K = 7: $2^7 = 128 > 80$? Yes!

And of course, one's professional judgment determines the number of classes.

**3. Determine the Class Size**

The Class Size C can be approximated (by rounding-up) by the following equation:

C ≈ (highest value – lowest value) / # of classes

C ≈ (75 – 16)/7 = 8.428571 ≈ 9 (using $2^k$ rule to get K = 7)

**4. Determine the Lowest Class Limit**

Note that the first class interval should include the smallest observation. On the other hand, the selection must not be too small so that the highest class interval accommodates the highest observation. In the example, let's say the smallest is 16, so that the first class interval is 12 - 20.

**5. Determine the Class Limits**

First determine the lower class limits by adding C to the lower class limits with the previous class. Since we have started with 15, the next number is 12 + 9 =21, followed by 21 + 9 = 30 and so on.

| CLASS | |
|---|---|
| 15 | 23 |
| 24 | 32 |
| 33 | 41 |
| 42 | 50 |
| 51 | 59 |
| 60 | 68 |
| 69 | 77 |

Before the actual tallying of frequencies, check if the last class frequency is nonzero. Otherwise, adjust the lowest class limit.

**6. Tally frequencies for each class**

| CLASS | | FREQ |
|---|---|---|
| 15 | 23 | 13 |
| 24 | 32 | 17 |
| 33 | 41 | 16 |
| 42 | 50 | 8 |
| 51 | 59 | 9 |
| 60 | 68 | 9 |
| 69 | 77 | 8 |
| **n** | = | 80 |

Here, it is a very good practice to add the frequencies and check if it is equal to the total number of observations.

**4. Complete the FDT**

**LCB/UCB:** *UCB* is halfway between the upper class limit of the same class and the lower class limit of the next class while the *LCB* is number halfway between the lower class limit of the same class and the upper class limit of the previous class.

**CM:** Class mark = (UCB + LCB) / 2

**RF:** Relative Frequency = frequency of the class / number of observations (2 decimal places)

**RFP:** Relative Frequency Percentage = RF X 100%
(may not be equal to 100 due to rounding-off errors)
**<CF:** Add all frequencies less than the UCB
**>CF:** Add all frequencies greater than the UCB

| CLASS | | FREQ | LCB | UCB | CM | RF | RFP | <CF | >CF |
|---|---|---|---|---|---|---|---|---|---|
| 15 | 23 | 13 | 14.5 | 23.5 | 19 | 0.16 | 16 | 13 | 80 |
| 24 | 32 | 17 | 23.5 | 32.5 | 28 | 0.21 | 21 | 30 | 67 |
| 33 | 41 | 16 | 32.5 | 41.5 | 37 | 0.20 | 20 | 46 | 50 |
| 42 | 50 | 8 | 41.5 | 50.5 | 46 | 0.10 | 10 | 54 | 34 |
| 51 | 59 | 9 | 50.5 | 59.5 | 55 | 0.11 | 11 | 63 | 26 |
| 60 | 68 | 9 | 59.5 | 68.5 | 64 | 0.11 | 11 | 72 | 17 |
| 69 | 77 | 8 | 68.5 | 77.5 | 73 | 0.10 | 10 | 80 | 8 |
| **n** | = | 80 | | | | | | | |

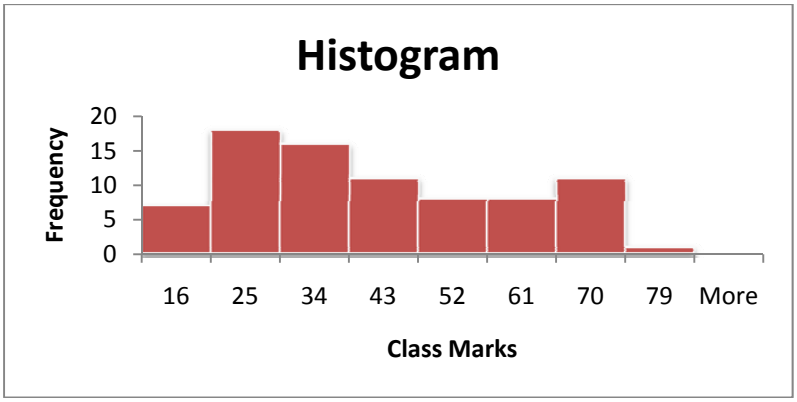
Table 2-5. The Frequency Distribution Table with 8 Class Intervals

Alternatively, one can also create a FDT with 10 class intervals, as shown in Table 2-4.

## HISTOGRAMS AND OGIVES

### The Frequency Histogram

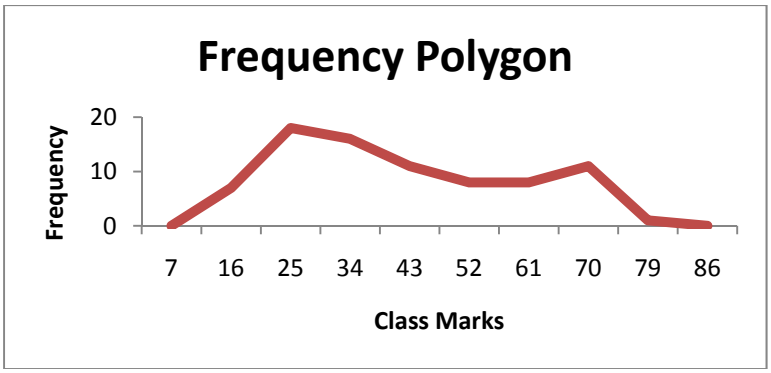
The frequency histogram is of course, a graph of frequencies. Every vertical column represents the frequency of observations in a particular class interval. It tells us about the shape of the frequency distribution.




Sometimes, instead of the actual frequencies, the magnitude of each bar represents relative frequencies. This histogram is called the *relative frequency histogram.*

### The Frequency Polygon

Instead of a vertical column chart, the frequency polygon uses a line chart in displaying frequencies. It is being done by simply creating a line graph wherein the class marks are its X values and the frequencies, the Y values. Extra class marks are added at the leftmost and rightmost part of the graph to close it (since these extra class marks have frequency = 0, hence a polygon).

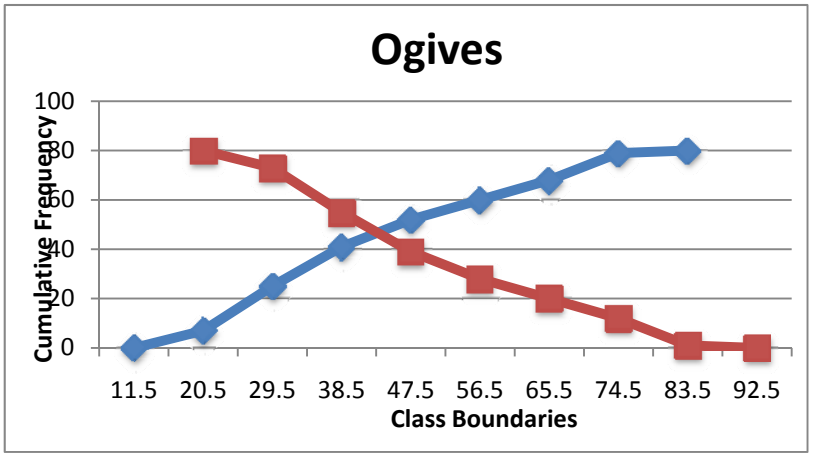



**The Ogives**

The term ogive is actually a the pointed, curved part of a ballistic like on bullets and missles. Ogives are also seen on church architecture, particularly the pointed arch windows of Gothic churches. Ogives in statistics also look very similar to that.

Ogives has two components:

1. < Ogive: <CF vs UCB → the increasing line.
2. > Ogive: >CF vs LCB → the decreasing line.




The Ogive has the same interpretation with the cumulative frequencies. The intersection of the two lines is actually where the median (the middle observation) falls.

# 3

# Measures of Central Tendency and Location

- The Summation Notation
- Mean
- Median
- Mode
- Mean Median Mode – Ungrouped Data
- Quantiles

*"I claim to be an average man of less than average ability. I have not the shadow of a doubt that any man or woman can achieve what I have, if he or she would make the same effort and cultivate the same hope and faith."*

-Mahatma Gandhi

**THE SUMMATION NOTATION AND ITS PROPERTIES**

The sum of n observations $X_i$ denoted by $\sum_{i=1}^{n} X_i$ is defined as

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + X_3 + ... + X_n$$

Wherein the summation is read as "the summation of X-sub-i where i is from 1 to n". The summation notation has three main properties. These properties can now be used for subsequent proofs.

**Property 1** : $\sum_{i=1}^{n}(X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$ The summation of a sum equals the sum of summations.

**Property 2** : $\sum_{i=1}^{n} cX_i = c\sum_{i=1}^{n} X_i$ Where c is a constant, Summation of c times a constant equals c times the summation.

**Property 3** : $\sum_{i=1}^{n} c = nc$ Where c is a constant, Summation of c times a constant equals n times c.

**MEASURES OF CENTRAL TENDENCY – UNGROUPED DATA**

When we say ungrouped data, it means that the data we are working with is not summarized, though it can be arranged into an array. In our case, ungrouped data is the raw data, while grouped data is the FDT.

Measures of central tendency are values which convey information of the centrality of the data set. The most common measures of central tendency are the mean, median and mode.

**The Arithmetic Mean**

The arithmetic mean is the most common average. Simply called as the mean, it is used far more frequently than any other means – the geometric mean and the harmonic mean. The mean is computed by adding all the values divided by the number of observations. Let us denote the (sample) mean by $\overline{X}$ and *n*, the number of observations. Then $\overline{X}$ is computed as

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

In some cases, the weights of the observations are not equal. One example is the computation of the GWA (General Weighted Average) wherein most of the subjects are worth 3 units while some, 4 units (Stat 131) or 5 units (Math 17 to Math 54). In such cases, the weighted average is used, denoted by $\overline{X}_W$.

$$\overline{X}_W = \frac{\sum_{i=1}^{n} W_i X_i}{\sum_{i=1}^{n} W_i}$$

**Properties of the Arithmetic Mean**

1. $\sum_{i=1}^{n}\left(X_i - \overline{X}\right) = 0$
2. $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ is minimum
3. Given $\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$, then $\frac{\sum_{i=1}^{n}(X_i + c)}{n} = \overline{X} + c$.
4. Given $\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}$, then $\frac{\sum_{i=1}^{n} cX_i}{n} = c\overline{X}$.

**The Median**

The median, also a measure of location is simply defined as the positional middle of an arranged data. If we denote the Median by $M_d$, and $X_{(i)}$ be the ith observation in the array. The median is calculated as follows:

| | |
|---|---|
| If n is ODD | $M_d = X_{([n+1]/2)}$ |
| If n is EVEN | $M_d = \frac{X_{(n/2)} + X_{([n/2]+1)}}{2}$ |

Unlike the mean, since the median is just a positional measure, it is only affected by the position of the values but not the magnitude of the observations.

**The Mode**

The mode, denoted by $M_o$, is the value that occurs most frequently in the data. It depends on the frequency of a values and hence unaffected by extreme observations. The mode may not exist, or if it does, it may not be unique. A data set with only one mode is unimodal, bimodal if it has two modes and trimodal if it has three.

**MEASURES OF CENTRAL TENDENCY – GROUPED DATA**

If our data is summarized in an FDT, the information on the ACTUAL value of the observations is lost. Nonetheless, approximations of such measures can still be computed in an FDT.

**Approximated Mean – Grouped Data**

The mean is approximated by the following formula:

$$\overline{X} = \frac{\sum_{i=1}^{k} f_i CM_i}{n}$$

Where $f_i$ = frequency of the ith class
$CM_i$ = class mark of the ith class
$n$ = number of observations
$k$ = number of classes

**Approximated Median – Grouped Data**

The median is approximated by the following formula:

$$M_d = LCB_{md} + c\left(\frac{n/2 - < CF_{md-1}}{f_{md}}\right)$$

Where $LCB_{md}$ = lower class boundary of the median class
$c$ = class size of the median class
$n$ = number of observations
$< CF_{md-1}$ = less than cumulative frequency of the class preceding the median class
$f_{md}$ = frequency of the median class

The median class is the class wherein its *<CF* is greater than or equal to *n/2* for the first time.

**Approximated Mode – Grouped Data**

The mode is approximated by the following formula:

$$M_o = LCB_{mo} + c\left(\frac{f_{mo} - f_1}{2f_{mo} - f_1 - f_2}\right)$$

Where $LCB_{mo}$ = lower class boundary of the modal class
$c$ = class size of the modal class
$f_{mo}$ = frequency of the modal class
$f_1$ = frequency of the class preceding the modal class
$f_2$ = frequency of the class following the modal class

**MEASURES OF LOCATION**

Also called as fractiles or quantiles, these are values below which a specified fraction or percentage of the observations must fall. Special forms of quantiles are the percentiles, deciles, quartiles and the median.

**Percentiles**

Percentiles are values that divide a set of observations in an array of 100 equal parts. For example, the first percentile, $p_1$ is the value below which 1% of the values fall. Similarly, $p_{95}$, or the 95$^{th}$ percentile is the value below which 95% of the values fall. The percentile is computed as follows:

$P_i$ = the value of the $\left[\frac{i(n+1)}{100}\right]th$ observation in the array.

In cases wherein the number of observations is less than 100 or the positioning value is not a whole number, one can round it off so that the positioning value becomes a whole number.

**Deciles, Quartiles and the Median**

Deciles, Quartiles and the Median can be expressed in terms of the percentile. For example, the deciles are measures of location which divide the array of observations into ten equal parts. Hence the ith decile, denoted by $D_j$ may be defined as the measure wherein jx10% of the observations fall below it. For example, the $4^{th}$ decile, or $D_4$ is the values wherein 40% of the observation falls below it.

In computing for the deciles and quartiles, the best way is to convert them first into its corresponding percentile value and then compute for the percentile. The table below presents the relationship between the special types of quantiles:

| Percentile | Decile | Quartile | Median |
|---|---|---|---|
| 10 | 1 | | |
| 20 | 2 | | |
| 25 | | 1 | |
| 30 | 3 | | |
| 40 | 4 | | |
| 50 | 5 | 2 | Md |
| 60 | 6 | | |
| 70 | 7 | | |
| 75 | | 3 | |
| 80 | 8 | | |
| 90 | 9 | | |
| 100 | 10 | | |

Table 3-1 Relationships of Special Types of Quantiles

For example, if one wants to find the $3^{rd}$ quartile, one has to compute for the $75^{th}$ percentile using the formula.

**PRACTICE EXERCISES – UNGROUPED DATA**

The table below presents 20 sets of randomly generated numbers. Using a calculator, try to practice getting the mean median and mode. You can also compute for some measures of dispersion.

| | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| DATA 1 | 84 | 91 | 22 | 40 | 78 | 40 | 51 | 71 | 16 | 56 | 23 | | | | | | |
| DATA 2 | 32 | 86 | 96 | 39 | 20 | 29 | 56 | 91 | 64 | | | | | | | | |
| DATA 3 | 92 | 86 | 85 | 73 | 25 | 95 | 61 | 36 | 17 | 97 | 45 | 34 | 98 | | | | |
| DATA 4 | 17 | 15 | 88 | 61 | 68 | 22 | 56 | | | | | | | | | | |
| DATA 5 | 56 | 55 | 22 | 50 | 11 | 15 | 44 | 94 | 75 | 85 | 58 | | | | | | |
| DATA 6 | 25 | 12 | 14 | 63 | 97 | 27 | 97 | 80 | 97 | 78 | 51 | 40 | 100 | | | | |
| DATA 7 | 19 | 60 | 25 | 23 | 14 | 88 | 19 | 47 | 61 | 71 | 81 | 48 | 67 | | | | |
| DATA 8 | 89 | 61 | 43 | 84 | 60 | 79 | 100 | 64 | 51 | | | | | | | | |
| DATA 9 | 72 | 61 | 43 | 98 | 91 | 84 | 61 | 18 | 92 | 27 | 100 | | | | | | |
| DATA 10 | 69 | 64 | 59 | 33 | 47 | 60 | 28 | 93 | 26 | 60 | 91 | | | | | | |
| DATA 11 | 96 | 57 | 62 | 31 | 48 | 13 | 71 | 53 | 86 | 12 | 44 | 74 | 59 | 14 | | | |
| DATA 12 | 12 | 88 | 59 | 16 | 100 | 90 | 12 | 92 | 62 | | | | | | | | |
| DATA 13 | 77 | 47 | 44 | 71 | 43 | 88 | 86 | 50 | 12 | 56 | 73 | | | | | | |
| DATA 14 | 26 | 60 | 60 | 95 | 26 | 72 | 19 | | | | | | | | | | |
| DATA 15 | 89 | 38 | 63 | 71 | 91 | 20 | 63 | 50 | 59 | 85 | 54 | | | | | | |
| DATA 16 | 12 | 94 | 82 | 66 | | | | | | | | | | | | | |
| DATA 17 | 78 | 20 | 25 | 68 | 11 | 97 | 14 | 60 | 91 | 47 | 48 | 100 | 43 | 26 | 91 | | |
| DATA 18 | 40 | 39 | 73 | 61 | 50 | 63 | 23 | 84 | 98 | 33 | 31 | 16 | 71 | 95 | 71 | 66 | 68 |
| DATA 19 | 22 | 96 | 85 | 88 | 22 | 14 | 25 | 43 | 43 | | | | | | | | |
| DATA 20 | 51 | 56 | 61 | 13 | 44 | 97 | 19 | 100 | | | | | | | | | |

| MEAN | MEDIAN | MODE | RANGE | STANDARD DEVIATION | CV |
|---|---|---|---|---|---|
| **52.000000** | 51 | 40 | 75 | 26.313495 | 0.506028743 |
| **57.000000** | 56 | | 76 | 28.874729 | 0.506574201 |
| **64.923077** | 73 | | 81 | 29.956807 | 0.461420006 |
| **46.714286** | 56 | | 73 | 28.715352 | 0.61470172 |
| **51.363636** | 55 | | 83 | 27.306676 | 0.531634397 |
| **60.076923** | 63 | 97 | 88 | 33.732925 | 0.561495557 |
| **47.923077** | 48 | 19 | 74 | 25.633512 | 0.534888688 |
| **70.111111** | 64 | | 57 | 18.857654 | 0.268968122 |
| **67.909091** | 72 | 61 | 82 | 28.644212 | 0.421802326 |
| **57.272727** | 60 | 60 | 67 | 22.680789 | 0.396013773 |
| **51.428571** | 55 | | 84 | 26.592984 | 0.517085806 |
| **59.000000** | 62 | 12 | 88 | 36.823905 | 0.624133988 |
| **58.818182** | 56 | | 76 | 22.710430 | 0.386112413 |
| **51.142857** | 60 | 26 | 76 | 28.322127 | 0.553784599 |
| **62.090909** | 63 | 63 | 71 | 21.769036 | 0.350599401 |
| **63.500000** | 74 | | 82 | 36.198527 | 0.570055538 |
| **54.600000** | 48 | 91 | 89 | 31.556752 | 0.577962483 |
| **57.764706** | 63 | 71 | 82 | 24.363214 | 0.421766427 |
| **48.666667** | 43 | 22 | 82 | 32.318725 | 0.664083395 |
| **55.125000** | 53.5 | | 87 | 31.674404 | 0.574592359 |

# 4

# *Measures of Dispersion and Skewness*

- *Measures of Absolute Dispersion*
- *Measures of Relative Dispersion*
- *Measure of Skewness and Kurtosis*

*"Variability is the law of life, and as no two faces are the same, so no two bodies are alike, and no two individuals react alike and behave alike under the abnormal conditions which we know as disease."*

-William Osler

## MEASURES OF DISPERSION

Measures of dispersion are values which use to describe the degree of spread in the data set, or its variation from the center. Measures of dispersion can either be absolute or relative. The common types of measures of dispersion are the standard deviation and the range.

### The Range

Before defining the range, let us first define what we call order statistics. Given an array with n observation we define an order statistic, denoted by $X_{(i)}$ = ith observation in the array

With order statistics and given an array with n observations, we can now define $X_{(1)}$ = minimum value and $X_{(n)}$ = the maximum value. The range, is now defined as

| | |
|---|---|
| Range = $X_{(n)} - X_{(1)}$ | (Ungrouped data) |
| Range = Highest Class Limit - Lowest Class Limit | (Data in FDT) |

### The Variance and Standard Deviation

Let us denote the population variance as $\sigma^2$ and the sample variance as $s^2$. Then the variance is defined as follows:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{X})^2}{n-1}$$

Taking the positive roots, we can now define the standard deviation.

- Note that the variance may be viewed as the "average squared deviations from the mean". It is not of the same unit as the original observations (for example if the observations are expressed in kms, then the variance has a unit in $kms^2$). Hence, getting the square root (standard deviation) brings it back to the original unit of measure.
- The sample standard deviation is a BIASED estimator of the population standard deviation. To correct for the bias, the sum is divided by n-1 instead of the usual n.
- To facilitate computation, the computational formula for the variance is used:

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

Getting the square root yields the computational formula for the standard deviation.

- The standard deviation is a measure of absolute dispersion.

### The Coefficient of Variation, CV

The Coefficient of Variation, denoted by *CV* is a measure of relative dispersion used to compare the scatter of one distribution with another distribution. Instead of the standard deviation, one should use

the coefficient of variation when comparing two data sets of different units or significantly different means. The *CV* is defined by the following formula:

$$CV = \frac{\sigma}{\mu} \times 100\% \qquad\qquad CV = \frac{s}{\overline{X}} \times 100\%$$

**MEASURES OF SKEWNESS**

Skewness is the measure of asymmetry of the distribution. It tells whether the distribution of observations symmetric, skewed to the right or skewed to the left. The Pearson's First and Second Coefficients of Skewness are as follows:

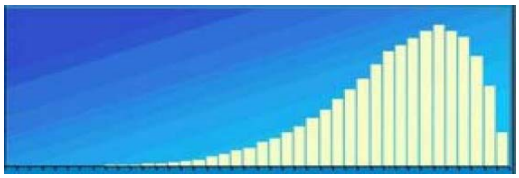
$$Skewness\ (Sk) = \frac{SampleMean\ - Mode}{stdev} \qquad Skewness\ (Sk) = \frac{3(SampleMean\ - Median\ )}{stdev}$$

Now, since the mode may not exist or may not be unique, the second formula is preferred. For appreciation's sake, the definitional formula for skewness is given as follows:

$$\frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3}{(n-1)s^3}$$

The definitional formula is tedious to compute that's why Pearson suggested the other two formulas.

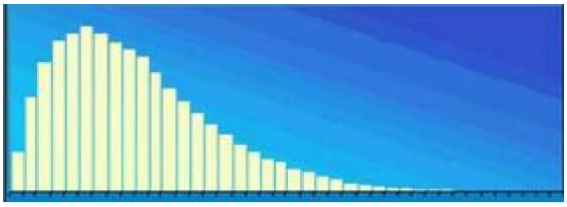
**Skewed to the Left**



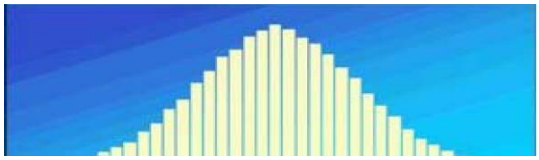
- Skewed to the LEFT
- NEGATIVELY Skewed
- Skewness < 0
- Mean < Median < Mode

**Skewed to the Right**

- Skewed to the RIGHT
- POSITIVELY Skewed
- Skewness > 0
- Mean > Median > Mode



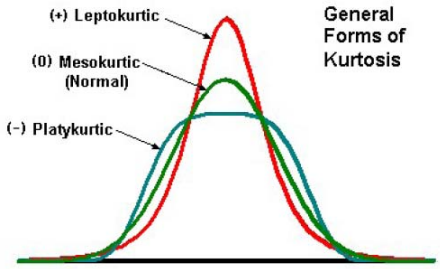
**Symmetric**



- Symmetric
- Skewness = 0
- Mean = Median = Mode

**MEASURES OF KURTOSIS**

Finally, the measure of Kurtosis is a measure of peakedness of the distribution. The definitional formula is the same as the skewness, only it is on the $4^{th}$ power instead of three.




Kurtosis > 3: Leptokurtic (Very Peaked)
Kurtosis = 3: Mesokurtic (Normal Peak)
Kurtosis < 3: Platykurtic (Flat Peaked)

# 5

# *Probability*

- *The Venn Diagram*
- *Random Experiment, Sample Spaces, Events*
- *Properties of Probability*
- *Conditional Probability*
- *Independence*
- *Counting Techniques*

*"Do not mistake coincidence for fate."*

*-Mr. Eko, Lost Season 2 Episode 9*

**VENN DIAGRAM**

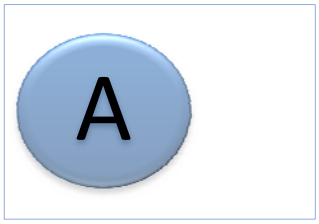
Given a finite collection of sets, the Venn diagram is a very effective tool in illustrating logical relationships among these sets. Also, since events can also be viewed as sets wherein its elements are outcomes, events can also be illustrated by Venn diagrams.

Venn diagrams were first introduced by John Venn during the 1880s.

The Sample Space, S

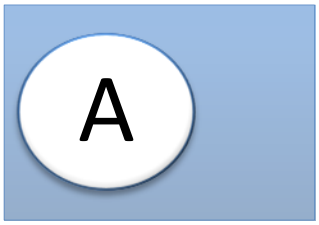
set of all outcomes

An event: A



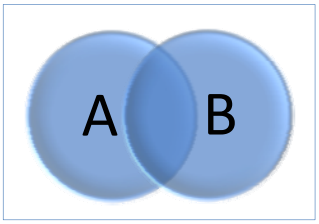

set of outcomes belonging in A

The complement of A: $A^C$



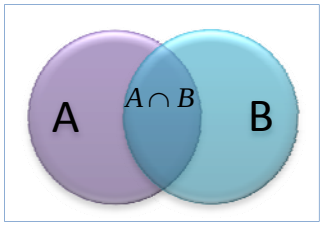

set of outcomes NOT belonging in A

Union of events A,B: $A \cup B$



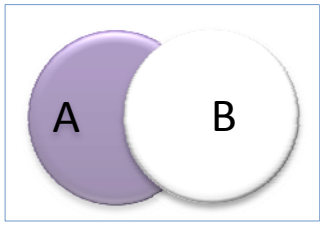

outcomes in A OR in B

Intersection of events A,B: $A \cap B$



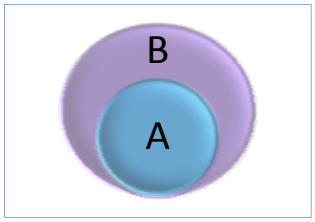

outcomes in A AND in B

Set Difference $A - B = A \cap B^C$



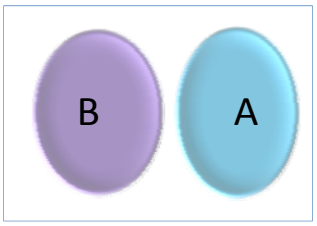

outcomes in A AND NOT in B

Subset $A \subset B$




all outcomes in A are in B

A and B are disjoint $A \cap B = \phi$




**SET OPERATIONS**

| | | |
|---|---|---|
| Idempotent Laws | $A \cup A = A$ | $A \cap A = A$ |
| Commutative Laws | $A \cup B = B \cup A$ | $A \cap B = B \cap A$ |
| Associative Laws | $(A \cup B) \cup C = A \cup (B \cup C)$ | |
| | $(A \cap B) \cap C = A \cap (B \cap C)$ | |
| Distributive Laws | $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$ | |
| | $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ | |
| Identity Laws | $A \cup \phi = A$ | $A \cap S = A$ |
| Bound Laws | $A \cup S = S$ | $A \cap \phi = \phi$ |
| Complement Laws | $A \cup A^C = S$ | $A \cap A^C = \phi$ |
| 0/1 Laws | $S^C = \phi$ | $\phi^C = S$ |
| Involution Law | $(A^C)^C = A$ | |
| De Morgan's Laws | $(A \cup B)^C = A^C \cap B^C$ | |
| | $(A \cap B)^C = A^C \cup B^C$ | |

Prove:

1. $A^C - B^C = B - A$
2. A-B and B-A are disjoint

*"Our probability of success depends on whether or not you commit them to memory"*
*-Nara Shikamaru, Naruto Shippuuden Episode 83*

**DEFINING THE SAMPLE SPACE – THE TREE DIAGRAM**

Suppose a random experiment has n stages, wherein the ith stage has $n_i$ outcomes. The following are the steps in constructing a tree diagram:

1. Start from a single point.
2. Construct the $m_1$ outcomes of the first stage, Construct the $m_2$ outcomes of the second stage…Construct the $m_n$ outcomes of the nth stage.
3. All elements of the sample space are the connected outcomes in the tree diagram.

**PROPERTIES OF PROBABILITY**

1. Probabilities of the sample and null space $P(S)=1 \qquad P(\phi)=0$
2. Given A, an event $0 \le P(A) \le 1$
3. Probability of a complement, $A^C$ $P(A)+P(A^C)=1 \Rightarrow P(A^C)=1-P(A)$
4. Probability of a union $A \cup B$ (A, B Events) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5. If A and B are disjoint $P(A \cup B) = P(A) + P(B))$
6. Probability of a difference, $A - B$ $P(A - B) = P(A) - P(AB)$
7. Probability of union of mutually exclusive events $A_1, A_2, A_3 ..., A_n$
   $P(A_1 \cup A_2 \cup A_3 \cup ... \cup A_n) = P(A_1) + P(A_2) + P(A_3) + ... + P(A_n)$
8. Probability of a union of three events, A B and C
   $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$
9. Monotonicity Property. Given A, B events: $A \subseteq B \Rightarrow P(A) \le P(B)$

**THE EVENT COMPOSITION METHOD**

1. Define the basic events.
2. List down the given probabilities of events in terms of the basic steps.
3. Write an equation expressing the event of interest.
4. Apply the appropriate property of probability to compute for the probability of the event of interest.

## Examples:

1. Suppose S is the set of all possible outcomes when a coin is tossed three times.
   a) What are all the possible outcomes of the random experiment?
   b) Let $B_i$ = event that the ith toss is a head, i = 1,2,3. Define the following in terms of $B_i$
      i. C = set of outcomes where all tosses result in heads
      ii. D = set of outcomes where only the 1st toss results in a head
      iii. E = set of outcomes where the 1st and 2nd tosses result in heads
      iv. F = set of outcomes where there is exactly one head
      v. G = set of outcomes where at least one of the tosses results in a head

2. Danny courts two girls, Anne and Bea. The chance that he will be Anne's boyfriend is 0.95; by Bea, 0.90; and by both girls, 0.88. Find the probability

a) that Danny will be the boyfriend of at least one of the girls.
b) that Danny will be dumped by both Anne and Bea.
c) that Danny will be the boyfriend of Anne alone.

**CONDITIONAL PROBABILITY**

Given A and B are events, the conditional probability of event A given event B, denoted by P(A|B) is defined by the following formula:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} \quad \text{if P(B) >0}$$

**SOME PROPERTIES OF THE CONDITIONAL PROBABILITY**

1. $P(\phi \mid B) = 0$
2. Conditional probability of the union of mutually exclusive events $A_1, A_2, A_3 \ldots, A_n$

$$P(A_1 \cup A_2 \cup A_3 \cup \ldots \cup A_n \mid B) = P(A_1 \mid B) + P(A_2 \mid B) + P(A_3 \mid B) + \ldots + P(A_n \mid B)$$

3. Conditional probability of a complement, $A^C$ $\quad P(A^C \mid B) = 1 - P(A \mid B)$
4. Given two events $A_1$ and $A_2$ $\quad P(A_1 \mid B) = P(A_1 A_2 \mid B) + P(A_1 A_2 \mid B)$
5. Conditional probability of a union $A_1 \cup A_2$

$$P(A_1 \cup A_2 \mid B) = P(A_1 \mid B) + P(A_2 \mid B) - P(A_1 A_2 \mid B)$$

6. Monotonicity. Given two events $A_1$ and $A_2$ $\quad A_1 \subseteq A_2 \Rightarrow P(A_1 \mid B) \leq P(A_2 \mid B)$
7. Theorem of Total Probabilities

$$P(A) = P(A \mid B)P(B) + P(A \mid B^c)P(B^c)$$

**INDEPENDENCE**

Two events A and be are said to be independent if any one of the following (equivalent) conditions is satisfied:

i. P(A|B) = P(A) P(B) > 0
ii. P(B|A) = P(B) P(A) > 0
iii. P(AB) = P(A)P(B)

The first two conditions of the definition show that it is consistent with the layman's notion of independence. That is, A and B are independent if the occurrence of event B has no effect on the probability of event A.

## The Medical Diagnosis Problem

Assume that a test to detect a disease whose prevalence is (1/1000) has a false positive rate of 5% and a true positive rate of 100%. What is the probability that a person found to have a positive result actually has the disease, assuming that you know nothing about the person's symptoms.

Define events D = the event that the person has the disease
T = the event that the test result is positive

*"But I was going to be a teacher my entire life, so I wasn't counting on money too much."*

- Clay Aiken

**FUNDAMENTAL COUNTING THEOREM**

Let's say we conduct a random experiment which is tossing a coin twice and a die once. We obviously know that the possible outcomes of a coin toss is a head (H) or a tail (T) and for a die toss, the number of dots (1, 2, 3, 4, 5, 6). Now, the next question is, how many possible outcomes are there in the random experiment? That is, what is the cardinality of our sample space, n(S)?

To answer this, one can list all the possible outcomes in the sample space and then count the number of elements:

S = {HH1, HH2, HH3, HH4, HH5, HH6, HT1, HT2, HT3, HT4, HT5, HT6,
TH1, TH2, TH3, TH4, TH5, TH6, TT1, TT2, TT3, TT4, TT5, TT6}

Hence, n(S) = number of elements in S = 24. This counting technique nonetheless becomes very tedious especially if we are going to consider a large sample size. Intuitively, one can do an alternate way:

n(S) = (# outcomes 1$^{st}$ Coin Toss) x (# outcomes 2$^{nd}$ Coin Toss) x (# outcomes die toss)
= 2 x 2 x 6 = **24**

Our random experiment has three tasks; each task can be done in 2, 2 and 6 ways respectively. The above technique can be generalized by the **Fundamental Counting Theorem**:

> If a job can be accomplished in *k* separate tasks, wherein the ith task can be done in $n_i$ ways, then the entire job can be done in $n_1 \times n_2 \times \ldots \times n_k$ ways

The FCT is a good foundation to start with. But what if there are more restrictions? For example, randomly choosing 2 cards from a box with cards labeled 1 to 6 yields 6 x 5 = 30 possible outcomes. But when the card we have chosen is being replaced by a card of the same label (i.e., we are allowed to chose the same number twice), the number of possible outcomes will become 6 x 6 = 36. Another example is that in lotto, ordering is unimportant whereas in sweepstakes, ordering is important.

Basically, in counting, we need to consider two restrictions. In solving counting problems, determining which rules apply is very important:

i. Is counting done *with* or *without* replacement?
ii. Is *ordering* important or not?

But before continuing, let us first define factorials, permutations and combinations.

**FACTORIALS**

Let n! denote the factorial operation where n is a nonnegative integer. n! is read as "n factorial". The factorial is a product of all positive integers from 1 to n. That is:

$$n! = n \times (n-1) \times (n-2) \times (n-3) \times \ldots \times 1$$

Also, we define 0! = 1.

**Factorial Examples:**

| | | | |
|---|---|---|---|
| 1! | 1 | 8! | 40,320 |
| 2! | 2 | 9! | 362,880 |
| 3! | 6 | 10! | 3,628,800 |
| 4! | 24 | 11! | 39,916,800 |
| 5! | 120 | 12! | 479,001,600 |
| 6! | 720 | 13! | 6,227,020,800 |
| 7! | 5,040 | 14! | 87,178,291,200 |

**PERMUTATION AND COMBINATION**

The permutation of n taken x at a time denoted by ${}_nP_x$ is defined as follows:

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

Combination on the other hand, denoted by ${}_nC_x$ is defined as follows:

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

It should be noted that in a permutation, ordering is taken into account, whereas in combination, ordering is unimportant. To understand more about this, let's say that in a set {A, B, C, D} of n = 4 elements, we need to form a permutation and a combination of size x = 3

How many possible permutations: ${}_4P_3 = 4!/(4-3)! = 4!/1! = 4x3x2x1/1 = 4x3x2 = 24$

How many possible combinations: ${}_4C_3 = 4!/3!(4-3)! = 4!/3!1! = 4x3x2x1/3x2x1x1 = 4$

The reason why there are more possible permutations than combinations is that, for example, the sequence (A, B, C) is distinct of (A, C, B) even though they have the same elements since in permutation, ordering is important. In combinations, the set {A, B, C} is the very same set {A, C, B}.

| PERMUTATION | COMBINATION | PERMUTATION | COMBINATION |
|---|---|---|---|
| (A,B,C) | {A,B,C} | (A,D,C) | {A,C,D} |
| (A,C,B) | | (A,C,D) | |
| (B,C,A) | | (D,C,A) | |
| (B,A,C) | | (D,A,C) | |
| (C,A,B) | | (C,A,D) | |
| (C,B,A) | | (C,D,A) | |
| (A,B,D) | {A,B,D} | (D,B,C) | {B,C,D} |
| (A,D,B) | | (D,C,B) | |
| (B,D,A) | | (B,C,D) | |
| (B,A,D) | | (B,D,C) | |
| (D,A,B) | | (C,D,B) | |
| (D,B,A) | | (C,B,D) | |

Note that for every possible combination, 6 permutations are possible.

## METHODS OF COUNTING

Now, we are ready to summarize the counting techniques, considering restrictions of order and replacement:

| | # of possible arrangements of size x from n objects | |
|---|---|---|
| | Without Replacement | With Replacement |
| Ordered | ${}_nP_x = \frac{n!}{(n-x)!}$ | $n^x$ |
| Unordered | ${}_nC_x = \frac{n!}{x!(n-x)!} = \binom{n}{x}$ | ${}_{n+x-1}C_x = \frac{(n+x-1)!}{x!(n-1)!} = \binom{n+x-1}{x}$ |

**Ordered, Without Replacement**

Suppose we have 6 cute Loco Rocos in different colors.

Y P Bk Bl R G

Now our task is to select 3 cutest Loco Rocos out of the 6, with our 1st choice as the cutest and the 3rd choice as the 3rd cutest. This task obviously is ordered, and once we have selected one Loco Roco, it will not be replaced. To illustrate:

**W/out Replacement:** Can choose  (Bk,P,Bl) but not  (R,G,G).

**Ordered:**  (G,Y,Bk) is different from  (G,Bk,Y).

Hence, the total number of permutations: ${}_6P_3 = \frac{6!}{(6-3)!} = \frac{6!}{3!} = 6 \times 5 \times 4 = 120$.

**Ordered, With Replacement**

Again, let us have the set of Loco Rocos.

Y P Bk Bl R G

As a gift for doing the first task time for us to get 4 cutest Loco Rocos! What's more, the Loco Roco that we have chosen will be replaced by another one of the same color.

This time, ordering still important since, even though you'd get 4 Loco Rocos regardless of the order of getting, you are still restricted to choose which is the cutest, the 2nd cutest and so on. Also, it is obvious that choosing is done with replacement. Hence a possible sequence would be:

 (Bk, Bl, Y, P), which is different from  (Bl, Y, P, Bk).

Another possible group would be:  (G, P, $Y_1$, $Y_2$),

or even this group:  ($R_1$, $R_2$, $R_3$, $R_4$)

Now, the total possible number of groups we can form: $6^4 = 1296$.

**Unordered, Without Replacement**

Let us modify the activity in the preceding section. This time, whatever Loco Roco is being chosen, it will not be replaced. That is:

If we choose  (G) from  , (6 choices)

our choice will be reduced to  (5 choices) and so on until we have chosen 4 Loco Rocos.

Now, the total number of possible combinations is: $\binom{6}{4} = \frac{6!}{4!(6-4)!} = \frac{6!}{4!2!} = 15$.

**Unordered, With Replacement**

This is the most difficult to compute without the formula. Let us say we adapt the activity in order, without replacement, but this time, ordering is unimportant (after all, you'd still get the 4 Loco Rocos regardless of the order of choosing).

Now, the total number of possible combinations is: $\binom{6+4-1}{4} = \binom{9}{4}\frac{9!}{4!(9-4)!} = \frac{6!}{4!5!} = 126$.

**More on Permutations**

The number of permutations of n objects at which x1 are identical, x2 are identical and so on is given by the following formula:

$$\frac{n!}{x_1!x_2!x_3!\ldots x_p!}$$

Example: How many different permutations can be made from the letters of the word MISSISSIPPI?
Solution: There are 4s, 4i, 2p and 1m. Hence the possible different permutations are:

$$\frac{11!}{4!\times 4!\times 2!\times 2!} = 34{,}650$$

The number of permutation of n objects arranged in a circle is (n-1)!

**PROBABILITY AND COUNTING RULES**

Recall the classical definition of probability:

$$P(A) = \frac{n(A)}{n(S)}$$

That is, the probability of an event A is the number of outcomes in A divided by the total number of outcomes.

**Practice Problems:**

1. A person can select eight different colors for an automobile body five different colors for the interior, and white or black sidewall tires. How many different color combinations are there for the automobile?
2. A person can select one of five different colors for brick borders, one type of six different ground coverings, and one of three different types of shrubbery. How many different types of landscape designs are there?
3. How many different types of identification cards consisting of 4 letters can be made from the first five letters of the alphabet if repetitions are allowed?
4. How many different types of identification cards consisting of 4 letters can be made from the first 5 letters of the alphabet if repetitions are not allowed?
5. A license plate consists of 2 letters and 3 digits. How many different plates can be made if repetitions are permitted? How many can be made if repetitions are not permitted?
6. How many different batting orders can a manager make with his starting team of 9 players?
7. In how many ways can a nurse select three patients from 8 patients to visit in the next hour? The order of visitation is important.
8. In how many different ways can a president, vice-president, secretary, and a treasurer be selected from a club with 15 members?
9. In how many different ways can an automobile repair shop owner select five automobiles to be repaired if there are 8 automobiles needing service? The order is important.
10. How many different signals using 6 flags can be made if 3 are red, 2 are blue, and 1 is white?
11. In how many ways can a large retail store select 3 sites on which to build a new store if it has 12 sites to choose from?
12. In how many ways can Mary select two friends to go to a movie with if she has 7 friends to choose from?
13. In how many ways can a real estate agent select 10 properties to place in an advertisement if she has 15 listings to choose from?
14. In how many ways can a committee of 3 elementary school teachers be selected from a school district which has 8 elementary schoolteachers?
15. In a box of 10 calculators, one is defective. In how many ways can four calculators be selected if the defective calculator is included in the group?
16. In a classroom, there are 10 men and 6 women. If 3 students are selected at random to give a presentation, find the probability that all 3 are women.
17. A carton contains 12 toasters, 3 of which are defective. If four toasters are sold at random, find the probability that exactly one will be defective.
18. If 100 tickets are sold for two prizes, and one person buys two tickets, find the probability that that person wins both prizes.
19. A committee of 3 people is formed from 6 nurses and 4 doctors. Find the probability that the committee contains 2 nurses and one doctor.
20. The committee members are selected at random. If 5 cards are dealt, find the probability of getting 4 of a kind.

Questions from: Probability Demystified by Allan G. Bluman. 2005, McGraw-Hill.

# 6

# Probability Distributions

- Concept of a Random Variable
- Discrete & Continuous Random Variables
- Expected Values
- Variance and Standard Deviation
- The Normal Distribution
- Other Common Distributions

*"The Random Variable is a function – it is not random and it is not a variable."*

## PROBABILISTIC MODEL AND THE RANDOM VARIABLE

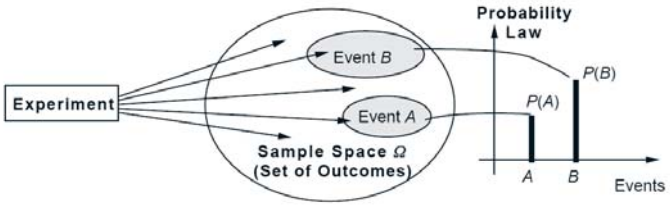



Figure 6-1. The Probabilistic Model in Terms of Events

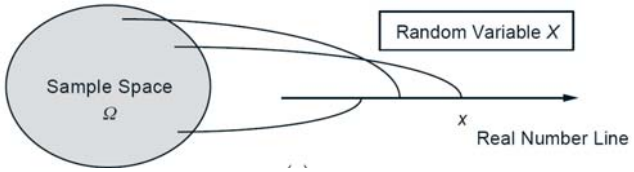



Figure 6-2 The Visualization of a Random Variable

## THE RANDOM VARIABLE

The random variable, usually denoted by X is a FUNCTION that assigns a numerical value to each possible outcome of the experiment. A random variable is denoted by a capital letter and a lower-case letter for one of its values.

## EXAMPLES OF RANDOM VARIABLES

1. In an experiment involving tossing a coin thrice, the number of heads in the sequence is considered a random variable. However the 3-long sequence of heads and tails (i.e, an outcome) is not considered a random variable because it does not have an explicit numerical value.
2. Consider an experiment of tossing a 4-sided die twice, and let the random variable X be the maximum of the two rolls. The table below shows the values of the random variables for each outcome in the sample space.

| Outcome | X | Outcome | X | Outcome | X | Outcome | X |
|---|---|---|---|---|---|---|---|
| **1,1** | 1 | 2,1 | 2 | 3,1 | 3 | 4,1 | 4 |
| **1,2** | 2 | 2,2 | 2 | 3,2 | 3 | 4,2 | 4 |
| **1,3** | 3 | 2,3 | 3 | 3,3 | 3 | 4,3 | 4 |
| **1,4** | 4 | 2,4 | 4 | 3,4 | 4 | 4,4 | 4 |

3. Given X and Y random variables, then V = X + Y and W = $X^2$ are also random variables
4. In a certain PUB loading bay along EDSA, the time needed for a bus to load and unload passengers, the number of passengers leaving the bus, and the delay with which a passenger experiences due to the loading time of the bus are all random variables.

## MAIN CONCEPTS RELATED TO RANDOM VARIABLES

1. A random variable is a real-valued function of the outcome of the experiment.
2. A function of a random variable defines another random variable.
3. We can associate with each r.v. certain statistics of interest, such as the mean and variance.
4. A random variable can be conditioned on an event or on another random variable.

5. There is a notion of independence of a random variable from an event or from another r.v.

### THE DISCRETE RANDOM VARIABLE AND DISCRETE PROBABILITY DISTRIBUTION

A discrete random variable is a real-valued function of the outcome of the experiment that can take a finite or countably infinite number of values. Each discrete random variable has an associated discrete probability distribution function (or a probability mass function, PMF), which gives the probability of each numerical value that the random variable can take.

The sum of the probabilities of **all** possible values of a discrete r.v. is 1.

### CALCULATION OF A DISCRETE PMF

1. Collect all the possible outcomes that give rise to the event {X = x}
2. Add their probabilities to obtain P(X = x)

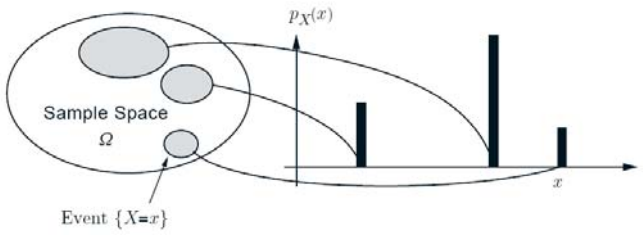



Figure 6-3. Illustration of the Method to Calculate the Discrete PMF

### THE CONTINUOUS RANDOM VARIABLE AND CONTINUOUS PROBABILITY DISTRIBUTION

A random variable defined on an infinite number of possibilities equal to the number of points on a line segment is a continuous r.v.. The probability that the value X falls within the interval, say [a,b] is the area under the curve of the continuous probability distribution function (or probability density function, PDF).

The entire area of the graph of the PDF above the X-axis must be equal to 1. And the probability of r.v. X at at a point x is zero (i.e. P(X=x)=0) since "there is no area below a point".

### CALCULATION OF A CONTINUOUS PDF

1. Find the region wherein X takes on the values in the interval [a,b]
2. Get the area of the region.

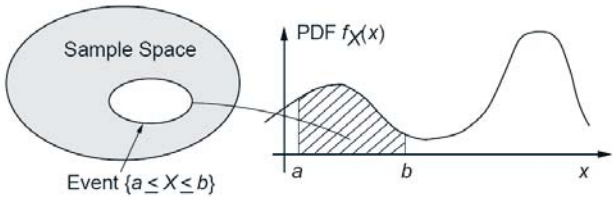



Figure 6-4. Illustration of the Method to Calculate the Continuous PDF

Remarks:

1. The probability distribution functions of discrete and continuous random variables are both valid probability functions. It can be shown that they satisfy the 3 postulates of a probability.
2. There are special types of probability distribution functions. One of them, for the discrete case is the Binomial distribution; for the continuous case, the normal (bell-shaped) distribution.

*"You have to expect things of yourself before you can do them."*

-Michael Jordan

## EXPECTATIONS

The PMF of a random variable X provides us with several possible values of X and the corresponding probabilities of all the possible values of X. It would then be desirable to **summarize** this information in a **single representative number**. Such can be accomplished by the ***expectation of X***, which is a weighted (in proportion to probabilities) average of the possible values of X.

We define the **expected value**, **expectation** or the **mean** of a random variable X and its function g(X) by

$$E[X] = \sum_{i=1}^{n} x_i f(x_i) \text{ and } E[g(X)] = \sum_{i=1}^{n} g(x_i) f(x_i)$$

When the values of the random variable X are plotted in the real line, E[X] is visually represented as the center of gravity.

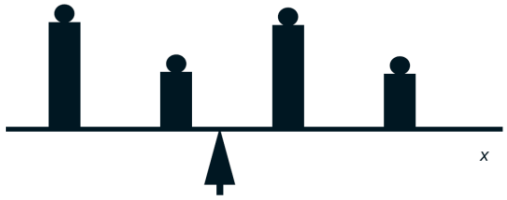



Figure 6-4 Interpretation of the Mean as the Center of Gravity

## EXAMPLES OF EXPECTATIONS

1. A player tosses a fair die. He can choose one of the two rules:
   a. If a prime number occurs he wins 10 times that number in Pesos, but if a non-prime number occurs, he loses 10 times that number in Pesos.
   b. If an even number occurs he wins 6 times that number in Pesos, but if an odd number occurs, he loses 7 times that number in Pesos.

   If he wants the game to be favorable to him, which rule should he choose?

   Solution:

   Define G = amount of money the person gains after tossing a die.

   Outcome – R.V. Mapping:

| Outcome | Prime? | G(a) | G(b) |
|---|---|---|---|
| 1 | N | -10 | -7 |
| 2 | Y | 20 | 12 |
| 3 | Y | 30 | -21 |
| 4 | N | -40 | 24 |
| 5 | Y | 50 | -35 |
| 6 | N | -60 | 36 |

   PMF Table with expectation computation:

| **G (rule a)** | -10 | 20 | 30 | -40 | 50 | -60 | |
|---|---|---|---|---|---|---|---|
| $f(g_i)$ **= P(G=g)** | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1 |
| $g_i f(g_i)$ | -10/6 | 20/6 | 30/6 | -40/6 | 50/6 | -60/6 | **E[X] = -10/6 = -1.67** |
| | | | | | | | |
| **W (rule b)** | -7 | 12 | -21 | 24 | -35 | 36 | |
| $f(g_i)$ **= P(G=g)** | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1 |
| $g_i f(g_i)$ | -7/6 | 12/6 | -21/6 | 24/6 | -35/6 | 36/6 | **E[X] = 9/6 = 1.5** |

2. Kuya Nards grills chicken isaws near Ilang-ilang Residence Hall. As agreed upon, his daily salary in Pesos, W is defined as W = 175 + 0.15X, where X = number of chicken isaws he has grilled in a day. Suppose that in any day, the PMF of X is given by the following table:

| X | 750 | 800 | 850 | 900 | 950 | 1000 |
|---|---|---|---|---|---|---|
| P(X=x) | 0.25 | 0.35 | 0.2 | 0.1 | 0.05 | 0.05 |

Find Kuya Nards' expected daily earning.

Solution:

| X | 750 | 800 | 850 | 900 | 950 | 1000 | |
|---|---|---|---|---|---|---|---|
| P(X=x) | 0.25 | 0.35 | 0.2 | 0.1 | 0.05 | 0.05 | |
| W | 287.5 | 295 | 302.5 | 310 | 317.5 | 325 | |
| wP(X=x) | 71.875 | 103.25 | 60.5 | 31 | 15.875 | 16.25 | **E[W]=298.75** |

Therefore, Kuya Nard's expected daily earning is 298.75 Pesos.

## THE VARIANCE AND THE STANDARD DEVIATION

The variance of the random variable X with mean E[X] is defined as

$$Var(X) = \sum_{i=1}^{n} (x_i - E[X])^2 f(x_i) \qquad \text{or} \qquad Var(X) = E(X^2) - [E(X)]^2$$

wherein the first equation is the definitional formula, and the second equation, the computational formula. The standard deviation is simply the square root of the variance of X.

**Example of Variance**

1. In example #2, find the variance of W using the two methods. Also, derive the standard deviation.

## PROPERTIES OF THE EXPECTATION AND VARIANCE

**Some properties of the expected value:**

b. E(c) = c for any constant c [the expected value of a constant is the constant itself]

c. E(aX + b) = aE(x) + b for constants a,b [expected value of a linear function of X]

d. E(X+Y) = E(X) + E(Y) [expectation(sum of rvs) = sum of expectations]
E(X-Y) = E(X) – E(Y)

e. If X, Y independent r.v, E(XY) = E(X)E(Y)

f. E[X-E(X)] = 0 [expected deviations of rv X from its mean is 0]

**Some properties of the variance:**

a. Var(c) = 0 for any constant c [variance of a constant is 0]

b. $Var(aX+b) = a^2 Var(X)$ [variance of a linear function of X]

c. If X, Y are independent then
Var(X+Y) = Var(X) + Var(Y)
Var(X-Y) = Var(X) + Var(Y)

*"Whenever I look through my notes on probability distributions, my eyes will look for her.*
*For I am always fascinated by the normal distribution –*
*the elegance she has formed from irrationalities,*
*the majestic curve that flows from a point in eternity towards the other unfathomable infinity,*
*the power emanating from her that challenges the impossible.*
*She has given me the power that love could only give: now I can defy probabilities."*

## HISTORY OF NORMAL DISTRIBUTION

Consider the question "If a fair coin is tossed 100 times, what is the probability of getting at least 60 heads?" To compute this, we need to compute the probability of getting 60 heads, the probability of getting 61 heads until the probability of getting all 100 heads, and then add up all these probabilities! 0_o

Abraham de Moivre, an 18th century statistician and consultant to gamblers was often called upon to make these lengthy computations. De Moivre noted that when the number of events (coin flips) increased, the shape of the binomial distribution approached a very smooth curve. Binomial distributions for 2, 4, and 12 flips are shown below.

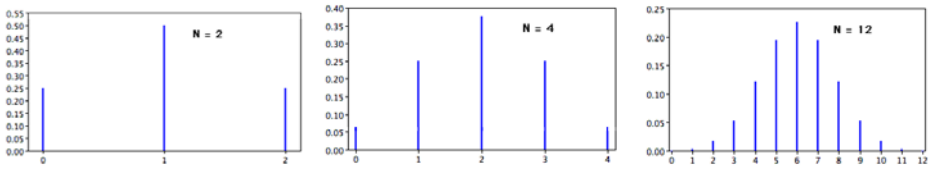



Figure 6-5 Graphs of PMFs of X~Bi(n,1/2) at n = 2,4 and 12

De Moivre reasoned that if he could find a mathematical expression for this curve, he would be able to solve problems such as finding the probability of 60 or more heads out of 100 coin flips much more easily. This is exactly what he did, and the curve he discovered is now called the normal curve.

The importance of the normal curve stems primarily from the fact that the distribution of many natural phenomena is at least approximately normally distributed. Independently the mathematicians Adrian in 1808 and Gauss in 1809 developed the formula for the normal distribution.

## THE NORMAL DISTRIBUTION

A random variable X is defined to follow a Normal Distribution, denoted by $X \sim N(\mu, \sigma^2)$ if its Probability Density Function is as follows:

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

Where: $E(X) = \mu$ and $Var(X) = \sigma^2$

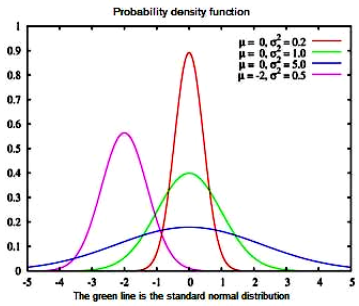



Figure 6-6 Graphs of the Normal Distribution at Different Values of E(X) and Var(X)

**PROPERTIES OF THE NORMAL DISTRIBUTION**

- The normal distribution function includes 3 famous irrational numbers: square root of two, pi and *e*.
- The normal PDF yields a bell-shaped curve symmetric with respect to a vertical axis formed through the mean.
- In both directions, the bell curve approaches the horizontal axis asymptotically.
- When X follows a normal distribution P( a≤X≤b) ≠ 0 for any real number, a ≠ b.
- The area under the curve and above the horizontal axis is equal to one.
- The mean μ defines the location of the curve while the variance $\sigma^2$ defines the shape of the bell curve.

**THE EMPIRICAL RULE (68-95-99 RULE)**

The empirical rule states that in a normally distributed population, 68.27% of the values are within 1 standard deviation of the mean, 95.45% of the values are within 2 standard deviations and 99.73% lie within 3 standard deviations.

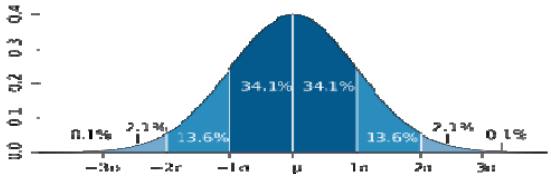



Figure 6-7. The Normal Curve showing the Empirical Rule

**THE STANDARD NORMAL DISTRIBUTION**

Given $X \sim N(\mu,\sigma^2)$. The normal r.v X can be transformed into a standard normal r.v. through the following transformation:

$$Z = \frac{X - \mu}{\sigma}$$

The standard normal r.v. Z is indeed a normal random variable with mean 0 and variance 1, that is $Z \sim N(0,1)$. Using the standard normal random variable, one can easily compute for the following probabilities by using the standard normal table:

| | | |
|---|---|---|
| $P(x_1<X<x_2) = P(z_1<Z<z_2)$ | Table value: | $f(z_2) - f(z_1)$ |
| $P(X<x_1) = P(Z<z_1)$ | Table value: | $f(z_1)$ |
| $P(X>x_2) = P(Z>z_2)$ | Table value: | $1- f(z_2)$ |

**OTHER TYPES OF PARAMETRIC CONTINUOUS PROBABILITY DISTRIBUTION FUNCTIONS**

THE UNIFORM DISTRIBUTION / RECTANGULAR DISTRIBUTION

$$X \sim U(a_1, a_2) \qquad f(x) = 1 \Big/ a_1 - a_2$$

THE EXPONENTIAL DISTRIBUTION

In the exponential distribution, X = length of time between successive happenings. The exponential distribution has been used as a model for lifetimes of various things.

$$X \sim Exp(\lambda) \qquad f(x) = \lambda e^{-\lambda x}$$

THE GAMMA DISTRIBUTION

Instead of the length of time between successive happenings, we are now interested in the random variable X = length of time until the rth occurrence of an event.

$$X \sim Gamma(r, \lambda) \qquad f(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$$

*"The most important questions of life are indeed, for the most part, really only problems of probability."*
-Pierre Simon Laplace, Théorie Analytique des Probabilités, 1812

## PARAMETERS

Parameters are constants that determine the specific form of a distribution function. It is the constant in the function / equation of a curve that can be varied to yield a family of similar curves. For example, an equation of a line y = mx + b has two parameters, m (slope) and b (intercept). Parameters play an important role in defining **special distribution functions**.

## THE UNIFORM DISCRETE DISTRIBUTION

The Uniform Discrete Distribution is a PMF that can be characterized by saying that all values of a finite set of possible values are equally probable. A random variable X is defined to have a discrete uniform distribution if its PMF is given by

$$P(X = x) = f(x) = 1/N$$ (f(x) form) and

| X | 1 | 2 | … | N |
|---|---|---|---|---|
| P(X=x) | 1/N | 1/N | … | 1/N |

(tabular form)

If X follows a Uniform Discrete Distribution, we write X~U(N), where N is the number of possible values. Also E[X] = (N+1)/2 and Var(X) = $(N^2-1)/12$

**Example**: In the movie Armageddon, 6 people drew lots to determine who is going to stay behind to set off the nuclear bomb. What is the probability that AJ (Ben Affleck) stays behind?
**Solution**: Let X = code of the person who will be left behind
(1-AJ, 2-Harry Stamper (Bruce Willis),…,6-$6^{th}$ person)
Now, X~U(6). Hence, P(AJ stays) = P(X=1) = 1/6

## THE BERNOULLI DISTRIBUTION

A Bernoulli trial is a random experiment whose outcomes can either be a "success" or a "failure". The random variable X is said to be a Bernoulli random variable if it takes only two values – 1 for a "success" and 0 for a "failure" and the probability of a "success" is given by *p* (i.e., P(X=1) = p). The PMF of X is given by

$$P(X = x) = f(x) = p^x(1-p)^{1-x}$$ and

| X | 1 | 0 |
|---|---|---|
| P(X=x) | p | 1-p |

If X follows a Bernoulli Distribution, we write X~Be(p). Also E[X] = p and Var(X) = p(1-p).

## EXAMPLES OF BERNOULLI TRIALS

For all its simplicity, the Bernoulli random variable is very important. In practice, it is used to model generic probabilistic situations with just two outcomes:

1. The state of a telephone at a given time can either be free or busy.
2. A person can either be healthy or sick with a certain disease.
3. Suppose that one Pinoy TV viewer can be classified ONLY as *Kapamilya* or *Kapuso* such that he/she is twice as likely to be *Kapuso* as *Kapamilya*. What is the probability that the TV viewer is a *Kapuso*? Solution: Consider the table:

| outcome | X | P(X=x) |
|---|---|---|
| ***Kapuso*** | 1 | 2/3 |
| ***Kapamilya*** | 0 | 1/3 |

Now X~Be(2/3). Hence P(person is *Kapuso*) = P(X=1) = 2/3.

**THE BINOMIAL DISTRIBUTION**

A binomial experiment possesses these four properties

1. The experiment consists of n repeated trials.
2. Each trial results in one of two outcomes: success or failure.
3. The probability of success for each trial is *p* and remains the same from trial to trial while the probability of failure is denoted by *q=1-p.*
4. The trials are independent.

The Binomial r.v. X is number the number of successes observed during the n trials. The PMF of X is given by

$$P(X = x) = f(x) = \left({}_nC_x\right)p^x(1-p)^{n-x}$$

- If X follows a Binomial Distribution, we write X~Bin(n,p). Also E[X] = np and Var(X) = np(1-p).
- The Binomial Distribution reduces to the Bernoulli Distribution when n = 1
- The Binomial Distribution is symmetric when p = 0.5

**EXAMPLES INVOLVING THE BINOMIAL DISTRIBUTION**

1. If we toss a biased coin five times, wherein it is twice as likely to come up heads as tails, what is the probability of the following:
   a. Observing two heads out of the five tosses
   b. Observing at least 2 heads
   c. Observing at most 2 heads
2. If a couple plans to have four children, which is more likely, having two boys and two girls or having three boys and 1 girl? Assume that the probability of having a boy is 0.5

**OTHER TYPES OF PARAMETRIC DISCRETE PROBABILITY DISTRIBUTION FUNCTIONS**

**HYPERGEOMETRIC DISTRIBUTION**

A hypergeometric experiment consists of selecting a sample size n using random sampling without replacement form a population of N elements, k of which can be classified as a "success" and N-k as "failure". The random variable X = the number of "successes" out of the n selected sample follows a

Hypergeometric Distribution with parameters N,k and n.

If X~Hyp(n,N,k) then $P(X = x) = f(x) = \binom{k}{x}\binom{N-k}{n-x}\Big/\binom{N}{n}$

E[X] = nk/N and $Var(X) = \left(\frac{N-n}{N-1}\right)\frac{nk}{N}\left(1-\frac{k}{N}\right)$

**POISSON DISTRIBUTION**

The Poisson Distribution provides a realistic model for many random phenomena. It is a probability distribution that expresses the probability of the random variable X = the number of occurrences in (i) a fixed period of time wherein the (ii) the average rate of occurrence, $\mu$ is known and (iii) the occurrences are independent with each other.

If X~Poi( $\mu$ ) then $P(X = x) = f(x) = \frac{e^{-\mu}\mu^x}{x!}$

$E(X) = \mu$ and $Var(X) = \mu$

**GEOMETRIC DISTRIBUTION**

The geometric experiment is similar to the binomial experiment. The only difference is that in the binomial experiment, the number of Bernoulli trials is fixed: n. whereas in the geometric experiment, it will only be concluded only after the first success is observed. It is a probability distribution which expresses the probability of the random variable X = the number of failures before the first success, wherein the probability of success is *p*.

If X~Geo(p) then $P(X = x) = f(x) = p(1-p)^x$

E(X) = (1-p)/p and Var(X) = $(1-p)/p^2$

**NEGATIVE BINOMIAL DISTRIBUTION**

The Negative Binomial Distribution is a generalization of the geometric distribution. Instead of waiting only for the first success, one may wish to observe the rth success. Hence it is a probability distribution which expresses the probability of the random variable X = the number of failures before the rth success.

If X~NB(r,p) then $P(X = x) = f(x) = \binom{r+x-1}{x} p^r (1-p)^x$

E(X) = r(1-p)/p and Var(X) = $r(1-p)/p^2$

# 7

# Sampling Distribution

- Sampling Variability
- Sampling Distribution
- Sampling Distribution of a Sample Mean
- The Central Limit Theorem
- The t-distribution

*"It is normal to give away a little of one's life in order not to lose it all."*

-Albert Camus

## SAMPLING VARIABILITY AND SAMPLING DISTRIBUTION

Consider a small population consisting of 20 employees in a certain company. The amount of money (in Pesos) each employee spent on a certain day is shown in the following table:

| Employee | Amount Spent | Employee | Amount Spent | Employee | Amount Spent |
|---|---|---|---|---|---|
| 1 | 362 | 8 | 263 | 15 | 325 |
| 2 | 353 | 9 | 355 | 16 | 377 |
| 3 | 439 | 10 | 496 | 17 | 408 |
| 4 | 356 | 11 | 296 | 18 | 427 |
| 5 | 453 | 12 | 315 | 19 | 380 |
| 6 | 332 | 13 | 357 | 20 | 444 |
| 7 | 418 | 14 | 336 | | |

Table 7-1. Amount spent by a company of 20 employees in a certain day.

For this population, the mean μ = 374.6 (Verify). Suppose that we do not know the true value of the parameter (as with the case of large populations), hence we just want to estimate it by computing the sample mean $\bar{x}$. That is, we get a sample of size 5 and we name it as sample 1.

| | | | | | |
|---|---|---|---|---|---|
| Sample 1: | 3 | 16 | 6 | 17 | 5 |
| X (amount spent): | 439 | 377 | 332 | 496 | 453 |

The sample mean for the 1st sample is 401.8, which is considerably greater than 374.6. Is this difference typical, or is this particular sample mean just unusually far away from the population mean μ? Let us furthermore investigate by getting 49 more samples, and for each sample we get their respective sample means:

| Sample | Sample Mean | Sample | Sample Mean | Sample | Sample Mean |
|---|---|---|---|---|---|
| **1** | 401.8 | **18** | 351.2 | **35** | 377.4 |
| **2** | 373.6 | **19** | 393.8 | **36** | 353.4 |
| **3** | 364.8 | **20** | 430 | **37** | 364 |
| **4** | 370.4 | **21** | 365.8 | **38** | 357 |
| **5** | 394.4 | **22** | 367.2 | **39** | 354 |
| **6** | 371.4 | **23** | 353.8 | **40** | 351.2 |
| **7** | 368.8 | **24** | 364.2 | **41** | 353.4 |
| **8** | 405.4 | **25** | 395.8 | **42** | 355.8 |
| **9** | 360.6 | **26** | 380.2 | **43** | 400.8 |
| **10** | 423.6 | **27** | 380 | **44** | 382 |
| **11** | 385.4 | **28** | 389.2 | **45** | 348.4 |
| **12** | 391 | **29** | 382.2 | **46** | 351 |
| **13** | 345.8 | **30** | 381.4 | **47** | 384.6 |
| **14** | 394.4 | **31** | 372.4 | **48** | 365.6 |
| **15** | 395.6 | **32** | 359.2 | **49** | 368 |
| **16** | 412.4 | **33** | 411.6 | **50** | 362.6 |
| **17** | 377.6 | **34** | 385.6 | | |

Table 7-2. Table of means of 50 possible samples.

Taking note that μ = 374.6, we remark the following:

- The value $\bar{x}$ varies from one sample to another (sampling variability).

- Some values of the sample mean $\bar{x}$ are larger than that of μ while some, smaller.
- Some values of the sample mean $\bar{x}$ are close than that of μ while some, considerably far.

But when we get the mean of all the 50 sample means, it is close to the parameter value: 376.6. In fact, if get all the possible samples of size 5 from the population (not just 50), the average of the sample means of all possible samples is the value of the population mean.

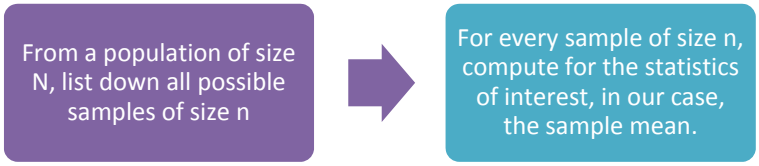
The sample mean is actually a function of the random variable X = amount spent by an employee on a particular day. A chain of implications is then formed:

- The sample mean which is a statistic, is a function of a random variable. Which implies that
- The sample mean being a function of a random variable is also a random variable. Which implies that
- The sample mean being a random variable also has a probability distribution.

Lastly, the probability distribution of a statistic has a special name: the sampling distribution of the sample mean. In general, any statistic has a corresponding sampling distribution.

**SAMPLING DISTRIBUTION OF THE SAMPLE MEAN**

Any statistic has a sampling distribution. But for this course, we will only discuss about the sampling distribution of $\bar{x}$. To get the sampling distribution of the sample mean, we perform two simple steps:




Process 7-1. Steps on How to Construct the Sampling Distribution of a Sample Mean.

For illustration purposes, let us consider a very small population of 5 individuals (N = 5) and we want to get a sample of size 2 (n=2). Suppose that X = age of the individual. The following are their corresponding ages.

| Individual: | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| Age: | 17 | 26 | 10 | 32 | 16 |

Population Mean E(X): 20.2 Population Variance Var(X): 60.96

**Scenario 1: Sampling with Replacement**

| Pair | Age | $\bar{X}$ | Pair | Age | $\bar{X}$ | Pair | Age | $\bar{X}$ |
|---|---|---|---|---|---|---|---|---|
| (1,1) | (17,17) | 17 | (3,1) | (10,17) | 13.5 | (5,1) | (16,17) | 16.5 |
| (1,2) | (17,26) | 21.5 | (3,2) | (10,26) | 18 | (5,2) | (16,26) | 21 |
| (1,3) | (17,10) | 13.5 | (3,3) | (10,10) | 10 | (5,3) | (16,10) | 13 |
| (1,4) | (17,32) | 24.5 | (3,4) | (10,32) | 21 | (5,4) | (16,32) | 24 |
| (1,5) | (17,16) | 16.5 | (3,5) | (10,16) | 13 | (5,5) | (16,16) | 16 |
| (2,1) | (26,17) | 21.5 | (4,1) | (32,17) | 24.5 | | | |
| (2,2) | (26,26) | 26 | (4,2) | (32,26) | 29 | | | |
| (2,3) | (26,10) | 18 | (4,3) | (32,10) | 21 | | | |
| (2,4) | (26,32) | 29 | (4,4) | (32,32) | 32 | | | |
| (2,5) | (26,16) | 21 | (4,5) | (32,16) | 24 | | | |

E($\bar{X}$) = 20.2 Var($\bar{X}$) = 30.48

**Scenario 2: Sampling without Replacement**

| Pair | Age | $\bar{X}$ | Pair | Age | $\bar{X}$ | Pair | Age | $\bar{X}$ |
|---|---|---|---|---|---|---|---|---|
| (1,2) | (17,26) | 21.5 | (3,1) | (10,17) | 13.5 | (5,1) | (16,17) | 16.5 |
| (1,3) | (17,10) | 13.5 | (3,2) | (10,26) | 18 | (5,2) | (16,26) | 21 |
| (1,4) | (17,32) | 24.5 | (3,4) | (10,32) | 21 | (5,3) | (16,10) | 13 |
| (1,5) | (17,16) | 16.5 | (3,5) | (10,16) | 13 | (5,4) | (16,32) | 24 |
| (2,1) | (26,17) | 21.5 | (4,1) | (32,17) | 24.5 | | | |
| (2,3) | (26,10) | 18 | (4,2) | (32,26) | 29 | | | |
| (2,4) | (26,32) | 29 | (4,3) | (32,10) | 21 | | | |
| (2,5) | (26,16) | 21 | (4,5) | (32,16) | 24 | | | |

E($\bar{x}$) = 20.2 Var($\bar{x}$) = 22.86

Based on the above information, we can now 3 properties for the sampling distribution of $\bar{x}$. A fourth and fifth rule is added to pave the way for the next section which is the very important Central Limit Theorem.

Property 1: E($\bar{X}$) = E(X) = Population Mean = μ.

Property 2: If sampling is done with replacement, $Var(\bar{X}) = \frac{Var(X)}{n} = \sigma^2/_n$ .

Property 3: If sampling is done without replacement, $Var(\bar{X}) = \frac{Var(X)}{n} \times \frac{N-n}{N-1} = \frac{\sigma^2}{n} \times \frac{N-n}{N-1}$.

Property 4: If the population is normally distributed (X follows a normal distribution), then the sampling distribution of $\bar{X}$ is also normal.

Property 5: Even if the population is not normally distributed, the sampling distribution of $\bar{X}$ can well be approximated by a normal curve, provided the sample size n is SUFFICIENTLY LARGE.

Property 1 states that the sampling distribution of $\bar{X}$ is always centered at the population mean. The second property implies that as n increases, the variance of the sample mean decreases as n increases. Also, it gives a proportional relationship between the variance of the sampling distribution and the variance of the population. This is the same with property 3, only that a finite population correction factor is multiplied. Properties 4 and 5 can be summarized as:

- Normal Sampling Distribution when population is normally distributed
- Approximately Normal Sampling Distribution when n is sufficiently large

Property 5 is known as the Central Limit Theorem. For clarity, the CLT will be discussed in the next section. As a conservative choice, the CLT can be safely applied if n is at least 30.

As an exercise:

1. Verify properties 2 and 3 in the example above.
2. Given X has an unknown probability distribution and E(X)=24, Var(X)=5.76, a sample of size n=37 is sampled from a population of size 100. Answer the following:
    a. Compute for the mean and variance of the sampling distribution of $\bar{X}$. Assume sampling is done with replacement.
    b. Compute for the mean and variance of the sampling distribution of $\bar{X}$. Assume sampling is done without replacement.
    c. Is the sampling distribution approximately normal? Why?
3. Express properties 2 and 3 in terms of stdev($\bar{X}$).

## THE CENTRAL LIMIT THEOREM

The Central Limit Theorem states that when the sample size n is sufficiently large, the sampling distribution of $\bar{X}$ is approximately normal no matter what the population distribution looks like.

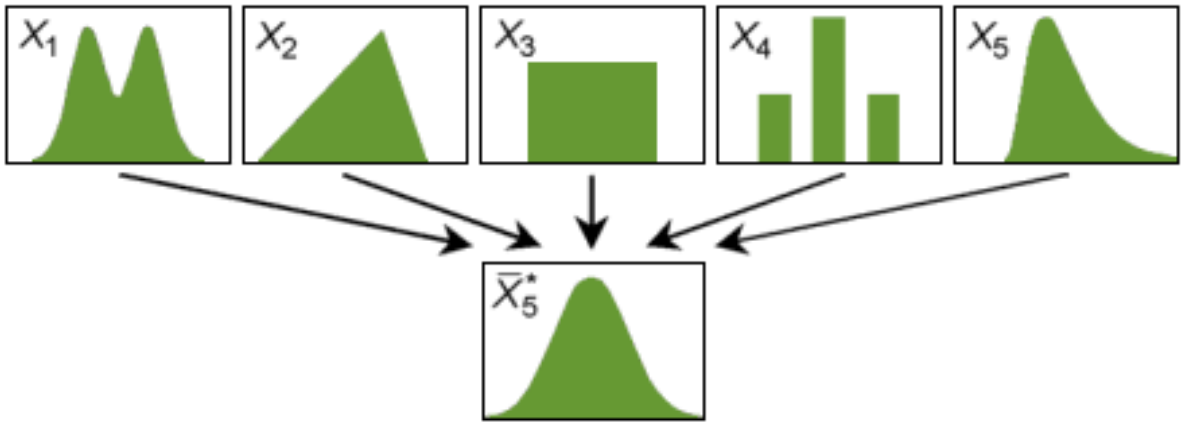



Figure 7-1. Visual Presentation of the Central Limit Theorem

As with the figure above, consider 5 random variables which have different shapes of probability distributions. When n is sufficiently large, that is, n is at least 30, the sampling distribution of the sample mean (the probability distribution of $\bar{X}$ is approximately normally distributed.

**The Standard Normal Random Variable Revisited**

Property 4 of the previous section and the CLT would then lead us to standardizing $\bar{X}$ which has 2 forms a and b:

a. Given that X follows a normal distribution, that is, the population is normally distributed
b. Given that n is at least 30

The standardized random variable Z defined as

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

a. Follows a standard normal distribution.
b. Approximately follows a standard normal distribution.

If σ is not known but n is sufficiently large, then the above still holds but this time, the sample variance is used instead of the population variance:

$$Z = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

As an exercise:

1. Given that X = the amount soda a softdrink machine dispenses in a can is normally distributed with mean 200oz and variance 0.16. There are 16 cans sampled. Find the probability that the sample mean soda volume is between 11.96 and 12.08.
2. Given the same experiment in #1 but this time the population variance is unknown. Nonetheless, s = 0.154. Would you use the formula above? If yes, compute the probability asked in #1. If no, why?

**THE T-DISTRIBUTION**

The t-distribution, like the standard normal distribution is also a bell shaped curve centered at 0. However, the t-distribution is more dispersed than the standard normal distribution. The parameter of the t-distribution is its degrees of freedom (df = n-1). Simply put, the degrees of freedom is the number of values in the calculation of a statistic which are free to vary.

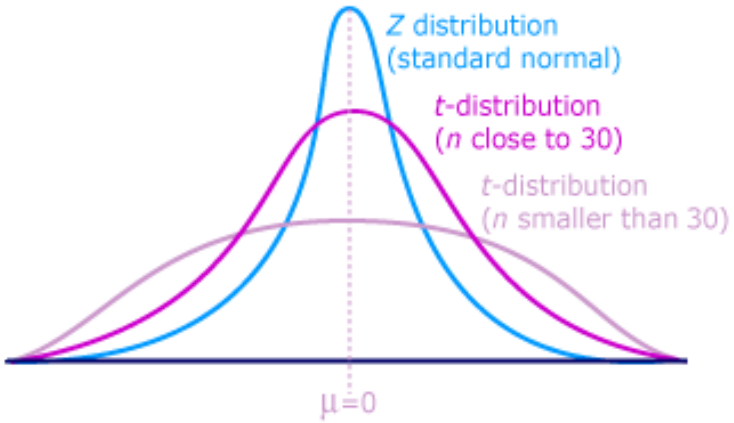



Figure 7-1. The Standard Normal and t-Distributions

To summarize, here are the key points in the t-distribution:

- Like the Z distribution, it is centered at zero.
- More dispersed than the standard normal distribution.
- The parameter is the degrees of freedom, df which is equal to n-1.
- As degrees of freedom increases, the t-distribution approaches the standard normal distribution.
- When n exceeds 30, then the t-distribution can be well approximated by the normal distribution.

The t-distribution is good in statistical inferences when the sample size is small and only the sample variance is given. The standardized random variable t defined as

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

is t-distributed with n-1 degrees of freedom. Of course this holds for any value of n (even large), but again, the t-distribution can be well approximated by the standard normal distribution when n exceeds 30.

# Statistical Inference

- *Estimation*
- *Hypothesis Testing*

# 8&9

*"Confidence comes not from always being right but from not fearing to be wrong."*
-Peter T. McIntyre

**CONFIDENCE INTERVAL ESTIMATION**

$(1-\alpha)100\%$ Confidence Interval for the population mean μ

A. $\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

B. $\left(\overline{X} - t_{\alpha/2,v}\frac{S}{\sqrt{n}}, \overline{X} + t_{\alpha/2,v}\frac{S}{\sqrt{n}}\right)$ where v = n - 1

C. $\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}}, \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right)$

$(1-\alpha)100\%$ Confidence Interval for the difference between two population means $\mu_1 - \mu_2$

Independent Samples

D. $\left((\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$

E. $\left((\overline{X}_1 - \overline{X}_2) - t_{\alpha/2,v}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2,v}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$

Where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ and $v = n_1 + n_2 - 2$

F. $\left((\overline{X}_1 - \overline{X}_2) - t_{\alpha/2,v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + t_{\alpha/2,v}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$

Where $v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$

G. $\left((\overline{X}_1 - \overline{X}_2) - z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$

Two Related/Paired Samples

H. $\left(\bar{d} - t_{\alpha/2,v}\frac{S_d}{\sqrt{n}}, \bar{d} + t_{\alpha/2,v}\frac{S_d}{\sqrt{n}}\right)$

$d_i = x_i - y_i$ $\quad \bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ $\quad S_d = \sqrt{\frac{n\sum_{i=1}^{n} d_i^{\,2} - \left(\sum_{i=1}^{n} d_i\right)^2}{n(n-1)}}$

$v = n - 1$ $\quad$ n = # of pairs

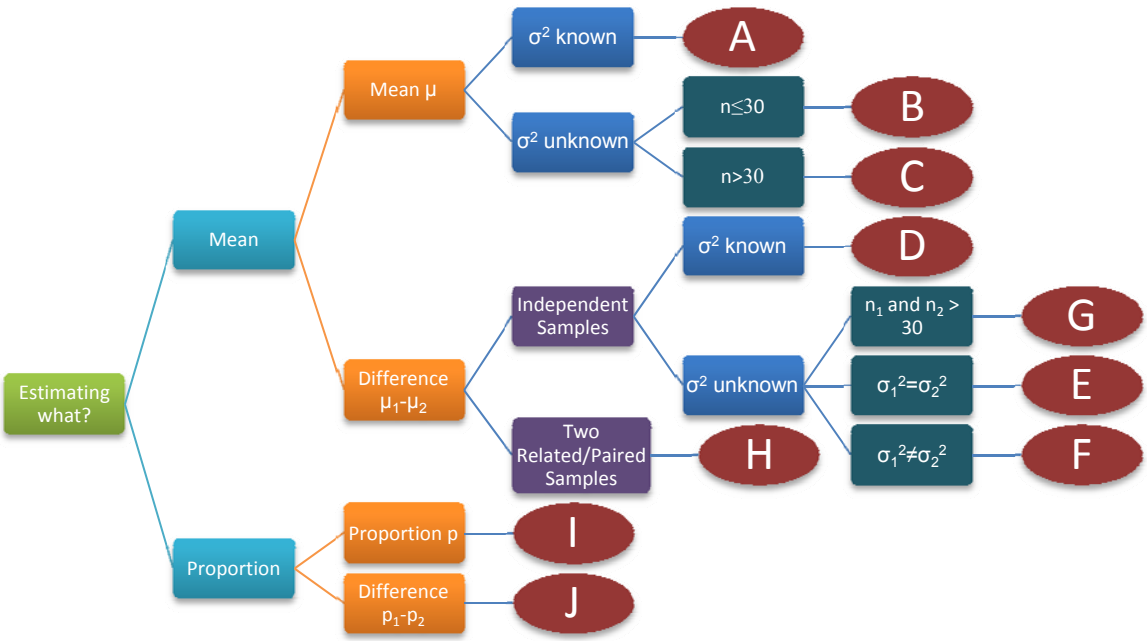
$(1-\alpha)100\%$ Confidence Interval for the population proportion p

I. $\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$

$(1-\alpha)100\%$ Confidence Interval for the difference between two population proportions $p_1$- $p_2$, where $n_1$ and $n_2$ are large

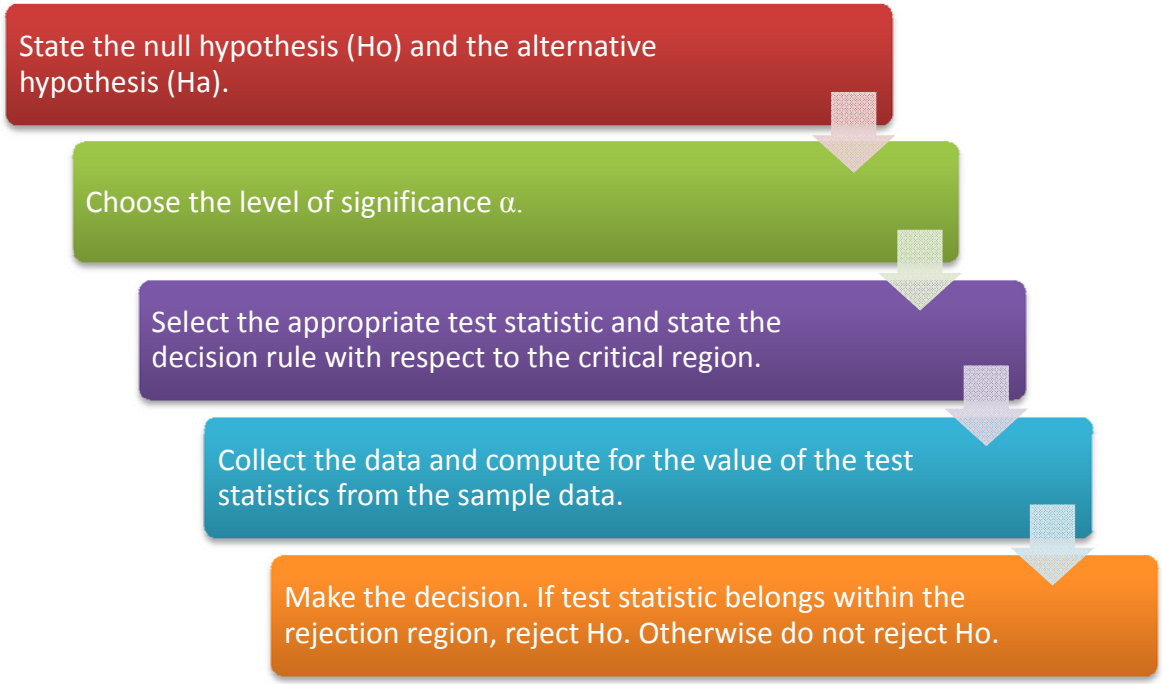
J. $\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}\right)$

In determining which formula should be used for a particular kind of data/given in a problem, one may refer to the flowchart below:



*"It is infinitely better to believe in a God which does not exist than not to believe in a God which exists."*

## HYPOTHESIS TESTING




On critical regions, remember the following:

Ha: $\theta>\theta_o$ → $Z>Z_\alpha$ or $t>t_\alpha$

Ha: $\theta<\theta_o$ → $Z<-Z_\alpha$ or $t<-t_\alpha$

Ha: $\theta\neq\theta_o$ → $|Z|>Z_{\alpha/2}$ or $|t|>t_{\alpha/2}$

Ha: the 2 variables are not independent

→ $\chi^2>\chi^2_{\alpha,(r-1)(c-1)}$

## LIST OF TEST STATISTICS

### Testing a hypothesis on the population mean

A. $$Z=\frac{\overline{X}-\mu_o}{\sigma/\sqrt{n}}$$

B. $$t=\frac{\overline{X}-\mu_o}{S/\sqrt{n}}$$ where v = n - 1

C. $$Z=\frac{\overline{X}-\mu_o}{S/\sqrt{n}}$$

**Testing the difference between two population means**

Independent Samples

D. $$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d_o}{\sqrt{\left(\sigma_1^2/n_1\right) + \left(\sigma_2^2/n_2\right)}}$$

E. $$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d_o}{S_p\sqrt{\left(1/n_1\right) + \left(1/n_2\right)}}$$

Where $S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$ and $v = n_1 + n_2 - 2$

F. $$t = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d_o}{\sqrt{\left(S_1^2/n_1\right) + \left(S_2^2/n_2\right)}}$$

Where $v = \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{\left(S_1^2/n_1\right)^2}{n_1-1} + \frac{\left(S_2^2/n_2\right)^2}{n_2-1}}$

G. $$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - d_o}{\sqrt{\left(S_1^2/n_1\right) + \left(S_2^2/n_2\right)}}$$

Two Related/Paired Samples

H. $$t = \frac{\overline{d} - d_o}{S_d/\sqrt{n}}$$

$d_i = x_i - y_i$ $\quad \overline{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ $\quad S_d = \sqrt{\frac{n\sum_{i=1}^{n} d_i^2 - \left(\sum_{i=1}^{n} d_i\right)^2}{n(n-1)}}$

$v = n - 1$ $\quad$ n = # of pairs

**Testing a hypothesis on proportions**

I. $$Z = \frac{x - np_o}{\sqrt{np_oq_o}}$$

**Testing the difference between two proportions**

J. $$Z = \frac{\bar{p}_1 - \bar{p}_2}{\sqrt{(\bar{p}_p)(\bar{q}_p)\left(\frac{1}{n_1 + n_2}\right)}}$$

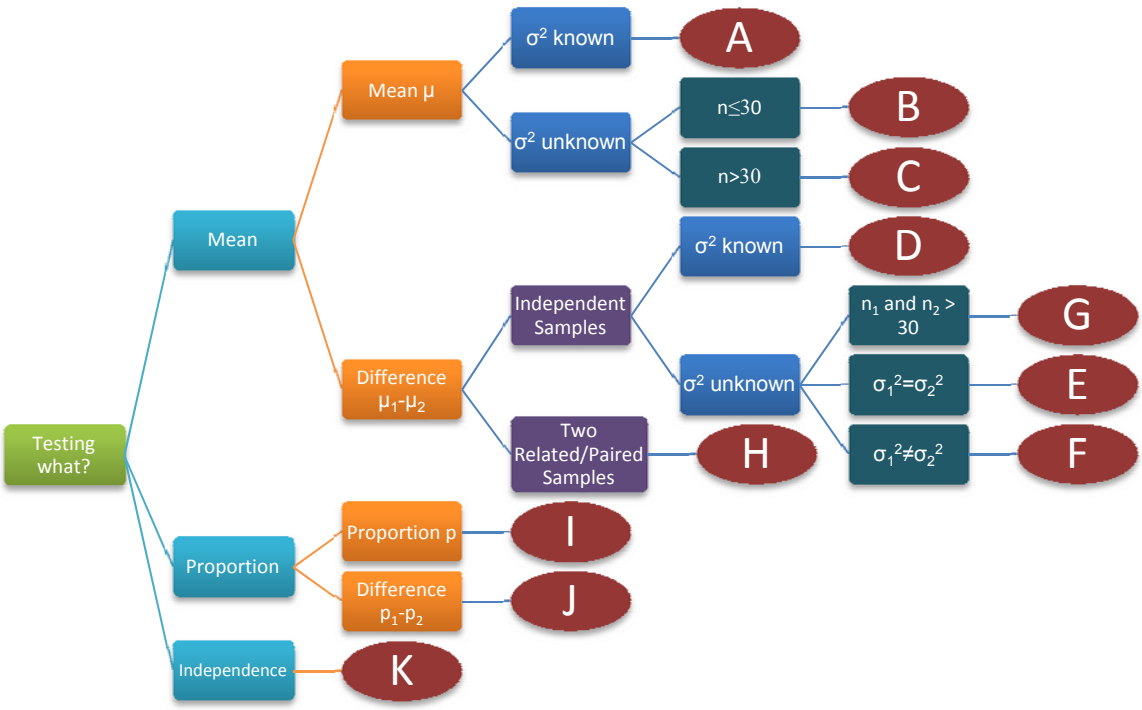
Where $\bar{p}_p = \frac{x_1 + x_2}{n_1 + n_2}$ and $\bar{q}_p = 1 - \bar{p}_p$

**Testing for independence**

K. $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Where $O_{ij}$ = observed number of cases in the ith row of the jth column
$E_{ij}$ = expected number of cases under Ho
= (column total) x (row total) / grand total

In determining which test statistic should be used for a particular kind of data/given in a problem, one may refer to the flowchart below:

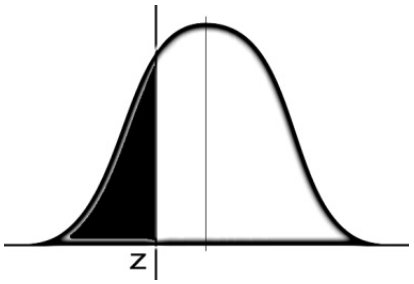


# Appendix

## Tables

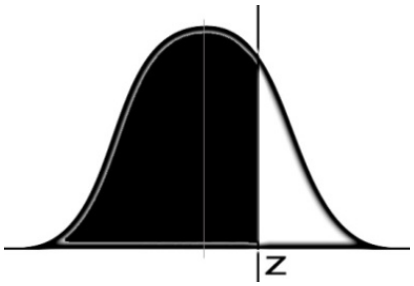
TABLE OF RANDOM NUMBERS

| | 00-04 | 05-09 | 10-14 | 15-19 | 20-24 | 25-29 | 30-34 | 35-39 | 40-44 | 45-49 |
|---|---|---|---|---|---|---|---|---|---|---|
| **00** | 13962 | 70992 | 65172 | 28053 | 02190 | 83634 | 66012 | 70305 | 66761 | 88344 |
| **01** | 43905 | 46941 | 72300 | 11641 | 43548 | 30455 | 07686 | 31840 | 3261 | 89139 |
| **02** | 00504 | 48658 | 38051 | 59408 | 16508 | 82979 | 92002 | 63606 | 41078 | 86326 |
| **03** | 61274 | 57238 | 47267 | 35303 | 29066 | 02140 | 60867 | 39847 | 50968 | 96719 |
| **04** | 43753 | 21159 | 16239 | 50595 | 62509 | 61207 | 86816 | 29902 | 23395 | 72640 |
| **05** | 83503 | 51662 | 21636 | 68192 | 84294 | 38754 | 84755 | 34053 | 94582 | 29215 |
| **06** | 36807 | 71420 | 35804 | 44862 | 23577 | 79551 | 42003 | 58684 | 9271 | 68396 |
| **07** | 19110 | 55680 | 18792 | 41487 | 16614 | 83053 | 00812 | 16749 | 45347 | 88199 |
| **08** | 82615 | 86984 | 93290 | 87971 | 60022 | 35415 | 20852 | 02909 | 99476 | 45568 |
| **09** | 05621 | 26584 | 36493 | 63013 | 68181 | 57702 | 49510 | 75304 | 38724 | 15712 |
| **10** | 06936 | 37293 | 55875 | 71213 | 83025 | 46063 | 74665 | 12178 | 10741 | 58362 |
| **11** | 84981 | 60458 | 16194 | 92403 | 80951 | 80068 | 47076 | 23310 | 74899 | 87929 |
| **12** | 66354 | 88441 | 96191 | 04794 | 14714 | 64749 | 43097 | 83976 | 83281 | 72038 |
| **13** | 49602 | 94109 | 36460 | 62353 | 00721 | 66980 | 82554 | 90270 | 12312 | 56299 |
| **14** | 78430 | 72391 | 96973 | 70437 | 97803 | 78683 | 4670 | 70667 | 58912 | 21883 |
| **15** | 33331 | 51803 | 15934 | 75807 | 46561 | 80188 | 78984 | 29317 | 27971 | 16440 |
| **16** | 62843 | 84445 | 56652 | 91797 | 45284 | 25842 | 96246 | 73504 | 21631 | 81223 |
| **17** | 19528 | 15445 | 77764 | 33446 | 41204 | 70067 | 33354 | 70680 | 66664 | 75486 |
| **18** | 16737 | 01887 | 50934 | 43306 | 75190 | 86997 | 56561 | 79018 | 34273 | 25196 |
| **19** | 99389 | 06685 | 45945 | 62000 | 76228 | 60645 | 87750 | 46329 | 46544 | 95665 |
| **20** | 36160 | 38196 | 77705 | 28891 | 12106 | 56281 | 86222 | 66116 | 39626 | 06080 |
| **21** | 05505 | 45420 | 44016 | 79662 | 92069 | 27628 | 50002 | 32540 | 19848 | 27319 |
| **22** | 85962 | 19758 | 92795 | 00458 | 71289 | 05884 | 37963 | 23322 | 73243 | 98185 |
| **23** | 28763 | 04900 | 54460 | 22083 | 89279 | 43492 | 00066 | 40857 | 86568 | 49336 |
| **24** | 42222 | 40446 | 82240 | 79159 | 44168 | 38213 | 46839 | 26598 | 29983 | 67645 |
| **25** | 43626 | 40039 | 51492 | 36488 | 70280 | 24218 | 14596 | 04744 | 89336 | 35630 |
| **26** | 97761 | 43444 | 95895 | 24102 | 07006 | 71923 | 04800 | 32062 | 41425 | 66862 |
| **27** | 49275 | 44270 | 52512 | 03951 | 21651 | 53867 | 73531 | 70073 | 45542 | 22831 |
| **28** | 15797 | 75134 | 39856 | 73527 | 78417 | 36208 | 59510 | 76913 | 22499 | 68467 |
| **29** | 04497 | 24853 | 43879 | 07613 | 26400 | 17180 | 18880 | 66083 | 02196 | 10638 |
| **30** | 95468 | 87411 | 30647 | 88711 | 01765 | 57688 | 60665 | 57636 | 36070 | 37285 |
| **31** | 01420 | 74218 | 71047 | 14401 | 74537 | 14820 | 45248 | 78007 | 65911 | 38583 |
| **32** | 74633 | 40171 | 97092 | 79137 | 30698 | 97915 | 36305 | 42613 | 87251 | 75608 |
| **33** | 46662 | 99688 | 59576 | 04887 | 02310 | 35508 | 69481 | 30300 | 94047 | 57096 |
| **34** | 10853 | 10393 | 03013 | 90372 | 89639 | 65800 | 88532 | 71789 | 59964 | 50681 |
| **35** | 68583 | 01032 | 67938 | 29733 | 71176 | 35699 | 10551 | 15091 | 52947 | 20134 |
| **36** | 75818 | 78982 | 24258 | 93051 | 02081 | 83890 | 66944 | 99856 | 87950 | 13952 |
| **37** | 16395 | 16837 | 00538 | 57133 | 89398 | 78205 | 72122 | 99655 | 25294 | 20941 |
| **38** | 53892 | 15105 | 40963 | 69267 | 85534 | 00533 | 27130 | 90420 | 72584 | 84576 |
| **39** | 66009 | 26869 | 91829 | 65078 | 89616 | 49016 | 14200 | 97469 | 88307 | 92282 |
| **40** | 45292 | 93427 | 92326 | 70206 | 15847 | 14302 | 60043 | 30530 | 57149 | 08642 |
| **41** | 34033 | 45008 | 41621 | 79437 | 98745 | 84455 | 66769 | 94729 | 17975 | 50963 |
| **42** | 13364 | 09937 | 00535 | 88122 | 47278 | 90758 | 23542 | 35273 | 67912 | 97670 |
| **43** | 03343 | 62593 | 93332 | 09921 | 25306 | 57483 | 98115 | 33460 | 55304 | 43572 |
| **44** | 46145 | 24476 | 62507 | 19530 | 41257 | 97919 | 02290 | 40357 | 38408 | 50031 |
| **45** | 37703 | 51658 | 17420 | 30593 | 39637 | 64220 | 45486 | 3698 | 80220 | 12139 |
| **46** | 12622 | 98083 | 17689 | 59677 | 56603 | 93316 | 79858 | 52548 | 67367 | 72416 |
| **47** | 56043 | 00251 | 70085 | 28067 | 78135 | 53000 | 18138 | 40564 | 77086 | 49557 |
| **48** | 43401 | 35924 | 28308 | 55140 | 07515 | 53854 | 23023 | 70268 | 80435 | 24269 |
| **49** | 18053 | 53460 | 32125 | 81357 | 26935 | 67234 | 78460 | 47833 | 20496 | 35645 |



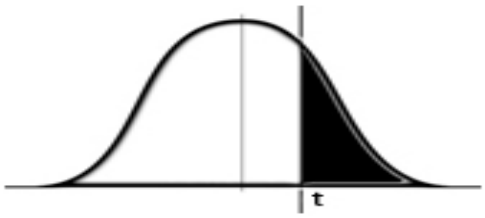

**STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.**

| Z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **-3.9** | 0.00005 | 0.00005 | 0.00004 | 0.00004 | 0.00004 | 0.00004 | 0.00004 | 0.00004 | 0.00003 | 0.00003 |
| **-3.8** | 0.00007 | 0.00007 | 0.00007 | 0.00006 | 0.00006 | 0.00006 | 0.00006 | 0.00005 | 0.00005 | 0.00005 |
| **-3.7** | 0.00011 | 0.00010 | 0.00010 | 0.00010 | 0.00009 | 0.00009 | 0.00008 | 0.00008 | 0.00008 | 0.00008 |
| **-3.6** | 0.00016 | 0.00015 | 0.00015 | 0.00014 | 0.00014 | 0.00013 | 0.00013 | 0.00012 | 0.00012 | 0.00011 |
| **-3.5** | 0.00023 | 0.00022 | 0.00022 | 0.00021 | 0.00020 | 0.00019 | 0.00019 | 0.00018 | 0.00017 | 0.00017 |
| **-3.4** | 0.00034 | 0.00032 | 0.00031 | 0.00030 | 0.00029 | 0.00028 | 0.00027 | 0.00026 | 0.00025 | 0.00024 |
| **-3.3** | 0.00048 | 0.00047 | 0.00045 | 0.00043 | 0.00042 | 0.00040 | 0.00039 | 0.00038 | 0.00036 | 0.00035 |
| **-3.2** | 0.00069 | 0.00066 | 0.00064 | 0.00062 | 0.00060 | 0.00058 | 0.00056 | 0.00054 | 0.00052 | 0.00050 |
| **-3.1** | 0.00097 | 0.00094 | 0.00090 | 0.00087 | 0.00084 | 0.00082 | 0.00079 | 0.00076 | 0.00074 | 0.00071 |
| **-3.0** | 0.00135 | 0.00131 | 0.00126 | 0.00122 | 0.00118 | 0.00114 | 0.00111 | 0.00107 | 0.00104 | 0.00100 |
| **-2.9** | 0.00187 | 0.00181 | 0.00175 | 0.00169 | 0.00164 | 0.00159 | 0.00154 | 0.00149 | 0.00144 | 0.00139 |
| **-2.8** | 0.00256 | 0.00248 | 0.00240 | 0.00233 | 0.00226 | 0.00219 | 0.00212 | 0.00205 | 0.00199 | 0.00193 |
| **-2.7** | 0.00347 | 0.00336 | 0.00326 | 0.00317 | 0.00307 | 0.00298 | 0.00289 | 0.00280 | 0.00272 | 0.00264 |
| **-2.6** | 0.00466 | 0.00453 | 0.00440 | 0.00427 | 0.00415 | 0.00402 | 0.00391 | 0.00379 | 0.00368 | 0.00357 |
| **-2.5** | 0.00621 | 0.00604 | 0.00587 | 0.00570 | 0.00554 | 0.00539 | 0.00523 | 0.00508 | 0.00494 | 0.00480 |
| **-2.4** | 0.00820 | 0.00798 | 0.00776 | 0.00755 | 0.00734 | 0.00714 | 0.00695 | 0.00676 | 0.00657 | 0.00639 |
| **-2.3** | 0.01072 | 0.01044 | 0.01017 | 0.00990 | 0.00964 | 0.00939 | 0.00914 | 0.00889 | 0.00866 | 0.00842 |
| **-2.2** | 0.01390 | 0.01355 | 0.01321 | 0.01287 | 0.01255 | 0.01222 | 0.01191 | 0.01160 | 0.01130 | 0.01101 |
| **-2.1** | 0.01786 | 0.01743 | 0.01700 | 0.01659 | 0.01618 | 0.01578 | 0.01539 | 0.01500 | 0.01463 | 0.01426 |
| **-2.0** | 0.02275 | 0.02222 | 0.02169 | 0.02118 | 0.02068 | 0.02018 | 0.01970 | 0.01923 | 0.01876 | 0.01831 |
| **-1.9** | 0.02872 | 0.02807 | 0.02743 | 0.02680 | 0.02619 | 0.02559 | 0.02500 | 0.02442 | 0.02385 | 0.02330 |
| **-1.8** | 0.03593 | 0.03515 | 0.03438 | 0.03362 | 0.03288 | 0.03216 | 0.03144 | 0.03074 | 0.03005 | 0.02938 |
| **-1.7** | 0.04457 | 0.04363 | 0.04272 | 0.04182 | 0.04093 | 0.04006 | 0.03920 | 0.03836 | 0.03754 | 0.03673 |
| **-1.6** | 0.05480 | 0.05370 | 0.05262 | 0.05155 | 0.05050 | 0.04947 | 0.04846 | 0.04746 | 0.04648 | 0.04551 |
| **-1.5** | 0.06681 | 0.06552 | 0.06426 | 0.06301 | 0.06178 | 0.06057 | 0.05938 | 0.05821 | 0.05705 | 0.05592 |
| **-1.4** | 0.08076 | 0.07927 | 0.07780 | 0.07636 | 0.07493 | 0.07353 | 0.07215 | 0.07078 | 0.06944 | 0.06811 |
| **-1.3** | 0.09680 | 0.09510 | 0.09342 | 0.09176 | 0.09012 | 0.08851 | 0.08691 | 0.08534 | 0.08379 | 0.08226 |
| **-1.2** | 0.11507 | 0.11314 | 0.11123 | 0.10935 | 0.10749 | 0.10565 | 0.10383 | 0.10204 | 0.10027 | 0.09853 |
| **-1.1** | 0.13567 | 0.13350 | 0.13136 | 0.12924 | 0.12714 | 0.12507 | 0.12302 | 0.12100 | 0.11900 | 0.11702 |
| **-1.0** | 0.15866 | 0.15625 | 0.15386 | 0.15151 | 0.14917 | 0.14686 | 0.14457 | 0.14231 | 0.14007 | 0.13786 |
| **-0.9** | 0.18406 | 0.18141 | 0.17879 | 0.17619 | 0.17361 | 0.17106 | 0.16853 | 0.16602 | 0.16354 | 0.16109 |
| **-0.8** | 0.21186 | 0.20897 | 0.20611 | 0.20327 | 0.20045 | 0.19766 | 0.19489 | 0.19215 | 0.18943 | 0.18673 |
| **-0.7** | 0.24196 | 0.23885 | 0.23576 | 0.23270 | 0.22965 | 0.22663 | 0.22363 | 0.22065 | 0.21770 | 0.21476 |
| **-0.6** | 0.27425 | 0.27093 | 0.26763 | 0.26435 | 0.26109 | 0.25785 | 0.25463 | 0.25143 | 0.24825 | 0.24510 |
| **-0.5** | 0.30854 | 0.30503 | 0.30153 | 0.29806 | 0.29460 | 0.29116 | 0.28774 | 0.28434 | 0.28096 | 0.27760 |
| **-0.4** | 0.34458 | 0.34090 | 0.33724 | 0.33360 | 0.32997 | 0.32636 | 0.32276 | 0.31918 | 0.31561 | 0.31207 |
| **-0.3** | 0.38209 | 0.37828 | 0.37448 | 0.37070 | 0.36693 | 0.36317 | 0.35942 | 0.35569 | 0.35197 | 0.34827 |
| **-0.2** | 0.42074 | 0.41683 | 0.41294 | 0.40905 | 0.40517 | 0.40129 | 0.39743 | 0.39358 | 0.38974 | 0.38591 |
| **-0.1** | 0.46017 | 0.45620 | 0.45224 | 0.44828 | 0.44433 | 0.44038 | 0.43644 | 0.43251 | 0.42858 | 0.42465 |
| **0.0** | 0.50000 | 0.49601 | 0.49202 | 0.48803 | 0.48405 | 0.48006 | 0.47608 | 0.47210 | 0.46812 | 0.46414 |



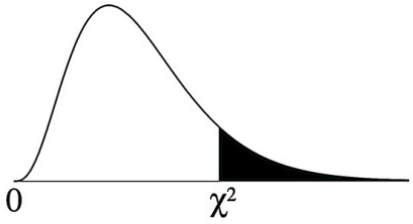

**STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.**

| Z | 0.00 | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 | 0.09 |
|---|---|---|---|---|---|---|---|---|---|---|
| **0.0** | 0.50000 | 0.50399 | 0.50798 | 0.51197 | 0.51595 | 0.51994 | 0.52392 | 0.52790 | 0.53188 | 0.53586 |
| **0.1** | 0.53983 | 0.54380 | 0.54776 | 0.55172 | 0.55567 | 0.55962 | 0.56356 | 0.56749 | 0.57142 | 0.57535 |
| **0.2** | 0.57926 | 0.58317 | 0.58706 | 0.59095 | 0.59483 | 0.59871 | 0.60257 | 0.60642 | 0.61026 | 0.61409 |
| **0.3** | 0.61791 | 0.62172 | 0.62552 | 0.62930 | 0.63307 | 0.63683 | 0.64058 | 0.64431 | 0.64803 | 0.65173 |
| **0.4** | 0.65542 | 0.65910 | 0.66276 | 0.66640 | 0.67003 | 0.67364 | 0.67724 | 0.68082 | 0.68439 | 0.68793 |
| **0.5** | 0.69146 | 0.69497 | 0.69847 | 0.70194 | 0.70540 | 0.70884 | 0.71226 | 0.71566 | 0.71904 | 0.72240 |
| **0.6** | 0.72575 | 0.72907 | 0.73237 | 0.73565 | 0.73891 | 0.74215 | 0.74537 | 0.74857 | 0.75175 | 0.75490 |
| **0.7** | 0.75804 | 0.76115 | 0.76424 | 0.76730 | 0.77035 | 0.77337 | 0.77637 | 0.77935 | 0.78230 | 0.78524 |
| **0.8** | 0.78814 | 0.79103 | 0.79389 | 0.79673 | 0.79955 | 0.80234 | 0.80511 | 0.80785 | 0.81057 | 0.81327 |
| **0.9** | 0.81594 | 0.81859 | 0.82121 | 0.82381 | 0.82639 | 0.82894 | 0.83147 | 0.83398 | 0.83646 | 0.83891 |
| **1.0** | 0.84134 | 0.84375 | 0.84614 | 0.84849 | 0.85083 | 0.85314 | 0.85543 | 0.85769 | 0.85993 | 0.86214 |
| **1.1** | 0.86433 | 0.86650 | 0.86864 | 0.87076 | 0.87286 | 0.87493 | 0.87698 | 0.87900 | 0.88100 | 0.88298 |
| **1.2** | 0.88493 | 0.88686 | 0.88877 | 0.89065 | 0.89251 | 0.89435 | 0.89617 | 0.89796 | 0.89973 | 0.90147 |
| **1.3** | 0.90320 | 0.90490 | 0.90658 | 0.90824 | 0.90988 | 0.91149 | 0.91309 | 0.91466 | 0.91621 | 0.91774 |
| **1.4** | 0.91924 | 0.92073 | 0.92220 | 0.92364 | 0.92507 | 0.92647 | 0.92785 | 0.92922 | 0.93056 | 0.93189 |
| **1.5** | 0.93319 | 0.93448 | 0.93574 | 0.93699 | 0.93822 | 0.93943 | 0.94062 | 0.94179 | 0.94295 | 0.94408 |
| **1.6** | 0.94520 | 0.94630 | 0.94738 | 0.94845 | 0.94950 | 0.95053 | 0.95154 | 0.95254 | 0.95352 | 0.95449 |
| **1.7** | 0.95543 | 0.95637 | 0.95728 | 0.95818 | 0.95907 | 0.95994 | 0.96080 | 0.96164 | 0.96246 | 0.96327 |
| **1.8** | 0.96407 | 0.96485 | 0.96562 | 0.96638 | 0.96712 | 0.96784 | 0.96856 | 0.96926 | 0.96995 | 0.97062 |
| **1.9** | 0.97128 | 0.97193 | 0.97257 | 0.97320 | 0.97381 | 0.97441 | 0.97500 | 0.97558 | 0.97615 | 0.97670 |
| **2.0** | 0.97725 | 0.97778 | 0.97831 | 0.97882 | 0.97932 | 0.97982 | 0.98030 | 0.98077 | 0.98124 | 0.98169 |
| **2.1** | 0.98214 | 0.98257 | 0.98300 | 0.98341 | 0.98382 | 0.98422 | 0.98461 | 0.98500 | 0.98537 | 0.98574 |
| **2.2** | 0.98610 | 0.98645 | 0.98679 | 0.98713 | 0.98745 | 0.98778 | 0.98809 | 0.98840 | 0.98870 | 0.98899 |
| **2.3** | 0.98928 | 0.98956 | 0.98983 | 0.99010 | 0.99036 | 0.99061 | 0.99086 | 0.99111 | 0.99134 | 0.99158 |
| **2.4** | 0.99180 | 0.99202 | 0.99224 | 0.99245 | 0.99266 | 0.99286 | 0.99305 | 0.99324 | 0.99343 | 0.99361 |
| **2.5** | 0.99379 | 0.99396 | 0.99413 | 0.99430 | 0.99446 | 0.99461 | 0.99477 | 0.99492 | 0.99506 | 0.99520 |
| **2.6** | 0.99534 | 0.99547 | 0.99560 | 0.99573 | 0.99585 | 0.99598 | 0.99609 | 0.99621 | 0.99632 | 0.99643 |
| **2.7** | 0.99653 | 0.99664 | 0.99674 | 0.99683 | 0.99693 | 0.99702 | 0.99711 | 0.99720 | 0.99728 | 0.99736 |
| **2.8** | 0.99744 | 0.99752 | 0.99760 | 0.99767 | 0.99774 | 0.99781 | 0.99788 | 0.99795 | 0.99801 | 0.99807 |
| **2.9** | 0.99813 | 0.99819 | 0.99825 | 0.99831 | 0.99836 | 0.99841 | 0.99846 | 0.99851 | 0.99856 | 0.99861 |
| **3.0** | 0.99865 | 0.99869 | 0.99874 | 0.99878 | 0.99882 | 0.99886 | 0.99889 | 0.99893 | 0.99896 | 0.99900 |
| **3.1** | 0.99903 | 0.99906 | 0.99910 | 0.99913 | 0.99916 | 0.99918 | 0.99921 | 0.99924 | 0.99926 | 0.99929 |
| **3.2** | 0.99931 | 0.99934 | 0.99936 | 0.99938 | 0.99940 | 0.99942 | 0.99944 | 0.99946 | 0.99948 | 0.99950 |
| **3.3** | 0.99952 | 0.99953 | 0.99955 | 0.99957 | 0.99958 | 0.99960 | 0.99961 | 0.99962 | 0.99964 | 0.99965 |
| **3.4** | 0.99966 | 0.99968 | 0.99969 | 0.99970 | 0.99971 | 0.99972 | 0.99973 | 0.99974 | 0.99975 | 0.99976 |
| **3.5** | 0.99977 | 0.99978 | 0.99978 | 0.99979 | 0.99980 | 0.99981 | 0.99981 | 0.99982 | 0.99983 | 0.99983 |
| **3.6** | 0.99984 | 0.99985 | 0.99985 | 0.99986 | 0.99986 | 0.99987 | 0.99987 | 0.99988 | 0.99988 | 0.99989 |
| **3.7** | 0.99989 | 0.99990 | 0.99990 | 0.99990 | 0.99991 | 0.99991 | 0.99992 | 0.99992 | 0.99992 | 0.99992 |
| **3.8** | 0.99993 | 0.99993 | 0.99993 | 0.99994 | 0.99994 | 0.99994 | 0.99994 | 0.99995 | 0.99995 | 0.99995 |
| **3.9** | 0.99995 | 0.99995 | 0.99996 | 0.99996 | 0.99996 | 0.99996 | 0.99996 | 0.99996 | 0.99997 | 0.99997 |




## Critical Values of the *t-distribution*

| | One-tail Probability | | | | | | |
|---|---|---|---|---|---|---|---|
| | 0.1 | 0.05 | 0.025 | 0.01 | 0.005 | 0.001 | |
| df | | | | | | | df |
| 1 | 3.078 | 6.314 | 12.706 | 31.821 | 63.656 | 318.289 | 1 |
| 2 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 22.328 | 2 |
| 3 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 10.214 | 3 |
| 4 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 7.173 | 4 |
| 5 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 5.894 | 5 |
| 6 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.208 | 6 |
| 7 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 4.785 | 7 |
| 8 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 4.501 | 8 |
| 9 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.297 | 9 |
| 10 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.144 | 10 |
| 11 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.025 | 11 |
| 12 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 3.930 | 12 |
| 13 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 3.852 | 13 |
| 14 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 3.787 | 14 |
| 15 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 3.733 | 15 |
| 16 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 3.686 | 16 |
| 17 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.646 | 17 |
| 18 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.610 | 18 |
| 19 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.579 | 19 |
| 20 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.552 | 20 |
| 21 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.527 | 21 |
| 22 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.505 | 22 |
| 23 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.485 | 23 |
| 24 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.467 | 24 |
| 25 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.450 | 25 |
| 26 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.435 | 26 |
| 27 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.421 | 27 |
| 28 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.408 | 28 |
| 29 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.396 | 29 |
| 30 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.385 | 30 |
| 40 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.307 | 40 |
| 50 | 1.299 | 1.676 | 2.009 | 2.403 | 2.678 | 3.261 | 50 |
| 60 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.232 | 60 |
| 70 | 1.294 | 1.667 | 1.994 | 2.381 | 2.648 | 3.211 | 70 |
| 80 | 1.292 | 1.664 | 1.990 | 2.374 | 2.639 | 3.195 | 80 |
| 90 | 1.291 | 1.662 | 1.987 | 2.368 | 2.632 | 3.183 | 90 |
| 100 | 1.290 | 1.660 | 1.984 | 2.364 | 2.626 | 3.174 | 100 |
| 110 | 1.289 | 1.659 | 1.982 | 2.361 | 2.621 | 3.166 | 110 |
| 120 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.160 | 120 |
| ∞ | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.090 | ∞ |




**Critical Values of the $\chi^2_{\alpha,v}$ distribution where the shaded region is α.**

| *df* | $\chi^2_{0.995}$ | $\chi^2_{0.990}$ | $\chi^2_{0.975}$ | $\chi^2_{0.950}$ | $\chi^2_{0.900}$ | $\chi^2_{0.100}$ | $\chi^2_{0.050}$ | $\chi^2_{0.025}$ | $\chi^2_{0.010}$ | $\chi^2_{0.005}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0.000 | 0.000 | 0.001 | 0.004 | 0.016 | 2.706 | 3.841 | 5.024 | 6.635 | 7.879 |
| **2** | 0.010 | 0.020 | 0.051 | 0.103 | 0.211 | 4.605 | 5.991 | 7.378 | 9.210 | 10.597 |
| **3** | 0.072 | 0.115 | 0.216 | 0.352 | 0.584 | 6.251 | 7.815 | 9.348 | 11.345 | 12.838 |
| **4** | 0.207 | 0.297 | 0.484 | 0.711 | 1.064 | 7.779 | 9.488 | 11.143 | 13.277 | 14.860 |
| **5** | 0.412 | 0.554 | 0.831 | 1.145 | 1.610 | 9.236 | 11.070 | 12.833 | 15.086 | 16.750 |
| **6** | 0.676 | 0.872 | 1.237 | 1.635 | 2.204 | 10.645 | 12.592 | 14.449 | 16.812 | 18.548 |
| **7** | 0.989 | 1.239 | 1.690 | 2.167 | 2.833 | 12.017 | 14.067 | 16.013 | 18.475 | 20.278 |
| **8** | 1.344 | 1.646 | 2.180 | 2.733 | 3.490 | 13.362 | 15.507 | 17.535 | 20.090 | 21.955 |
| **9** | 1.735 | 2.088 | 2.700 | 3.325 | 4.168 | 14.684 | 16.919 | 19.023 | 21.666 | 23.589 |
| **10** | 2.156 | 2.558 | 3.247 | 3.940 | 4.865 | 15.987 | 18.307 | 20.483 | 23.209 | 25.188 |
| **11** | 2.603 | 3.053 | 3.816 | 4.575 | 5.578 | 17.275 | 19.675 | 21.920 | 24.725 | 26.757 |
| **12** | 3.074 | 3.571 | 4.404 | 5.226 | 6.304 | 18.549 | 21.026 | 23.337 | 26.217 | 28.300 |
| **13** | 3.565 | 4.107 | 5.009 | 5.892 | 7.042 | 19.812 | 22.362 | 24.736 | 27.688 | 29.819 |
| **14** | 4.075 | 4.660 | 5.629 | 6.571 | 7.790 | 21.064 | 23.685 | 26.119 | 29.141 | 31.319 |
| **15** | 4.601 | 5.229 | 6.262 | 7.261 | 8.547 | 22.307 | 24.996 | 27.488 | 30.578 | 32.801 |
| **16** | 5.142 | 5.812 | 6.908 | 7.962 | 9.312 | 23.542 | 26.296 | 28.845 | 32.000 | 34.267 |
| **17** | 5.697 | 6.408 | 7.564 | 8.672 | 10.085 | 24.769 | 27.587 | 30.191 | 33.409 | 35.718 |
| **18** | 6.265 | 7.015 | 8.231 | 9.390 | 10.865 | 25.989 | 28.869 | 31.526 | 34.805 | 37.156 |
| **19** | 6.844 | 7.633 | 8.907 | 10.117 | 11.651 | 27.204 | 30.144 | 32.852 | 36.191 | 38.582 |
| **20** | 7.434 | 8.260 | 9.591 | 10.851 | 12.443 | 28.412 | 31.410 | 34.170 | 37.566 | 39.997 |
| **21** | 8.034 | 8.897 | 10.283 | 11.591 | 13.240 | 29.615 | 32.671 | 35.479 | 38.932 | 41.401 |
| **22** | 8.643 | 9.542 | 10.982 | 12.338 | 14.041 | 30.813 | 33.924 | 36.781 | 40.289 | 42.796 |
| **23** | 9.260 | 10.196 | 11.689 | 13.091 | 14.848 | 32.007 | 35.172 | 38.076 | 41.638 | 44.181 |
| **24** | 9.886 | 10.856 | 12.401 | 13.848 | 15.659 | 33.196 | 36.415 | 39.364 | 42.980 | 45.559 |
| **25** | 10.520 | 11.524 | 13.120 | 14.611 | 16.473 | 34.382 | 37.652 | 40.646 | 44.314 | 46.928 |
| **26** | 11.160 | 12.198 | 13.844 | 15.379 | 17.292 | 35.563 | 38.885 | 41.923 | 45.642 | 48.290 |
| **27** | 11.808 | 12.879 | 14.573 | 16.151 | 18.114 | 36.741 | 40.113 | 43.195 | 46.963 | 49.645 |
| **28** | 12.461 | 13.565 | 15.308 | 16.928 | 18.939 | 37.916 | 41.337 | 44.461 | 48.278 | 50.993 |
| **29** | 13.121 | 14.256 | 16.047 | 17.708 | 19.768 | 39.087 | 42.557 | 45.722 | 49.588 | 52.336 |
| **30** | 13.787 | 14.953 | 16.791 | 18.493 | 20.599 | 40.256 | 43.773 | 46.979 | 50.892 | 53.672 |
| **40** | 20.707 | 22.164 | 24.433 | 26.509 | 29.051 | 51.805 | 55.758 | 59.342 | 63.691 | 66.766 |
| **50** | 27.991 | 29.707 | 32.357 | 34.764 | 37.689 | 63.167 | 67.505 | 71.420 | 76.154 | 79.490 |
| **60** | 35.534 | 37.485 | 40.482 | 43.188 | 46.459 | 74.397 | 79.082 | 83.298 | 88.379 | 91.952 |
| **70** | 43.275 | 45.442 | 48.758 | 51.739 | 55.329 | 85.527 | 90.531 | 95.023 | 100.425 | 104.215 |
| **80** | 51.172 | 53.540 | 57.153 | 60.391 | 64.278 | 96.578 | 101.879 | 106.629 | 112.329 | 116.321 |
| **90** | 59.196 | 61.754 | 65.647 | 69.126 | 73.291 | 107.565 | 113.145 | 118.136 | 124.116 | 128.299 |
| **100** | 67.328 | 70.065 | 74.222 | 77.929 | 82.358 | 118.498 | 124.342 | 129.561 | 135.807 | 140.169 |

## REFERENCES:

*Introduction to Statistics by Ronald E. Walpole*
*Introduction to Probability by Charles M. Grinstead and Laurie Snell*
*Introduction to Probability by Dimitri P.Bertsekas and John N. Tsitsiklis*
*Probability Deyistified by Alan G. Bluman*
*Schaum's Outline of Theory and Problems of Statistics by Murray R. Spiegel and Larry J. Stephens*
*Statistics for Dummies by Deborah Ramsey*

## Thank you so much for your comments and contributions

***Angelo Alberto***
***Roselle Alteria***
***Rosemarie Anga-on***
***Regyn Avena***
***Katherine Calderon***
***Virgil Gabriel Garcia***
***Elise Christina Lasco***
***Anna Mendoza***
***Gary Mondejar***
***Primadonna Prima***
***Maria Ana Tacbad***
***Crisille Villaluna***
***Tomas Jorge Maddela***

*Samantha Anis*
*Jerome Bautista*
*Gladys Mae Chavez*
*Kharla Corales*
*Ma. Agnes Isabel Malihan*
*Michelle Magalino*
*Giancarlo Miguel Pantaleon*
*Alan Rutaquio*
*Charmaine Gulanes*
*Anna Leah Cabigas*
*Maria Margarita Aspi*
*Gil Kenneth San Miguel*
*Andrea Carmelli Mendoza*
*Vheejay Tampol*
*Kathleen Grace De Castro*
*Christine Javier*
*Hosanna Krizia Leus*
*Karen Orticio*
*Jessica Mae Balanquit*
*Yrish Mae Estoce*
*Diana Vie Fernandez*
*Larissa Anna Bagadion*
*Leirald Reyes*
*Paul John Martinet*
*Ritz Angelica Alejandro*
*Michelle Leanda Lugtu*
*Jennifer Cristobal*
*Beatrice Marie Achacoso*
*Terrence Erard Teh*
*Josephine Gamos*
*Ernst John Humawid*