MATH 153

STATISTICAL METHODS 1

Semester 1, 2017

Lecture Notes

Compiled by

Sampson Twumasi-Ankrah (Ph.D)

Department of Mathematics

KNUST, Ghana

Content

Basic concepts of statistics; descriptive statistics: organization and presentation of data; measures of central tendency; dispersion; percentiles and Box-and-Whisker plots; skewness; kurtosis; probability: random experiments; probability laws; computation of probability of single events; counting techniques in probability, random variables and probability distribution: binomial distribution; Poisson distribution; normal distribution; and application of permutation and combination; test of hypothesis.

Contents

1. BASIC CONCEPTS OF STATISTICS

1.1 Nature of Statistics

Statistics plays a major role in many different areas of our lives. In other words, ***Statistics*** is used in all disciplines.

"Statisticians get to play in everyone else's back yard." (John Tukey)

Examples:

1. A pharmacist is concerned that administering caffeine to premature babies will increase the incidence of necrotizing enterocolitis.

2. In a genetics study involving patients with Alzheimer's disease (AD), researchers wish to identify genes that are differentially expressed (when compared to non-AD patients).

3. In a clinical trial, physicians want to determine which of two drugs is more effective for treating HIV in the early stages of the disease.

4. In a public health study involving "at-risk" teenagers, epidemiologists want to know whether smoking is more common in a particular demographic class.

5. In an agricultural experiment, researchers want to know which of four fertilizers (which vary in their nitrogen contents) produces the highest corn yield.

6. A food scientist is interested in determining how different feeding schedules (for pigs) could affect the spread of salmonella during the slaughtering process.

7. A research dietician wants to determine if academic achievement is related to body mass index (BMI) among African American students in the fourth grade.

Used appropriately, statistics can provide an understanding of the world around us. Used inappropriately, it can lend support to inaccurate beliefs. Understanding the methodologies of statistics will provide you with the ability to analyze and critique studies. With this ability, you will be an informed consumer of information, which will enable you to distinguish solid analysis from the bogus presentation of numerical "facts."

Insisting on the use of statistical analyses to draw conclusions is an extension of the argument that objectivity is critical in science. Without the use of

statistics, little can be learnt from most research studies. And because of the increasing use of statistics in so many areas of our lives, it has become very desirable to understand and practice statistical thinking. This is important even if you do not use statistical methods directly.

1.1.1 What is statistics?

When asked this question, many people respond that statistics is numbers. After all, we are bombarded by numbers that supposedly represent how we feel and who we are.

> *For example, we hear on the radio that 50% of first marriages, 67% of second marriages, and 74% of third marriages end in divorce (Forest Institute of Professional Psychology, Springfield, MO).*

Certainly, statistics has a lot to do with numbers, but this definition is only partially correct. Statistics is also about where the numbers come from (that is, how they were obtained) and how closely the numbers reflect reality.

- ***Definition***

  Statistics is the science of collecting, organizing, summarizing, and analyzing information to draw conclusions or answer questions. In addition, statistics is about providing a measure of confidence in any conclusions.

It is helpful to consider this definition in four parts. The first part of the definition states that statistics involves the collection of information. The second refers to the organization and summarization of information. The third states that the information is analyzed to draw conclusions or answer specific questions. The fourth part states that results should be reported with some measure that represents how convinced we are that our conclusions reflect reality.

In simple terms, *statistics is the science of data;* how to interpret data, analyze data, and design studies to collect data.

1.1.2 Branches of Statistics

1. **Descriptive statistics** is the branch of statistics that involves the organization, summarization, and display of data. Two general techniques are used to accomplish this goal.

a. Organize the entire set of scores into a table or a graph that allows researchers (and others) to see the whole set of scores. (summarizing data graphically)
b. Compute one or two summary values (such as the average) that describe the entire group. (Summarizing data numerically).

In summary, *descriptive statistics* describe data through numerical summaries, tables, and graphs.

2. **Inferential statistics** is the branch of statistics that uses methods that take a result from a sample, extend it to the population, and measure the reliability of the result.

1.1.3 Variables and Variation

- A **variable** is any attribute, characteristic, or measurable property that can vary from one observation to another. Any observation could have an infinite number of variables, such as height, weight, color, or density. For example, consider biochemistry/pharmacy students in a specific graduating class. Just a few of the numerous variables that could be associated with each student include: gender, height, weight, marital status, systolic blood pressure, blood type (A, B, AB, O), blood glucose level, etc. The number of possible variables is limited only by our imagination.

  Because these measurements may take on different values, repeat measurements observed under apparently identical conditions do not, in general, give the identical results (i.e., they are usually not exactly reproducible).

  Example 1

  Duplicate determinations of serum concentration of a drug 1 hr after an injection will not be identical no matter if the duplicates come from (a) the same blood sample or (b) from separate samples from two different persons or (c) from the same person on two different occasions.

- Variation is an inherent characteristic of experimental observations. To isolate and to identify particular causes of variability requires special experimental designs and analysis. Variation in observations is due to a number of causes. For example, an assay will vary depending on:

    1. The instrument used for the analysis
    2. The analyst performing the assay
    3. The particular sample chosen
    4. Unidentified, uncontrollable background error, commonly known as ''noise''

  This inherent variability in observation and measurement is a principal reason for the need of statistical methodology in experimental design and data analysis. In the absence of variability, scientific experiments would be short and simple: interpretation of experimental results from well-designed experiments would be unambiguous. In fact, without variability, single observations would often be sufficient to define the properties of an object or a system. Since few, if any, processes can be considered absolutely invariant, statistical treatment is often essential for summarizing and defining the nature of data, and for making decisions or inferences based on these variable experimental observations.

❖ ***TYPES OF VARIABLES***

Variables can be classified as *qualitative* (aka, categorical) or *quantitative* (aka, numeric).

a. Qualitative or categorical

Qualitative variables take on values that are names or labels. In other words, they allow for classification of individuals based on some attribute or characteristic.

Example:

i. Different side effects resulting from different drug treatments or the presence or absence of a defect in a finished product. These kinds of data are frequently observed in clinical and pharmaceutical experiments and processes.

ii. *A finished tablet classified in quality control as ''defective'' or ''not defective''.*
iii. *In clinical studies, the categorization of a patient by sex (male or female) or race is a classification according to attributes.*
iv. *When calculating ED50 or LD50, animals are categorized as ''responders'' or ''non-responders'' to various levels of a therapeutic agent, a categorical response.*

These examples describe variables that cannot be ordered. A male is not associated with a higher or lower numerical value than a female.

b. Quantitative

Quantitative variables are numeric. They represent a measurable quantity of individual.

For example, when we speak of the population of a city, we are talking about the number of people in the city, a measurable attribute of the city. Therefore, population would be a quantitative variable.

Types of Quantitative Variables

We can further classify quantitative variables into two types: *discrete* or *continuous*.

a. Discrete Variable

A discrete variable is a quantitative variable that has either a finite number of possible values or a countable number of possible values. The term countable means that the values result from counting, such as 0, 1, 2, 3, and so on.

These kinds of variables are commonly observed in biological and pharmaceutical experiments and are exemplified by measurements such as the number of anginal episodes in 1 week or the number of side effects of different kinds after drug treatment.

b. Continuous Variable

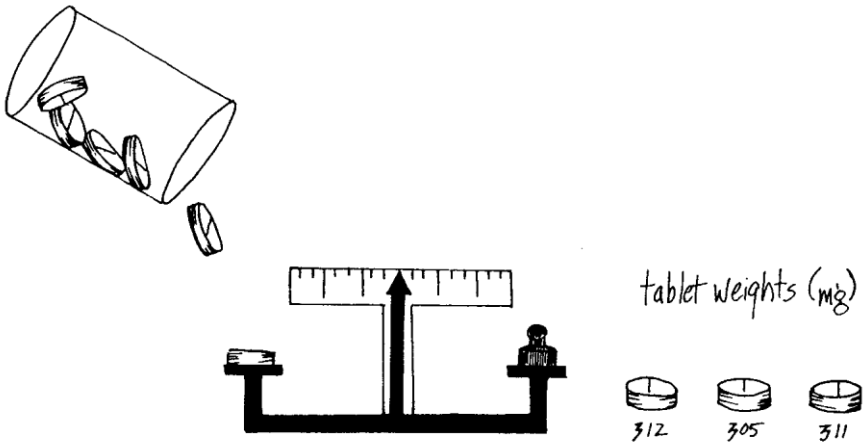
A continuous variable is one that can take on *any* value within some range or interval (i.e., within a specified lower and upper limit).

Illustration

The limiting factor for the total number of possible observations or results is the sensitivity of the measuring instrument. When weighing tablets or making blood pressure measurements, there are an infinite number of possible values that can be observed if the measurement could be made to an unlimited number of decimal places.

However, if the balance, for example, is sensitive only to the nearest milligram, the data will appear as discrete values. For tablets targeted at 1 g and weighed to the nearest milligram, the tablet weights might range from 900 to 1100 mg, a total of 201 possible integral values (900, 901, 902, 903, ⋯, 1098, 1099, 1100).

For the same tablet weighed on a more sensitive balance, to the nearest 0.1 mg, values from 899.5 to 1100.4 might be possible, a total of 2010 possible values, and so on.



Tablet weights: an example of a variable measurement (a random variable).

Exceptional Cases

Often, continuous variables cannot be easily measured but can be ranked in order of magnitude.

In the assessment of pain in a clinical study of analgesics, a patient can have a continuum of pain. To measure pain on a continuous numerical scale would be difficult. On the other hand, a patient may be able to differentiate slight pain from moderate pain, moderate pain from severe pain, and so on.

In analgesic studies, scores are commonly assigned to pain severity, such as no pain, 0, slight pain, 1, moderate pain, 2, and severe pain, 3. Although the scores cannot be thought of as an exact characterization of pain, the value 3 does represent more intense pain than the values 0, 1, or 2.

The scoring system above is a representation of a continuous variable by discrete ''scores'' which can be rationally ordered or ranked from low to high. This is commonly known as a rating scale, and the ranked data are on an ordinal scale. The rating scale is an effort to quantify a continuous, but subjective, variable.

Summary

If you count to get the value of a quantitative variable, it is ***discrete***. If you measure to get the value of a quantitative variable, it is ***continuous***. When deciding whether a variable is discrete or continuous, ask yourself if it is counted or measured.

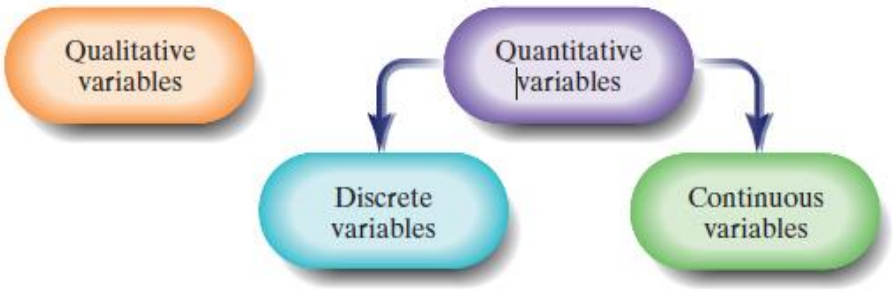


Illustration of the relationship among qualitative, quantitative, discrete, and continuous variables.

❖ ***DATA VRS VARIABLE***

The list of observed values for a *variable* is *data*. Example, gender is a variable; the observations male or female are data. Qualitative data are observations corresponding to a qualitative variable. Quantitative data are observations corresponding to a quantitative variable. Discrete data are observations corresponding to a discrete variable, and continuous data are observations corresponding to a continuous variable.

Univariate vs. Bivariate Data

Statistical data are often classified according to the number of variables being studied.

Univariate data

When we conduct a study that looks at only one variable, we say that we are working with univariate data. Suppose, for example, that we conducted a survey to estimate the average weight of patients under treatment. Since we are only working with one variable (weight), we would be working with univariate data.

Bivariate data

When we conduct a study that examines the relationship between two variables, we are working with bivariate data. Suppose we conducted a study to see if there were a relationship between the height and weight of patients under treatment. Since we are working with two variables (height and weight), we would be working with bivariate data.

1.1.4 Populations and Samples

Samples are usually a relatively small number of observations taken from a relatively large population or universe. The sample values are the observations, the data, obtained from the population.

The population consists of data with some clearly defined characteristic(s). For example, a population may consist of ***all*** patients with a particular disease, or tablets from a production batch. The sample in these cases could consist of a selection of patients to participate in a clinical study, or tablets chosen for a weight determination.

The sample is only part of the available data. In the usual experimental situation, we make observations on a relatively small sample in order to make inferences about the characteristics of the whole, the population. The totality of available data is the ***population or universe***.

When designing an experiment, the population should be clearly defined so that samples chosen are representative of the population. This is important in clinical trials, for example, where inferences to the treatment of disease states are crucial. The exact nature or character of the population is rarely known, and often impossible to ascertain, although we can make assumptions about its properties.

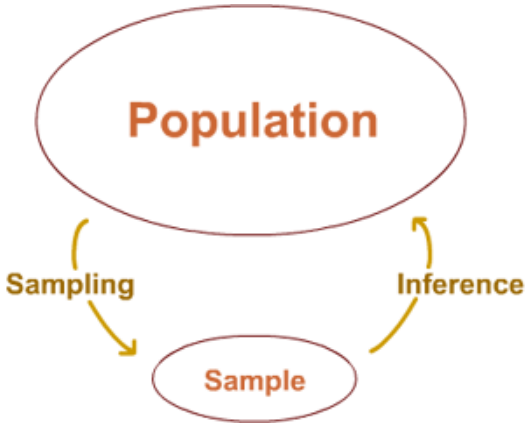
Population Parameters and Sample Statistics

''Any measurable characteristic of the population is called a parameter''.

For example, the average weight of a batch of tablets or the average blood pressure of hypertensive persons in Ghana are parameters of the respective populations. Parameters are generally denoted by Greek letters; for example, the mean of the population is denoted as $\boldsymbol{\mu}$. Note that parameters are characteristic of the population, and are values that are usually unknown to us.

Quantities derived from the sample are called *sample statistics*. Corresponding to the true average weight of a batch of tablets is the average weight for the small sample taken from the population of tablets. We should be very clear about the nature of samples.

A parameter, for example, the mean weight of a batch of tablets, is a fixed value; it does not vary. Sample statistics are variable. Their values depend on the particular sample chosen and the variability of the measurement.



Procedures for Inferential Statistical Tests

There are several important parts to completing an appropriate statistical test.

1. **Establish a research question**.

   It is impossible to acquire new knowledge and to conduct research without a clear idea of what you wish to explore. For example, we would like to know if three batches of a specific drug are the same regarding their content uniformity. Simply stated: are these three batches equal?

2. **Formulate a hypothesis**.

   Although covered in a later chapter, we should formulate a hypothesis that will be either rejected or not rejected based on the results of the statistical test. In this case, the hypothesis that is being tested is that Batch A equals Batch B equals Batch C. The only alternative to this hypothesis is that the batches are not all equal to each other.

3. **Select an appropriate test**.

   Using information about the data (identifying the dependent and independent variables) the correct test is selected based on whether these variables are discrete or continuous. For example, batches A, B, and C represent an independent variable with three discrete levels and the assay result for the drug's contents is a continuous variable (%) dependent upon the batch from which it was selected.

   Therefore, the most appropriate statistical test would be one that can handle a continuous dependent variable and a discrete independent variable with three categories. A common mistake is to collect the data first, without consideration of these first three requirements for statistical tests, only to realize that a statistical judgment cannot be made because of the arbitrary format of the data.

4. **Sample correctly**.

   The sample should be randomly selected from each batch. An appropriate sample size should be selected to provide the most accurate results.

5. **Collect data**.

   The collection should ensure that each observed result is independent of any other assay.

6. **Perform test**.

   Only this portion of the statistical process actually involves the number crunching associated with statistical analysis. Many commercially available computer packages are available to save us the tedium of detailed mathematical manipulations.

7. **Make a decision**.

   Based on the data collected and statistical manipulation of the sample data, a statement (inference) is made regarding the entire population from which the sample was drawn. In our example, based on the results of the test statistics, the hypothesis that all three batches are equal (based on content uniformity), is either rejected or the sample does not provide enough information to reject the hypothesis.

Illustration

Caffeine is commonly used to treat newborn infants for apnea of prematurity and to prevent the onset of other acute conditions. Known as "the silver bullet" in the treatment of prematurely born infants, caffeine is widely regarded within the neonatal care community to be safe and cost effective. It has also been approved by the US Food and Drug Administration for use with preterm infants due to its history of providing beneficial outcomes with no long-term adverse side effects.

Research Question: *Does treating premature infants with caffeine increase the chances of developing necrotizing enterocolitis?*

Necrotizing enterocolitis (NEC) is a serious disease characterized by infection and inflammation of the intestine. It is most commonly observed in premature infants. Left untreated, NEC can lead to serious health complications and even death.

In an 18-month period during 2008-2009, there were 615 premature infants admitted to the neonatal intensive care unit at Palmetto Richland Hospital in Columbia, SC. Infants were assigned to either receive caffeine or not.

- 35 out of 137 patients (about 26 percent) receiving caffeine developed NEC
- 10 out of 478 patients (about 2 percent) not receiving caffeine developed NEC.

Reference: *Cox et al. (2015). Evaluation of caffeine and the development of necrotizing enterocolitis. Journal of Neonatal-Perinatal Medicine 8, 339-347.*

1.1.5 Scales of Measurement

Four levels of measurement scales are generally distinguished.

- ***Nominal or Categorical Measurements***

  Nominal measurements allow patients to be classified with respect to some characteristic. Examples of such measurements are marital status, sex, and blood group. The following are properties of a nominal scale:

  - The categories are mutually exclusive (an individual can belong to only one category).
  - The categories have no logical order—numbers may be assigned to categories but merely as convenient labels.

- ***Ordinal Scale Measurements***

  The next level of measurement is the ordinal scale. This scale has one additional property over those of a nominal scale—a logical ordering of the categories. With such measurements, the numbers assigned to the categories indicate the amount of a characteristic possessed.

  A psychiatrist may, for example, grade patients on an anxiety scale as 'not anxious', 'mildly anxious', 'moderately anxious', or 'severely anxious' and use the numbers 0, 1, 2, and 3 to label the categories, with lower numbers indicating less anxiety.

  The psychiatrist cannot infer, however, that the difference in anxiety between patients with scores of, say, 0 and 1 is the same as the difference

between patients assigned scores of 2 and 3. The scores on an ordinal scale, however, do allow patients to be ranked with respect to the characteristic being assessed. The following are the properties of an ordinal scale:

- The categories are mutually exclusive.
- The categories have some logical order.
- The categories are scaled according to the amount of a particular characteristic that they indicate.

❖ ***Interval Scales***

The third level of measurement is the interval scale. Such scales possess all the properties of an ordinal scale plus the additional property that equal differences between category levels, on any part of the scale, reflect equal differences in the characteristic being measured.

An example of such a scale is temperature on the Celsius (C) or Fahrenheit (F) scale; the difference between temperatures of 80°F and 90°F represents the same difference in heat as that between temperatures of 30° and 40° on the Fahrenheit scale. An important point to make about interval scales is that the zero point is simply another point on the scale; it does not represent the starting point of the scale or the total absence of the characteristic being measured. The properties of an interval scale are as follows:

- The categories are mutually exclusive.
- The categories have a logical order.
- The categories are scaled according to the amount of the characteristic that they indicate.
- Equal differences in the characteristic are represented by equal differences in the numbers assigned to the categories.
- The zero point is completely arbitrary.

❖ ***Ratio Scales***

The final level of measurement is the ratio scale. This type of scale has one property in addition to those listed for interval scales—namely, the possession of a true zero point that represents the absence of the characteristic being measured. Consequently, statements can be made about both the differences on the scale and the ratio of points on the scale.

An example is weight, where not only is the difference between 100 and 50 kg the same as that between 75 and 25 kg, but an object weighing 100 kg can also be said to be twice as heavy as one weighing 50 kg. This is not true of, say, temperature on the Celsius or Fahrenheit scales, where a reading of 100° on either scale does not represent twice the warmth of a temperature of 50°. If, however, two temperatures are measured on the Kelvin scale, which does have a true zero point (absolute zero or –273°C), then statements about the ratio of the two temperatures can be made. The properties of a ratio scale are the following:

- The categories are mutually exclusive.
- The data categories have a logical order.
- The categories are scaled according to the amount of the characteristic that they possess.
- Equal differences in the characteristic being measured are represented by equal differences in the numbers assigned to the categories.
- The zero point represents an absence of the characteristic being measured

Exercise 1.1

For each of the variables listed below from the line listing in Table 2.1, identify what type of variable it is.

A. Nominal

B. Ordinal

C. Interval

D. Ratio

_____ 1. Date of diagnosis

_____ 2. Town of residence

_____ 3. Age (years)

_____ 4. Sex

_____ 5. Highest alanine aminotransferase (ALT)

1.2 PRECISION, ACCURACY, AND BIAS

When dealing with variable measurements, the definitions of *precision* and *accuracy* should be clearly defined from a statistical point of view.

1.2.1 Precision

In the vocabulary of statistics, precision refers to the extent of variability of a group of measurements observed under similar experimental conditions. Observations, relatively close in magnitude, are considered to be precise as reflected by a small standard deviation. (Note that means are more precisely measured than individual observations according to this definition). An important, sometimes elusive concept is that a precise set of measurements may have the same mean as an imprecise set.

In most experiments with which we will be concerned, the mean and standard deviation of the data are independent (i.e., they are unrelated). Fig. 1.3 shows the results of two assay methods, each performed in triplicate. Both methods have an average result of 100□, but method II is more precise.

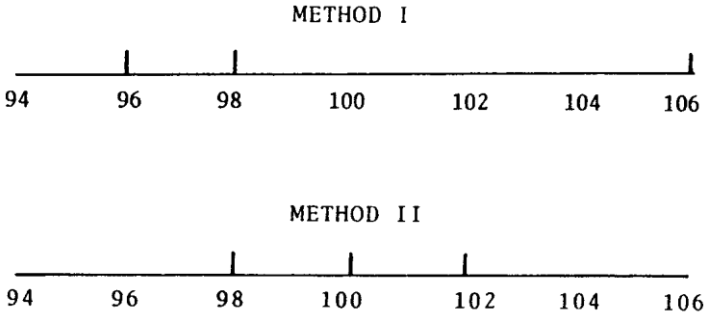


Figure 1.2.1

Representation of two analytical methods with the same accuracy but different precisions.

1.2.2 Accuracy

Accuracy refers to the closeness of an individual observation or mean to the true value. The ‘‘true’’ value is that result which would be observed in the

absence of error (e.g., the true mean tablet potency or the true drug content of a preparation being assayed). In the example of the assay results shown in *Figure 1.2.1*, both methods are apparently equally accurate (or inaccurate).

Figure 1.2.2 shows the results of two dissolution methods for two formulations of the same drug, each formulation replicated four times by each method. The objective of the in vitro dissolution test is to simulate the in vivo oral absorption of the drug from the two dosage-form modifications.

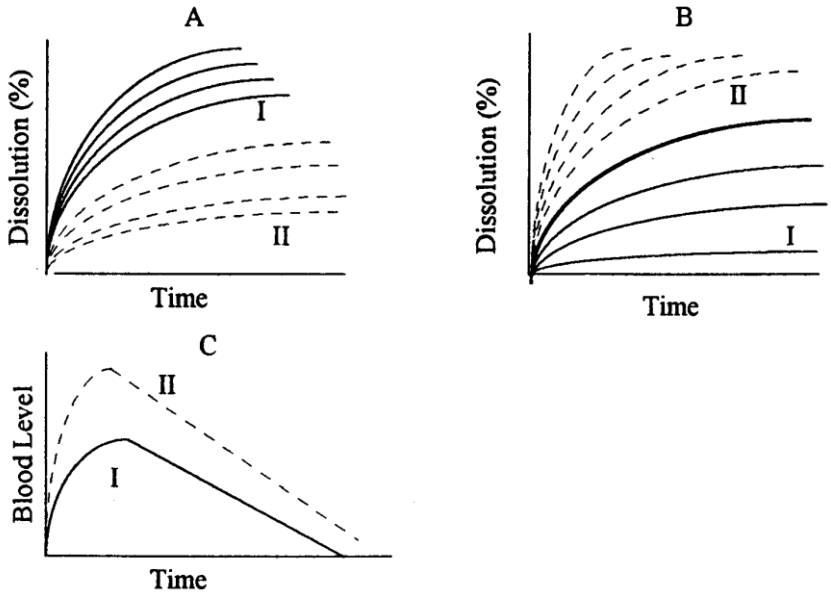


Figure 1.2.2 In vitro dissolution results for two formulations using two different methods and in vivo blood level versus time results. Methods A and B, in vitro; C, in vivo.

The first dissolution method, A, is very precise but does not give an accurate prediction of the in vivo results. According to the dissolution data for method A, we would expect that formulation I would be more rapidly and extensively absorbed in vivo. The actual in vivo results depicted in *Fig. 1.2.2* show the contrary result. The less precise method, method B in this example, is a more accurate predictor of the true in vivo results. This example is meant to show that a precise measurement need not be accurate, nor an accurate measurement precise.

Of course, the best circumstance is to have data that are both precise and accurate. If possible, we should make efforts to improve both the accuracy and precision of experimental observations. For example, in drug analysis,

advanced electronic instrumentation can greatly increase the accuracy and precision of assay results.

1.2.3 Bias

Accuracy can also be associated with the term *bias*. The notion of bias has been discussed in Sect. 1.4 in relation to the concept of unbiased estimates (e.g., the mean and variance). The meaning of bias in statistics is similar to the everyday definition in terms of ''fairness.''

An accurate measurement, no matter what the precision, can be thought of as unbiased, because an accurate measurement is a ''fair'' estimate of the true result. A biased estimate is systematically either higher or lower than the true value. A biased estimate can be thought of as giving an ''unfair'' notion of the true value.

For example, when estimating the average result of experimental data, the mean, $\bar{X}$, represents an estimate of the true population parameter, $\boldsymbol{\mu}$, and in this sense is considered accurate and unbiased.

An average blood pressure reduction of 10 mmHg due to an antihypertensive agent, derived from data from a clinical study of 200 patients, can be thought of as an unbiased estimate of the true blood pressure reduction due to the drug, provided that the patients are appropriately selected at ''random.'' The true reduction in this case is the average reduction that would be observed if the antihypertensive effect of the drug were known for all members of the population (e.g., all hypertensive patients).

The outcome of a single experiment, such as the 10 mmHg reduction observed in the 200 patients above, will in all probability not be identical to the true mean reduction. But the mean reduction as observed in the 200 patients is an accurate and unbiased assessment of the population average.

A biased estimate is one which, on the average, does not equal the population parameter. In the example cited above for hypertensives, a biased estimate

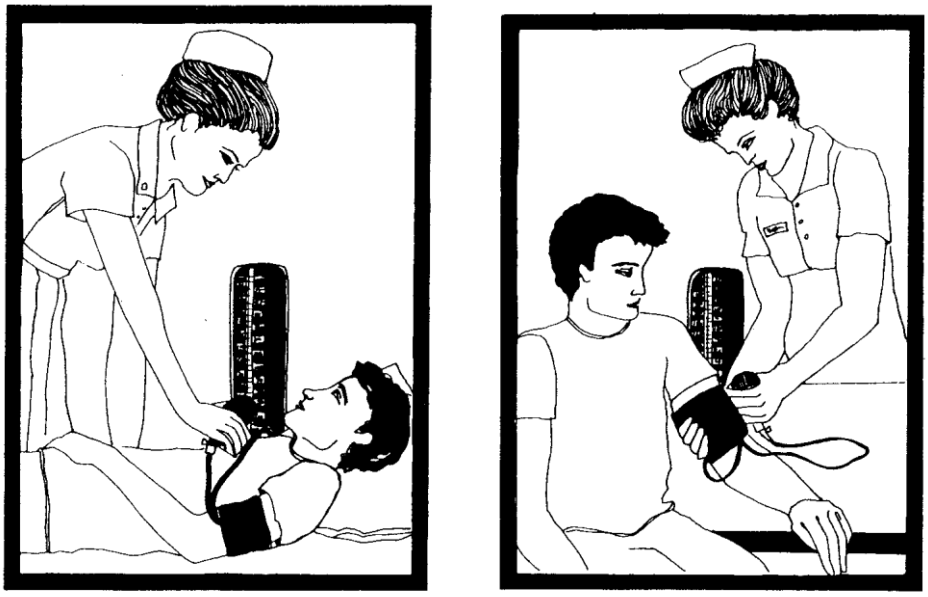


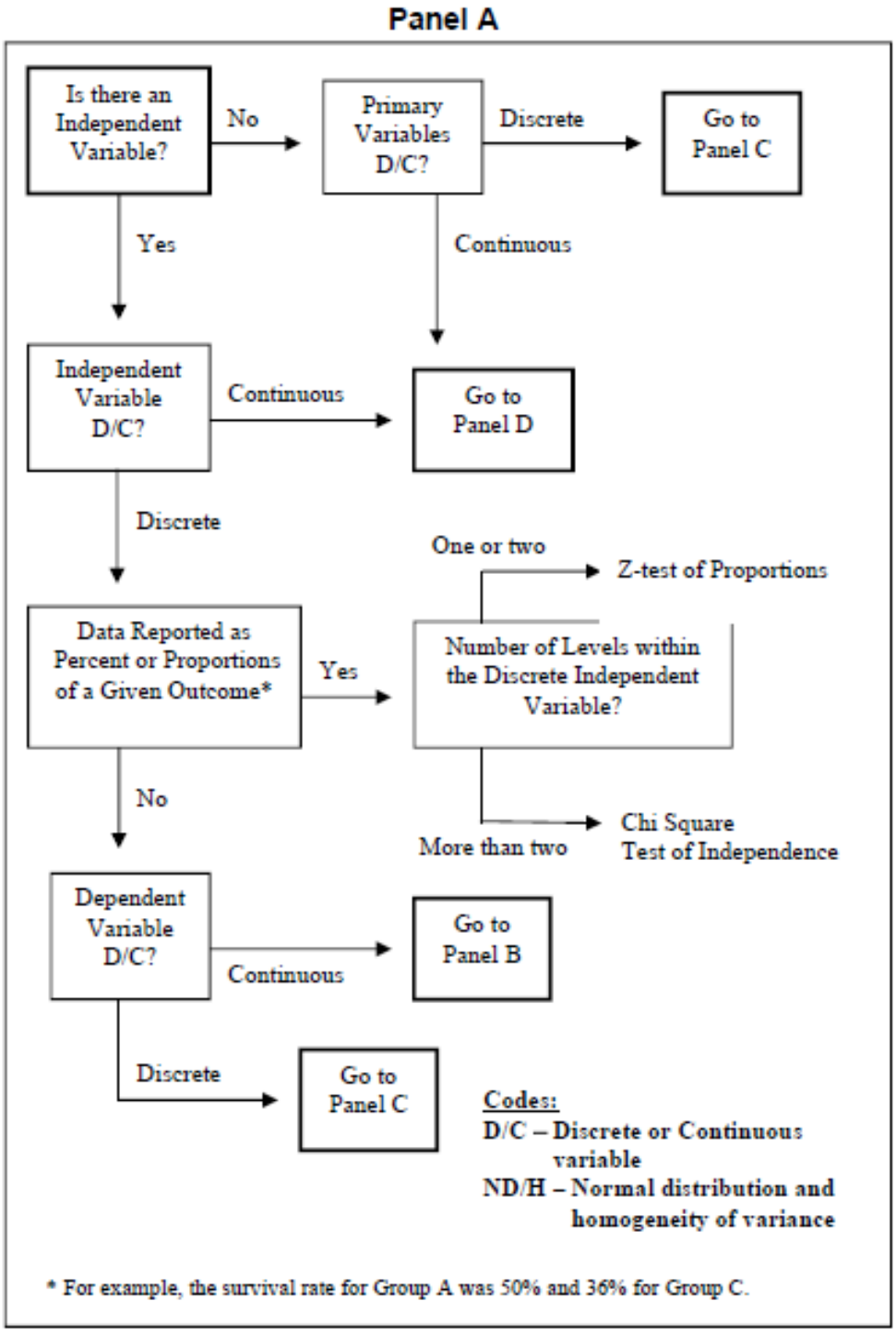
Figure 1.2.3 Bias in determining the effect of an antihypertensive drug.

would result if for all patients one nurse took all the measurements before therapy and another nurse took all measurements during therapy, and each nurse had a different criterion or method for determining blood pressure.

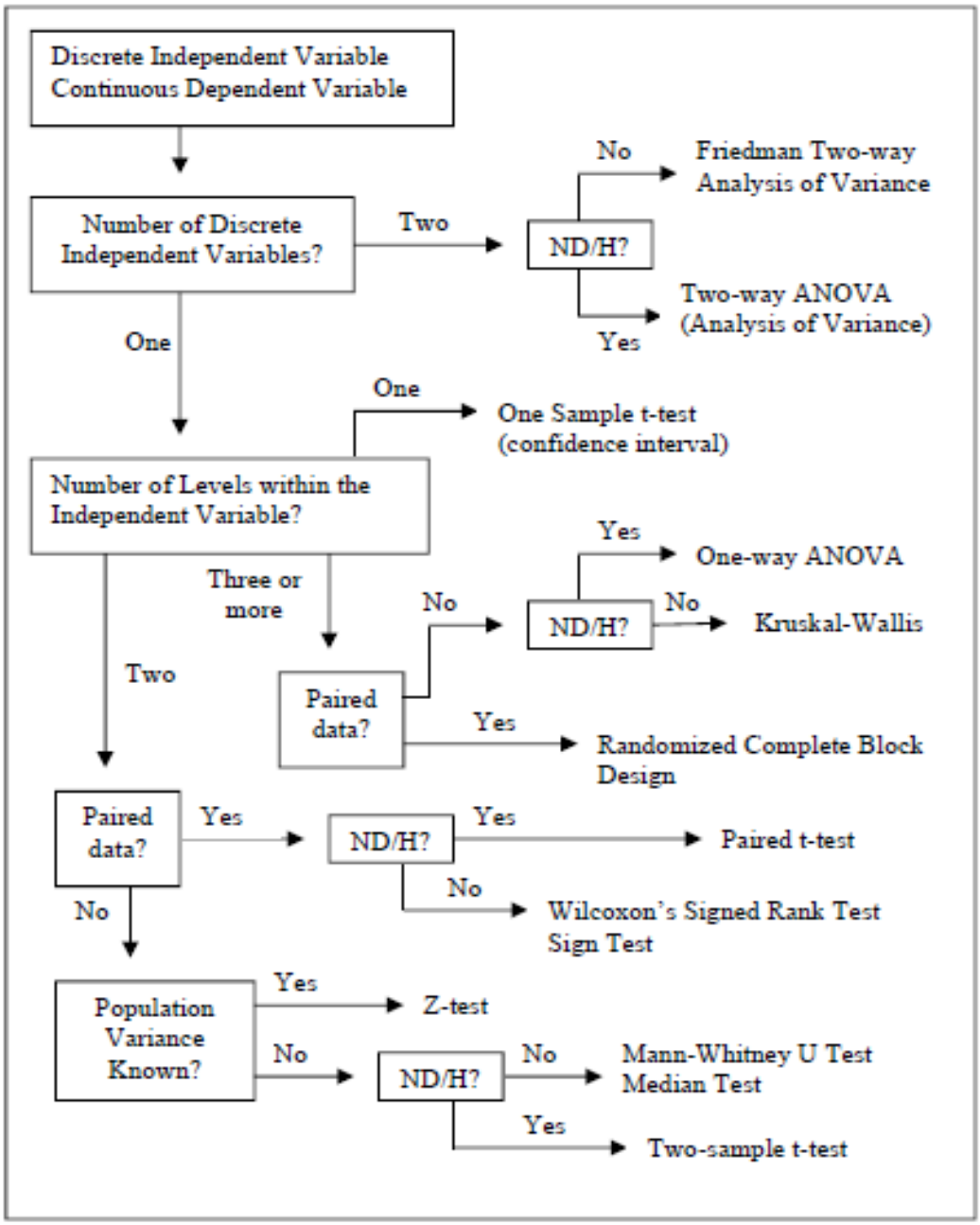
See *Figure 1.2.3* for a clarification as to why this procedure leads to a biased estimate of the drug's effectiveness in reducing blood pressure. If the supine position results in higher blood pressure than the sitting position, the results of the study will tend to show a bias in the direction of too large a blood pressure reduction.

The statistical estimates that we usually use, such as the mean and variance, are unbiased estimates. Bias often results from **(a)** the improper use of experimental design; **(b)** improper choice of samples; **(c)** unconscious bias, due to lack of blinding, for example; or **(d)** improper observation and recording of data, such as that illustrated in Fig. 1.2.3.

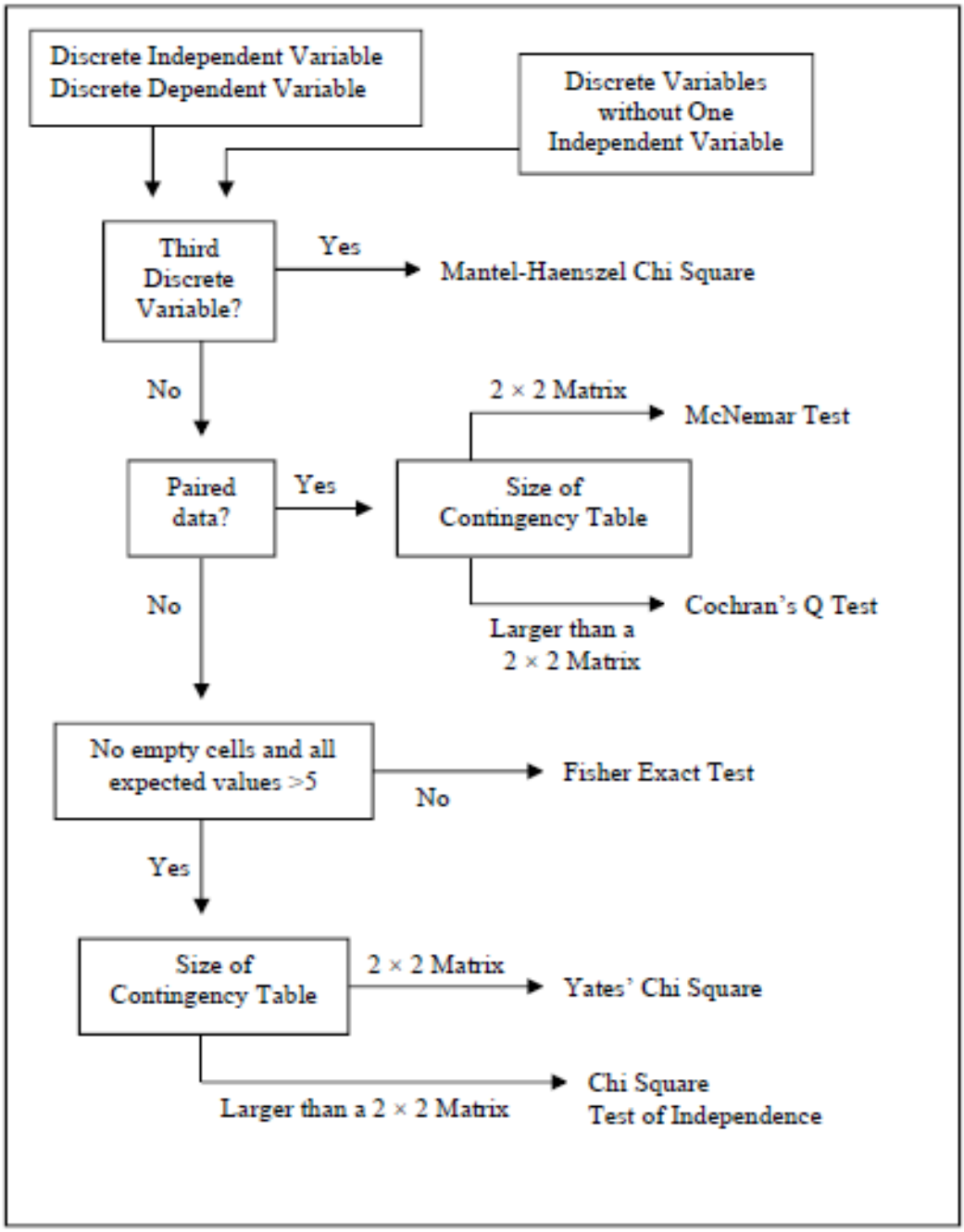
Flow Charts for Selection of Appropriate Inferential Tests



Panel B


Discrete Independent Variable
Continuous Dependent Variable
Number of Discrete Independent Variables?
Two
ND/H?
No
Friedman Two-way Analysis of Variance
Yes
Two-way ANOVA (Analysis of Variance)
One
Number of Levels within the Independent Variable?
One
One Sample t-test (confidence interval)
Three or more
Yes
One-way ANOVA
No
ND/H?
No
Kruskal-Wallis
Two
Paired data?
Yes
Randomized Complete Block Design
Paired data?
Yes
ND/H?
Yes
Paired t-test
No
Wilcoxon's Signed Rank Test
Sign Test
No
Population Variance Known?
Yes
Z-test
No
ND/H?
No
Mann-Whitney U Test
Median Test
Yes
Two-sample t-test

Panel C


Discrete Independent Variable
Discrete Dependent Variable
Discrete Variables without One Independent Variable
Third Discrete Variable?
Yes
Mantel-Haenszel Chi Square
No
2 × 2 Matrix
McNemar Test
Paired data?
Yes
Size of Contingency Table
No
Cochran's Q Test
Larger than a 2 × 2 Matrix
No empty cells and all expected values >5
Fisher Exact Test
No
Yes
Size of Contingency Table
2 × 2 Matrix
Yates' Chi Square
Larger than a 2 × 2 Matrix
Chi Square Test of Independence

Assignment 1

1. Which of the following selected variables associated with a random sample of 50,000 tablets, mentioned earlier in this chapter, are discrete variables and which are continuous?

Amount of active ingredient (content uniformity)

Dissolution test − pass or fail criteria

Disintegration rate

Change in manufacturing process − old process versus new

Friability − pass or fail criteria

Hardness

Impurities − present or absent

Size − thickness/diameter

Tablet weight

Immediate release or sustained release

Formulation A, B, or C

2 Summarizing Data

This unit is categorized into two sections: summarizing data graphically and numerically.

2.1 Summarizing Data Graphically

The purpose of this *section* is to learn how to organize raw data in tables or graphs, which allow for a quick overview of the information collected. The procedures used in describing data in this section depend on whether the data are qualitative, discrete, or continuous.

2.1.1 Organizing Qualitative Data

Here, we will concentrate on tabular and graphical summaries of qualitative data.

Organize Qualitative Data in Tables

Recall that qualitative (or categorical) data provide measures that categorize or classify an individual or observation. When qualitative data are collected; we are often interested in determining the number of individuals observed within each category.

Definition

A ***frequency distribution*** lists each category of data and the number of occurrences for each category of data.

Example

Organizing Qualitative Data into a Frequency Distribution

Problem: A physical therapist wants to get a sense of the types of rehabilitation required by her patients. To do so; she obtains a simple random sample of 30 of her patients and records the body part requiring rehabilitation. See Table 1.

Construct a frequency distribution of location of injury.

Table 1

Back	Back	Hand	Neck	Knee	Knee
Wrist	Back	Groin	Shoulder	Shoulder	Back
Elbow	Back	Back	Back	Back	Back
Back	Shoulder	Shoulder	Knee	Knee	Back
Hip	Knee	Hip	Hand	Back	Wrist

Source: Krystal Catton, student at Joliet Junior College

- ✓ **Approach**: To construct a frequency distribution, we create a list of the body parts (categories) and tally each occurrence. Finally, we add up the number of tallies to determine the frequency.

- ✓ **Solution**: See Table 2. From the table, we can see that the back is the most common body part requiring rehabilitation, with a total of 12.

Table 2

Body Part	Tally	Frequency
Back	𝍸 𝍸 ||	12
Wrist	||	2
Elbow	|	1
Hip	||	2
Shoulder	||||	4
Knee	𝍸	5
Hand	||	2
Groin	|	1
Neck	|	1

With frequency distributions, it is a good idea to add up the frequency column to make sure that it sums to the number of observations. In the case of the data in Example 1, the frequency column adds up to 30, as it should.

Often, rather than being concerned with the frequency with which categories of data occur, we want to know the *relative frequency* of the categories.

Definition

The ***relative frequency*** is the proportion (or percent) of observations within a category and is found using the formula:

$$\text{Relative frequency} = \frac{frequency}{sum - of - all - frequencies}$$

A ***relative frequency distribution*** lists each category of data together with the relative frequency.

Example

Constructing a Relative Frequency Distribution of Qualitative Data

Problem: Using the data in Table 2, construct a relative frequency distribution.

- ✓ **Approach**: Add all the frequencies, and then use Formula (1) to compute the relative frequency of each category of data.
- ✓ **Solution**: We add the values in the frequency column in Table 2: Sum of all frequencies = 12+2+1+2+4+5+2+1+1=30

  We now compute the relative frequency of each category.

  For example, the relative frequency of the category *Back* is:

  $$12/30 = 0.4$$

  After computing the relative frequency for the remaining categories, we obtain the relative frequency distribution shown in Table 3.

Table 3

Body Part	Frequency	Relative Frequency
Back	12	$\frac{12}{30} = 0.4$
Wrist	2	$\frac{2}{30} \approx 0.0667$
Elbow	1	0.0333
Hip	2	0.0667
Shoulder	4	0.1333
Knee	5	0.1667
Hand	2	0.0667
Groin	1	0.0333
Neck	1	0.0333

From the table, we can see that the most common body part for rehabilitation is the back.

It is a good idea to add up the relative frequencies to be sure they sum to 1. In fraction form, the sum should be exactly 1. In decimal form, the sum may differ slightly from 1 due to rounding.

Exercise 2.3

Using the same vaccination data as in Exercise 2.2, find the mode. (If you answered Exercise 2.2, find the mode from your frequency distribution.)

2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

Construct Bar Graphs

Once raw data are organized in a table, we can create graphs. Graphs allow us to see the data and get a sense of what the data are saying about the individuals in the study.

The cliché, ―A picture is worth a thousand words,‖ has a similar application when dealing with data. In general, pictures of data result in a more powerful message than tables.

Try the following exercise for yourself:

Open a newspaper and look at a table and a graph. Study each. Now put the paper away and close your eyes. What do you see in your mind's eye? Can you recall information more easily from the table or the graph? In general, people are more likely to recall information obtained from a graph than they are from a table.

One of the most common devices for graphically representing qualitative data is a bar graph. Both nominal and ordinal data can easily be displayed with this type of graph.

Definition

A ***bar graph*** is constructed by labeling each category of data on either the horizontal or vertical axis, and the frequency or relative frequency of the category on the other axis. Rectangles of equal width are drawn for each

category. The height of each rectangle represents the category's frequency or relative frequency.

Example

Constructing a Frequency and Relative Frequency Bar Graph

Problem: Use the data summarized in Table 3 to construct the following:

(a) Frequency bar graph

(b) Relative frequency bar graph

- ✓ **Approach**: We will use a horizontal axis to indicate the categories of the data (body parts, in this case) and a vertical axis to represent the frequency or relative frequency. Rectangles of equal width are drawn to the height that is the frequency or relative frequency for each category. The bars do not touch each other.
- ✓ **Solution**
  - Figure 2.1 (a) shows the frequency bar graph.
  - Figure 2.1 (b) shows the relative frequency bar graph.

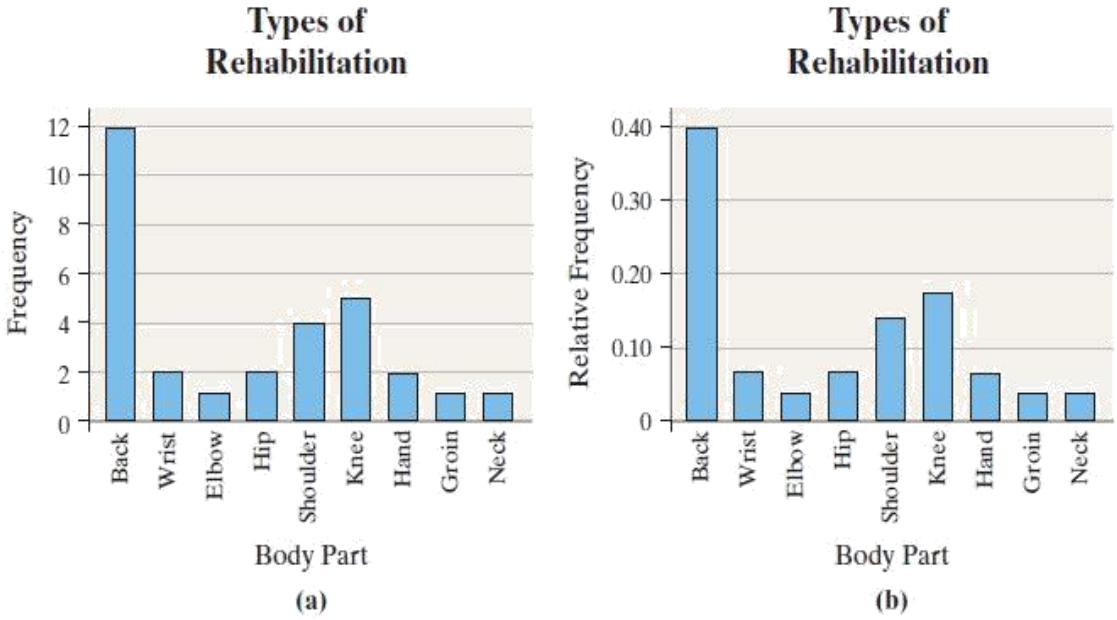


Figure 2.1

Some statisticians prefer to create bar graphs with the categories arranged in decreasing order of frequency. Such graphs help prioritize categories for decision making purposes in areas such as quality control, human resources, and marketing.

Definition

A ***Pareto chart*** is a bar graph whose bars are drawn in decreasing order of frequency or relative frequency.

Figure 2.2 illustrates a relative frequency Pareto chart for the data in Table 3.

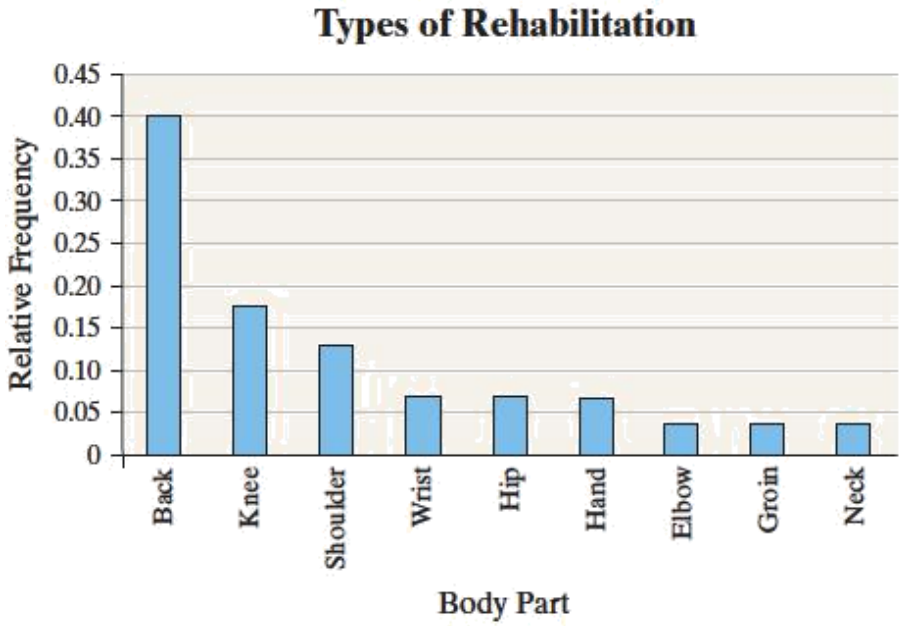


Figure 2.2

Side-by-Side Bar Graphs

Graphics can provide insight when you are comparing two sets of data. For example, suppose we wanted to know if more people are finishing college today than in 1990.

We could draw a **side-by-side bar graph** to compare the two data sets. Data sets should be compared by using relative frequencies, because different sample or population sizes make comparisons using frequencies difficult or misleading. However, when making comparisons, relative frequencies alone are not sufficient.

Example

Problem: The data in Table 4 represent the educational attainment in 1990 and 2006 of adults 25 years and older who are residents of the United States. The data are in thousands. So 16,502 represent 16,502,000.

a Draw a side-by-side relative frequency bar graph of the data.

b Are a greater proportion of Americans dropping out of college before earning a degree?

Table 4

Educational Attainment	1990	2006
Less than 9th grade	16,502	11,742
9th to 12th grade, no diploma	22,842	16,154
High school diploma	47,643	60,898
Some college, no degree	29,780	32,611
Associate's degree	9,792	16,760
Bachelor's degree	20,833	35,153
Graduate or professional degree	11,478	18,567
Totals	**158,870**	**191,885**

Source: U.S. Census Bureau

✓ **Approach**: First, we determine the relative frequencies of each category for each year. To construct the side-by-side bar graphs, we draw two bars for each category of data. One of the bars will represent 1990 and the other will represent 2006.

✓ **Solution**: Table 5 shows the relative frequency for each category. The side-by-side bar graph is shown in Figure 4.

Figure 4

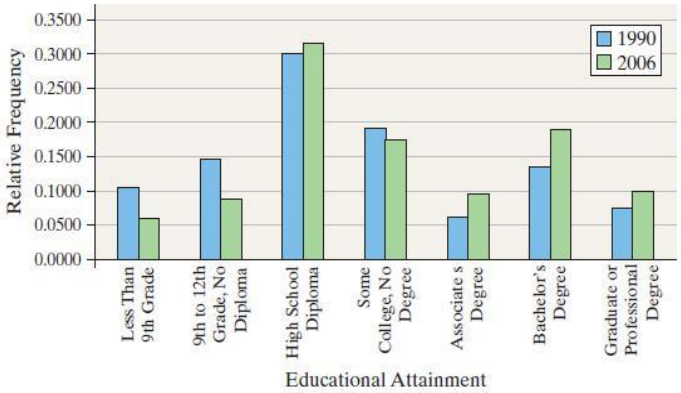


Table 5

Educational Attainment	1990	2006
Less than 9th grade	0.1039	0.0612
9th to 12th grade, no diploma	0.1438	0.0842
High school diploma	0.2999	0.3174
Some college, no degree	0.1874	0.1700
Associate's degree	0.0616	0.0873
Bachelor's degree	0.1311	0.1832
Graduate or professional degree	0.0722	0.0968

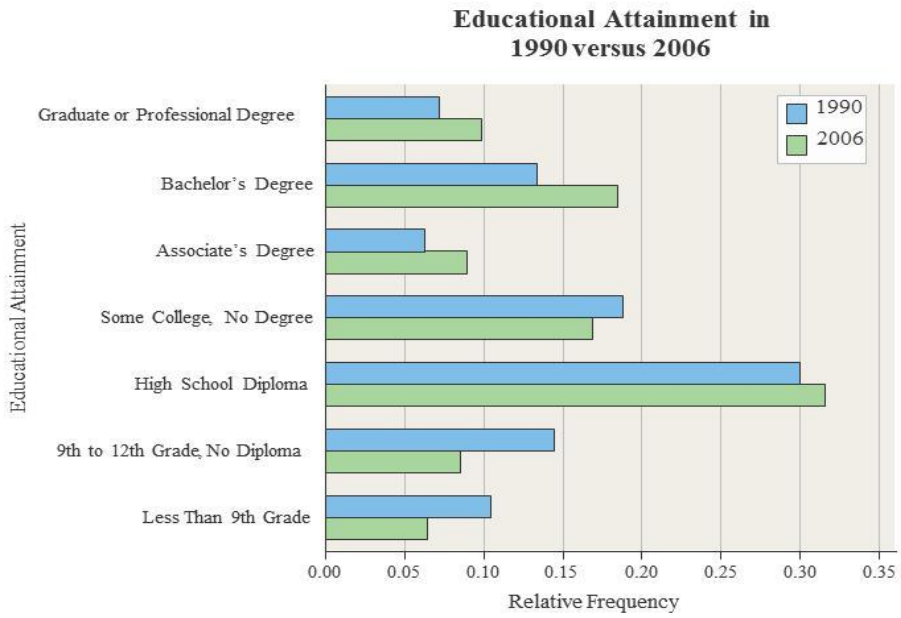
From the graph, we can see that the proportion of Americans 25 years and older who had some college, but no degree, was higher in 1990. This information is not clear from the frequency table, because the sizes of the populations are different. Increases in the number of American who

did not complete a degree are due partly to increases in the sizes of the populations.

Horizontal Bars

So far we have only looked at bar graphs with vertical bars. However, the bars may also be horizontal. Horizontal bars may be preferred when the category names are lengthy. For example, Figure 5 uses horizontal bars to display the same data as in Figure 4.

Figure 5



Construct Pie Charts

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

Definition

A **pie chart** is a circle divided into sectors. Each sector represents a category of data. The area of each sector is proportional to the frequency of the category.

Example

Problem: The data presented in Table 6 represent the educational attainment of residents of the United States 25 years or older in 2006, based on data obtained from the U.S. Census Bureau. The data are in thousands. Construct a pie chart of the data.

Table 6

Educational Attainment	2006
Less than 9th grade	11,742
9th to 12th grade, no diploma	16,154
High school diploma	60,898
Some college, no degree	32,611
Associate's degree	16,760
Bachelor's degree	35,153
Graduate or professional degree	18,567
Totals	**191,885**

✓ **Approach**: The pie chart will have seven parts, or sectors, corresponding to the seven categories of data. The area of each sector is proportional to the frequency of each category. For example,

$$11{,}742/191{,}885 = 0.0612$$

of all U.S. residents 25 years or older have less than a 9th-grade education. The category —less than 9th grade‖ will make up 6.12% of the area of the pie chart. Since a circle has 360 degrees, the degree measure of the sector for the category —less than 9th-grade‖ will be $(0.0612)360° \approx 22°$. Use a protractor to measure each angle.

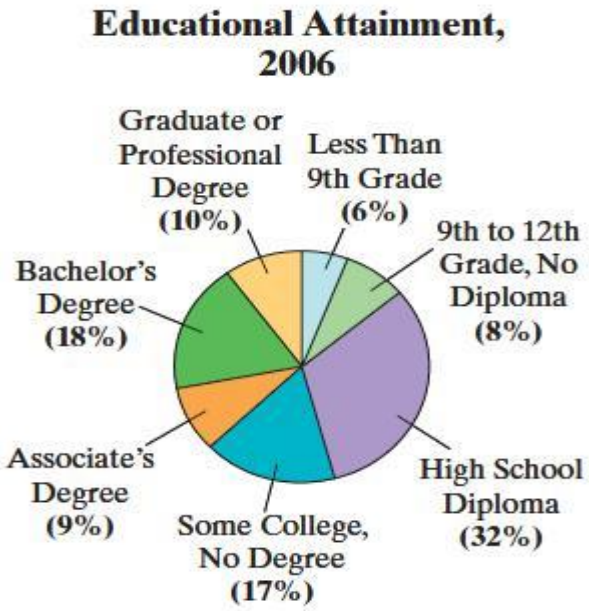
✓ **Solution**: We follow the approach presented for the remaining categories of data to obtain Table 7.

Table 7

Educational Attainment	Frequency	Relative Frequency	Degree Measure of Each Sector
Less than 9th grade	11,742	0.0612	22
9th to 12th grade, no diploma	16,154	0.0842	30
High school diploma	60,898	0.3174	114
Some college, no degree	32,611	0.1700	61
Associate's degree	16,760	0.0873	31
Bachelor's degree	35,153	0.1832	66
Graduate or professional degree	18,567	0.0968	35

To construct a pie chart by hand, we use a protractor to approximate the angles for each sector. See Figure 6.

Figure 6



2.1.2 ORGANIZING QUANTITATIVE DATA

In summarizing quantitative data, we first determine whether the data are discrete or continuous. If the data are discrete and there are relatively few different values of the variable, then the categories of data (called **classes**) will be the observations (as in qualitative data). If the data are discrete, but there are many different values of the variables or if the data are continuous, then the categories of data (the *classes*) must be created using intervals of numbers.

We will first present the techniques required to organize discrete quantitative data when there are relatively few different values and then proceed to organizing continuous quantitative data.

Organize Discrete Data in Tables

We use the values of a discrete variable to create the classes when the number of distinct data values is small.

Example

Constructing Frequency and Relative Frequency Distributions from Discrete Data

Problem: The administrator of Wendy's pharmacy is interested in studying the typical number of customers who arrive during the lunch hour. The data in Table 8 represent the number of customers who arrive at Wendy's for 40 randomly selected 15-minute intervals of time during lunch. For example, during one 15-minute interval, seven customers arrived. Construct a frequency and relative frequency distribution.

Table 8

Number of Arrivals at Wendy's							
7	6	6	6	4	6	2	6
5	6	6	11	4	5	7	6
2	7	1	2	4	8	2	6
6	5	5	3	7	5	4	6
2	2	9	7	5	9	8	5

- ✓ **Approach**: The number of people arriving could be 0, 1, 2, 3, …. From Table 8, we see that there are 11 categories of data from this study: 1, 2, 3, … , 11. We tally the number of observations for each category, count each tally, and create the frequency and relative frequency distributions.

- ✓ **Solution**: The frequency and relative frequency distributions are shown in Table 9.

Table 9

Number of Customers	Tally	Frequency	Relative Frequency
1	|	1	$\frac{1}{40} = 0.025$
2	𝍸 |	6	0.15
3	|	1	0.025
4	||||	4	0.1
5	𝍸 ||	7	0.175
6	𝍸 𝍸 |	11	0.275
7	𝍸	5	0.125
8	||	2	0.05
9	||	2	0.05
10		0	0.0
11	|	1	0.025

On the basis of the relative frequencies, 27.5% of the 15-minute intervals had 6 customers arrive at Wendy's during the lunch hour.

Construct Histograms of Discrete Data

As with qualitative data, quantitative data may be represented graphically. We begin our discussion with a graph called the *histogram*, which is similar to the bar graph drawn for qualitative data.

Definition

A **histogram** is constructed by drawing rectangles for each class of data. The height of each rectangle is the frequency or relative frequency of the class. The width of each rectangle is the same and the rectangles touch each other.

Example

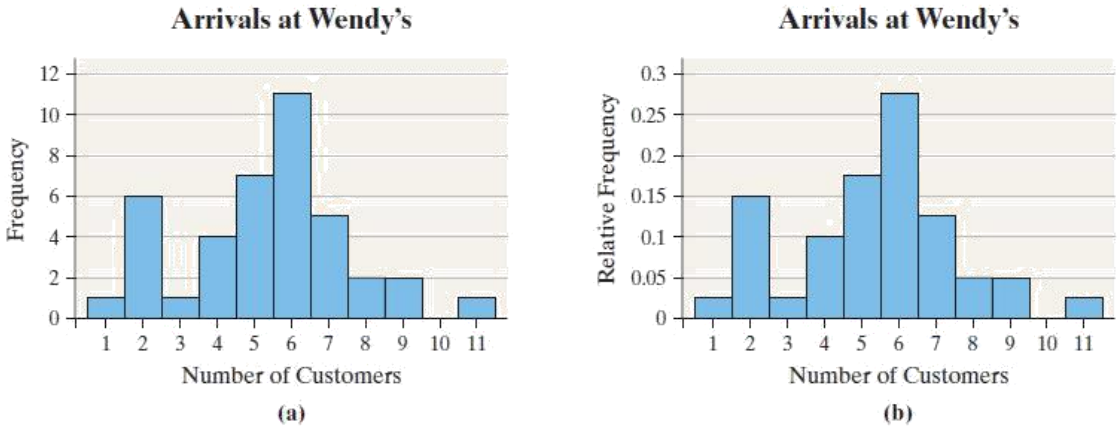
Drawing a Histogram for Discrete Data

Problem: Construct a frequency histogram and a relative frequency histogram using the data summarized in Table 9.

- ✓ **Approach**: On the horizontal axis, we place the value of each category of data (number of customers). The vertical axis will be the frequency or relative frequency of each category. Rectangles of equal width are drawn, with the center of each rec- tangle located at the value of each category. For example, the first rectangle is centered at 1. For the frequency histogram, the height of the rectangle will be the frequency of the category. For the relative frequency histogram, the height of the rectangle will be the relative frequency of the category. Remember, the rectangles touch for histograms.

- ✓ **Solution**: Figure 7(a) on the next page shows the frequency histogram. Figure 7(b) shows the relative frequency histogram.

Figure 7



Organize Continuous Data in Tables

Classes are the categories by which data are grouped.When a data set consists of a relatively small number of different discrete data values, the classes for the corresponding frequency distribution are predetermined to be those data values (as in Example 1). However, when a data set consists of a large number of different discrete data values or when a data set consists of continuous data, then no such predetermined classes exist. Therefore, the classes must be created by using intervals of numbers.

Table 10 is a typical frequency distribution created from continuous data. The data represent the number of U.S. residents between the ages of 25 and 74 who have earned a bachelor's degree or higher.The data are based on the Current Population Survey conducted in 2006.

In the table, we notice that the data are categorized, or grouped, by intervals of numbers. Each interval represents a class. For example, the first class is 25-to 34-year-old residents of the United States who have a bachelor's degree or higher.

We read this interval as follows: —The number of residents of the United States in 2006 who were between 25 and 34 years of age and have a bachelor's degree or higher was 11,806,000.‖ There are five classes in the table, each with a *lower class limit* and an *upper class limit*. The **lower class limit** of a class is the smallest

value within the class, while the **upper class limit** of a class is the largest value within the class. The lower class limit for the first class in Table 10 is 25; the upper class limit is 34. The **class width** is the difference between consecutive lower class limits. The class width for the data in Table 10 is Notice that the classes in Table 10 do not overlap. This is necessary to avoid confusion as to which class a data value belongs. Notice also that the class widths are equal for all classes.

Example

Organizing Continuous Data into a Frequency and Relative Frequency Distribution

Problem: Suppose you are considering investing in a Roth IRA. You collect the data in Table 12, which represent the 3-year rate of return (in percent, adjusted for sales charges) for a simple random sample of 40 small-capitalization growth mutual funds. Construct a frequency and relative frequency distribution of the data.

- ✓ **Approach**: To construct a frequency distribution, we first create classes of equal width. There are 40 observations in Table 12, and they range from 10.06 to 23.76, so we decide to create the classes such that the lower class limit of the first class is 10 (a little smaller than the smallest data value) and the class width is 2.

Table 12

Three-Year Rate of Return of Mutual Funds (as of 10/31/07)

13.50	13.16	10.53	14.74	13.20	12.24	12.61	19.11
14.47	12.29	13.92	16.16	12.07	10.99	15.07	10.06
14.14	12.77	19.74	12.76	13.34	11.32	15.41	17.37
13.51	15.44	15.10	17.13	12.37	16.34	11.34	10.57
15.70	13.28	23.76	22.68	14.81	23.54	19.65	14.07

Source: TD Ameritrade

There is nothing magical about the choice of 2 as a class width. We could have selected a class width of 8 (or any other class width, as well).We choose a class width that we think will nicely summarize the data. If our choice

doesn't accomplish this, we can always try another one. The lower class limit of the second class will be 10 + 2 = 12. Because the classes must not overlap, the upper class limit of the first class is 11.99. Continuing in this fashion, we obtain the following classes:

10 - 11.99

12 - 13.99

.

.

.

22 - 23.99

This gives us seven classes. We tally the number of observations in each class, count the tallies, and create the frequency distribution. The relative frequency distribution would be created by dividing each class's frequency by 40, the number of observations.

✓ **Solution:** We tally the data as shown in the second column of Table 13. The third column in the table shows the frequency of each class. From the frequency distribution, we conclude that a 3-year rate of return between 12% and 13.99% occurs with the most frequency. The fourth column in the table shows the relative frequency of each class. So, 35% of the small-capitalization growth mutual funds had a 3-year rate of return between 12% and 13.99%.

Table 13

Class (3-year rate of return)	Tally	Frequency	Relative Frequency
10–11.99	𝍸 |	6	6/40 = 0.15
12–13.99	𝍸 𝍸 ||||	14	14/40 = 0.35
14–15.99	𝍸 𝍸	10	10/40 = 0.25
16–17.99	||||	4	0.1
18–19.99	|||	3	0.075
20–21.99		0	0
22–23.99	|||	3	0.075

Three mutual funds had 3-year rates of return between 22% and 23.99%.We might consider these mutual funds worthy of our investment. This type of information would be more difficult to obtain from the raw data.

Guidelines for Determining the Lower Class Limit of the First Class and Class Width

Choosing the Lower Class Limit of the First Class

Choose the smallest observation in the data set or a convenient number slightly lower than the smallest observation in the data set. For example, in Table 12, the smallest observation is 10.06. A convenient lower class limit of the first class is 10.

Determining the Class Width

(ϕ) Decide on the number of classes. Generally, there should be between 5 and 20 classes. The smaller the data set, the fewer classes you should have. For example, we might choose 8 classes for the data in Table 12.

(γ) Determine the class width by computing

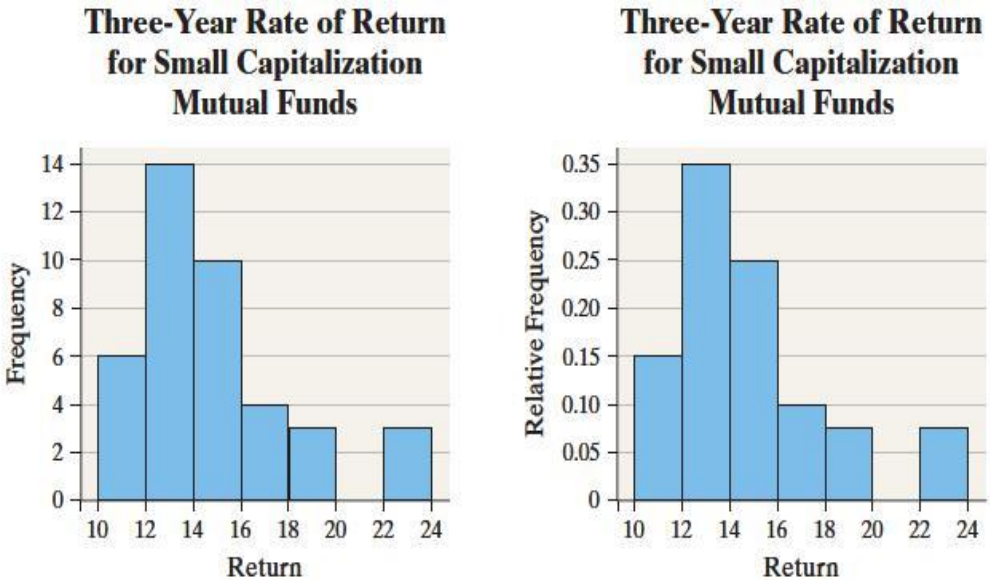
$$class - width \approx \frac{Largest\ data\ value - smallest\ data\ value}{number\ of\ classes}$$

Round this value *up* to a convenient number. For example, using the data in Table 12, we obtain class width $\approx \frac{23.76 - 10.06}{8} = 1.7125$ .We would round this up to 2 because this is an easy number to work with. Rounding up may result in fewer classes than were originally intended.

To draw the frequency histogram, we will use the frequency distribution in Table 13. We label the lower class limits of each class on the horizontal axis. Then, for each class, we draw a rectangle whose width is the class width and whose height is the frequency. To construct the relative frequency histogram, we let the height of the rectangle be the relative frequency, instead of the frequency.

Figure 8(a) represents the frequency histogram, and Figure 8(b) represents the relative frequency histogram.

Figure 8



Draw Stem-and-Leaf Plots

A **stem-and-leaf plot** is another way to represent quantitative data graphically. In a stem-and-leaf plot (sometimes called simply a *stem plot*), we use the digits to the left of the rightmost digit to form the **stem**. Each rightmost digit forms a **leaf**. For example, a data value of 147 would have 14 as the stem and 7 as the leaf.

Example
Constructing a Stem-and-Leaf Plot

Problem: The data in Table 16 represent the two-year average percentage of persons living in poverty, by state for the years 2005–2006. Draw a stem-and-leaf plot of the data.

✓ **Approach**

Step 1: We will treat the integer portion of the number as the stem and the decimal portion as the leaf. For example, the stem of Alabama will be 15 and the leaf will be 5.The stem of 15 will include all data from 15.0 to 15.9.

Step 2: Write the stems vertically in ascending order, and then draw a vertical line to the right of the stems.

Step 3: Write the leaves corresponding to the stem.

Step 4: Within each stem, rearrange the leaves in ascending order.Title the plot and provide a legend to indicate what the values represent.

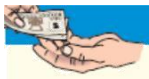


Table 16

Two-Year Average Percentage of Persons Living in Poverty (2005–2006)

State	Percent	State	Percent	State	Percent
Alabama	15.5	Kentucky	15.8	North Dakota	11.3
Alaska	9.4	Louisiana	17.6	Ohio	12.2
Arizona	14.8	Maine	11.4	Oklahoma	15.4
Arkansas	15.8	Maryland	9.1	Oregon	11.9
California	12.7	Massachusetts	11.1	Pennsylvania	11.3
Colorado	10.6	Michigan	12.6	Rhode Island	11.3
Connecticut	8.7	Minnesota	8.1	South Carolina	13.1
Delaware	9.3	Mississippi	20.4	South Dakota	11.3
D.C.	19.8	Missouri	11.5	Tennessee	14.9
Florida	11.3	Montana	13.7	Texas	16.3
Georgia	13.5	Nebraska	9.9	Utah	9.2
Hawaii	8.9	Nevada	10.1	Vermont	7.7
Idaho	9.7	New Hampshire	5.5	Virginia	8.9
Illinois	11.0	New Jersey	7.8	Washington	9.1
Indiana	11.6	New Mexico	17.4	West Virginia	15.3
Iowa	10.8	New York	14.3	Wisconsin	10.2
Kansas	12.7	North Carolina	13.5	Wyoming	10.3

Source: U.S. Census Bureau, *Current Population Survey*, 2006

Figure 10

Percentage of Persons Living in Poverty

5	5 | 5	5 | 5
6	6 |	6 |
7	7 | 8 7	7 | 7 8
8	8 | 7 9 1 9	8 | 1 7 9 9
9	9 | 4 3 7 1 9 2 1	9 | 1 1 2 3 4 7 9
10	10 | 6 8 1 2 3	10 | 1 2 3 6 8
11	11 | 3 0 6 4 1 5 3 9 3 3 3	11 | 0 1 3 3 3 3 3 4 5 6 9
12	12 | 7 7 6 2	12 | 2 6 7 7
13	13 | 5 7 5 1	13 | 1 5 5 7
14	14 | 8 3 9	14 | 3 8 9
15	15 | 5 8 8 4 3	15 | 3 4 5 8 8
16	16 | 3	16 | 3
17	17 | 6 4	17 | 4 6
18	18 |	18 |
19	19 | 8	19 | 8
20	20 | 4	20 | 4
		Legend: 5/5 represennts 5.5%
(a)	(b)	(c)

✓ **Solution**

Step 1: The stem from Alabama is 15 and the corresponding leaf is 5.The stem from Alaska is 9 and its leaf is 4, and so on.

Step 2: Since the lowest data value is 5.5 and the highest data value is 20.4, we need the stems to range from 5 to 20.We write the stems vertically in Figure 10(a), along with a vertical line to the right of the stem.

Step 3: We write the leaves corresponding to each stem. See Figure 10(b).

Step 4: We rearrange the leaves in ascending order, give the plot a title, and provide a legend. See Figure 10(c).

The following summarizes the method for constructing a stem-and-leaf plot.

Construction of a Stem-and-Leaf Plot

Step 1: The stem of a data value will consist of the digits to the left of the right-most digit. The leaf of a data value will be the rightmost digit.

Step 2: Write the stems in a vertical column in increasing order. Draw a vertical line to the right of the stems.

Step 3: Write each leaf corresponding to the stems to the right of the vertical line.

Step 4: Within each stem, rearrange the leaves in ascending order, title the plot, and provide a legend to indicate what the values represent.

Draw Dot Plots

A **dot plot** is drawn by placing each observation horizontally in increasing order and placing a dot above the observation each time it is observed. Though limited in usefulness, dot plots can be used to quickly visualize the data.

Example

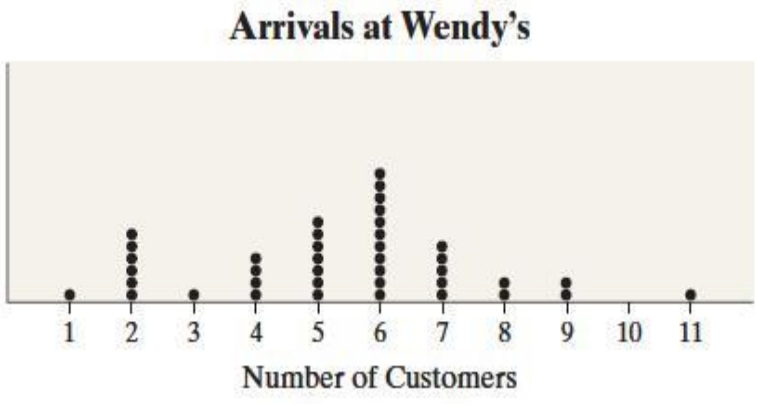
Drawing a Dot Plot

Problem: Draw a dot plot for the number of arrivals at Wendy's data from Table 8 on page 82.

Approach: The smallest observation in the data set is 1 and the largest is 11.We write the numbers 1 through 11 horizontally. For each observation, we place a dot above the value of the observation.

Solution: Figure 15 shows the dot plot

Figure 15



49

NB: A statistical software will be used in class to show how these graphs are drawn.

Assignment 2

(i) During clinical trials, observed adverse effects are often classified by the following scale:

Mild: Experience was trivial and did not cause any real problem.

Moderate: Experience was a problem but did not interfere significantly with patient's daily activities or clinical status.

Severe: Experience interfered significantly with the normal daily activities or clinical status.

Based on 1109 patients involved in the Phase I and II clinical trials for bigomycin, it was observed that 810 experienced no adverse effects, while 215, 72, and 12 subjects suffered from mild, moderate, and severe adverse effects, respectively. **Prepare visual and tabular presentations for this data.**

(a)The following assay results (percentage of label claim) were observed in 50 random samples during a production run.

102 100 96 99 101 102 100 105 97 100

92 103 101 100 99 102 96 100 101 98

107 95 98 100 100 99 97 104 101 103

98 101 100 105 99 101 102 100 87 98

101 103 93 99 101 97 100 102 99 104

Report these results as a stem plot, and histogram.

2.2 Summarizing Data Numerically

When we look at a distribution of data, we should consider three characteristics of the distribution: shape, center, and spread. In Section 1, we discussed methods for organizing raw data into tables and graphs.

These graphs (such as the histogram) allow us to identify the shape of the distribution. The center and spread are numerical summaries of the data.

The center of a data set is commonly called the *average*. There are many ways to describe the *average* value of a distribution. In addition, there are many ways to measure the spread of a distribution. The most appropriate measure of center and spread depends on the shape of the distribution.

Once these three characteristics of the distribution are known, we can analyze the data for interesting features, including unusual data values, called *outliers*.

MEASURING CENTER

Measures of center are numerical values that tend to report in some sense the middle of a set of data, we will focus on the *mean* and the *median*. If the data are a sample, the mean and median would be called *statistics*. If the data form an entire population then these measures of center would be called *parameters*.

Mean

DEFINITION:

The mean of a set of n observations is simply the sum of the observations divided by the number of observations, n.

Special notation:

If $x_1, x_2, \ldots, x_n$ denote a sample of n observations, then the *mean of the sample* is called "x-bar" and is denoted by:

$$\bar{x} = \frac{\sum x_i}{n} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The *mean of a population* is denoted by the Greek letter μ.

Properties and uses of the arithmetic mean

- The mean has excellent statistical properties and is commonly used in additional statistical manipulations and analyses. One such property is called the *centering property of the mean.* When the mean is subtracted from each observation in the data set, the sum of these differences is zero (i.e., the negative sum is equal to the positive sum). This demonstrates that the mean is the arithmetic center of the distribution.
- Because of this centering property, the mean is sometimes called the *center of gravit*y of a frequency distribution. If the frequency distribution is plotted on a graph, and the graph is balanced on a fulcrum, the point at which the distribution would balance would be the mean.
- The arithmetic mean is the best descriptive measure for data that are normally distributed.
- On the other hand, the mean is not the measure of choice for data that are severely skewed or have extreme values in one direction or another. Because the arithmetic mean uses all of the observations in the distribution, it is affected by any extreme value. Suppose that the last value in the previous distribution was 131 instead of 31. The mean would be 225 / 5 = 45.0 rather than 25.0. As a result of one extremely large value, the mean is much larger than all values in the distribution except the extreme value (the "outlier").

Example

Mean Number of Children per Household

Problem

Suppose that the number of children in a simple random sample of 10 households is as follows: 2, 3, 0, 2, 1, 0, 3, 0, 1, 4

a. Calculate the sample mean number of children per household.

b. Interpret your answer.

c. Suppose that the observation for the last household in the above list was incorrectly recorded as 40 instead of 4.What would happen to the mean?

✓ **Solution**

a. The sample mean number of children per household is given by:

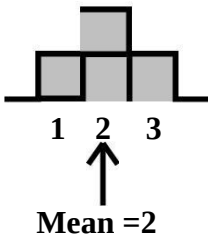
$$\bar{x} = \frac{2+3+0+2+1+0+3+0+1+4}{10} = \frac{16}{10} = 16.$$

b. We expect about 1.6 children per household, on average. We report 1.6 even though it is not possible to have 1.6 children in any one given household; that is, the 1.6 is not rounded up to say 2.We are reporting a value that we would expect *on average,* over many samples of 10 households.

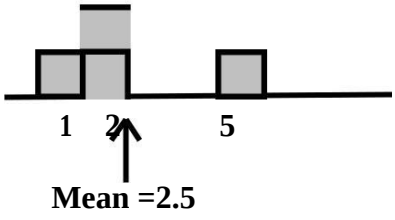
c. The sample mean would now be given by:

$$\bar{x} = \frac{2+3+0+2+1+0+3+0+1+40}{10} = \frac{52}{10} = 5.2$$

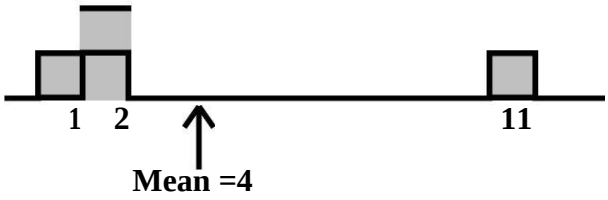
Note that 9 of the 10 observations are less than the mean. The mean is *sensitive to extreme observations*. Most graphical displays would have detected this out-lying observation. The *mean* = the *point of equilibrium,* the point where the distribution would balance.



If the distribution is *symmetric*, as in the first picture at the left, the mean would be exactly at the center of the distribution.



As the largest observation is moved further to the right, making this observation somewhat extreme, the *mean shifts towards the extreme observation*.



If a distribution appears to be skewed, we may wish also to report a more *resistant* measure of center.

Exercise 2.5

Determine the mean for the same set of vaccination data.

2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

Other Means

Weighted Mean

With the *arithmetic* mean, we usually give each observation equal weight. If we believe that the values to be averaged do not carry the same weight, then we should use a weighted average.

Illustration

The average of 3 cholesterol readings, 210, 180 and 270 is (660)/3 = 220. Suppose that the value of 210 is really the average of two values (200 and 220), we might want to consider giving this value twice as much weight as the other two values, resulting in an average

$$[210 + 210 + 180 + 270] / 4 = 217.5$$
or
$$[2 \times 210 + 180 + 270] /[2 + 1 + 1] = 217.5$$

The formula for a weighted average, $\bar{X}w$ is

$$\bar{X}_w = \frac{\sum(w_i X_i)}{\sum(w_i)}$$

where *Wi* is the weight assigned to the value *Xi*.

Harmonic mean

The harmonic mean is defined as

$$\frac{N}{\sum 1/\ x_i}$$

Geometric Mean

Properties and uses of the geometric mean

The geometric mean is the average of logarithmic values, converted back to the base. The geometric mean tends to dampen the effect of extreme values and is always smaller than the corresponding arithmetic mean. In that sense, the geometric mean is less sensitive than the arithmetic mean to one or a few extreme values.

• The geometric mean is the measure of choice for variables measured on an exponential or logarithmic scale, such as dilutional titers or assays.

• The geometric mean is often used for environmental samples, when levels can range over several orders of magnitude. For example, levels of coliforms in samples taken from a body of water can range from less than 100 to more than 100,000.

Exercise 2.6

Using the dilution titers shown below, calculate the geometric mean titer of convalescent antibodies against tularemia among 10 residents of Martha's Vineyard. [Hint: Use only the second number in the ratio, i.e., for 1:640, use 640.]

ID #	Acute	Convalescent
1	1:16	1:512
2	1:16	1:512
3	1:32	1:128
4	not done	1:512
5	1:32	1:1024
6	"negative"	1:1024
7	1:256	1:2048
8	1:32	1:128
9	"negative"	1:4096
10	1:16	1:1024

A measure of center that is more resistant to extreme values is the *median*.

Median

DEFINITION:

The *median* of a set of *n* observations, ordered from smallest to largest, is a value such that half of the observations are less than or equal to that value and half the observations are greater than or equal to that value.

If the number of observations is *odd*, the median is the middle observation. If the number of observations is *even*, the median is any number between the two middle observations, including either of the two middle observations.
To be consistent, we will define the median as the mean or average of the two middle observations.

Location of the median: $(n+1)/2$, where n is the number of observations.

Properties and uses of the median

• The median is a good descriptive measure, particularly for data that are skewed, because it is the central point of the distribution.

• The median is relatively easy to identify. It is equal to either a single observed value (if odd number of observations) or the average of two observed values (if even number of observations).

• The median, like the mode, is not generally affected by one or two extreme values (outliers). For example, if the values on the previous page had been 4, 23, 28, 31, and 131 (instead of 31), the median would still be 28.

• The median has less-than-ideal statistical properties. Therefore, it is not often used in statistical manipulations and analyses.

Do It!

Median Number of Children per Household

Find the *median* number of children in a household from this sample of 10 households, that is, find the median of

Observation Number:	1	2	3	4	5	6	7	8	9	10
Number of Children:	2	3	0	1	4	0	3	0	1	2

(a) Order the observations from smallest to largest:

(b) Calculate $(n+1)/2$ = _____

(c) Median = _____

(d) What happens to the median if the fifth observation in the first list was incorrectly recorded as 40 instead of 4?

(e) What happens to the median if the third observation in the first list was incorrectly recorded as -20 instead of 0?

Note: The *median is resistant*—that is, it does not change, or changes very

little, in response to extreme observation

Exercise 2.4

Determine the median for the same vaccination data used in Exercises 2.2. and 2.3.

2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

The Mode

DEFINITION:

The *mode* of a set of observations is the most frequently occurring value; it is the value having the highest frequency among the observations.

The *mode* of the values: { 0, 0, 0, 0, 1, 1, 2, 2, 3, 4 } is 0

For { 0, 0, 0, 1, 1, 2, 2, 2, 3, 4 } two modes, 0 and 2 (*bimodal)*

The *mode* is not often used as a measure of center for quantitative data. The *mode* can be computed for *qualitative* data. The modal race category is ―white.‖ If categories were given coded as:

1=White,

2=Asian,

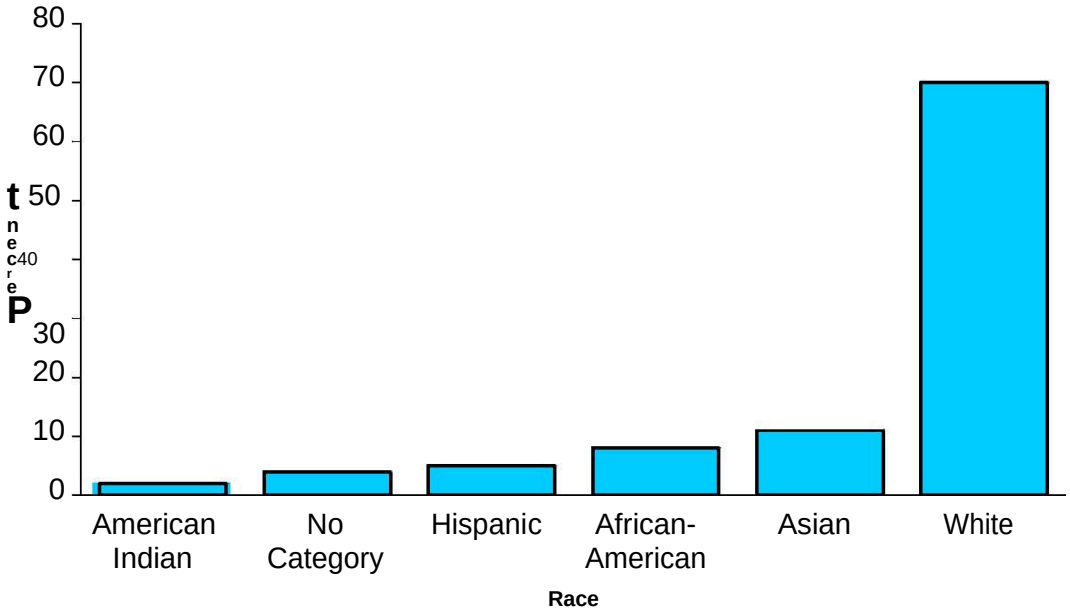
3=African-American,

4=Hispanic,

5=American Indian,

6=No category listed,

then the mode would be the value 1.



Properties and uses of the mode

The mode is the easiest measure of central location to understand and explain. It is also the easiest to identify, and requires no calculations.

• The mode is the preferred measure of central location for addressing which value is the most popular or the most common. For example, the mode is used to describe which day of the week people most prefer to come to the influenza vaccination clinic, or the "typical" number of doses of DPT the children in a particular community have received by their second birthday.

• As demonstrated, a distribution can have a single mode. However, a distribution has more than one mode if two or more values tie as the most frequent values. It has no mode if no value appears more than once.

• The mode is used almost exclusively as a "descriptive" measure. It is almost never used in statistical manipulations or analyses.

• The mode is not typically affected by one or two extreme values (outliers).

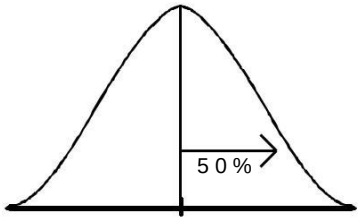
Exercise 2.3

Using the same vaccination data as in Exercise 2.2, find the mode. (If you answered Exercise 2.2, find the mode from your frequency distribution.)

2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

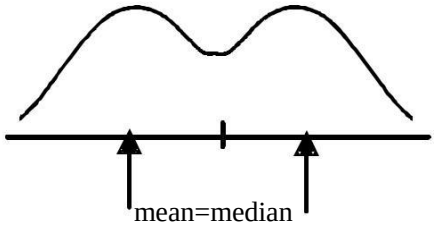
Which Measure of Center to Use?

Bell-shaped, Symmetric



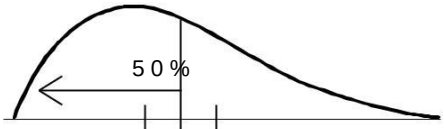
mean=median=mode

Bimodal



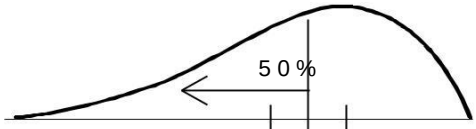
two modes

Skewed Right



mode mean median

Skewed Left



mean mode median

Mean, Median, and Mode

The most common measure of center is the mean, which locates the balancing point of the distribution. The mean equals the sum of the observations, divided by how many there are. The mean is also affected by extreme observations (outliers and values which are far in the tail of a distribution that is skewed). So the mean tends to be a good choice for locating the center of a distribution that is unimodal and roughly symmetric, with no outliers.

The median is a more robust measure of center, that is, it is not influenced by extreme values. The median is the middle observation when the data are ordered from smallest to largest. If you have an odd number of values, the median is the one in the middle. If you have an even number of values, the median is the mean of the two middle values, and fall exactly half way between them. If you have n observations, then (n+1)/2 tells you the location or position of the median. For skewed distributions or distributions with outliers, the median tends to be the better choice for locating the center.

The mode is the value(s) that occurs most often. For a distribution, the mode is the value associated with the highest peak. The most frequent value can be far from the center of the distribution, so the mode is not really a measure of center. However, the mode is the only measure of the three that can be used for qualitative data.

Tips:

☎☋☽ When you see or hear an —averageǁ reported, ask which average was really computed -- the mean or the median.

☎♌☽ Think about or examine the distribution of values to assess if the measure of center used is appropriate.

Exercise 2.7

For each of the variables listed below from the line listing in Table 2.9, identify which measure of central location is best for representing the data.

A. Mode
B. Median
C. Mean
D. Geometric mean
E. No measure of central location is appropriate

_____ 1. Year of diagnosis

_____ 2. Age (years)

_____ 3. Sex

_____ 4. Highest IFA titer

_____ 5. Platelets x 10^6/L

_____ 6. White blood cell count x 10^9/L

Table 2.9 Line Listing for 12 Patients with Human Monocytotropic Ehrlichiosis — Missouri, 1998–1999

Patient ID	Year of Diagnosis	Age (years)	Sex	Highest IFA* Titer	Platelets x 10^6/L	White Blood Cell Count x 10^9/L
01	1999	44	M	1:1024	90	1.9
02	1999	42	M	1:512	114	3.5
03	1999	63	M	1:2048	83	6.4
04	1999	53	F	1:512	180	4.5
05	1999	77	M	1:1024	44	3.5
06	1999	43	F	1:512	89	1.9
10	1998	22	F	1:128	142	2.1
11	1998	59	M	1:256	229	8.8
12	1998	67	M	1:512	36	4.2
14	1998	49	F	1:4096	271	2.6
15	1998	65	M	1:1024	207	4.3
18	1998	27	M	1:64	246	8.5
Mean:	1998.5	50.92	na	1:976.00	144.25	4.35
Median:	1998.5	51	na	1:512	128	3.85
Geometric Mean:	1998.5	48.08	na	1:574.70	120.84	3.81
Mode:	none	none	M	1:512	none	1.9, 3.5

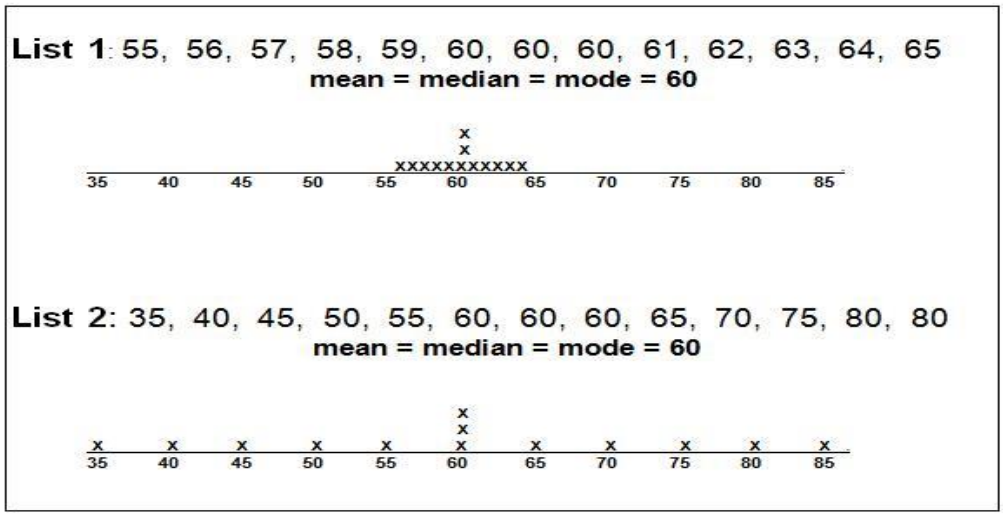
* Immunofluorescence assay

Data Source: Olano JP, Masters E, Hogrefe W, Walker DH. Human monocytotropic ehrlichiosis, Missouri. Emerg Infect Dis 2003;9:1579-86.

MEASURING VARIATION OR SPREAD

Both sets of data have the same mean, median and mode but the values obviously differ in another respect, the variation or spread of the values.

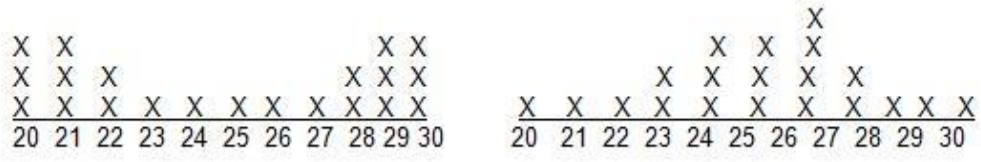
The values in List 1 are much more tightly clustered around the center value of 60. The values in List 2 are much more dispersed or spread out.



Range

The *range* is the simplest measure of variability or spread. Range is just the difference between the largest value and the smallest value. Range can give a distorted picture of the actual pattern of variation.

Two distributions: same range but different patterns of variation. The first distribution has most of its values far from the center, while the second distribution has most of its values closer to the center.



Interquartile Range

The interquartile range measures the spread of the middle 50% of the data. You first find the median (represented by Q2, the value that divides the data into two halves), and then find the median for each half.The three values that divide the data into four parts are called the *quartiles*, represented by Q1, Q2, and Q3. The difference between the third quartile and the first quartile is called the *interquartile range*, denoted by IQR=Q3-Q1.

Properties and uses of the interquartile range

• The interquartile range is generally used in conjunction with the median. Together, they are useful for characterizing the central location and spread of any frequency distribution, but particularly those that are skewed.

• For a more complete characterization of a frequency distribution, the 1st and 3rd quartiles are sometimes used with the minimum value, the median, and the maximum value to produce a five-number summary of the distribution. For example, the five-number summary for the length of stay data is:

Minimum value = 0,

Q1 = 6.75,

Median = 10,

Q3 = 14.5, and

Maximum value = 49.

Together, the five values provide a good description of the center, spread, and shape of a distribution. These five values can be used to draw a graphical illustration of the data, as in the boxplot in Figure 2.8.

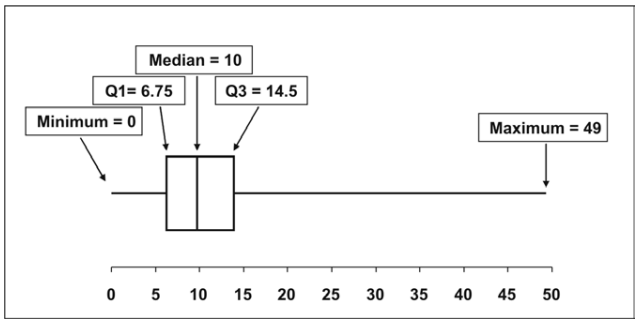


Figure 2.8

Finding the Quartiles

(a) Find the median of all of the observations.
(b) First Quartile = Q1 = median of observations that fall below the median.
(c) Third Quartile = Q3 = median of observations that fall above the median.

Notes

(a) When the number of observations is odd, the middle observation is the median. This observation is not included in either of the two halves when computing **Q1** and **Q3**.

(b) Although different books, calculators, and computers may use slightly different ways to compute the quartiles, they are all based on the same idea.

(c) In a left-skewed distribution, the first quartile will be farther from the median than the third quartile is. If the distribution is symmetric, the quartiles should be the same distance from the median.

Example Quartiles for Age

The ages of the 20 subjects in the medical study are listed below in order.

32, 37, 39, 40, 41, 41, 41, 42, 42, 43,

44, 45, 45, 45, 46, 47, 47, 49, 50, 51

The histogram of the ages is also provided.

a. Calculate the median age.

b. Calculate the first Quartile Q1 for this age data.

c. Calculate the third Quartile Q3 for this age data.

d. Calculate the range for this age data.

Exercise 2.8

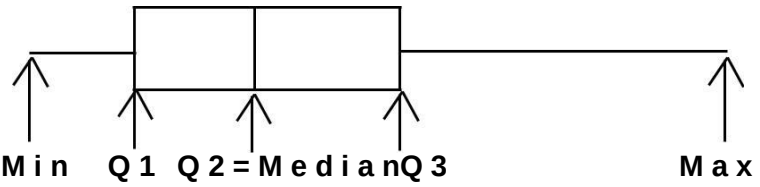
Determine the first and third quartiles and interquartile range for the same vaccination data as in the previous exercises.
2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

Five-Number Summary

Five-number summary:

Minimum, Q1, Median, Q3, Maximum

Box plot:



To Build a Basic Boxplot

List the data values in order from smallest to largest.

☎♓☽ Find the five number summary: minimum, Q1, median, Q3, and maximum.

☎♓♓☽ Locate the values for Q1, the median and Q3 on the scale. These values determine the —box‖ part of the boxplot. The quartiles determine the ends of the box, and a line is drawn inside the box to mark the value of the median.

☎♓♓♓☽ Draw lines (called whiskers) from the midpoints of the ends of the box out to the minimum and maximum.

Using the 1.5 x IQR Rule to Identify Outliers and Build a Modified Boxplot

☎♋☽ List the data values in order from smallest to largest.

☎♌☽ Find the five number summary: minimum, Q1, median, Q3, and maximum.

☎♍☽ Locate the values for Q1, the median and Q3 on the scale. These values determine the —box‖ part of the boxplot. The quartiles determine the ends of the box, and a line is drawn inside the box to mark the value of the median.

☎♎☽ Find the IQR = Q3 – Q1.

☎♌☽ Compute the quantity STEP = 1.5 x (IQR

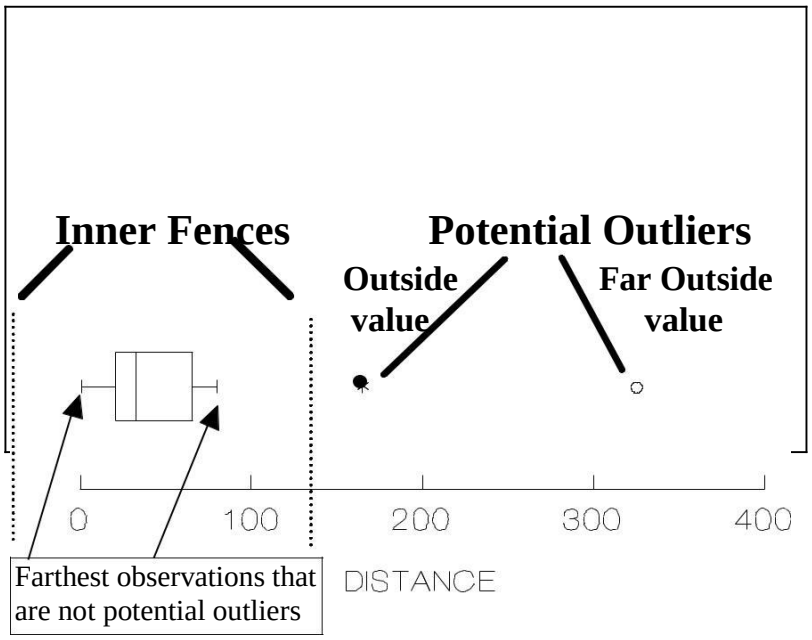
☎♍☽ Find the location of the *inner fences* by taking 1 step out from each of the quartiles

lower inner fence = Q1 – STEP;

upper inner fence = Q3 + STEP.

☎♎☽ Draw the lines (whiskers) from the midpoints of the ends of the box out to the smallest and largest values WITHIN the inner fences.

☎♏☽ Observations that fall OUTSIDE the inner fences are considered potential outliers. If there are any outliers, plot them individually along the scale using a solid dot.



Five-number summary:

min=1

Q1=21

median=32

Q3=66

max=325

Example: Any Age Outlier?

Let's apply the "rule of thumb" to our age data set to assess if there are any outliers.

a. Construct the fences for the modified boxplot based on the 1.5 * IQR rule.

b. Are there any outliers using the 1.5 * IQR rule?

Deviations: -4, 1, 3

Squared Deviations: 16, 1, 9

Observation x	Deviation $x - \bar{x}$	Squared Deviation $(x - \bar{x})^2$
0	0 - 4 = -4	16
5	5 - 4 = 1	1
7	7 - 4 = 3	9

mean = 4 sum always = 0 sum = 26

$$\text{sample variance} = \frac{(-4)^2 + (1)^2 + (3)^2}{3-1} = \frac{16+1+9}{2} = \frac{26}{2} = 13$$

$$\text{sample standard deviation} = \sqrt{13} \cong 3.6$$

Interpretation of the Standard Deviation

Think of the standard deviation as roughly an average distance of the observations from their mean. If all of the observations are the same, then the standard deviation will be 0 (i.e. no spread). Otherwise the standard deviation is positive and the more spread out the observations are about their mean, the larger the value of the standard deviation.

Special Notation

If $x_1, x_2, \ldots, x_n$ denote a sample of n observations, the sample variance is denoted by:

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{n\sum(x_i^2) - (\sum x_i)^2}{n(n-1)}$$

Sample standard deviation, denoted by s, is the square root of the variance:

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

The population standard deviation, denoted by the Greek letter σ (sigma), is the square root of the population variance and is computed as:

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}.$$

Remarks:

(ι) *The variance is measured in squared units. By taking the square root of the variance we bring this measure of spread back into the original units.*

(ιι) *Just as the mean is not a resistant measure of center, since the standard deviation used the mean in its definition, it is not a resistant measure of spread. It is heavily influenced by extreme values.*

- *There are statistical arguments that support why we divide by n –1 instead of n in the denominator of the sample standard deviation.*

Exercise 2.9
Calculate the standard deviation for the same set of vaccination data.
2, 0, 3, 1, 0, 1, 2, 2, 4, 8, 1, 3, 3, 12, 1, 6, 2, 5, 1

IQR and Standard Deviation

The interquartile range, IQR, is the distance between the first and third quartiles (Q3 - Q1), and measures the spread of the middle 50% of the data. When the median is used as a measure of center, the IQR is often used as a measure of spread. For skewed distributions, or distributions with outliers, the IQR tends to be a better measure of spread if your goal is to summarize that distribution.

Adding the minimum and maximum values to the median and quartiles results in the five-number summary. A graphical display of the five-number summary is a *boxplot*, and the length of the box corresponds to the IQR.

The standard deviation is roughly the average distance of the observed values from their mean. The mean and the standard deviation are most useful for approximately symmetric distributions with no outliers.

Standard error of the mean

Definition of standard error

The standard deviation is sometimes confused with another measure with a similar name — the standard error of the mean. However, the two are not the same. The standard deviation describes variability in a set of data. The standard error of the mean refers to variability we might expect in the arithmetic means of repeated samples taken from the same population.

The standard error assumes that the data you have is actually a sample from a larger population. According to the assumption, your sample is just one of an infinite number of possible samples that could be taken from the source population. Thus, the mean for your sample is just one of an infinite number of other sample means. The standard error quantifies the variation in those sample means.

Method for calculating the standard error of the mean

Step 1. Calculate the standard deviation.

Step 2. Divide the standard deviation by the square root of the number of observations (n).

Properties and uses of the standard error of the mean

• The primary practical use of the standard error of the mean is in calculating confidence intervals around the arithmetic mean. (Confidence intervals are addressed in the next section.)

Exercise 2.10

When the serum cholesterol levels of 4,462 men were measured, the mean cholesterol level was 213, with a standard deviation of 42. Calculate the standard error of the mean for the serum cholesterol level of the men studied.

> ***Tip****: The numerical summaries presented in this chapter provide information about the center and spread of a distribution, but a graph, such as a histogram or stem-and-leaf plot, provides the best picture of the overall shape of the distribution.*

Exercise 2.11

The data in Table 2.13 (on page 2-57) are from an investigation of an outbreak of severe abdominal pain, persistent vomiting, and generalized weakness among residents of a rural village. The cause of the outbreak was eventually identified as flour unintentionally contaminated with lead dust.

1. Summarize the blood level data with a frequency distribution.

2. Calculate the arithmetic mean. [Hint: Sum of known values = 2,363]

3. Identify the median and interquartile range.

4. Calculate the standard deviation. [Hint: Sum of squares = 157,743]

5. Calculate the geometric mean using the log lead levels provided. [Hint: Sum of log lead levels = 68.45]

Table 2.13 Age and Blood Lead Levels (BLLs) of Ill Villagers and Family Members — Country X, 1996

ID	Age (Years)	BLL†	Log_{10}BLL	ID	Age (Years)	BLL	Log_{10}BLL
1	3	69	1.84	22	33	103	2.01
2	4	45	1.66	23	33	46	1.66
3	6	49	1.69	24	35	78	1.89
4	7	84	1.92	25	35	50	1.70
5	9	48	1.68	26	36	64	1.81
6	10	58	1.77	27	36	67	1.83
7	11	17	1.23	28	38	79	1.90
8	12	76	1.88	29	40	58	1.76
9	13	61	1.79	30	45	86	1.93
10	14	78	1.89	31	47	76	1.88
11	15	48	1.68	32	49	58	1.76
12	15	57	1.76	33	56	?	?
13	16	68	1.83	34	60	26	1.41
14	16	?	?	35	65	104	2.02
15	17	26	1.42	36	65	39	1.59
16	19	78	1.89	37	65	35	1.54
17	19	56	1.75	38	70	72	1.86
18	20	54	1.73	39	70	57	1.76
19	22	73	1.86	40	76	38	1.58
20	26	74	1.87	41	78	44	1.64
21	27	63	1.80				

† Blood lead levels measured in micrograms per deciliter (mcg/dL)
? Missing value

Data Source: Nasser A, Hatch D, Pertowski C, Yoon S. Outbreak investigation of an unknown illness in a rural village, Egypt (case study). Cairo: Field Epidemiology Training Program, 1999.

Coefficient of Variation

The variability of data may often be better described as a relative variation rather than as an absolute variation, such as that represented by the standard deviation or range. One common way of expressing the variability, which takes into account its relative magnitude, is the ratio of the standard deviation to the mean, s.d./$\bar{X}$. This ratio, often expressed as a percentage, is called the *coefficient of variation*, abbreviated as C.V., or R.S.D., the relative standard deviation.

A coefficient of variation of 0.1 or 10% means that the s.d. is one- tenth of the mean. This way of expressing variability is useful in many situations. It puts the variability in perspective relative to the magnitude of the measurements and allows a comparison of the variability of different kinds of measurements.

For example, a group of rats of average weight 100 g and s.d. of 10 g has the same relative variation (C.V.) as a group of animals with average weight 70 g and standard deviation of 7 g. Many measurements have an almost constant C.V., the magnitude of the s.d. being proportional to the mean.

In biological data, the coefficient of variation is often between 20 and 50%, and one would not be surprised to see an occasional C.V. as high as 100% or more. The relatively large C.V. observed in biological experiments is due mostly to ‗biological variation,'' the lack of reproducibility in living material. On the other hand, the variability in chemical and instrumental analyses of drugs is usually relatively small. Thus it is not unusual to find a C.V. of less than 1% for some analytical procedures.

2.3 How to Describe Data Patterns in Statistics

Graphic displays are useful for seeing patterns in data. Patterns in data are commonly described in terms of:

Center

Spread: The **spread** of a distribution refers to the variability of the data. If the observations cover a wide range, the spread is larger. If the observations are clustered around a single value, the spread is smaller.

Shape: The shape of a distribution is described by the following characteristics:

> **Symmetry**. When it is graphed, a symmetric distribution can be divided at the center so that each half is a mirror image of the other.
>
> **Number of peaks**. Distributions can have few or many peaks. Distributions with one clear peak are called **unimodal**, and distributions with two clear peaks are called **bimodal**. When a symmetric distribution has a single peak at the center, it is referred to as **bell-shaped**.
>
> **Skewness**. When they are displayed graphically, some distributions have many more observations on one side of the graph than the other. Distributions with fewer observations on the right (toward higher values) are said to be **skewed right**; and distributions with fewer observations on the left (toward lower values) are said to be **skewed left**.
>
> **Uniform**. When the observations in a set of data are equally spread across the range of the distribution, the distribution is called a **uniform distribution**. A uniform distribution has no clear peaks.

Unusual Features

Sometimes, statisticians refer to unusual features in a set of data. The two most common unusual features are gaps and outliers.

> **Gaps**. Gaps refer to areas of a distribution where there are no observations. The first figure below has a gap; there are no observations in the middle of the distribution.
>
> **Outliers**. Sometimes, distributions are characterized by extreme values that differ greatly from the other observations. These extreme values are

called outliers. The second figure below illustrates a distribution with an outlier. Except for one lonely observation (the outlier on the extreme right), all of the observations fall between 0 and 4. As a "rule of thumb", an extreme value is often considered to be an outlier if it is at least 1.5 interquartile ranges below the first quartile (Q1), or at least 1.5 interquartile ranges above the third quartile (Q3).

How to Compare Data Sets

Common graphical displays (e.g., dot plots, boxplots, stem plots, bar charts) can be effective tools for comparing data from two or more data sets.

Four Ways to Describe Data Sets

When you compare two or more data sets, focus on four features:

Center. Graphically, the center of a distribution is the point where about half of the observations are on either side.

Spread. The spread of a distribution refers to the variability of the data. If the observations cover a wide range, the spread is larger. If the observations are clustered around a single value, the spread is smaller.

Shape. The shape of a distribution is described by symmetry, skewness, number of peaks, etc.

Unusual features. Unusual features refer to gaps (areas of the distribution where there are no observations) and outliers.

Examples/Illustration [using statistical software]

1. Calculate the measures of central tendency for prolactin levels (ng/L) obtained during a clinical trial involving 10 subjects.

   9.4 7.0 7.6 6.3 6.7
   6.8 10.6 8.9 9.4

2. Listed below are the results of a first time in human clinical trial of a new agent with 90 mg/tablet administered to six healthy male volunteers. Report the measures of central tendency for these Cmax results.

Cmax for Initial Pharmacokinetic with New Agent

Subject Number	Cmax (ng/ml)
001	60
002	71
003	111
004	46
005	81
006	96

SELF-ASSESSMENT QUIZ

Now that you have read Lesson 2 and have completed the exercises, you should be ready to take the self-assessment quiz. This quiz is designed to help you assess how well you have learned the content of this lesson. You may refer to the lesson text whenever you are unsure of the answer.

Unless instructed otherwise, choose ALL correct answers for each question.

Use Table 2.16 for Questions 1 and 2, and for Questions 10–13.

Table 2.16 Admitting Clinical Characteristics of Patients with Severe Acute Respiratory Syndrome — Singapore, March–May, 2003

ID	Date of Diagnosis	Sex	Age (Years)	How Acquired	Symptoms†	Temp (°C)	Lymphocyte Count (x 10^9/L)‡	Outcome
01	*	Female	71	Community	F, confusion	38.7	0.78	Survived
02	3/16	Female	43	Community	C,D,S,H,F	38.9	0.94	Died
03	3/29	Male	40	HCW¶	C,H,M,F	36.8	0.71	Survived
04	*	Female	78	Community	D	36.0	1.02	Died
05	*	Female	53	Community	C,D,F	39.6	0.53	Died
06	4/6	Male	63	Community	C,M,F,dizziness	35.1	0.63	Died
07	*	Male	84	Inpatient	D,F	38.0	0.21	Died
08	*	Male	63	Inpatient	F	38.5	0.83	Survived
09	*	Female	74	Inpatient	F	38.0	1.34	Died
10	*	Male	72	Inpatient	F	38.5	1.04	Survived
11	*	Female	28	HCW	H,M,F	38.2	0.30	Survived
12	*	Female	24	HCW	M,F	38.0	0.84	Survived
13	*	Female	28	HCW	M,F	38.5	1.13	Survived
14	*	Male	21	HCW	H,M,F	38.8	0.97	Survived

* Date of onset not provided in manuscript
† C=cough, D=dyspnea, F=fever, H=headache, M=myalgia, S=sore throat
‡ Normal > 1.50 x 10^9/L
¶ HCW = health-care worker

Data Source: Singh K, Hsu L-Y, Villacian JS, Habib A, Fisher D, Tambyah PA. Severe acute respiratory syndrome: lessons from Singapore. Emerg Infect Dis 2003;9:1294–8.

1. Table 2.16 is an example of a/an _____.

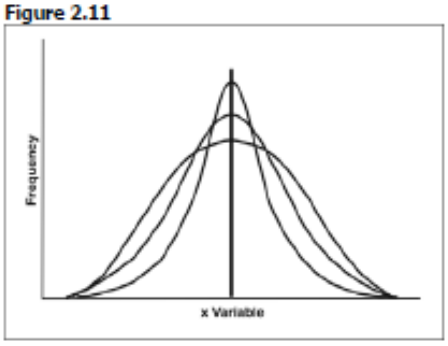
2. For each of the following variables included in Table 2.16, identify if it is:

   A. Categorical
   B. Continuous
   C. Interval
   D. Nominal
   E. Ordinal
   F. Qualitative
   G. Quantitative
   H. Ratio

   _____ Sex
   _____ Age
   _____ Lymphocyte Count

3. Which of the following best describes the similarities and differences in the three distributions shown in Figure 2.11?

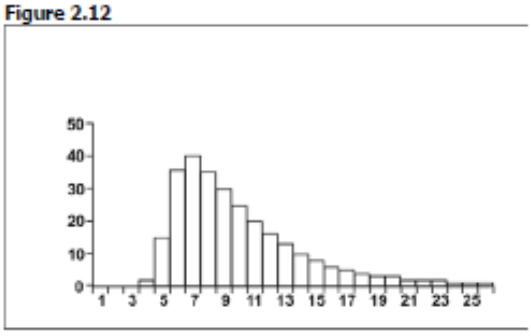
Figure 2.11



A. Same mean, median, mode; different standard deviation
B. Same mean, median, mode; same standard deviation
C. Different mean, median, mode; different standard deviation
D. Different mean, median, mode; same standard deviation

4. Which of the following terms accurately describe the distribution shown in Figure 2.12?

Figure 2.12



A. Negatively skewed
B. Positively skewed
C. Skewed to the right
D. Skewed to the left
E. Asymmetrical

5. What is the likely relationship between mean, median, and mode of the distribution shown in Figure 2.12?
A. Mean < median < mode
B. Mean = median = mode
C. Mean > median > mode
D. Mode < mean and median, but cannot tell relationship between mean and median

6. The mode is the value that:
   A. Is midway between the lowest and highest value
   B. Occurs most often
   C. Has half the observations below it and half above it
   D. Is statistically closest to all of the values in the distribution

7. The median is the value that:
   A. Is midway between the lowest and highest value
   B. Occurs most often
   C. Has half the observations below it and half above it
   D. Is statistically closest to all of the values in the distribution

8. The mean is the value that:
   A. Is midway between the lowest and highest value
   B. Occurs most often
   C. Has half the observations below it and half above it
   D. Is statistically closest to all of the values in the distribution

9. The geometric mean is the value that:
   A. Is midway between the lowest and highest value on a log scale
   B. Occurs most often on a log scale
   C. Has half the observations below it and half above it on a log scale
   D. Is statistically closest to all of the values in the distribution on a log scale

Use Table 2.16 for Questions 10-13. Note that the sum of the 14 temperatures listed in Table 2.16 is 531.6.

10. The mode of the temperatures listed in Table 2.16 is:
    A. 37.35°C
    B. 37.9°C
    C. 38.0°C
    D. 38.35°C
    E. 38.5°C

11. The median of the temperatures listed in Table 2.16 is:
    A. 37.35°C
    B. 37.9°C
    C. 38.0°C
    D. 38.35°C
    E. 38.5°C

12. The mean of the temperatures listed in Table 2.16 is:
    A. 37.35°C
    B. 37.9°C
    C. 38.0°C
    D. 38.35°C
    E. 38.5°C

13. The midrange of the temperatures listed in Table 2.16 is:
    A. 37.35°C
    B. 37.9°C
    C. 38.0°C
    D. 38.35°C
    E. 38.5°C

14. In epidemiology, the measure of central location generally preferred for summarizing skewed data such as incubation periods is the:
    A. Mean
    B. Median
    C. Midrange
    D. Mode

15. The measure of central location generally preferred for additional statistical analysis is the:
    A. Mean
    B. Median
    C. Midrange
    D. Mode

16. Which of the following are considered measures of spread?
    A. Interquartile range
    B. Percentile
    C. Range
    D. Standard deviation

17. The measure of spread **most** affected by one extreme value is the:
    A. Interquartile range
    B. Range
    C. Standard deviation
    D. Mean

18. The interquartile range covers what proportion of a distribution?
    A. 25%
    B. 50%
    C. 75%
    D. 100%

19. The measure of central location most commonly used with the interquartile range is the:
    A. Arithmetic mean
    B. Geometric mean
    C. Median
    D. Midrange
    E. Mode

20. The measure of central location most commonly used with the standard deviation is the:
    A. Arithmetic mean
    B. Median
    C. Midrange
    D. Mode

21. The algebraic relationship between the variance and standard deviation is that:
    A. The standard deviation is the square root of the variance
    B. The variance is the square root of the standard deviation
    C. The standard deviation is the variance divided by the square root of n
    D. The variance is the standard deviation divided by the square root of n

22. Before calculating a standard deviation, one should ensure that:
    A. The data are somewhat normally distributed
    B. The total number of observations is at least 50
    C. The variable is an interval-scale or ratio-scale variable
    D. The calculator or software has a square-root function

23. Simply by scanning the values in each distribution below, identify the distribution with the largest standard deviation.
    A. 1, 10, 15, 18, 20, 20, 22, 25, 30, 39
    B. 1, 3, 8, 10, 20, 20, 30, 32, 37, 39
    C. 1, 15, 17, 19, 20, 20, 21, 23, 25, 39
    D. 41, 42, 43, 44, 45, 45, 46, 47, 48, 49

24. Given the area under a normal curve, which two of the following ranges are the same? (Circle the TWO that are the same.)
    A. From the 2.5^{th} percentile to the 97.5^{th} percentile
    B. From the 5^{th} percentile to the 95^{th} percentile
    C. From the 25^{th} percentile to the 75^{th} percentile
    D. From 1 standard deviation below the mean to 1 standard deviation above the mean
    E. From 1.96 standard deviations below the mean to 1.96 standard deviations above the mean

25. The primary use of the standard error of the mean is in calculating the:
    A. confidence interval
    B. error rate
    C. standard deviation
    D. variance

3 Probability

3.1 Basic Concepts of Probability

Introduction

Probability forms the basis of inferential statistics. We can think of the probability of an outcome as the likelihood of observing that outcome. If something has a high likelihood of happening, it has a high probability (close to 1). If something has a small chance of happening, it has a low probability (close to 0). If something occurs that has a low probability, we investigate to find out "what's up."

Probability is a measure of the likelihood of a random phenomenon or chance behavior. Probability describes the long-term proportion with which a certain **outcome** will occur in situations with short-term uncertainty.

Illustration

Consider an experiment in which only one of two possible outcomes can occur. For example, the result of treatment with an antibiotic is that an infection is either *cured* or *not cured* within 5 days.

The probability of a cure is not easily ascertained a priori, i.e., prior to performing an experiment. If the antibiotic were widely used, based on his or her own experience, a physician prescriber of the product might be able to give a good estimate of the probability of a cure for patients treated with the drug. For example, in the physician's practice, he or she may have observed that approximately three of four patients treated with the antibiotic are cured. For this physician, the probability that a patient will be cured when treated with the antibiotic is 75%.

NB: *The exact probability can be determined only by treating the total population and observing the proportion cured, a practical impossibility in this case. In this context, it would be fair to say that exact probabilities are nearly always unknown.*

Definitions

(c) Experiment:

An experiment is any process that generates a set of data or well-defined outcomes. There are two types of experiments, namely Deterministic and Random (or Chance) Experiment. In the deterministic experiments the observed results are not subject to chance while the outcomes of random experiments cannot be predicted with certainty. A random experiment could be as simple as tossing a coin or die and observing an outcome or complex as choosing 50 people from a population and testing them for the AIDS disease.

(d) Trial: Each repetition of an experiment is called a trial. That is, a trial is a single performance of an experiment.

(e) Outcome: The possible result of each trial of an experiment is called an outcome. When an outcome of an experiment has equal chance of occurring as the others the outcomes are said to be *equally likely*. For example, the toss of a coin and a die yield the possible outcomes in the sets, {H, T} and {1, 2, 3, 4, 5, 6} and a play of a football match yields {win (W), loss (L), draw (D)}.

(f) Sample Space:

Sample space is the collection of all possible outcomes at a probability experiment. We use the notation ***S*** for sample space. Each element or outcome of the experiment is called sample point. For example,

(i) The results of two and three tosses of a coin give the following sample spaces:

S = {HH, HT, TH, TT}

S = {HHH, HHT, HTH, HTT, THH, THT, TTH,TTT}

(ii) A toss of a die and a coin simultaneously give the results. S = {H1, H2, H3, H4, H5, H6, T1, T2, T3, T4, T5, T6}

(iii) The outcomes of two tosses of a die are as shown in the table

T1 \ T2	1	2	3	4	5	6
1	(1,1)	(1,2)	(1,3)	(1,4)	(1,5)	(1,6)
2	(2,1)	(2,2)	(2,3)	(2,4)	(2,5)	(2,6)
3	(3,1)	(3,2)	(3,30	(3,4)	(3,5)	(3,6)
4	(4,1)	(4,2)	(4,3)	(4,4)	(4,5)	(4,6)
5	(5,1)	(5,2)	(5,3)	(5,4)	(5,5)	(5,6)
6	(6,1)	(6,2)	(6,3)	(6,4)	(6,5)	(6,6)

where T1 and T2 represent the first and second tosses respectively.

(iv) Drawing a card from a packet of playing cards has sample space with 52 cards made up 13 Heart, 13 Spade, 13 Diamond and 13 Club cards.

(e) Event: An event is a collection of one or more outcomes from an experiment. That is, it is a subset of a sample space. It is denoted by a capital letter. For example we may have:

(i) The event of observing a head (H) in three tosses of a coin,
A = {HTT, TTH}

(ii) The event of obtaining a total score of 8 on two tosses of a die,

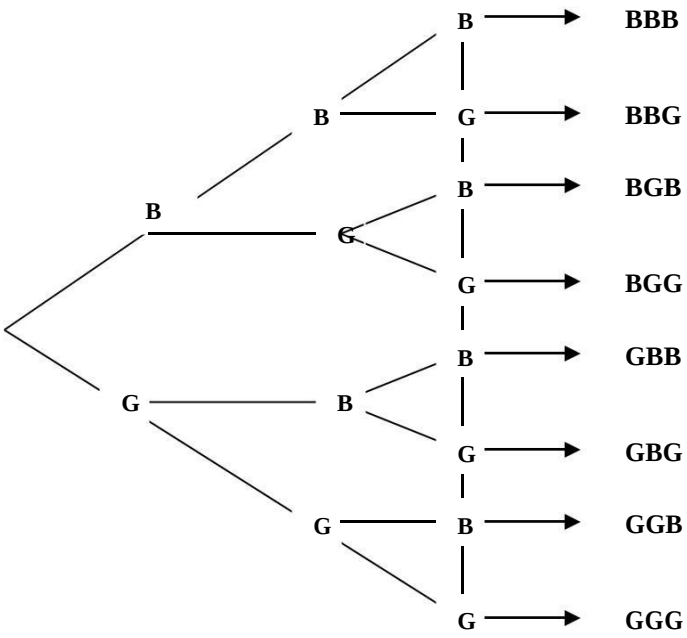
B = {2,6), (3,5), (4,4), (5,3), (6,2)}

(iii) Consider a newly married couple planning to have three children. The event of the family having two girls is:

D = {BGG, GBG, GGB}

(h) Tree Diagram: The tree diagram represents pictorially the outcomes of random experiment. The probability of an outcome which is a sequence of trials, is represented by any path of the tree. For example,

(ii) Consider a couple planning to have three children, assuming each child born is equally likely to be a boy (B) or girl (G).



Determination of Probability of an Event

The probability of an event A, denoted, P(A), gives the numerical measure of the likelihood of the occurrence of event A which is such that $0 \leq P(A) \leq 1$.

If P (A) = 0, the event A is said to be impossible to occur and if P(A) = 1, A is said to be certain. If $A^{/}$ is the complement of the event A, then $P(A^{/}) = 1 - P(A)$, called the probability that event A will not occur.

There are three main schools of thought in defining and interpreting the probability of an event. These are the Classical Definition, Empirical Concept and the Subjective Approach. The first two are referred to as the Objective Approach.

a. **The Classical Definition:** This is based on the assumption that the outcomes of an experiment are equally likely. For example, if an experiment can lead to n mutually exclusive and equally likely outcomes, then the probability of the event A is defined by

$$\boldsymbol{P(A) = \frac{n(A)}{n(S)} = \frac{Number\ of\ succesful\ outcomes}{Number\ of\ possible\ outcomes}}$$

The classical definition of probability of event A is referred to as priori probability because it is determined before any experiment is performed to observe the outcomes of event A.

b. **The Empirical Concept:** This concept uses the relative frequencies of past occurrences to develop probabilities for future. The probability of an event A happening in future is determined by observing what fraction of the time similar events happened in the past. That is,

$$\boldsymbol{P(A) = \frac{number\ of\ times\ A\ occured\ in\ the\ past}{Total\ number\ of\ observations}}$$

The relative frequency of the occurrence of the event A used to estimate P(A) becomes more accurate if trials are largely repeated. The relative frequency

approach of defining P(A) is sometimes called posteriori probability because P(A) is determined only after event A is observed.

c. The Subjective Definition: The subjective concept of probability is based on the degree of belief through the evidence available. The probability of an event A may therefore be assessed through experience, intuitive, judgment or expertise. For example, determining the probability of getting a cure of a disease or going to rain today. This approach to probability has been developed relatively recently and is related to *Bayesian Decision Analysis*. Although the subjective view of probability has enjoyed increased attention over the years, it has not been fully accepted by statisticians who have traditional orientations.

Example 1:.

Consider the problem of a couple planning to have three children, assuming each child born is equally likely to be a boy (B) or a girl (G).

a. List the possible outcomes in this experiment

b. What is the probability of the couple having exactly two girls?

Solution:

(a) The sample space for this experiment is

S = {BBB, BBG, BGB, BGG, GBG, GGB, GGG}

(b) Let A be the event of the couple having exactly two girls. Then,

A = {BGG, GBG, GGB}

$$\boldsymbol{P(A) = \frac{n(A)}{n(S)} = \frac{3}{8}}$$

Example 3:

A die is tossed twice. List all the outcomes in each of the following events and compute the probability of each event.

(a) The sum of the scores is less than 4

(b) Each toss results in the same score

(c) The sum of scores on both tosses is a prime number

(d) The product of the scores is at least 20

Solution:

The sample space for the experiment is the set of ordered paired (m, n), where *m* and *n* each takes the values 1, 2, 3, 4, 5 and 6. Thus,

S = {(1, 1), (1, 2), (1, 3), ... , (6, 6)}, where n(S) = 36

a) A = {(1, 1), (1, 2), (2, 1)}

$$P(A) = \frac{3}{36} = \frac{1}{12}$$

b) B = {each toss results in the same score}

{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)}

$$P(B) = \frac{6}{36} = \frac{1}{6}$$

c) D = {sum of scores on both tosses is prime}

D = {(1, 1), (1, 2), (1, 4), (1, 6), (2, 1), (2, 3), (2, 5), (3, 2), (3, 4), (4, 1), (4, 3), (5, 2), (5, 6), (6, 1), (6, 5)}

$$P(D) = \frac{15}{36} = \frac{5}{12}$$

d) E = {product of the scores is at least 20}

{(4, 5), (4, 6), (5, 4), (5, 5), (5, 6), (6, 4), (6 5), (6, 6)}

$$P(E) = \frac{8}{36} = \frac{2}{9}$$

Probability of Compound Events

Two or more events are combined to form a single event using the set operations, $\cup$ and $\cap$. The event

(i) (A $\cup$ B) occurs if either A or B both occur(s).

(ii) (A $\cap$ B) occurs if both A and B occur.

Definitions:

(a) **Mutually Exclusive Events:** Two or more events which have no common outcome(s) (i.e. never occur at the same time) are said to be mutually exclusive. If A and B are mutually exclusive events of an experiment, then $A \cap B = \phi$ and $P(A \cup B) = P(A) + P(B)$, since $P(A \cap B) = 0$.

(b) **Independent Events:** Two or more events are said to be independent if the probability of occurrence of one is not influenced by the occurrence or non-occurrence of the other(s). Mathematically, the two events, A and B are said to be independent, if and only if $P(A \cap B) = P(A) \bullet P(B)$. However, if A and B are such that, $P(A \cap B) = P(A) \bullet P(B/A)$, they are said to be conditionally independent.

(c) **Conditional Probability:** Let A and B be two events in the sample space, S with P(B) > 0. The probability that an event A occurs given that event B has already occurred, denoted P(A/B), is called the conditional probability of A given B. The conditional probability of A given B is defined as.

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

In particular, if S is a finite equiprobable space, then

$$P(A \cap B) = \frac{n(A \cap B)}{n(S)}, \; P(B) = \frac{n(B)}{n(S)} \; and \; P(A/B) = \frac{n(A \cap B)}{n(B)}$$

(d) **Exhaustive Events:** Two or more events defined on the same sample space are said to be exhaustive if their union is equal to the sample space *S* (thus, if they partition the sample space mutually exclusively).

Eg: if

$A_1, A_2, A_3 \in S$, $\quad A_1 \cup A_2 \cup A_3 = S.$

Definition (partition of sample space): The events $A_1, A_2, A_3, \quad, A_n$ form a partition of the same sample space S if the following hold:

(a) $A_i \neq \phi$ for all $i = 1, 2, 3,, n$

(b) $A_{i \atop n} \cap A_j = \phi$ for all $i \neq j;\ i, j = 1, 2, 3,, n$

c. $\bigcup_{i=1} A_i = S$

In other words, the n - events $A_1, A_2, A_3, \quad, A_n$ form a partition of the sample space S if the n - events are (a) nonempty, (b) mutually exclusive and (c) collectively exhaustive.

Example

1.(a) In a certain population of women, 40% have had breast cancer, 20% are smokers and 13% are smokers and have had breast cancer. If a women is selected at random from the population, what is the probability that she had breast cancer, smokes or both?

(b) Let A and B be events such that $P(A) = 0.6$, $P(B) = 0.5$ and $(A \cup B) = 0.8$

(i) Find $P(A / B)$

(ii) Are A and B independent?

Solution:

(a) Let B be the event of women with breast cancer and W the event of women who smoke. Then,

$P(B) = 0.4$, $P(W) = 0.2$ and $(B \cap W) = 0.13$

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

$= 0.4 + 0.20 - 0.13$

$= 0.47$

(b) Given that $P(A) = 0.6$, $P(B) = 0.5$ and $(A \cup B) = 0.8$

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

$= 0.6 + 0.5 - 0.8 = 0.3$

$$P(A/B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) > 0$$

$$= \frac{\mathbf{0.3}}{\mathbf{0.5}} = \frac{\mathbf{3}}{\mathbf{5}} = \mathbf{0.6}$$

(ii) A and B are independent if $P(A) \bullet P(B) = P(A \cap B)$

$P(A) \bullet P(B) = (0.6)(0.5) = 0.3 = P(A \cap B)$

Which means that A and B are independent.

Example

Complex components are assembled in a plant that uses two different assembly lines, A and A'. Line A uses older equipment than A', so it is somewhat slower and less reliable. Suppose on a given day line A has assembled 8 components, of which 2 have been identified as defective (B) and 6 as nondefective (B'), whereas A' has produced 1 defective and 9 nondefective components. This information is summarized in the accompanying table.

		Condition		Total
		B	B′	
Line	**A**	2	6	8
	A′	1	9	10
		3	15	18

Unaware of this information, the sales manager randomly selects 1 of these 18 components for a demonstration. Prior to the demonstration

P(line A component selected) = P(A) = $\frac{N(A)}{N} = \frac{8}{18} = 0.44$

However, if the chosen component turns out to be defective, then the event B has occurred, so the component must have been 1 of the 3 in the B column of the table. Since these 3 components are equally likely among themselves after B has occurred,

$$\boldsymbol{P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{2/18}{3/18} = \frac{2}{3}}$$

The Multiplication Rule for $P(A \cap B)$

The definition of conditional probability yields the following result, obtained by multiplying both sides of the conditional probability equation by $P(B)$.

$$P(A / B) = \frac{P(A \cap B)}{P(B)}$$

$$P(A / B) * P(B) = \frac{P(A \cap B)}{P(B)} * P(B)$$

$$P(A / B) * P(B) = P(A \cap B)$$

This rule is important because it is often the case that $P(A \cap B)$ is desired, whereas both $P(B)$ and $P(A / B)$ can be specified from the problem description.

The Law of Total Probability

Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = P(B / A_1) * P(A_1) + \ldots + P(B / A_k) * P(A_k)$$

$$= \sum_{i=1}^{k} P(B / A_i) * P(A_i)$$

Bayes' Rule

The power of Bayes' rule is that in many situations where we want to compute $P(A \mid B)$ it turns out that it is difficult to do so directly, yet we might have direct information about $P(B \mid A)$. Bayes' rule enables us to compute $P(A \mid B)$ in terms of $P(B \mid A)$.

$$P(A / B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B / A)P(A)}{P(B)}$$

Bayes' Theorem

Let A and A^c constitute a partition of the sample space S such that with P(A) > 0 and $P(A^c) > 0$, then for any event B in S such that P(B) > 0,

$$P(A/B) = \frac{P(B/A)P(A)}{P(B/A)P(A) + P(B/A^c)P(A^c)}$$

Example

A paint-store chain produces and sells latex and semigloss paint. Based on long-range sales, the probability that a customer will purchase latex paint is 0.75. Of those that purchase latex paint, 60% also purchase rollers. But only 30% of semigloss pain buyers purchase rollers. A randomly selected buyer purchases a roller and a can of paint. What is the probability that the paint is latex?

Solution

L = {The customer purchases latex paint.}, P(L) = 0.75

S = {The customer purchases semigloss paint.}, P(S) = 0.25

R = {The customer purchases roller.}

P(R | L) =0.6

P(R | S) =0.3

$$P(R) = P(R|L)P(L) + P(R|S)P(S) = 0.6 \times 0.75 + 0.3 \times 0.25 = 0.525$$

$$P(L/R) = \frac{P(L \cap R)}{P(R)} = \frac{P(R/L)P(L)}{P(R)} = \frac{0.6 \times 0.75}{0.6 \times 0.75 + 0.3 \times 0.25} \approx 0.857$$

Axioms of Probability

Given an experiment and a sample space, S , the objective of probability is to assign to each event A a number P(A), called the probability of the event A, which will give a precise measure of the chance that A will occur. To ensure that the probability assignments will be consistent with our intuitive notions of probability, all assignments should satisfy the following axioms (basic properties) of probability.

A.1: For every event A, $0 \leq P(A) \leq 1$

A.2: $P(S) = 1$

A.3: If A and B are mutually exclusive events, i.e $A \cap B = \phi$ then

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

A.4: If A_1 , A_2 , A_3 ,, A_n is a sequence of *n* mutually exclusive events, then,

$$P(A_1 \cup A_2 \cup A_3 \cup A_n) = P(A_1) + P(A_2) + P(A_3) + \ldots + P(A_n) \text{ or}$$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{N} P(A_i) .$$

The following theorems arise directly from the above axioms:

(i) **Theorem 1:** If ϕ is the empty set, then $P(\phi) = 0$.

Proof:

Let A be any event, then A and ϕ are mutually exclusive and $A = A \cup \phi$

Then by A.3 $P(A) = P(A \cup \phi) = P(A) + P(\phi)$ and $P(\phi) = 0$

(ii) **Theorem 2:** If A' is the complement of an event A , then

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

Proof

$$S = A \cup A'$$

$$P(S) = P(A) + P(A')$$

$$1 = P(A) + P(A^{/}), \text{ by A.2}$$
$$P(A^{/}) = 1 - P(A)$$

Some Rules of Probability

(a) **The Addition Rule:**

Let A_1 , A_2 , A_3 ,, A_n be events of the sample space, S. Then

(i) $P(A_1 \cup A_2)) = P(A_1) + P(A_2) - P(A_1 \cap A_2)$

(ii) $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3) - P(A_1 \cap A_2) - P(A_1 \cap A_3) - P(A_2 \cap A_3) + P(A_1 \cap A_2 \cap A_3)$.

If the events A_1 , A_2 , A_3 , , A_n are mutually exclusive, then

(i) $P(A_1 \cup A_2) = P(A_1) + P(A_2)$

(ii) $P(A_1 \cup A_2 \cup A_3) = P(A_1) + P(A_2) + P(A_3)$

(iii) $P(A_1 \cup A_2 \cup ..... \cup A_n)) = p(A_1) + P(A_2) + P(A_3) ..... + P(A_n)$

(b) **The Multiplication Theorem:**

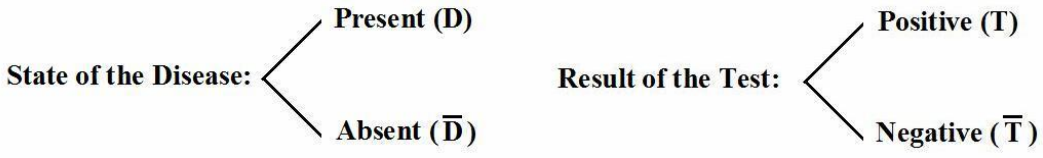
If A_1 , A_2 , A_3 ,, A_n are events of the same sample space, S, then

(i) $P(A_1 \cap A_2) = P(A_1) \bullet .P(A_2 / A_1)$

(ii) $P(A_1 \cap A_2 \cap A_3) = P(A_1) \bullet .P(A_2 / A_1) \bullet .P(A_3 / A_1 \cap A_2)$

3.2 Bayes' Theorem, Screening Tests, Sensitivity, Specificity, and Predictive Value Positive and Negative:

There are two states regarding the disease and two states regarding the result of the screening test:



We define the following events of interest:

D: the individual has the disease (presence of the disease)

$\overline{D}$: the individual does not have the disease (absence of the disease)

T: the individual has a positive screening test result

$\overline{T}$: the individual has a negative screening test result

There are 4 possible situations:

		True status of the disease	
		+ve (D: Present)	−ve ($\overline{D}$:Absent)
Result of the test	+ve (T)	Correct diagnosing	false positive result
	−ve ($\overline{T}$)	false negative result	Correct diagnosing

Definitions of False Results:

There are two false results:

1. **A false positive** result:

   This result happens when a test indicates a positive status when the true status is negative. Its probability is:

   $P(T \mid \overline{D}) = P($ positive result $\mid$ absence of the disease $)$

2. **A false negative** result:

   This result happens when a test indicates a negative status when the true status is positive. Its probability is:

   $P(\overline{T} \mid D) = P($ negative result $\mid$ presence of the disease $)$

Definitions of the Sensitivity and Specificity of the test:

1. **The Sensitivity:**

The sensitivity of a test is the probability of a positive test result given the presence of the disease.

$P(T \mid D) = P($ positive result of the test $\mid$ presence of the disease $)$

2. **The specificity:**

The specificity of a test is the probability of a negative test result given the absence of the disease.

$P(\overline{T} \mid \overline{D}) = P($ negative result of the test $\mid$ absence of the disease $)$

To clarify these concepts, suppose we have a sample of (n) subjects who are cross-classified according to Disease Status and Screening Test Result as follows:

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

For example, there are (a) subjects who have the disease and whose screening test result was positive.

From this table, we may compute the following conditional probabilities:

1. The probability of the false positive result:

$$P(T \mid \overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{b}{b+d}$$

2. The probability of false negative result:

$$P(\overline{T} \mid D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{c}{a+c}$$

3. The sensitivity of the screening test:

$$P(T \mid D) = \frac{n(T \cap D)}{n(D)} = \frac{a}{a+c}$$

4. The specificity of the screening test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{d}{b+d}$$

Definitions of the Predictive Value Positive and Predictive Value Negative of a Screening Test:

1. The predictive value positive of a screening test:

The predictive value positive is the probability that a subject has the disease, given that the subject has a positive screening test result:

$$P(D \mid T) = P(\text{ the subject has the disease} \mid \text{ positive result})$$

$$= P(\text{ presence of the disease} \mid \text{ positive result})$$

2. The predictive value negative of a screening test:

The predictive value negative is the probability that a subject does not have the disease, given that the subject has a negative screening test result:

$$P(\overline{T} \mid \overline{D}) = P(\text{ the subject does not have the disease} \mid \text{ negative result})$$

$$= P(\text{ absence of the disease} \mid \text{ negative result})$$

Calculating the predictive Value Positive and Predictive Value Negative:

(How to calculate $P(D \mid T)$ and $P(\overline{T} \mid \overline{D})$):

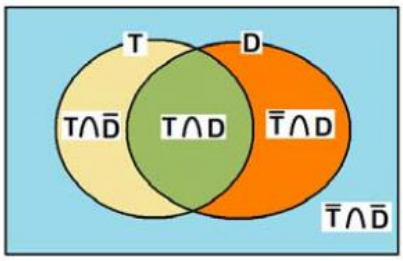
We calculate these conditional probabilities using the knowledge of:

1. The sensitivity of the test = $P(D \mid T)$
2. the specificity of the test = $P(\overline{T} \mid \overline{D})$
3. the probability of the relevant disease in the general population, $P(D)$. (It is usually obtained from another independent study).

Calculating the Predictive Value Positive, $P(D \mid T)$:

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

But we know that:



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

$P(T \cap D) = P(T \mid D)P(D)$ … multiplication rule.

$P(T \cap \overline{D}) = P(T \mid \overline{D})P(\overline{D})$ … multiplication rule.

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

Therefore, we reach the following version of Bayes' Theorem:

$$P(D \mid T) = \frac{P(T \mid D)P(D)}{P(T \mid D)P(D) + P(T \mid \overline{D})P(\overline{D})} \quad \text{…….. (1)}$$

NOTE:

$P(T \mid D)$ = sensitivity

$P(T \mid \overline{D}) = 1 - P(\overline{T} \mid \overline{D}) = 1 -$ specificity

$P(D) =$ The probability of the relevant disease in the general population.

$P(\overline{D}) = 1 - P(D)$

Calculating the Predictive Value Negative, $P(\overline{D} \mid \overline{T})$:

To obtain the predictive value negative of a screening test, we use the following statement of Bayes' theorem:

$$P(\overline{D} \mid \overline{T}) = \frac{P(\overline{D} \mid \overline{T})P(\overline{D})}{P(\overline{D} \mid \overline{T})P(\overline{D}) + P(T \mid D)P(D)} \quad \text{…… (2)}$$

NOTE:

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

$P(\overline{T} \mid D) = 1 - P(T \mid D) = 1 -$ sensitivity

Example:

A medical research team wished to evaluate a proposed screening test for Alzheimer's disease. The test was given to a random sample of 450 patients with Alzheimer's disease and an independent random sample of 500 patients without symptoms of the disease. The two samples were drawn from populations of subjects who were 65 years of age or older. The results are as follows:

	Alzheimer Disease		
Test Result	**Present (D)**	**Absent ($\overline{D}$)**	**Total**
Positive (T)	436	5	441
Negative ($\overline{T}$)	14	495	509
Total	450	500	950

Based on another independent study, it is known that the percentage of patients with Alzheimer's disease (the rate of prevalence of the disease) is 11.3% out of all subjects who were 65 years of age or older.

Solution:

Using the data, we estimate the following quantities:

1. the sensitivity of the test:

$$P(T \cap D) = \frac{n(T \cap D)}{n(D)} = \frac{436}{450} = 0.9689$$

2. The specificity of the test:

$$P(\overline{T} \mid \overline{D}) = \frac{n(\overline{T} \cap \overline{D})}{n(\overline{D})} = \frac{495}{500} = 0.99$$

3. The probability of the disease in the general population, $P(D)$: The rate of disease in the relevant general population, $P(D)$, cannot be computed from the sample data given in the table. However, it is given that the percentage of patients with Alzheimer's disease is 11.3% out of all subjects who were 65 years of age or older. Therefore $P(D)$ can be computed to be:

$$P(D) = \frac{11.3\%}{100\%}$$

4. The predictive value positive of the test:

   We wish to estimate the probability that a subject who is positive on the test has Alzheimer disease. We use the Bayes' formula of Equation (1):

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

From the tabulated data, we compute:

$$P(T \mid D) = \frac{436}{450} = 0.9689 \quad \text{(From part no.1)}$$

$$P(T \mid \overline{D}) = \frac{n(T \cap \overline{D})}{n(\overline{D})} = \frac{5}{500} = 0.01$$

Substituting of these results into Equation (1), we get:

$$P(D \mid T) = \frac{(0.9689)P(D)}{(0.9689)P(D) + (0.01)P(\overline{D})}$$

$$= \frac{(0.9689)(0.113)}{(0.9689)(0.113) + (0.01)(1-0.113)} = 0.93$$

As we see, in this case, the predictive value positive of the test is very high.

5. The predictive value negative of the test:

   We wish to estimate the probability that a subject who is negative in the test does not have Alzheimer disease. We use the Bayes' formula of Equation (2):

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

To compute $P(\overline{D} \mid \overline{T})$, we first compute the following probabilities:

$$P(\overline{T} \mid \overline{D}) = \frac{495}{500} = 0.99 \qquad \text{(from part no. 2)}$$

$$P(\overline{D}) = 1 - P(D) = 1 - 0.113 = 0.887$$

$$P(\overline{T} \mid D) = \frac{n(\overline{T} \cap D)}{n(D)} = \frac{14}{450} = 0.0311$$

Substituting in Equation (2) gives:

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

$$= \frac{(0.99)(0.887)}{(0.99)(0.887) + (0.0311)(0.113)}$$

$$= 0.996$$

As we see, the predictive value negative is also very high.

3.3 BINOMIAL DISTRIBUTION

Bernoulli Trial is an experiment with only two possible outcomes;

S = success and F = failure (Boy or girl, dead or alive, cured or not cured).

Binomial distribution is a discrete distribution. It is used to model an experiment for which:

1. The experiment has a sequence of n Bernoulli trials.
2. The probability of success is $P(S) = p$, and the probability of failure is $P(F) = 1 - p = q$.
3. The probability of success $P(S) = p$ is constant for each trial.
4. The trials are independent; that is the outcome of one trial has no effect on the outcome of any other trial.

In this type of experiment, we are interested in the discrete r. v. representing the number of successes in the n trials.

X = the number of success in the n trials

The possible values of X (number of success in n trials) are:

X = 1, 2, 3, …, n

The r.v X has a binomial distribution with parameters n and p, and we write:

X ~ Binomial (n,p)

The probability distribution of X is given by:

$$P(X = x) = \begin{cases} {}^{n}C_{x}\, p^{x} q^{n-x} \\ 0 \end{cases}$$

Where: ${}^{n}C_{x} = \dfrac{n!}{x!(n-x)!}$

We can write the probability distribution of X as a table as follows.

x	$P(X=x)$
0	${}^{n}C_{0}\, p^{0} q^{n-0} = q^{n}$
1	${}^{n}C_{1}\, p^{1} q^{n-1}$
2	${}^{n}C_{2}\, p^{2} q^{n-2}$
⋮	⋮
n-1	${}^{n}C_{n-1}\, p^{n-1} q^{1}$
n	${}^{n}C_{n}\, p^{n} q^{0} = p^{n}$
total	1.00

Result: (Mean and Variance for normal distribution)

If X ~ Binomial (n, p), then

- The mean: $\mu = np$ (expected value)
- The variance: $\sigma^{2} = npq$

Example:

Suppose that the probability that a man has high blood pressure is 0.15. suppose that we randomly select a sample of 6 men.

1) Find the probability distribution of the random variable (X) representing the number of men with high blood pressure in the sample.
2) Find the expected number of men with high blood pressure in the sample (mean of X).
3) Find the variance X.

4) What is the probability that there will be exactly 2 men with high blood pressure?
5) What is the probability that there will be at most 2 men with high blood pressure?
6) What is the probability that there will be at least 4 men with high blood pressure?

Solution:

We are interested in the following random variable:

X = the number of men with high blood pressure in the sample of 6 men.

Notes:

Bernoulli trial: diagnosing whether a man has a high blood pressure or not. There are two outcomes for each trial:

- S = success: the man has high blood pressure
- F = failure: the man does not have high blood pressure
- Number of trials = 6(we need to check 6 men)
- Probability of success: P(S)= p = 0.15
- Probability of failure: P(F) = q = 1 – p = 0.85
- Number of trials: n = 6
- The trials are independent because of the fact that the result of each man does not affect the result of any other man since the selection was made are random.

The random variable X has a binomial distribution with parameters: n = 6 and p = 0.15, that is:

X ~ Binomial (n, p)

X ~ Binomial (6, 0.15)

The possible values of X are:

x = 0, 1, 2, 3, 4, 5, 6

1) The probability distribution of X is:

$$P(X = x) = \begin{cases} {}^{6}C_{x}(0.15)^{x}(0.85)^{6-x}; x = 0,1,2,3,4,5,6 \\ 0 \end{cases}$$

The probabilities of all values of X are:

$$P(X = 0) = {}^{6}C_{0}(0.15)^{0}(0.85)^{6} = (1)(1)(0.85)^{6} = 0.37715$$
$$P(X = 1) = {}^{6}C_{1}(0.15)^{1}(0.85)^{5} = (6)(0.15)(0.85)^{5} = 0.39933$$
$$P(X = 2) = {}^{6}C_{2}(0.15)^{2}(0.85)^{4} = (15)(0.15)^{2}(0.85)^{4} = 0.17618$$
$$P(X = 3) = {}^{6}C_{3}(0.15)^{3}(0.85)^{3} = (20)(0.15)^{3}(0.85)^{3} = 0.04145$$
$$P(X = 4) = {}^{6}C_{4}(0.15)^{4}(0.85)^{2} = (15)(0.15)^{4}(0.85)^{2} = 0.00549$$
$$P(X = 5) = {}^{6}C_{5}(0.15)^{5}(0.85)^{1} = (6)(0.15)^{5}(0.85)^{1} = 0.00039$$
$$P(X = 6) = {}^{6}C_{6}(0.15)^{6}(0.85) = (1)(0.15)^{6}(1) = 0.00001$$

The probability distribution of X can by presented by the following table:

2) The mean of the distribution (the expected number of men out of 6 with high blood pressure) is:

$$\mu = np = (6)(0.15) = 0.9$$

3) The variance is:

$$\sigma^{2} = npq = (6)(0.15)(0.85) = 0.765$$

4) The probability that there will be exactly 2 men with high blood pressure:

$$P(X = 2) = 0.17618$$

5) The probability that there will be at most 2 men with high blood pressure is:

$$P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2)$$
$$= 0.37715 + 0.39933 + 0.17618$$
$$= 0.95266$$

6) The probability that there will be at least 4 men with high blood pressure is:

$$P(X \ge 4) = P(X = 4) + P(X = 5) + P(X = 6)$$
$$= 0.00549 + 0.00039 + 0.00001$$
$$= 0.00589$$

Example: Reading Assignment

Suppose that 25% of the people in a certain population have low hemoglobin levels. The experiment is to choose 5 people at random from this population. Let the discrete random variable X be the number of people out of 5 with low hemoglobin levels.

1) Find the probability distribution of X.
2) Find the probability that at least 2 people have low hemoglobin levels.
3) Find the probability that at most 3 people have low hemoglobin levels.
4) Find the expected number of people with low hemoglobin levels out of the 5 people.
5) Find the variance of the number of people with low hemoglobin levels out of the 5 people

Solution:

X = the number of people out of 5 with low hemoglobin levels

The Bernoulli trial is the process of diagnosing the person

Success = the person has low hemoglobin

Failure = the person does not have low hemoglobin

$n = 5$ (number of trials)

$p = 0.25$ (probability of success)

$q = 1 - p = 0.75$ (probability of failure)

a) X has a binomial distribution with parameter n = 5 and p =0.25

X ~ Binomial (n, p)

X ~ Binomial (5, 0.25)

The possible values of X are:

x = 0, 1, 2, 3, 4, 5

the probability distribution is

$$P(X = x) = \begin{cases} {}^{n}C_{x} p^{x} q^{n-x}; \\ 0 \end{cases}$$

x	$P(X=x)$
0	${}^{n}C_{0} \times 0.25^{0} \times 0.75^{5-0} = 0.23730$
1	${}^{n}C_{1} \times 0.25^{1} \times 0.75^{5-1} = 0.39551$
2	${}^{n}C_{2} \times 0.25^{2} \times 0.75^{5-2} = 0.26367$
3	${}^{n}C_{3} \times 0.25^{3} \times 0.75^{5-3} = 0.08789$
4	${}^{n}C_{4} \times 0.25^{4} \times 0.75^{5-4} = 0.01465$
5	${}^{n}C_{5} \times 0.25^{5} \times 0.75^{5-5} = 0.00098$
total	$\sum P(X = x) = 1$

b) The probability that at least 2 people have low hemoglobin levels:

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5)$$
$$= 0.26367 + 0.08789 + 0.01465 + 0.0098$$
$$= 0.36719$$

c) The probability that at most 3 people have low hemoglobin levels:

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$
$$= 0.23730 + 0.39551 + 0.26367 + 0.08789$$
$$= 0.98437$$

d) The expected number of people with low hemoglobin levels out of the 5 people (the mean of X):

$$\mu = np = 5 \times 0.25 = 1.25$$

e) The variance of the number of people with low hemoglobin levels out of the 5 people (variance of X) is:

$$\sigma^2 = npq = 5 \times 0.25 = 0.9375$$

3.4 POISSON DISTRIBUTION

It is a discrete distribution. The Poisson distribution is used to model a discrete random variable representing the number of occurrences of some random event in an interval of time or space (or some volume of matter).

The possible values of X are:

$$x = 0, 1, 2, 3, \ldots$$

the discrete random variable X is said to have a Poisson distribution with parameter (average or mean) λ if the probability distribution of X is given by

where e = 2.71828 (the natural number). We write:

$$X \sim Poisson(\lambda)$$

Mean and Variance of Poisson Distribution

If $X \sim Poisson(\lambda)$, then:

The mean (average) of X is: $\mu = \lambda$ (Expected value)

The variance of X is: $\sigma^2 = \lambda$

Example:

Some random quantities that can be modeled by Poisson distribution:

- Number of patients in a waiting room in an hour.
- Number of surgeries performed in a month.
- Number of rats in each house in a particular city.

Note:

- λ is the average (mean) of the distribution.
- If X = the number of patients seen in the emergency unit in a day, and if $X \sim Poisson(\lambda)$, then:
    1. The average (mean) of patients seen every day in the emergency unit $= \lambda$, then:
    2. The average (mean) of patients seen every month in the emergency unit $= 30\lambda$

3. The average (mean) of patients seen every year in the emergency unit $= 365\lambda$
4. The average (mean) of patients seen every hour in the emergency unit $= \lambda / 24$

Also, notice that:

i. If $Y =$ the number of patients seen every month, then:

$$Y \sim Poisson(\lambda^*), \text{ where } \lambda^* = 30\lambda$$

ii. $W =$ the number of patients seen every year, then:

$$W \sim Poisson(\lambda^*), \text{ where } \lambda^* = 365\lambda$$

iii. $V =$ the number of patients seen every hour, then:

$$V \sim Poisson(\lambda^*), \text{ where } \lambda^* = \frac{\lambda}{24}$$

Example:

Suppose that the number of snake bites cases seen at KATH in a year has a Poisson distribution with average 6 bite cases.

1. What is the probability that in a year:
   (i) The number of snake bite cases will be 7?
   (ii) The number of snake bite cases will be less than 2?
2. What is the probability that there will be 10 snake bite cases in 2 years?
3. What is the probability that there will be no snake bite cases in a month?

Solution:

i. $X =$ number of snake bite cases in a year

$$X \sim Poisson(6)$$

$$P(X = x) = \frac{e^{-6}6^x}{x!}; \quad x = 0, 1, 2, \ldots$$

(i) $$P(X = 7) = \frac{e^{-6}6^7}{7!} = 0.13768$$

(ii) $$P(X < 2) = P(X = 0) + P(X = 1)$$

$$= \frac{e^{-6}6^0}{0!} + \frac{e^{-6}6^1}{1!} = 0.00248 + 0.01487 = 0.01735$$

ii. Y = number of snake bite cases in 2 years

Y ~ Poisson (12) $\left(\lambda^* = 2\lambda = (2)(6) = 12\right)$

$$P(Y = y) = \frac{e^{-12}12^y}{y!}: \text{ y = 0, 1, 2, …}$$

$$\therefore P(Y = 10) = \frac{e^{-12}12^0}{10!} = 0.1048$$

iii. W = number of snake bite cases in a month.

W ~ Poisson (0.5) $\left(\lambda^* = \frac{\lambda}{12} = \frac{6}{12} = 0.5\right)$

3.5 THE NORMAL DISTRIBUTION

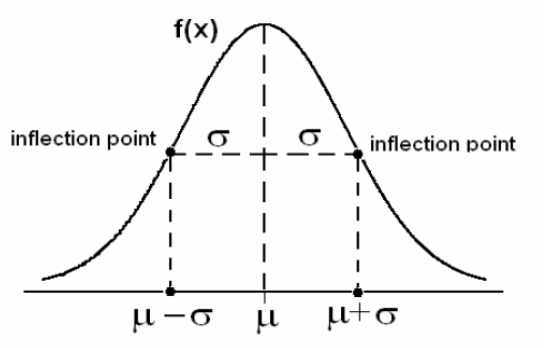
- One of the most important continuous distributions.
- Many measurable characteristics are normally or approximately normal distributed. Examples: Height, weight, …
- The probability density function of the normal distribution is given by:

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

where $(e = 2.71828)$ and $(\pi = 3.14159)$.

The parameters of the distribution are the mean (μ) and the standard deviation (σ).

- The continuous random variable X which has a normal distribution has several important characteristics:
  1. $-\infty < X < \infty$,
  2. The density function of X, $f(x)$, has a bell-shaped curve.



mean = μ
standard deviation = σ
variance = σ^2

3. The highest point of the curve of $f(x)$ at the mean μ. $(Mode = \mu)$
4. The curve of $f(x)$ is symmetric about the mean μ.

   $\mu = mean = \text{mod}e = median$
5. The normal distribution depends on two parameters:

   $mean = \mu$ (determines the location)

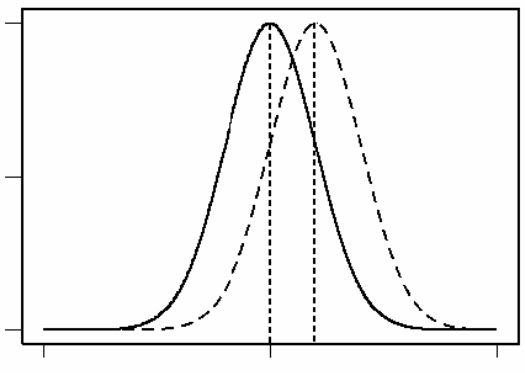
   Standard deviation = σ (determines the shape)
6. If the random variable is normally distributed with mean μ and standard deviation σ (variance σ^2), we write:

   $X \sim Normal(\mu, \sigma^2)$ or $X \sim N\left(\mu, \sigma^2\right)$
7. The location of the normal distribution depends on μ. The shape of the normal distribution depends on σ.

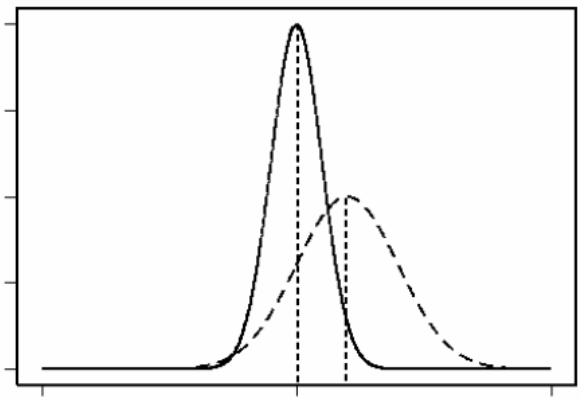
Note: The location of the normal distribution depends on μ and its shape depends on σ. Suppose we have two normal distributions:

_____ $N(\mu_1, \sigma_1)$

– – – – – $N(\mu_2, \sigma_2)$



$\mu_1 < \mu_2$, $\sigma_1 = \sigma_2$

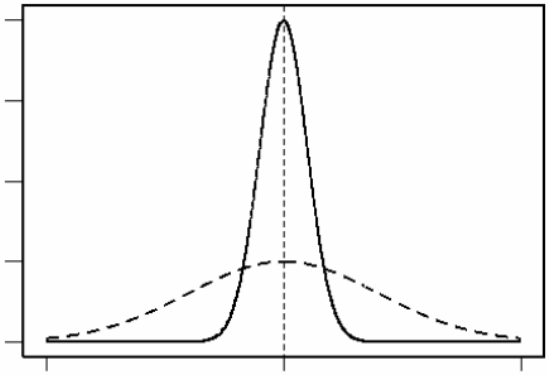


$\mu_1 < \mu_2,\ \sigma_1 < \sigma_2$

STANDARD NORMAL

The normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ is called the standard normal distribution and is denoted by $Normal(0,1)$ or $N(0,1)$. The standard normal random variable is denoted by (Z), and write:

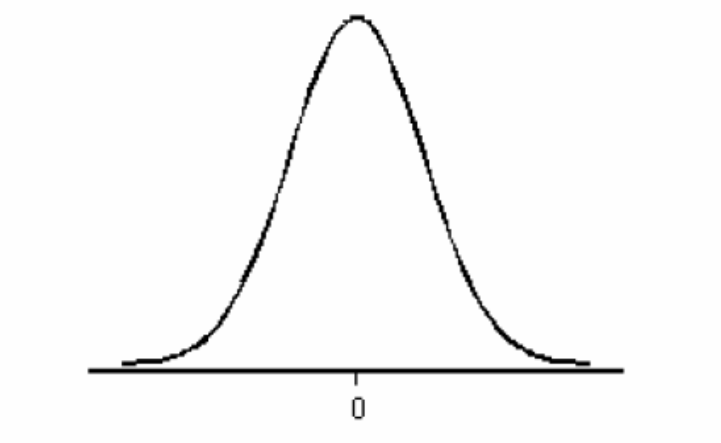
$$Z \sim N(0,1)$$



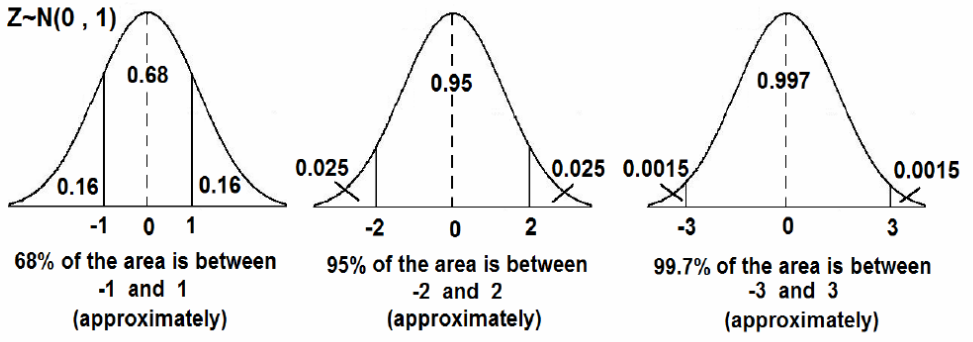
$\mu_1 = \mu_2,\ \sigma_1 < \sigma_2$

The probability density function (pdf) of $Z \sim N(0,1)$ is given by:

$$f(z) = n(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$



The standard normal distribution, $Normal(0,1)$, is very important because probabilities of any normal distribution can be calculated from the probabilities of the standard normal distribution.



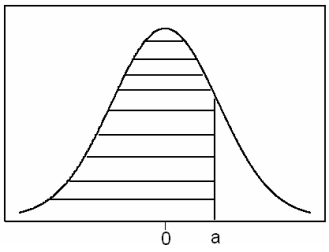
Result:

If $X \sim Normal(\mu, \sigma^2)$, then $Z = \dfrac{X-\mu}{\sigma} \sim Normal(0,1)$.

Calculating Probabilities of Normal (0,1):

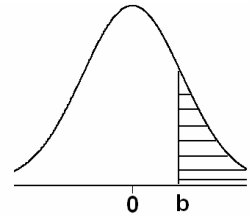
Suppose $Z \sim Normal(0,1)$. For the standard normal distribution $Z \sim N(0,1)$, there is a special table used to calculate probabilities if the form, $P(Z \le a)$:

i. $P(Z \le a) =$ From the table



ii. $P(Z \ge b) = 1 - P(Z \le b)$:

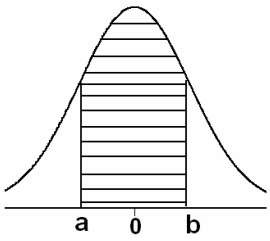
Where $P(Z \le b) =$ from the table



iii. $P(a \le Z \le b) = P(Z \le b) - P(Z \le a)$

Where: $P(Z \le b) =$ from the table

$P(Z \le a) =$ from the table

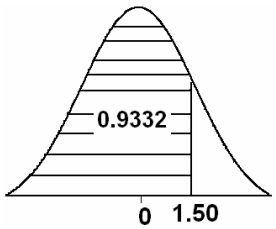


iv. $P(Z = a) = 0$ for every a.

Example:

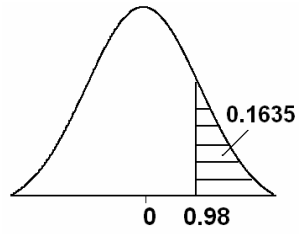
Suppose that $Z \sim N(0,1)$

1. $P(Z \le 1.50) = 0.9332$

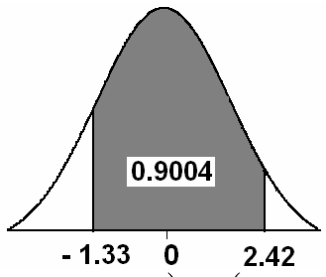


Z	0.00	0.01	…
:	⇓		
1.50 ⇒	0.9332		
:			

2. $P(Z \ge 0.98) = 1 - P(Z \le 0.98) = 1 - 0.8365 = 0.1635$



Z	0.00	…	0.08
:	:	:	⇓
:	…	…	⇓
0.90⇒	⇒	⇒	0.8365

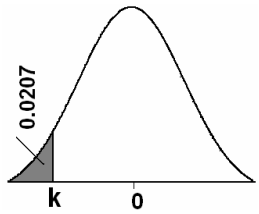


Z	…		−0.03
:	:		⇓
−1.30	⇒		0.0918
:			

3. $P(-1.33 \le Z \le 2.42) = P(Z \ge 2.42) - P(Z \le -1.33) = 0.9922 - 0.0918 = 0.9004$

Example:

Suppose that $Z \sim N(0,1)$. Find the value of k such that $P(Z \le k) = 0.0207$.



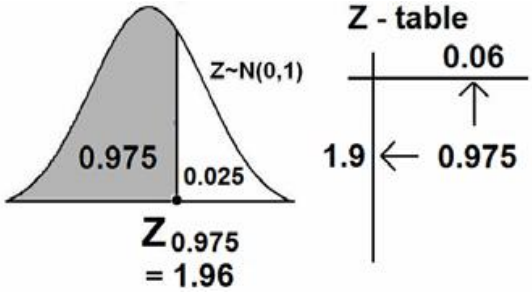
Z	...	−0.04	
:	:	⇑ ⇑	
−2.0	⇐⇐	0.0207	
:			

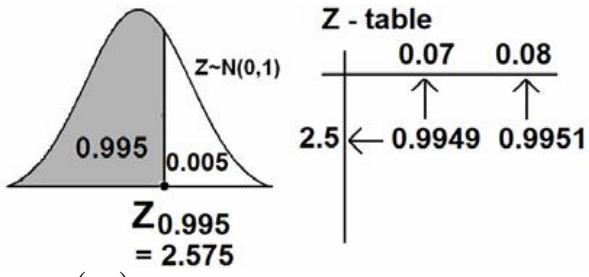
Solution:

$k = -2.04$. Notice that $k = Z_{0.0207} = -2.04$

Example:

If





$Z \sim N(0,1)$, then:

$Z_{0.90} = 1.285$

$Z_{0.95} = 1.645$

$Z_{0.975} = 1.96$

$Z_{0.99} = 2.325$

Using the result: $Z_A = -Z_{1-A}$

$$Z_{0.10} = -Z_{0.90} = -1.285$$
$$Z_{0.05} = -Z_{0.95} = -1.645$$
$$Z_{0.025} = -Z_{0.975} = -1.96$$
$$Z_{0.01} = -Z_{0.99} = -2.325$$

Calculating Probabilities of Normal (μ, σ^2):

- Recall the result:

$$X \sim Normal(\mu, \sigma^2) \Leftrightarrow Z = \frac{X-\mu}{\sigma} \sim Normal(0,1)$$

- $X \le a \Leftrightarrow \frac{X-\mu}{\sigma} \le \frac{a-\mu}{\sigma} \Leftrightarrow z \le \frac{a-\mu}{\sigma}$

1. $P(X \le a) = P\left(z \le \frac{a-\mu}{\sigma}\right) =$ from the table.
2. $P(X \ge a) = 1 - P(X \le a) = 1 - P\left(z \le \frac{a-\mu}{\sigma}\right)$
3. $P(a \le X \le b) = P(X \le b) - P(X \le a)$
$$= P\left(Z \le \frac{b-\mu}{\sigma}\right) - P\left(Z \le \frac{a-\mu}{\sigma}\right)$$
4. $P(X = a) = 0,$ for every a.

NORMAL DISTRIBUTION APPLICATION

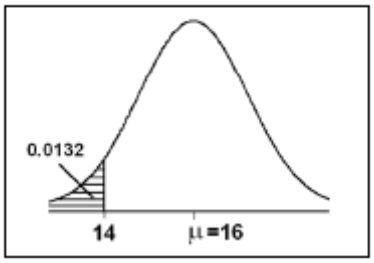
Example

Suppose that the hemoglobin levels of healthy adult males are approximately normally distributed with a mean of 16 and a variance of 0.81.

(a) Find that probability that a randomly chosen healthy adult male has a hemoglobin level less than 14.

(b) What is the percentage of healthy adult males who have hemoglobin level less than 14?

(c) In a population of 10,000 healthy adult males, how many would you expect to have hemoglobin level less than 14?



Solution:

$X =$ hemoglobin level for healthy adult males

Mean: $\mu = 16$

Variance: $\sigma^2 = 0.81$

Standard deviation: $\sigma = 0.9$

$X \sim \text{Normal}\ (16, 0.81)$

(a) The probability that a randomly chosen healthy adult male has hemoglobin level less than 14 is $P(X \le 14)$.

$$P(X \le 14) = P\left(Z \le \frac{14-\mu}{\sigma}\right)$$

$$= P\left(Z \le \frac{14-16}{0.9}\right)$$

$$= P(Z \le -2.22)$$

$$= 0.0132$$

(b) The percentage of healthy adult males who have hemoglobin level less than 14 is:

$$P(X \le 14) \times 100\% = 0.0132 \times 100\% = 1.32\%$$

(c) In a population of 10000 healthy adult males, we would expect that the number of males with hemoglobin level less than 14 to be:

$$P(X \le 14) \times 10000 = 0.0132 \times 10000 = 132 \text{ males}$$

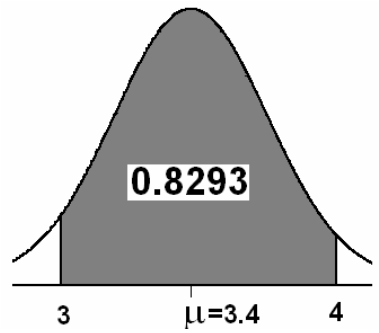
Example:

Suppose that the birth weight of Saudi babies has a normal distribution with mean $\mu = 3.4$ and standard deviation $\sigma = 0.35$.

(a) Find the probability that a randomly chosen Saudi baby has a birth weight between 3.0 and 4.0 kg.

(b) What is the percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg.

(c) In a population of 100000 Saudi babies, how many would you expect to have birth weight between 3.0 and 4.0 kg?



Solution:

$X =$ birth weight of Saudi babies

Mean: $\mu = 3.4$

Standard deviation: $\sigma = 0.35$

Variance: $\sigma^2 = (0.35)^2 = 0.1225$

$X \sim \text{Normal}\ (3.4, 0.1225)$

(a) The probability that a randomly chosen baby has a birth weight between 3.0 and 4.0 kg is

$$P(3.0 \le X \le 4.0) = P(X \le 4.0) - P(X \le 3.0)$$

$$= P\left(Z \le \frac{4.0 - \mu}{\sigma}\right) - P\left(Z \le \frac{3.0 - \mu}{\sigma}\right)$$

$$= P\left(Z \le \frac{4.0 - 3.4}{0.35}\right) - P\left(Z \le \frac{3.0 - 3.4}{0.35}\right)$$

$$= P(Z \le 1.71) - P(Z \le -1.14)$$

$$= 0.9564 - 0.1271 = 0.8293$$

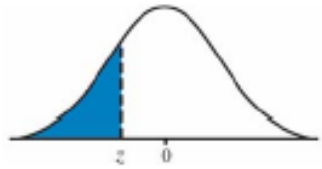
(b) The percentage of Saudi babies who have a birth weight between 3.0 and 4.0 kg is

$$P(3.0 \le X \le 4.0) \times 100\% = 0.8293 \times 100\% = 82.93\%$$

(c) In a population of 100, 000 Saudi babies, we would expect that the number of babies with birth weight between 3.0 and 4.0 kg to be

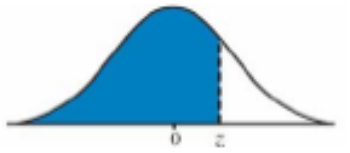
$$P(3.0 \le X \le 4.0) \times 100000 = 0.8293 \times 100000 = 82930$$

Standard Normal Table
Areas Under the Standard Normal Curve



z	-0.09	-0.08	-0.07	-0.06	-0.05	-0.04	-0.03	-0.02	-0.01	-0.00	z
-3.50	0.00017	0.00017	0.00018	0.00019	0.00019	0.00020	0.00021	0.00022	0.00022	0.00023	**-3.50**
-3.40	0.00024	0.00025	0.00026	0.00027	0.00028	0.00029	0.00030	0.00031	0.00032	0.00034	**-3.40**
-3.30	0.00035	0.00036	0.00038	0.00039	0.00040	0.00042	0.00043	0.00045	0.00047	0.00048	**-3.30**
-3.20	0.00050	0.00052	0.00054	0.00056	0.00058	0.00060	0.00062	0.00064	0.00066	0.00069	**-3.20**
-3.10	0.00071	0.00074	0.00076	0.00079	0.00082	0.00084	0.00087	0.00090	0.00094	0.00097	**-3.10**
-3.00	0.00100	0.00104	0.00107	0.00111	0.00114	0.00118	0.00122	0.00126	0.00131	0.00135	**-3.00**
-2.90	0.00139	0.00144	0.00149	0.00154	0.00159	0.00164	0.00169	0.00175	0.00181	0.00187	**-2.90**
-2.80	0.00193	0.00199	0.00205	0.00212	0.00219	0.00226	0.00233	0.00240	0.00248	0.00256	**-2.80**
-2.70	0.00264	0.00272	0.00280	0.00289	0.00298	0.00307	0.00317	0.00326	0.00336	0.00347	**-2.70**
-2.60	0.00357	0.00368	0.00379	0.00391	0.00402	0.00415	0.00427	0.00440	0.00453	0.00466	**-2.60**
-2.50	0.00480	0.00494	0.00508	0.00523	0.00539	0.00554	0.00570	0.00587	0.00604	0.00621	**-2.50**
-2.40	0.00639	0.00657	0.00676	0.00695	0.00714	0.00734	0.00755	0.00776	0.00798	0.00820	**-2.40**
-2.30	0.00842	0.00866	0.00889	0.00914	0.00939	0.00964	0.00990	0.01017	0.01044	0.01072	**-2.30**
-2.20	0.01101	0.01130	0.01160	0.01191	0.01222	0.01255	0.01287	0.01321	0.01355	0.01390	**-2.20**
-2.10	0.01426	0.01463	0.01500	0.01539	0.01578	0.01618	0.01659	0.01700	0.01743	0.01786	**-2.10**
-2.00	0.01831	0.01876	0.01923	0.01970	0.02018	0.02068	0.02118	0.02169	0.02222	0.02275	**-2.00**
-1.90	0.02330	0.02385	0.02442	0.02500	0.02559	0.02619	0.02680	0.02743	0.02807	0.02872	**-1.90**
-1.80	0.02938	0.03005	0.03074	0.03144	0.03216	0.03288	0.03362	0.03438	0.03515	0.03593	**-1.80**
-1.70	0.03673	0.03754	0.03836	0.03920	0.04006	0.04093	0.04182	0.04272	0.04363	0.04457	**-1.70**
-1.60	0.04551	0.04648	0.04746	0.04846	0.04947	0.05050	0.05155	0.05262	0.05370	0.05480	**-1.60**
-1.50	0.05592	0.05705	0.05821	0.05938	0.06057	0.06178	0.06301	0.06426	0.06552	0.06681	**-1.50**
-1.40	0.06811	0.06944	0.07078	0.07215	0.07353	0.07493	0.07636	0.07780	0.07927	0.08076	**-1.40**
-1.30	0.08226	0.08379	0.08534	0.08691	0.08851	0.09012	0.09176	0.09342	0.09510	0.09680	**-1.30**
-1.20	0.09853	0.10027	0.10204	0.10383	0.10565	0.10749	0.10935	0.11123	0.11314	0.11507	**-1.20**
-1.10	0.11702	0.11900	0.12100	0.12302	0.12507	0.12714	0.12924	0.13136	0.13350	0.13567	**-1.10**
-1.00	0.13786	0.14007	0.14231	0.14457	0.14686	0.14917	0.15151	0.15386	0.15625	0.15866	**-1.00**
-0.90	0.16109	0.16354	0.16602	0.16853	0.17106	0.17361	0.17619	0.17879	0.18141	0.18406	**-0.90**
-0.80	0.18673	0.18943	0.19215	0.19489	0.19766	0.20045	0.20327	0.20611	0.20897	0.21186	**-0.80**
-0.70	0.21476	0.21770	0.22065	0.22363	0.22663	0.22965	0.23270	0.23576	0.23885	0.24196	**-0.70**
-0.60	0.24510	0.24825	0.25143	0.25463	0.25785	0.26109	0.26435	0.26763	0.27093	0.27425	**-0.60**
-0.50	0.27760	0.28096	0.28434	0.28774	0.29116	0.29460	0.29806	0.30153	0.30503	0.30854	**-0.50**
-0.40	0.31207	0.31561	0.31918	0.32276	0.32636	0.32997	0.33360	0.33724	0.3409	0.34458	**-0.40**
-0.30	0.34827	0.35197	0.35569	0.35942	0.36317	0.36693	0.37070	0.37448	0.37828	0.38209	**-0.30**
-0.20	0.38591	0.38974	0.39358	0.39743	0.40129	0.40517	0.40905	0.41294	0.41683	0.42074	**-0.20**
-0.10	0.42465	0.42858	0.43251	0.43644	0.44038	0.44433	0.44828	0.45224	0.45620	0.46017	**-0.10**
-0.00	0.46414	0.46812	0.47210	0.47608	0.48006	0.48405	0.48803	0.49202	0.49601	0.50000	**-0.00**

Standard Normal Table (continued)
Areas Under the Standard Normal Curve



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09	z
0.00	0.50000	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.52790	0.53188	0.53586	**0.00**
0.10	0.53983	0.54380	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535	**0.10**
0.20	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409	**0.20**
0.30	0.61791	0.62172	0.62552	0.62930	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173	**0.30**
0.40	0.65542	0.65910	0.66276	0.66640	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793	**0.40**
0.50	0.69146	0.69497	0.69847	0.70194	0.70540	0.70884	0.71226	0.71566	0.71904	0.72240	**0.50**
0.60	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.75490	**0.60**
0.70	0.75804	0.76115	0.76424	0.76730	0.77035	0.77337	0.77637	0.77935	0.78230	0.78524	**0.70**
0.80	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327	**0.80**
0.90	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891	**0.90**
1.00	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214	**1.00**
1.10	0.86433	0.86650	0.86864	0.87076	0.87286	0.87493	0.87698	0.87900	0.88100	0.88298	**1.10**
1.20	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147	**1.20**
1.30	0.90320	0.90490	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774	**1.30**
1.40	0.91924	0.92073	0.92220	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189	**1.40**
1.50	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408	**1.50**
1.60	0.94520	0.94630	0.94738	0.94845	0.94950	0.95053	0.95154	0.95254	0.95352	0.95449	**1.60**
1.70	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.96080	0.96164	0.96246	0.96327	**1.70**
1.80	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062	**1.80**
1.90	0.97128	0.97193	0.97257	0.97320	0.97381	0.97441	0.97500	0.97558	0.97615	0.97670	**1.90**
2.00	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.98030	0.98077	0.98124	0.98169	**2.00**
2.10	0.98214	0.98257	0.98300	0.98341	0.98382	0.98422	0.98461	0.98500	0.98537	0.98574	**2.10**
2.20	0.98610	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.98840	0.98870	0.98899	**2.20**
2.30	0.98928	0.98956	0.98983	0.99010	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158	**2.30**
2.40	0.99180	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361	**2.40**
2.50	0.99379	0.99396	0.99413	0.99430	0.99446	0.99461	0.99477	0.99492	0.99506	0.99520	**2.50**
2.60	0.99534	0.99547	0.99560	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643	**2.60**
2.70	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.99720	0.99728	0.99736	**2.70**
2.80	0.99744	0.99752	0.99760	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807	**2.80**
2.90	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861	**2.90**
3.00	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.9990	**3.00**
3.10	0.99903	0.99906	0.99910	0.99913	0.99916	0.99918	0.99921	0.99924	0.99926	0.99929	**3.10**
3.20	0.99931	0.99934	0.99936	0.99938	0.99940	0.99942	0.99944	0.99946	0.99948	0.99950	**3.20**
3.30	0.99952	0.99953	0.99955	0.99957	0.99958	0.99960	0.99961	0.99962	0.99964	0.99965	**3.30**
3.40	0.99966	0.99968	0.99969	0.99970	0.99971	0.99972	0.99973	0.99974	0.99975	0.99976	**3.40**
3.50	0.99977	0.99978	0.99978	0.99979	0.99980	0.99981	0.99981	0.99982	0.99983	0.99983	**3.50**

Critical Values of the t-distribution (t_α)

ν=df	$t_{0.90}$	$t_{0.95}$	$t_{0.975}$	$t_{0.99}$	$t_{0.995}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.3062	1.6896	2.0301	2.4377	2.7238
40	1.3030	1.6840	2.0210	2.4230	2.7040
45	1.3006	1.6794	2.0141	2.4121	2.6896
50	1.2987	1.6759	2.0086	2.4033	2.6778
60	1.2958	1.6706	2.0003	2.3901	2.6603
70	1.2938	1.6669	1.9944	2.3808	2.6479
80	1.2922	1.6641	1.9901	2.3739	2.6387
90	1.2910	1.6620	1.9867	2.3685	2.6316
100	1.2901	1.6602	1.9840	2.3642	2.6259
120	1.2886	1.6577	1.9799	2.3578	2.6174
140	1.2876	1.6558	1.9771	2.3533	2.6114
160	1.2869	1.6544	1.9749	2.3499	2.6069
180	1.2863	1.6534	1.9732	2.3472	2.6034
200	1.2858	1.6525	1.9719	2.3451	2.6006
∞	1.282	1.645	1.960	2.326	2.576

3.6 ESTIMATION ABOUT POPULATION PARAMETERS

Introduction:

Statistical Inferences: (Estimation and Hypotheses Testing)

It is the procedure by which we reach a conclusion about a population on the basis of the information contained in a sample drawn from that population.

There are two main purposes of statistics;

- Descriptive Statistics: Organization & summarization of the data
- Statistical Inference: Answering research questions about some unknown population parameters.

1. **Estimation:**

Approximating (or estimating) the actual values of the unknown parameters:

- **Point Estimate:** A point estimate is single value used to estimate the corresponding population parameter.

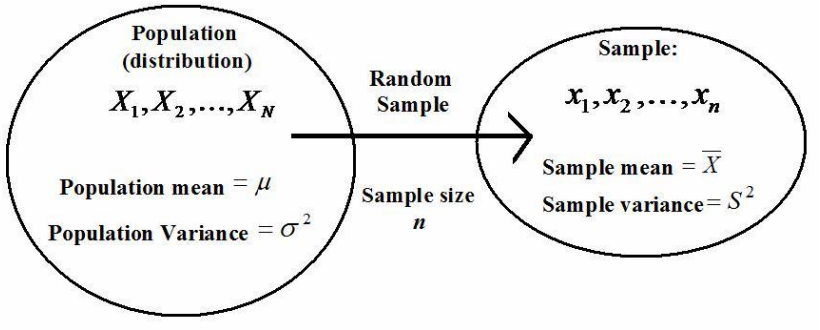
- **Interval Estimate (or Confidence Interval):** An interval estimate consists of two numerical values defining a range of values that most likely includes the parameter being estimated with a specified degree of confidence.

2. **Hypothesis Testing:**

Answering research questions about the unknown parameters of the population (confirming or denying some conjectures or statements about the unknown parameters).

Confidence Interval for a Population Mean (μ):

In this section, we are interested in estimating the mean of a certain population (μ).



Population	Sample
Population size $= N$	Sample size $= n$
Population values: $X_1, X_2, ..., X_N$	Sample values: $x_1, x_2, ..., x_n$
Population mean: $\mu = \dfrac{\sum_{i=1}^{N} X_i}{N}$	Sample mean: $\bar{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$
Population mean: $\sigma^2 = \dfrac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$	Sample mean: $s^2 = \dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$

i. Point Estimation of μ:

A point estimate of the mean is a single number used to estimate (or approximate) the true value of μ.

- Draw a random sample of size n from the population:

$$x_1, x_2, ..., x_n$$

- Compute the sample mean:

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

Result:

The sample mean $\bar{x} = \dfrac{1}{n} \sum_{i=1}^{n} x_i$ is a "good" point estimate of the population mean (μ).

ii. Confidence Interval (Interval Estimate) of μ:

An interval estimate of μ is an interval (L,U) containing the true value of μ "with a probability of $1-\alpha$".

- $1-\alpha =$ is called the confidence coefficient (level)
- $L =$ lower limit of the confidence interval
- $U =$ upper limit of the confidence interval

Result: (For the case when σ is known)

1. If $X_1, X_2, ..., X_N$ is a random sample of size n from a normal distribution with mean μ and known variance σ^2, then: A $(1-\alpha)\times 100\%$ confidence interval for μ is:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \sigma_{\overline{X}}$$

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

1. If $X_1, X_2, ..., X_N$ is a random sample of size n from a non-normal distribution with mean μ and known variance σ^2, and if the sample size n is large $(n \geq 30)$, then: An approximate $(1-\alpha)\times 100\%$ confidence interval for μ is:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \sigma_{\overline{X}}$$

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right)$$

$$\overline{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Note that:

1. We are $(1-\alpha)\times 100\%$ confident that the true value of μ belongs to the interval $\left(\overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$
2. Upper limit of the confident interval $= \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
3. Lower limit of the confidence interval $= \overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$
4. $Z_{1-\frac{\alpha}{2}}$ = Reliability Coefficient
5. $Z_{1-\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ = margin of error = precision of the estimate
6. In general, the interval estimate (confidence interval) may be expressed as follows:

$$\overline{X} \pm Z_{1-\frac{\alpha}{2}}\sigma_{\overline{X}}$$

Estimator $\pm$ (reliability coefficient) $\times$ standard error

Estimator $\pm$ margin of error

3.6.1 THE T DISTRIBUTION:

Confidence Interval using t: We have already introduced and discussed the t distribution.

Result: For the case when σ is known normal + population.

If $X_1, X_2, ..., X_N$ is a random sample of size n from a normal distribution with mean μ and unknown variance σ^2, then: A $(1-\alpha)\times 100\%$ confidence interval for μ is:

$$\overline{X} \pm t_{1-\frac{\alpha}{2}}\hat{\sigma}_{\overline{X}}$$

$$\overline{X} \pm t_{1-\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$$

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$$

Where the degree of freedom is:

$$df = v = n - 1$$

Note that:

1. We are $(1-\alpha)\times 100\%$ confident that the true value of μ belongs to the interval

   $$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right)$$

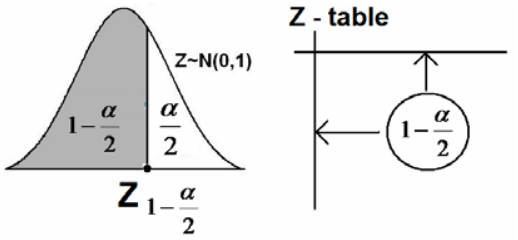
2. $\hat{\sigma}_{\overline{X}} \frac{S}{\sqrt{n}}$ (estimate of the standard error of $\overline{X}$)
3. $t_{1-\frac{\alpha}{2}}$ = Reliability Coefficient
4. In this case, we replace σ by S and Z by t.
5. In general, the interval estimate (confidence interval may be expressed as follows:)

   Estimator $\pm$ (Reliability Coefficient) $\times$ (Estimate of the Standard Error)

   $$\overline{X} \pm t_{1-\frac{\alpha}{2}} \sigma_{\overline{X}}$$

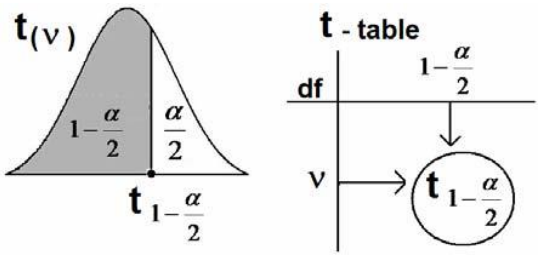
Notes: Finding Reliability Coefficient

1. We find the reliability coefficient $Z_{1-\frac{\alpha}{2}}$ from the $Z-$ table as follows:



2. We find the reliability coefficient $t_{1-\frac{\alpha}{2}}$ from the $t-$table as follows:

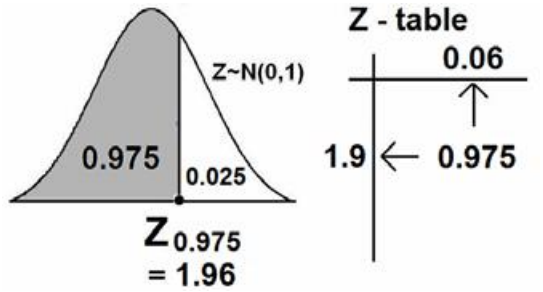
$(df = v = n-1)$

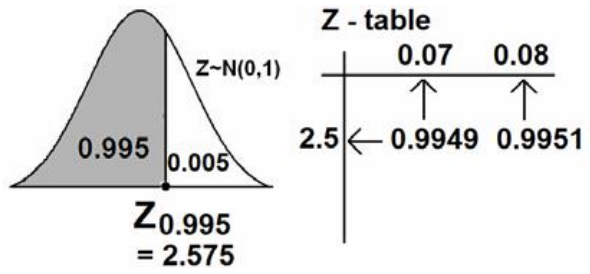


Example:

Suppose that $Z \sim N(0,1)$. Find $Z_{1-\frac{\alpha}{2}}$ for the following cases:

1. $\alpha = 0.1$
2. $\alpha = 0.05$
3. $\alpha = 0.01$





Solution:

1. For $\alpha = 0.1$:

$$1-\frac{\alpha}{2} = 1-\frac{0.1}{2} = 0.95 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$$

2. For $\alpha = 0.05$:

$$1-\frac{\alpha}{2} = 1-\frac{0.05}{2} = 0.975 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96$$

3. For $\alpha = 0.01$:

$$1-\frac{\alpha}{2} = 1-\frac{0.01}{2} = 0.995 \quad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.995} = 2.575$$

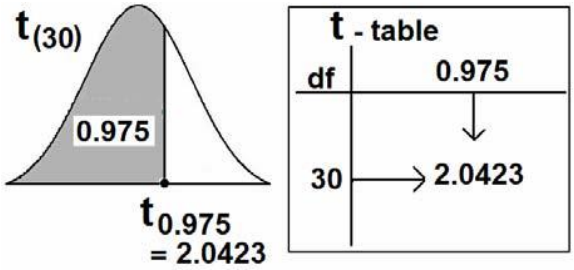
Example:

Suppose that $t \sim t(30)$. Find $t_{1-\frac{\alpha}{2}}$ for $\alpha = 0.05$

Solution:

$df = v = 30$

$$1-\frac{\alpha}{2} = 1-\frac{0.01}{2} = 0.995 \qquad \Rightarrow \quad Z_{1-\frac{\alpha}{2}} = Z_{0.995} = 2.575$$



Example: (the case where σ^2 is known)

Diabetic ketoacidosis is a potential fatal complication of diabetes mellitus throughout the world and is characterized in part by very high blood glucose levels. In a study on 123 patients living in Saudi Arabia of age 15or more who were admitted for diabetic ketoacidosis, the mean blood glucose level was 26.2 mmol/l. Suppose that the blood glucose levels for such patients have a normal distribution with a standard deviation of

3.3 mmol/l.

1. Find a point estimate for the mean blood glucose level of such diabetic ketoacidosis patients.
2. Find a 90% confidence interval for the mean blood glucose level of such diabetic ketoacidosis patients.

Solution

Variable = X = blood glucose level (quantitative variable).

Population = diabetic ketoacidosis patients in Saudi Arabia of age 15 or more.

Parameter of interest is: μ = the mean blood glucose level.

Distribution is normal with standard deviation $\sigma = 3.3$.

σ^2 is known $(\sigma^2 = 10.89)$

X ~ Normal $(\mu, 10.89)$

μ = ?? (unknown – we need to estimate μ)

Sample size: $n = 123$ (large)

Sample Mean: $\overline{X} = 26.2$

1. **Point Estimation:**

   We need to find a point estimate for μ.

   $\overline{X} = 26.2$ is a point estimate for μ.

   $\mu \approx 26.2$

2. **Interval Estimation (Confidence Interval = CI):**

   We need to find 90% C. I for μ.

   $90\% = (1-\alpha)100\%$

   $1-\alpha = 0.9 \Leftrightarrow \alpha = 0.1 \Leftrightarrow \frac{\alpha}{2} = 0.5 \Leftrightarrow 1-\frac{\alpha}{2} = 0.95$

   The reliability coefficient is: $Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$

   90% confidence interval for μ. is:

$$\left(\overline{X} - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right)$$

$$\left(26.2 - (1.645)\frac{3.3}{\sqrt{123}}, 26.2 + (1.645)\frac{3.3}{\sqrt{123}}\right)$$

$$(26.2 - 0.489471, 26.2 + 0.4894714)$$

$$(25.710529, 26.6894714)$$

   We are 90% confident that the true value of the mean μ. lies in the interval $(25.71, 26.69)$, that is:

$$25.71 < \mu < 26.69$$

   Note: for this example, even if the distribution is not normal, we may use the same solution because the sample size $n = 123$is large.

Example: (The case where σ^2 is unknown)

A study was conducted to study the age characteristics of Saudi women having breast lump. A sample of 121 Saudi women gave a mean of 37 years with a standard deviation of 10 years. Assume that the ages of Saudi women having breast lumps are normally distributed.

a. Find a point estimate for the mean age of Saudi women having breast lumps.
b. Construct a 99% confidence interval for the mean age of Saudi women having breast lumps

Solution

X = Variable = Age of Saudi women having breast lumps (quantitative variable).

Population = All Saudi women having breast lumps.

Parameter of interest is: μ = the age mean of Saudi women having breast lumps.

$X \sim \text{Normal}\left(\mu, \sigma^2\right)$

$\mu = ??$ (unknown – we need to estimate μ)

$\sigma^2 = ??$ (unknown)

Sample size: $n = 121$

Sample Mean: $\overline{X} = 37$

Sample standard deviation: $S = 10$

Degrees of freedom: $df = v = 121 - 1 = 120$

a. Point Estimation: We need to find a point estimate for μ.

   $\overline{X} = 37$ is a "good" point estimate for μ.

   $\mu \approx 37$ years

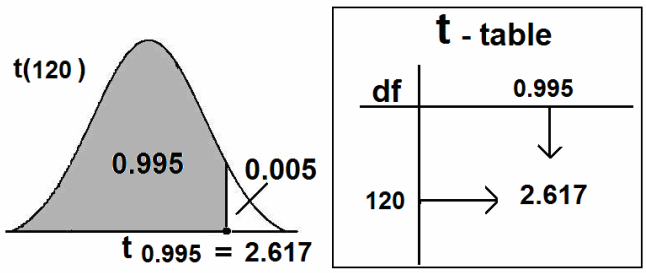
b. Interval Estimation (Confidence Interval = C. I): We need to find 99% C. I for μ.

   $$99\% = (1-\alpha)100\%$$

   $$1-\alpha = 0.99 \Leftrightarrow \alpha = 0.01 \qquad \Leftrightarrow \frac{\alpha}{2} = 0.005 \qquad \Leftrightarrow 1-\frac{\alpha}{2} = 0.995$$

   $$v = df = 120$$

   The reliability coefficient is: $t_{1-\frac{\alpha}{2}} = t_{0.995} = 2.617$



99% confidence interval for μ is:

$$\overline{X} \pm t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$$

$$37 \pm (2.617) \frac{10}{\sqrt{121}}$$

$$37 \pm 2.38$$

$$(37-2.38, 37+2.38)$$

$$(34.62, 39.38)$$

Another way:

$$\left(\overline{X} - t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}, \overline{X} + t_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \right)$$

$$\left(37 - (2.617) \frac{10}{\sqrt{121}}, 37 + (2.617) \frac{10}{\sqrt{121}} \right)$$

$$(37-2.38, 37+2.38)$$

$$(34.62, 39.38)$$

We are 99% confident that the true value of the mean μ lies in the interval (34.61, 39.39), that is:

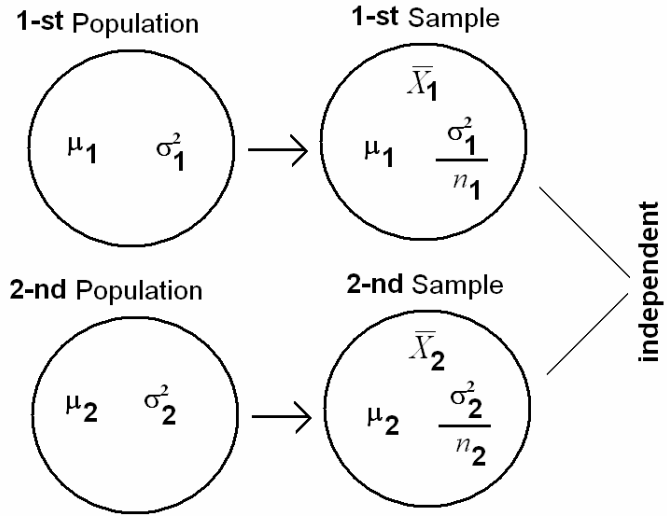
$$(34.62 < \mu < 39.38)$$

Confidence Interval for the Difference between Two Population Means (μ_1, μ_2):

Suppose that we have two populations:

- 1st population with mean μ_1 and variance σ_1^2

- 2nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means $(\mu_1 - \mu_2)$
- We independently select a random sample of size n_1 from the 1st population and another random sample of size n_2 from the 2nd population:
- Let $\overline{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1st sample.
- Let $\overline{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2nd sample.
- The sampling distribution of $\overline{X}_1 - \overline{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.



Recall:

1. Mean of $\overline{X}_1 - \overline{X}_2$ is: $\mu_{\overline{X}_1 - \overline{X}_2} = \mu_1 - \mu_2.$
2. Variance of $\overline{X}_1 - \overline{X}_2$ is: $\sigma^2_{\overline{X}_1 - \overline{X}_2} = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$
3. Standard error of $\overline{X}_1 - \overline{X}_2$ is: $\sigma_{\overline{X}_1 - \overline{X}_2} = \sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$
4. If the two random samples were selected from normal distributions (or non-normal distributions with large sample sizes) with known variances σ_1^2 and

σ_2^2, then the difference between the sample means $(\overline{X}_1 - \overline{X}_2)$ has a normal distribution with mean $(\mu_1 - \mu_2)$ and variance $((\sigma_1^2/n_1)+(\sigma_2^2/n_2))$, that is:

- $\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$
- $Z = \dfrac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Point Estimation of $\mu_1 - \mu_2$:

Result:

$\overline{X}_1 - \overline{X}_2$ is a "good" point estimate for $\mu_1 - \mu_2$.

Interval Estimation (Confidence Interval) of $\mu_1 - \mu_2$:

we will consider two cases.

1. First Case: σ_1^2 and σ_2^2 are known:

   If σ_1^2 and σ_2^2 are known, we use the following result to find an interval estimate for $\mu_1 - \mu_2$.

   Result:

   A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\overline{X}_1 - \overline{X}_2) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\left((\overline{X}_1 - \overline{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\overline{X}_1 - \overline{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$(\overline{X}_1 - \overline{X}_2) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X}_1 - \overline{X}_2) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Estimator ± (Reliability Coefficient) × (Standard Error)

2. **Second Case:**

   Unknown equal variances: ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown):

If σ_1^2 and σ_2^2 are equal but unknown ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the pooled estimate of the common variance σ^2 is

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

Where S_1^2 is the variance of the 1st sample and S_1^2 is the variance of the 2nd sample. The degree of freedom of S_p^2 is

$$df = v = n_1 + n_2 - 2$$

We use the following result to find an interval estimate for $\mu_1 - \mu_2$ when we have normal populations with unknown and equal variances.

Result:

A $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

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

Where reliability coefficient $t_{1-\frac{\alpha}{2}}$ is the t-value with $df = v = n_1 + n_2 - 2$ degrees of freedom.

Example: (1st Case: σ_1^2 and σ_2^2 are known)

An experiment was conducted to compare time length (duration time) of two types of surgeries (A) and (B). 75 surgeries of type (A) and 50 surgeries of type (B) were performed. The average time length for (A) was 42 minutes and average for (B) was 36 minutes.

1. Find a point estimate for $\mu_A - \mu_B$ where μ_A and μ_B are population means of the time length of surgeries of type (A) and (B), respectively.
2. Find a 96% confidence interval for $\mu_A - \mu_B$. Assume that the population standard deviations are 8 and 6 for type (A) and (B), respectively.

Solution:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 75$	$n_B = 50$
Sample Mean	$\bar{X}_A = 42$	$\bar{X}_B = 36$
Population Standard Deviation	$\sigma_A = 8$	$\sigma_B = 6$

1. A point estimate for $\mu_A - \mu_B$ is:

$$\bar{X}_A - \bar{X}_B = 42 - 36 = 6$$

2. Finding a 96% confidence interval for $\mu_A - \mu_B$:

$\alpha = ??$

$96\% = (1-\alpha)100\% \Leftrightarrow 0.96 = (1-\alpha) \Leftrightarrow \alpha = 0.04 \Leftrightarrow \alpha/2 = 0.02$

Reliability Coefficient: $Z_{1-\frac{\alpha}{2}} = Z_{0.98} = 2.055$

A 96% C.I for $\mu_A - \mu_B$ is:

$$(\bar{X}_A - \bar{X}_B) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}$$

$$6 \pm Z_{0.98} \sqrt{\frac{8^2}{75} + \frac{6^2}{50}}$$

$$6 \pm (2.055) \sqrt{\frac{64}{75} + \frac{36}{50}}$$

$$6 \pm 2.578$$

$$3.422 < \mu_A - \mu_B < 8.58$$

We are 96% confident that $\mu_A - \mu_B \in (3.42, 8.58)$.

Note: since the confidence interval does not include zero, we conclude that the two population means are not equal $(\mu_A - \mu_B \neq 0 \Leftrightarrow \mu_A \neq \mu_B)$. Therefore, we may conclude that the mean time length is not the same for the two types if surgeries.

Example: (2nd Case: $\sigma_1^2 = \sigma_2^2$ unknown)

To compare the time length (duration time) pf two types surgeries (A) and (B), an experiment shows the following results based on two independent samples:

Type A: 140, 138, 143, 142, 144, 137

Type B: 135, 140, 136, 142, 138, 140

1. Find a point estimate for $\mu_A - \mu_B$ where $\mu_A(\mu_B)$ is the mean time length of type A(B).
2. Assuming normal populations with equal variances, find a 95% confidence interval for $\mu_A - \mu_B$.

Solution:

First, we calculate the mean and the variances of the two samples, and we get:

Surgery	Type (A)	Type (B)
Sample Size	$n_A = 6$	$n_B = 6$
Sample Mean	$\bar{X}_A = 140.67$	$\bar{X}_B = 138.50$
Sample Variance	$S^2_A = 7.87$	$S^2_B = 7.10$

1. A point estimate for $\mu_A - \mu_B$ is:

$$\bar{X}_A - \bar{X}_B = 140.67 - 138.50 = 2.17$$

2. Finding a 95% confidence interval for $\mu_A - \mu_B$:

$95\% = (1-\alpha)100\% \Leftrightarrow 0.95 = (1-\alpha) \Leftrightarrow \alpha = 0.05 \Leftrightarrow \alpha/2 = 0.025$

Reliability Coefficient: $t_{1-\frac{\alpha}{2}} = Z_{0.95} = 2.228$

The pooled estimate of the common variance is:

$$S_p^2 = \frac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}$$

$$= \frac{(6-1)(7.87) + (6-1)(7.1)}{6+6-2} = 7.485$$

A 95% C.I for $\mu_A - \mu_B$ is:

$$(\bar{X}_A - \bar{X}_B) \pm t_{1-\frac{\alpha}{2}} \sqrt{\frac{S_p^2}{n_A} + \frac{S_p^2}{n_B}}$$

$$2.17 \pm (2.228)\sqrt{\frac{7.485}{6}+\frac{7.485}{6}}$$

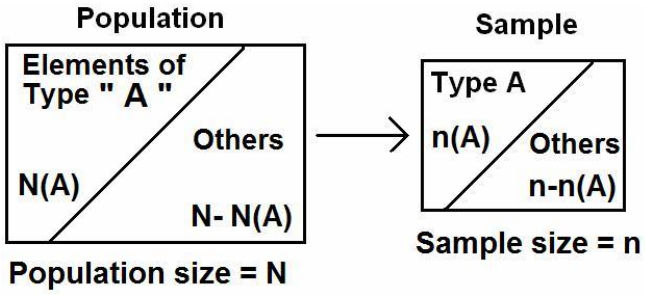
$$2.17 \pm 3.519$$

$$-1.35 < \mu_A - \mu_B < 5.69$$

We are 95% confident that $\mu_A - \mu_B \in (-1.35, 5.69)$.

Note: since the confidence interval includes zero, we conclude that the two population means may be equal $(\mu_A - \mu_B = 0 \Leftrightarrow \mu_A = \mu_B)$. Therefore, we may conclude that the mean time length is the same for both types of surgeries.

Confidence Interval for a Population Proportion (p):



Recall:

1. For the population:

N (A) = number of elements in the population with a specified characteristic "A"

N = total number of elements in the population (population size)

The population proportion is:

$p = \frac{N(A)}{N}$ (p is a parameter)

2. For the sample:

$n(A)$ = number of elements in the sample with the same characteristic "A".

$n = sample\ size$

The sample proportion is:

$$\hat{p} = \frac{n(A)}{n} \qquad (\hat{p}\ is\ a\ statistic)$$

3. The sampling distribution of the sample proportion $(\hat{p})$ is used to make inferences about the population proportion *(p)*.

4. The mean of $(\hat{p})$ is: $\mu_{\hat{p}} = p$

5. The variance of $(\hat{p})$ is: $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$

6. The standard error (standard deviation) of $(\hat{p})$ is:

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

7. For large sample size $(n \geq 30, np > 5, n(1-p) > 5)$, the sample proportion $(\hat{p})$ has approximately a normal distribution with mean $\mu_{\hat{p}} = p$ and a variance $\sigma_{\hat{p}}^2 = p(1-p)/n$, that is:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right) \qquad (approximately)$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1) \qquad (approximately)$$

(i) Point Estimate for (p):

Result:

A good point estimate for the population proportion (p) is the sample proportion ($\hat{p}$) .

(ii) Interval Estimation (Confidence Interval) for (p):

Result:

For large sample size (n $\geq$ 30, np > 5, n(1-p) > 5), an approximate (1-α)100% confidence interval for (p) is:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\left(\hat{p} - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} , \qquad \hat{p} + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

$$\text{Estimator} \pm (\text{Reliability Coefficient}) x \ (Standard\ Error)$$

Example: In a study on the obesity of women, a random sample of 950 women was taken. It was found that 611 of these women were obese (overweight by a certain percentage).

(1) Find a point estimate for the true proportion of women who are obese.

(2) Find a 95% confidence interval for the true proportion of Saudi women who are obese.

Solution:

Variable: whether or not a women is obese (qualitative variable)

Population: all women

Parameter: p =the proportion of women who are obese.

Sample: n = 950 (950 women in the sample)

$n(A)$ = 611 (611 women in the sample who are obese)

The sample proportion (the proportion of women who are obese in the sample.) is:

$$\hat{p} = \frac{n(A)}{n} = \frac{611}{950} = 0.643$$

(1) A point estimate for p is: $\hat{p} = 0.643$

(2) We need to construct 95% C.I. for the proportion (p).

$$95\% = (1-\alpha)100\% \Leftrightarrow 0.95 = 1-\alpha \Leftrightarrow \alpha = 0.05 \Leftrightarrow \frac{\alpha}{2} = 0.025 \Leftrightarrow 1-\frac{\alpha}{2} = 0.975$$

The reliability coefficient: $Z_{1-\frac{\alpha}{2}} = Z_{0.975} = 1.96$

A 95% C.I. for the proportion (p) is:

$$\hat{p} \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$0.643 \pm (1.96)\sqrt{\frac{(0.643)(1-0.643)}{950}}$$

$$0.643 \pm (1.96)(0.01554)$$

$$0.643 \pm 0.0305$$

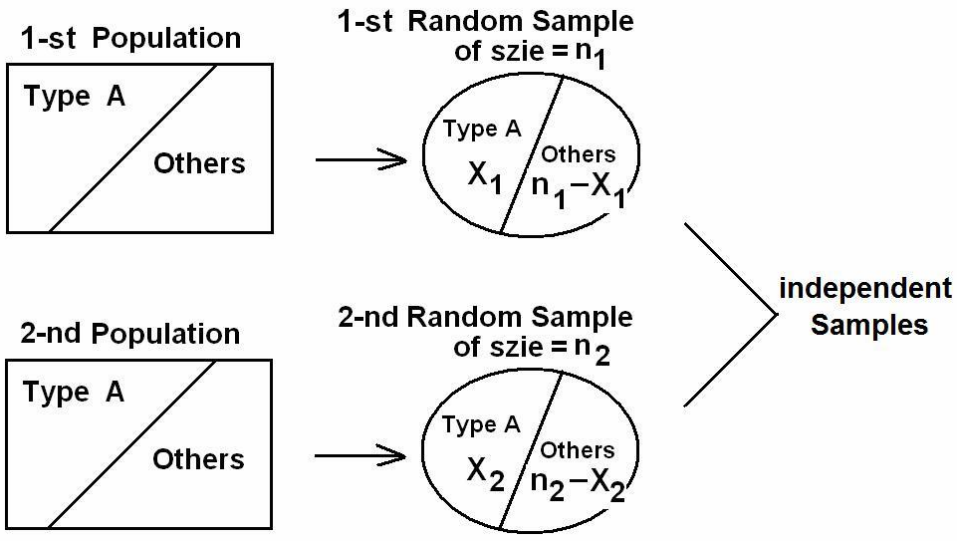
$$(0.6127\,,0.6735)$$

We are 95% confident that the true value of the population proportion of obese women, p, lies in the interval (0 .61, 0.67),

that is:

0.61 < p < 0.67

Confidence Interval for the Difference Between Two Population Proportions ($p_1 - p_2$):



Suppose that we have two populations with:

- p_1 = population proportion of elements of type (A) in the 1-st population.
- p_2 = population proportion of elements of type (A) in the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about $p_1 - p_2$.
- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population :

- • Let X_1 = no. of elements of type (A) in the 1-st sample.
- • Let X_2 = no. of elements of type (A) in the 2-nd sample.

- $\widehat{p_1} = \frac{X_1}{n_1} = the\ sample\ proportion\ of\ the\ 1st\ sample$

- $\widehat{p_2} = \frac{X_2}{n_2} = the\ sample\ proportion\ of\ the\ 2nd\ sample$
- The sampling distribution of $\widehat{p_1} - \widehat{p_2}$ is used to make inferences about $p_1 - p_2$.

Recall:

1. Mean of $\widehat{p_1} - \widehat{p_2}$ is: $\mu_{\widehat{p_1}-\widehat{p_2}} = p_1 - p_2$

2. Variance of $\widehat{p_1} - \widehat{p_2}$ is: $\sigma^2_{\widehat{p_1}-\widehat{p_2}} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

3. Standard error (standard deviation) of $\widehat{p_1} - \widehat{p_2}$ is:

$$\sigma_{\widehat{p_1}-\widehat{p_2}} = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

4. For large samples sizes

$(n_1 \geq 30,\ n_2 \geq 30,\ n_1 p_1 > 5,\ n_2 p_2 > 5,\ n_2 q_2 > 5)$, we have that $\widehat{p_1} - \widehat{p_2}$ has approximately normal distribution with mean $\mu_{\widehat{p_1}-\widehat{p_2}} = p_1 - p_2$ and variance $\sigma^2_{\widehat{p_1}-\widehat{p_2}} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$, that is:

$$\widehat{p_1} - \widehat{p_2} \sim N\left(p_1 - p_2\ , \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right) \qquad (Approximately)$$

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1) \quad (Approximately)$$

Note: $q_1 = 1 - p_1 \quad and \quad q_2 = 1 - p_2$

Point Estimation for $p_1 - p_2$:

Result:

A good point estimator for the difference between the two proportions, $p_1 - p_2$, is:

$$\widehat{p_1} - \widehat{p_2} = \frac{X_1}{n_1} + \frac{X_2}{n_2}$$

Interval Estimation (Confidence Interval) for $p_1 - p_2$:

Result:

For large n_1 and n_2, an approximate $(1-\alpha)100\%$ confidence interval for $p_1 - p_2$ is:

$$(\widehat{p_1} - \widehat{p_2}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$$

$$\left((\widehat{p_1} - \widehat{p_2}) - Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}} \ , \qquad (\widehat{p_1} - \widehat{p_2}) + Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}} \right)$$

$$\text{Estimator} \ \pm \ (\text{Reliability Coefficient}) x \ (Standard \ Error)$$

Example:

A researcher was interested in comparing the proportion of people having cancer disease in two cities (A) and (B). A random sample of 1500 people was taken from the first city (A), and another independent

random sample of 2000 people was taken from the second city (B). It was found that 75 people in the first sample and 80 people in the second sample have cancer disease.

(1) Find a point estimate for the difference between the proportions of people having cancer disease in the two cities.

(2) Find a 90% confidence interval for the difference between the two proportions.

Solution:

p_1 = population proportion of people having cancer disease in the first city (A)

p_2 = population proportion of people having cancer disease in the second city (B)

$\widehat{p_1} = \ sample\ proportion\ of\ the\ 1st\ sample\ \widehat{p_2} =$
$sample\ proportion\ of\ the\ 2nd\ sample$

X_1= number of people with cancer in the first sample

X_2= number of people with cancer in the second sample For the first sample we have:

n_1 = 1500 , X_1=75

$$\widehat{p_1} = \frac{X_1}{n_1} = \frac{75}{1500} = 0.05 \qquad \widehat{q_1} = 1 - 0.05 = 0.95$$

For the second sample we have:

$$n_2 = 2000 \quad , \quad X_2 = 80$$

$$\widehat{p_2} = \frac{X_2}{n_2} = \frac{80}{2000} = 0.04 \quad , \quad \widehat{q_2} = 1 - 0.04 = 0.96$$

(1) Point Estimation for $p_1 - p_2$:

A good point estimate for the difference between the two proportions, $p_1 - p_2$, is:

$$\widehat{p_1} - \widehat{p_2} = 0.05 - 0.04 = 0.01$$

(2) Finding 90% Confidence Interval for $p_1 - p_2$:

90% = (1−α)100% ⇔ 0. 90 = (1−α) ⇔ α=0.1 ⇔ α/2 = 0.05

The reliability coefficient: $Z_{1-\frac{\alpha}{2}} = Z_{0.95} = 1.645$

A 90% confidence interval for $p_1 - p_2$ is:

$$(\widehat{p_1} - \widehat{p_2}) \pm Z_{1-\frac{\alpha}{2}} \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$$

$$(\widehat{p_1} - \widehat{p_2}) \pm Z_{0.95} \sqrt{\frac{\widehat{p_1}\widehat{q_1}}{n_1} + \frac{\widehat{p_2}\widehat{q_2}}{n_2}}$$

$$0.01 \pm 1.645 \sqrt{\frac{(0.05)(0.95)}{1500} + \frac{(0.04)(0.96)}{2000}}$$

$$0.01 \pm 0.01173$$

$$-0.0017 < p_1 - p_2 < 0.0217$$

We are 90% confident that $p_1 - p_2 \in (-0.0017, 0.0217)$.

Note: Since the confidence interval includes zero, we may conclude that the two population proportions are equal ($p_1 - p_2 = 0 \Leftrightarrow p_1 = p_2$). Therefore,

we may conclude that the proportion of people having cancer is the same in both cities.

4 Test Hypotheses About Population Parameters:

In this chapter, we are interested in testing some hypotheses about the unknown population parameters.

4.1 Introduction

Consider a population with some unknown parameter θ. We are interested in testing (confirming or denying) some conjectures about θ. For example, we might be interested in testing the conjecture that $\theta > \theta_o$, where θ_o is a given value.

- A hypothesis is a statement about one or more populations.
- A research hypothesis is the conjecture or supposition that motivates the research.
- A statistical hypothesis is a conjecture (or a statement) concerning the population which can be evaluated by appropriate statistical technique.
- For example, if θ is an unknown parameter of the population, we might be interested in testing the conjecture sating that $\theta \geq \theta_o$ against $\theta < \theta_o$ (for some specific value θ_o).
- We usually test the null hypothesis (H_o) against the alternative (or the research) hypothesis (H_1 or H_A) by choosing one of the following situations:
  - (i) $H_o: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$
  - (ii) $H_o: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$
  - (iii) $H_o: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$

- Equality sign must appear in the null hypothesis.
- H_o is the null hypothesis and H_A is the alternative hypothesis. (H_o and H_A are complement of each other)
- The null hypothesis (H_o) is also called "the hypothesis of no difference".
- The alternative hypothesis (H_A) is also called the research hypothesis.
- There are 4 possible situations in testing a statistical hypothesis:

Condition of Null Hypothesis H_o

(Nature/reality)

Possible Action (Decision)		H_o is true	H_o is false
	Accepting H_o	Correct Decision	Type II error (β)
	Rejecting H_o	Type I error (α)	Correct Decision

- There are two types of Errors:

  - Type I error = Rejecting H_o when H_o is true

$P(\text{Type I error}) = P(\text{Rejecting Ho} \mid \text{Ho is true}) = \alpha$

  - Type II error = Accepting Ho when Ho is false $P(\text{Type II error}) = P(\text{Accepting Ho} \mid \text{Ho is false}) = \beta$

- The level of significance of the test is the probability of rejecting true H_o:

$\alpha = P(\text{Rejecting } H_o \mid H_o \text{ is true}) = P(\text{Type I error})$

- There are 2 types of alternative hypothesis:

One-sided alternative hypothesis:

-$H_O: \theta \geq \theta_o$ against $H_A: \theta < \theta_o$

-$H_O: \theta \leq \theta_o$ against $H_A: \theta > \theta_o$

o Two-sided alternative hypothesis:

- $H_O: \theta = \theta_o$ against $H_A: \theta \neq \theta_o$

- We will use the terms "accepting" and "not rejecting" interchangeably. Also, we will use the terms "acceptance" and "nonrejection" interchangeably.

- We will use the terms "accept" and "fail to reject" interchangeably

4.2 The Procedure of Testing H_o (against H_A):

The test procedure for rejecting H_o (accepting H_A) or accepting H_o (rejecting H_A) involves the following steps:

1. Determining a test statistic (T.S.)

We choose the appropriate test statistic based on the point estimator of the parameter.

The test statistic has the following form:

$$Test\ statistic = \frac{Estimate - hypothesized\ parameter}{Standard\ Error\ of\ the\ Estimate}$$

2. Determining the level of significance (α):

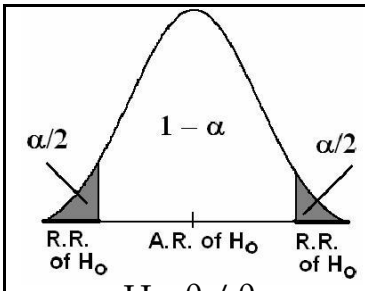
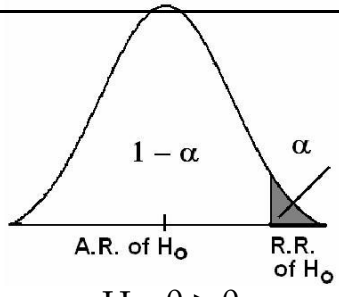
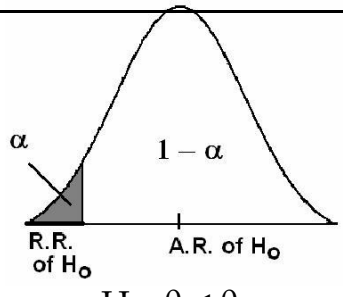
α = 0.01, 0.025, 0.05, 0.10

3. Determining the rejection region of H_o (R.R.) and the acceptance region of H_o (A.R.).

The R.R. of H_o depends on H_A and α:

- H_A determines the direction of the R.R. of H_o
- α determines the size of the R.R. of H_o

(α = the size of the R.R. of H_o = shaded area)

 $H_A: \theta \neq \theta_o$	 $H_A: \theta > \theta_o$	 $H_A: \theta < \theta_o$
Two-sided alternative	One-sided alternative	One-sided alternative

4. Decision:

We reject H_o (and accept H_A) if the value of the test statistic (T.S.) belongs to the R.R. of H_o , and vice versa.

Notes:

1. The rejection region of Ho (R.R.) is sometimes called "the critical region".

2. The values which separate the rejection region (R.R.) and the acceptance region (A.R.) are called "the critical values".

Hypothesis Testing: A Single Population Mean (μ)

Suppose that $X_1, X_2, \ldots, X_n$ is a random sample of size n from a distribution (or population) with mean μ and variance σ^2.

We need to test some hypotheses (make some statistical inference) about the mean (μ).

Recall:

1. $\bar{X}$ is a "good" point estimate for μ.
2. Mean of $\bar{X}$ is: $\mu_{\bar{x}} = \mu$.
3. Variance of $\bar{X}$ is: $\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$
4. Standard error (standard deviation) of $\bar{X}$ is : $\sigma_{\bar{X}} = \sqrt{\sigma_{\bar{X}}^2} = \frac{\sigma}{\sqrt{n}}$
5. For the case of normal distribution with any sample size or the case of non-normal distribution with large sample size, and for known variance σ^2 , we have:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

6. For the case of normal distribution with unknown variance σ^2 and with any sample size, we have:

$$t = \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t(n-1)$$

Where $s = \sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)}$ and $df = v = n-1$

The Procedure for hypotheses testing about the mean (μ):

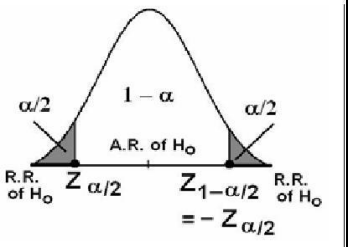
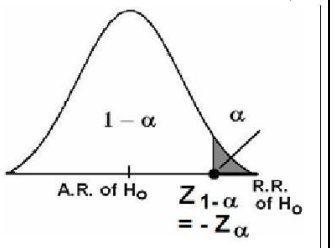
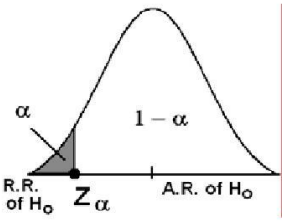
Let μ_o be a given known value.

(1) First case:

Assumptions:

- The variance σ^2 is known.
- Normal distribution with any sample size, or
- Non-normal distribution with large sample size.

- Test Procedures:

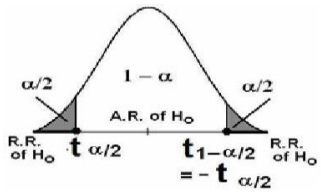
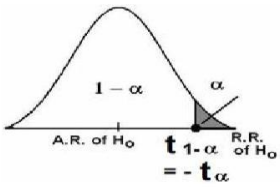
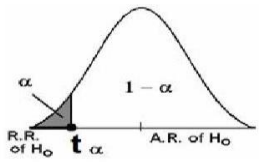
Hypothesis	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$
Test Statistic (T.S.)	Calculate the value of: $Z = \frac{\bar{\bar{X}} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Critical Value (s)	$Z_{\alpha/2}$ and $-Z_{\alpha/2}$	$Z_{1-\alpha} = -Z_\alpha$	Z_α
Decision:	We reject H_0 (and acceptH_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2}$ $= -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_\alpha$ One-Sided Test	$Z < Z_\alpha$ One-Sided Test

(2) Second case

Assumptions:

- The variance σ^2 is unknown.
- Normal distribution.
- Any sample size.

Test Procedures:

Hypothesis	$H_0: \mu = \mu_0$ $H_A: \mu \neq \mu_0$	$H_0: \mu \leq \mu_0$ $H_A: \mu > \mu_0$	$H_0: \mu \geq \mu_0$ $H_A: \mu < \mu_0$
Test Statistic (T.S.)	Calculate the value of: $t = \frac{\bar{\bar{X}} - \mu_0}{s/\sqrt{n}} \sim t(n-1) \quad df = v = n-1$		
R.R. & A.R. of H_0			
Critical Value (s)	$t_{\alpha/2}$ and $-t_{\alpha/2}$	$t_{1-\alpha} = -t_\alpha$	t_α
Decision:	We reject H_0 (and acceptH_A) at the significance level α if:		
	$t < t_{\alpha/2}$ or $t > t_{1-\alpha/2} = -t_{\alpha/2}$ Two-Sided Test	$t > t_{1-\alpha} = -t_\alpha$ One-Sided Test	$t < t_\alpha$ One-Sided Test

Example: (first case: variance σ^2 is known)

A random sample of 100 recorded deaths in the United States during the past year showed an average of 71.8 years. Assuming a population standard deviation of 8.9 year, does this seem to indicate that the mean life span today is greater than 70 years? Use a 0.05 level of significance.

Solution:

$n = 100\ \ (large),$ $\bar{X} = 71.8\,,$ $\sigma = 8.9\ \ (\sigma\ \ is\ known)$

$\mu = average\ (mean)\ life\ span$

$\mu_0 = 70$

$\alpha = 0.05$

Hypothesis:

$H_0: \mu \leq 70$ $(\mu_0 = 70)$

$H_A: \mu > 70$ $(research\ hypothesis)$

Test statistics (T.S.) :

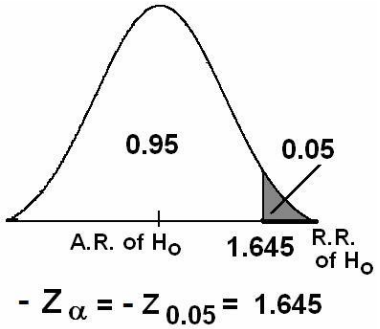
$$Z = \frac{\bar{\bar{X}} - \mu_0}{\sigma/\sqrt{n}} = \frac{71.8-70}{8.9/\sqrt{100}} = 2.02$$

Level of significance:

α=0.05

Rejection Region of H_o (R.R.): (critical region)

$-Z_{\alpha} = -Z_{0.05} = 1.645$ (critical value)



We should reject H_o if:

$Z > -Z_{\alpha} = -Z_{0.05} = 1.645$

Decision:

Since $Z=2.02 \in$R.R., i.e., $Z= 2.02 > -Z_{0.05}$, we reject $H_o: \mu \leq 70$ at α=0.05 and accept $H_A: \mu > 70$. Therefore, we conclude that the mean life span today is greater than 70 years.

Note: Using P- Value as a decision tool:

P-value is the smallest value of α for which we can reject the null hypothesis H_o.

Calculating P-value:

- Calculating P-value depends on the alternative hypothesis H_A
- Suppose that $Z_c = \frac{\bar{\bar{X}} - \mu_0}{\sigma/\sqrt{n}}$ is the computed value of the test statistic.
- The following table illustrates how to compute P-value, and how to use P-value for testing the null hypothesis:

Alternative Hypothesis:	$H_A: \mu \neq \mu_0$	$H_A: \mu > \mu_0$	$H_A: \mu < \mu_0$
P-Value =	2xP(Z > | Z_C |)	P(Z > Z_C)	P(Z < Z_C)
Significance Level =	α		
Decision:	Reject H_0 if P-value < α		

Example:

For the previous example, we have found that:

$Z = \frac{\bar{\bar{X}} - \mu_0}{\sigma/\sqrt{n}} = 2.02$

The alternative hypothesis was $H_A: \mu > 70$

P-Value = P(Z > Z_C)

$= P(Z > 2.02) = 1 - P(Z < 2.02) = 1 - 0.9783$

$= 0.0217$

The level of significance was $\alpha = 0.05$.

Since P-value < α, we reject H_o

Example: (second case: variance σ^2 is unknown)

The manager of a private clinic claims that the mean time of the patient-doctor visit in his clinic is 8 minutes. Test the hypothesis that $\mu=8$ minutes against the alternative that $\mu \neq 8$ minutes if a random sample of 50 patient-doctor visits yielded a mean time of 7.8 minutes with a standard deviation of 0.5 minutes. It is assumed that the distribution of the time of this type of visits is normal. Use a 0.01 level of significance.

Solution:

The distribution is normal

n= 50 (large) mean= 7.8

s= 0.5 (sample standard deviation) σ is unknown

μ= mean time of the visit

$\mu_o=8$

$\alpha=0.01$ $(\alpha/2 = 0.005)$

Hypotheses:

Ho: $\mu = 8$ $(\mu o=8)$

HA: $\mu \neq 8$ (research hypothesis)

Test statistic (T.S.):

$$t = \frac{\bar{\bar{X}} - \mu_0}{s/\sqrt{n}} = \frac{7.8-8}{0.5/\sqrt{50}} = -2.83$$

$$df = v = n - 1 = 50 - 1 = 49$$

Level of significance:

$\alpha = 0.01$

Rejection Region of H_o (R.R.): (critical region)

$t_{\alpha/2} = t_{0.005} (= -t_{0.995}) = -2.678$ (1^{st} critical value)

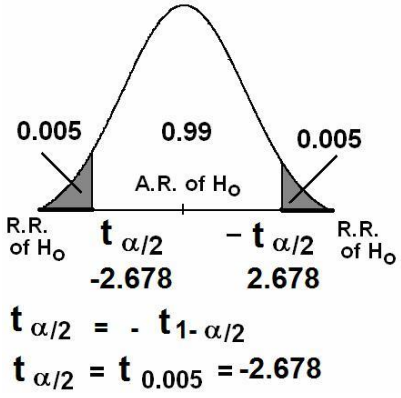
$-t_{\alpha/2} = -t_{0.005} = 2.678$ (2^{nd} critical value)

We should reject H_o if:

$t < t_{\alpha/2} = t_{0.005} = -2.678$

or

$t > -t_{\alpha/2} = -t_{0.005} = 2.678$



Decision:

Since $t = -2.83 \in$ R.R., i.e., $t = -2.83 < t_{0.005}$, we reject H_o: $\mu = 8$ at $\alpha = 0.01$ and accept H_A: $\mu \neq 8$. Therefore, we conclude that the claim is not correct.

Note:

For the case of non-normal population with unknown variance, and when the sample size is large (n≥30), we may use the following test statistic:

$$Z = \frac{\bar{\bar{X}} - \mu_0}{s/\sqrt{n}} \sim N(0,1)$$

That is, we replace the population standard deviation (σ) by the sample standard deviation (S), and we conduct the test as described for the first case.

Hypothesis Testing: The Difference Between Two Population Means: (Independent Populations)

Suppose that we have two (independent) populations:

- 1-st population with mean μ_1 and variance σ_1^2
- 2-nd population with mean μ_2 and variance σ_2^2
- We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about the difference between the means ($\mu_1 - \mu_2$).

- We independently select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:

- Let $\bar{X}_1$ and S_1^2 be the sample mean and the sample variance of the 1-st sample.

- Let $\bar{X}_2$ and S_2^2 be the sample mean and the sample variance of the 2-nd sample.
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about $\mu_1 - \mu_2$.

We wish to test some hypotheses comparing the population means.

Hypotheses:

We choose one of the following situations:

(i) $H_o: \mu_1 = \mu_2$ against $H_A: \mu_1 \neq \mu_2$

(ii) $H_o: \mu_1 \geq \mu_2$ against $H_A: \mu_1 < \mu_2$

(iii) $H_o: \mu_1 \leq \mu_2$ against $H_A: \mu_1 > \mu_2$

or equivalently,

(i) $H_o: \mu_1 - \mu_2 = 0$ against $H_A: \mu_1 - \mu_2 \neq 0$

(ii) $H_o: \mu_1 - \mu_2 \geq 0$ against $H_A: \mu_1 - \mu_2 < 0$

(iii) $H_o: \mu_1 - \mu_2 \leq 0$ against $H_A: \mu_1 - \mu_2 > 0$

Test Statistic:

(1) First Case:

For normal populations (or non-normal populations with large sample sizes), and if σ_1^2 and σ_2^2 are known, then the test statistic is:

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

(2) Second Case:

For normal populations, and if σ_1^2 and σ_2^2 are unknown but equal ($\sigma_1^2 = \sigma_2^2 = \sigma^2$), then the test statistic is:

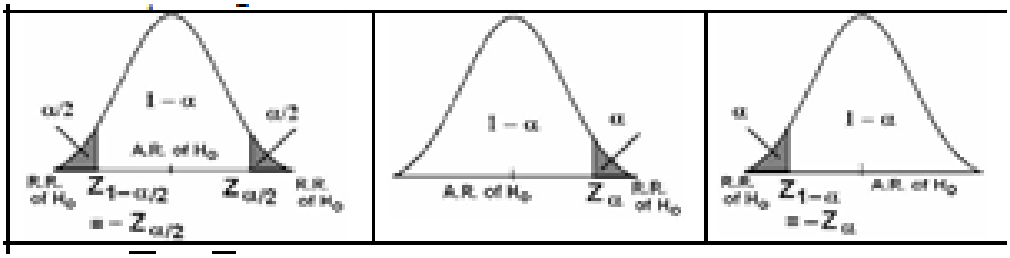
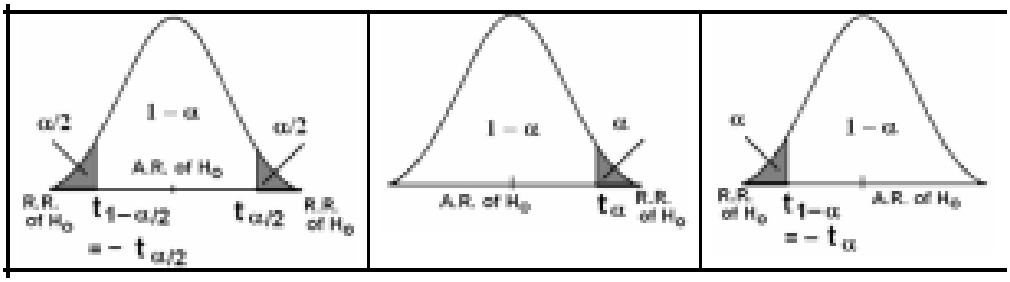
$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \sim t(n_1 + n_2 - 2)$$

where the pooled estimate of σ^2 is

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

and the degrees of freedom of S_p^2 is df= $v = n_1 + n_2 - 2$.

Summary of Testing Procedure:

Hypothesis	$H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$	$H_0: \mu_1 - \mu_2 \leq 0$ $H_A: \mu_1 - \mu_2 > 0$	$H_0: \mu_1 - \mu_2 \geq 0$ $H_A: \mu_1 - \mu_2 < 0$
Test Statistic For the First Case:	$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$ {if σ_1^2 and σ_2^2 are known}		
R.R. & A.R. of H_0 (For the First Case)			
Test Statistic For the Second Case:	$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} \sim t(n_1 + n_2 - 2)$ {if $\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown}		
R.R. & A.R. of H_0 (For the Second Case)			
Decision:	We reject H_0 (and acceptH_A) at the significance level α if:		
	T.S. ∈ R.R. Two-Sided Test	T.S. ∈ R.R. One-Sided Test	T.S. ∈ R.R. One-Sided Test

Example: (σ_1^2 and σ_2^2 are known)

Researchers wish to know if the data they have collected provide sufficient evidence to indicate the difference in mean serum uric acid levels between individuals with Down's syndrome and normal individuals. The data consist of serum uric acid on 12 individuals with Down's syndrome and 15 normal individuals. The sample means are $\overline{X_1} = 4.5\ mg/100ml$ and $\overline{X_2} = 3.4\ mg/100ml$. Assume the populations are normal with variances $\sigma_1^2 = 1$ and $\sigma_2^2 = 1.5$. Use significance level $\alpha = 0.5$.

Solution:

μ_1 = mean serum uric acid levels for the individuals with Down's syndrome.

μ_2 = mean serum uric acid levels for the normal individuals.

$$n_1 = 12 \quad \overline{X_1} = 4.5 \quad \sigma_1^2 = 1$$
$$n_2 = 15 \quad \overline{X_2} = 3.4 \quad \sigma_2^2 = 1.5$$

Hypotheses:

H_o: $\mu_1 = \mu_2$ against H_A: $\mu_1 \neq \mu_2$

or

H_o: $\mu_1 - \mu_2 = 0$ against H_A: $\mu_1 - \mu_2 \neq 0$

Calculation:

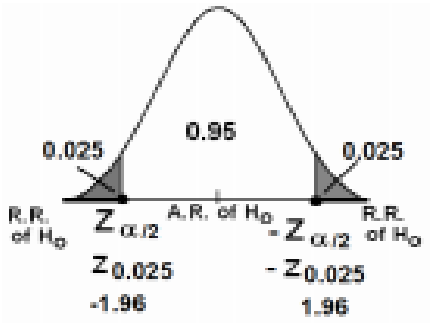
α=0.05

$Z_{0.75}$ = 1.96 (1st critical value)

$-Z_{0.75}$ = -1.96 (2nd critical value)

Test Statistic (T.S.):

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{4.5 - 3.4}{\sqrt{\frac{1}{12} + \frac{1.5}{15}}} = 2.569$$



Decision:

Since Z=2.569 ∈R.R. we reject H_o: $\mu_1=\mu_2$ and we accept (do not reject) H_A: $\mu_1 \neq \mu_2$ at α=0.05. Therefore, we conclude that the two population means are not equal.

Notes:

1. We can easily show that a 95% confidence interval for (μ_1-μ_2) is (0.26, 1.94) ,that is:

$$0.26<\mu_1-\mu_2<1.94$$

Since this interval does not include 0, we say that 0 is not a candidate for the difference between the population means (μ_1-μ_2), and we conclude that $\mu_1-\mu_2\neq 0$, i.e., $\mu_1\neq\mu_2$. Thus we arrive at the same conclusion by means of a confidence interval.

2. $P-Value = 2 \times P(Z > |\ Z_C\ |)$

$=2P(Z > 2.57) = 2[1- P(Z < 2.57)] = 2(1- 0.9949)$

=0.0102

The level of significance was $\alpha = 0.05$.

Since P-value < α, we reject H_o.

Example: ($\sigma_1^2 = \sigma_2^2 = \sigma^2$ is unknown**)**

An experiment was performed to compare the abrasive wear of two different materials used in making artificial teeth. 12 pieces of material 1 were tested by exposing each piece to a machine measuring wear. 10 pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average wear of 85 units with a sample standard deviation of 4, while the samples of materials 2 gave an average wear of 81 and a sample standard deviation of 5. Can we conclude at the 0.05 level of significance that the mean abrasive wear of material 1 is greater than that of material 2? Assume normal populations with equal variances.

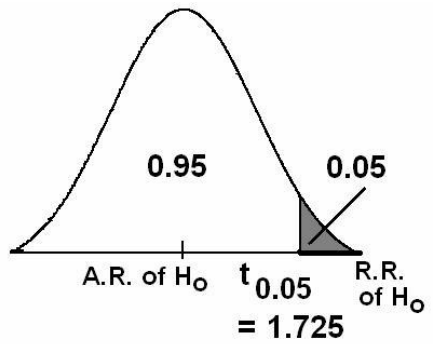
Solution:

Material 1	**Material 2**
$n_1 = 12$	$n_2 = 10$
$\overline{X_1} = 85$	$\overline{X_2} = 81$
$S_1 = 4$	$S_2 = 5$

Hypotheses:

H_o: $\mu_1 \leq \mu_2$

H_A: $\mu_1 > \mu_2$



Or equivalently,

H_o: $\mu_1 - \mu_2 \leq 0$

H_A: $\mu_1 - \mu_2 > 2$

Calculation:

α=0.05

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)(4)^2 + (10 - 1)(5)^2}{12 + 10 - 2} = 20.05$$

$$df = v = n_1 + n_2 - 2 = 12 + 10 - 2 = 20$$

$t_{0.05}$ = 1.725 (critical value)

Test Statistic (T.S.):

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} = \frac{85 - 81}{\sqrt{\frac{20.05}{12} + \frac{20.05}{10}}} = 1.04$$

Decision:

Since T=1.04 ∈A.R. (T=1.04< $t_{0.05}$ = 1.725), we accept (do not reject) H_o and we reject H_A:μ_1 μ_2> 0 (H_A:$\mu_1 > \mu_2$) at α=0.05. Therefore, we conclude that the mean abrasive wear of material 1 is <u>not</u> greater than that of material 2.

7.4 Paired Comparisons:

- In this section, we are interested in comparing the means of two related (non-independent/dependent) normal populations.

- In other words, we wish to make statistical inference for the difference between the means of two related normal populations.
- Paired t-Test concerns about testing the equality of the means of two related normal populations.

Examples of related populations are:

1. Height of the father and height of his son.
2. Mark of the student in MATH and his mark in STAT.
3. Pulse rate of the patient before and after the medical treatment.
4. Hemoglobin level of the patient before and after the medical treatment.

Example: (effectiveness of a diet program)

Suppose that we are interested in studying the effectiveness of a certain diet program. Let the random variables X and Y are as follows:

X = the weight of the individual before the diet program
Y= the weight of the same individual after the diet program

We assume that the distributions of these random variables are normal with means μ_1 and μ_2 , respectively.

These two variables are related (dependent/non-independent) because they are measured on the same individual.

Populations:

1-st population (X): weights before a diet program

mean = μ_1

2-nd population (Y): weights after the diet program

mean = μ_2

Question:

Does the diet program have an effect on the weight?

Answer is:

No if $\mu_1 = \mu_2$ $(\mu_1-\mu_2 = 0)$

Yes if $\mu_1 \neq \mu_2$ $(\mu_1-\mu_2 \neq 0)$

Therefore, we need to test the following hypotheses:

Hypotheses:

H_o: $\mu_1 = \mu_2$ (H_o: the diet program has no effect on weight)

H_A: $\mu_1 \neq \mu_2$ (H_A: the diet program has an effect on weight)

Equivalently we may test:

H_o: $\mu_1 - \mu_2 = 0$

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

Testing procedures:

- We select a random sample of n individuals. At the beginning of the study, we record the individuals' weights before the diet program (X). At the end of the diet program, we record the individuals' weights

after the program (Y). We end up with the following information and calculations:

Individual **i**	**Weight before** $\boldsymbol{X_i}$	**Weight after** $\boldsymbol{Y_i}$	**Difference** $\boldsymbol{D_i = X_i - Y_i}$
1	X_1	Y_1	$D_1 = X_1 - Y_1$
2	X_2	Y_2	$D_2 = X_2 - Y_2$
. .	. .	. .	. .
n	X_n	Y_n	$D_n = X_n - Y_n$

- Hypotheses:

H_o: the diet program has no effect on weight

H_A: the diet program has an effect on weight

Equivalently,

H_o: $\mu_1 = \mu_2$

H_A: $\mu_1 \neq \mu_2$

Equivalently,

H_o: $\mu_1 - \mu_2 = 0$

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

Equivalently,

H_o: $\mu_D = 0$

H_A: $\mu_D \neq 0$

where: $\mu_D = \mu_1 - \mu_2$

- We calculate the following quantities:

- The differences (D-observations):

$D_i = X_i - Y_i$ (i=1, 2, …, n)

- Sample mean of the D-observations (differences):

$$\overline{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{D_1 + D_2 + \cdots + D_n}{n}$$

- Sample Variance of the D-observations (differences):

$$SD^2 = \frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n-1} = \frac{(D_1 - \overline{D})^2 + (D_2 - \overline{D})^2 + \cdots + (D_n - \overline{D})^2}{n-1}$$

- Sample standard deviation of the D-observations:

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

- Test Statistic:

We calculate the value of the following test statistic:

$$t = \frac{\overline{D}}{S_D/\sqrt{n}} \sim t(n-1)$$

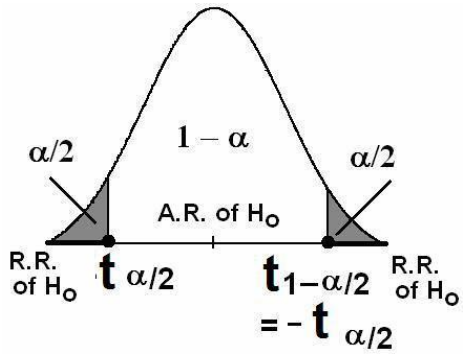
This statistic has a t-distribution with df = v = n -1.

- Rejection Region of H_o:

Critical values are: $t_{\alpha/2}$ and $t_{1-\alpha/2} = -t_{\alpha/2}$.

The rejection region (critical region) at the significance level α is:

$$t < t_{\alpha/2} \quad or \quad t > t_{1-\alpha/2} = -t_{\alpha/2}$$



- Decision:

We reject H_o and accept H_A at the significance level α if T∈R.R. ,i.e., if:

$$t < t_{\alpha/2} \qquad or \qquad t > t_{1-\alpha/2} = -t_{\alpha/2}$$

Numerical Example:

In the previous example, suppose that the sample size was 10 and the data were as follows:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Does these data provide sufficient evidence to allow us to conclude that the diet program is effective? Use α=0.05 and assume that the populations are normal.

Solution:

μ1 = the mean of weights before the diet program

μ2 = the mean of weights after the diet program

Hypotheses:

Ho: $\mu_1 = \mu_2$ (Ho: the diet program is not effective)

H_A: $\mu_1 \neq \mu_2$ (HA: the diet program is effective)

Equivalently,

Ho: $\mu_D = 0$

H_A: $\mu_D \neq 0$ (where: $\mu_D = \mu_1 - \mu_2$)

Calculations:

i	X_i	Y_i	$D_i = X_i - Y_i$
1	86.6	79.7	6.9
2	80.2	85.9	-5.7
3	91.5	81.7	9.8
4	80.6	82.5	-1.9
5	82.3	77.9	4.4
6	81.9	85.8	-3.9
7	88.4	81.3	7.1
8	85.3	74.7	10.6
9	83.1	68.3	14.8
10	82.1	69.7	12.4
sum	$\sum X = 842$	$\sum Y = 787.5$	$\sum D = 54.5$

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} = \frac{54.5}{10} = 5.45$$

$$SD^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1} = \frac{(6.9 - 5.45)^2 + \cdots + (12.4 - 5.45)^2}{10 - 1} = 50.3283$$

$$SD = \sqrt{SD^2} = \sqrt{50.3283} = 7.09$$

Test Statistic:

$$t = \frac{\bar{D}}{S_D/\sqrt{n}} = \frac{5.45}{7.09/\sqrt{10}} = 2.431$$

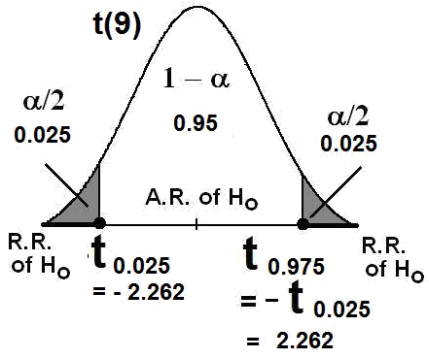
Degrees of freedom: df= v= n -1 = 10-1= 9

Significance level: α=0.05

Rejection Region of H_o:

Critical values: $t_{0.025} = -2.262$ and $t_{0.975} = -t_{0.025} = 2.262$

Critical Region: $t < -2.262$ or $t > 2.262$



Decision:

Since t= 2.43 ∈R.R., i.e., $t=2.43 > t_{0.975} = -t_{0.025} = 2.262$, we reject:

Ho: $\mu_1 = \mu_2$ (the diet program is not effective)

and we accept:

H_1: $\mu_1 \neq \mu_2$ (the diet program is effective)

Consequently, we conclude that the diet program is effective at α=0.05.

Note:

- The sample mean of the weights before the program is $\bar{X} = 84.2$.
- The sample mean of the weights after the program is $\bar{Y} = 78.75$
- Since the diet program is effective and since $\bar{X} = 84.2 > \bar{Y} = 78.75$, we can conclude that the program is effective in reducing the weight.

Confidence Interval for the Difference between the Means of Two Related Normal Populations ($\mu_D = \mu_1 - \mu_2$):

In this section, we consider constructing a confidence interval for the difference between the means of two related (non-independent) normal populations. As before, let us define the difference between the two means as follows:

$$\mu_D = \mu_1 - \mu_2$$

where μ_1 is the mean of the first population and μ_2 is the mean of the second population. We assume that the two normal populations are not independent.

Result:

A (1-α) 100% confidence interval for $\mu_D = \mu_1 - \mu_2$ is:

$$\bar{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

$$\bar{D} - t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}} < \mu_D < \bar{D} + t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

Where

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} \quad , \quad S_D = \sqrt{S_D^2} \quad , \quad S_D^2 = \frac{\sum_{i=1}^{n} (D_i - \bar{D})^2}{n-1} \quad , \quad df = v = n - 1.$$

Example:

Consider the data given in the previous numerical example:

Individual (i)	1	2	3	4	5	6	7	8	9	10
Weight before (X_i)	86.6	80.2	91.5	80.6	82.3	81.9	88.4	85.3	83.1	82.1
Weight after (Y_i)	79.7	85.9	81.7	82.5	77.9	85.8	81.3	74.7	68.3	69.7

Find a 95% confidence interval for the difference between the mean of weights before the diet program (μ_1) and the mean of weights after the diet program (μ_2).

Solution:

We need to find a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$:

$$\bar{D} \pm t_{1-\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

We have found:

$$\bar{D} = 5.45 \quad , \quad S_D^2 = 50.3283 \quad , \quad S_D = \sqrt{S_D^2} = 7.09$$

The value of the reliability coefficient $t_{1-\frac{\alpha}{2}}$ $(df = v = n - 1 = 9)$ is

$$t_{1-\frac{\alpha}{2}} = t_{0.975} = 2.262$$

Therefore, a 95% confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$5.45 \pm (2.262)\frac{7.09}{\sqrt{10}}$$

$$5.45 \pm 5.0715$$

$$0.38 < \mu_D < 10.52$$

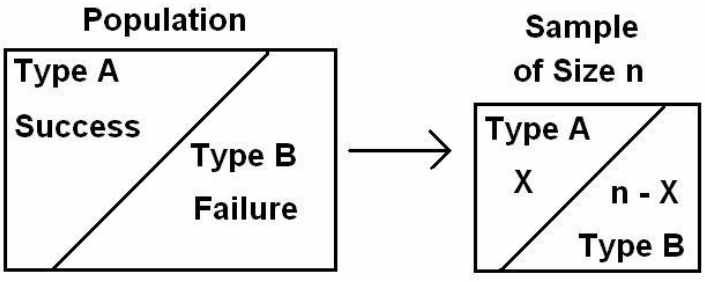
$$0.38 < u_1 - u_2 < 10.52$$

We are 95% confident that $\mu_D = \mu_1 - \mu_2 \in (0.38, 10.52)$.

Note: Since this interval does not include 0, we say that 0 is not a candidate for the difference between the population means $(u_1 - u_2)$, and we conclude that $u_1 - u_2 \neq 0,\ i.e.\ u_1 \neq u_2$. Thus we arrive at the same conclusion by means of a confidence interval.

Hypothesis Testing: A Single Population Proportion (p):

In this section, we are interested in testing some hypotheses about the population proportion (p).



Recall:

• p = Population proportion of elements of Type A in the population.

$$p = \frac{no.of\ elements\ of\ type\ A\ in\ the\ population}{Total\ no.of\ elements\ in\ the\ population}$$

$$p = \frac{A}{N} \qquad (N = population\ size)$$

• n = sample size

• X = no. of elements of type A in the sample of size n.

• $\hat{p} = \ Sample\ proportion\ elements\ of\ Type\ A\ in\ the\ sample$

$$\hat{p} = \frac{no.of\ elements\ of\ type\ A\ in\ the\ sample}{no.of\ elements\ in\ the\ sample}$$

$$\hat{p} = \frac{X}{n} \qquad (n = sample\ size = no.of\ elements\ in\ the\ sample$$

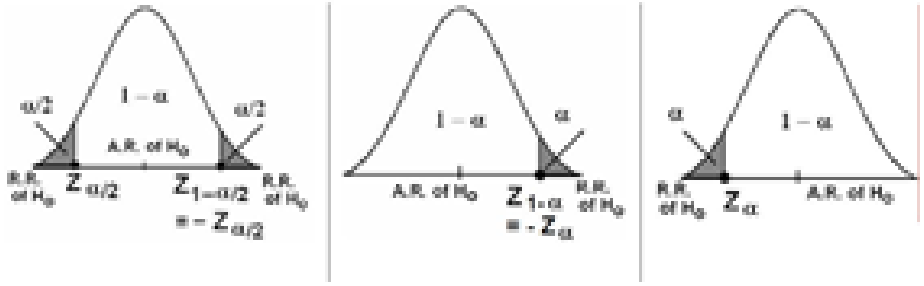
• $\hat{p}$ is a "good" point estimate for p.

• For large n, ($n \geq 30$, $np > 5$), we have

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

• Let p_0 be a given known value.

• Test Procedure:

Hypothesis	$H_0: p = p_0$ $H_A: p \neq p_0$	$H_0: p \leq p_0$ $H_A: p > p_0$	$H_0: p \geq p_0$ $H_A: p < p_0$
Test Statistic (T.S.)	Calculate the value of: $Z = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Decision:	We reject H_0 (and acceptH_A) at the significance level α if:		
	$Z < Z_{\alpha/2}$ or $Z > Z_{1-\alpha/2} = -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_{1-\alpha} = -Z_{\alpha}$ One-Sided Test	$Z < Z_{\alpha}$ One-Sided Test

Example:

A researcher was interested in the proportion of females in the population of all patients visiting a certain clinic. The researcher claims that 70% of all patients in this population are females. Would you agree with this claim if a random survey shows that 24 out of 45 patients are females? Use a 0.10 level of significance.

Solution:

p = Proportion of female in the population.

n=45 (large)

X= no. of female in the sample = 24

$\hat{p} = \ proportion\ of\ females\ in\ the\ sample$

$\hat{p} = \frac{X}{n} = \frac{24}{45} = 0.5333$

$p_0 = \frac{70}{100} = 0.7 \qquad\qquad \alpha = 0.10$

Hypotheses:

Ho: $p = 0.7$ ($p_o = 0.7$)

H_A: $p \neq 0.7$

Level of significance:

α=0.10

Test Statistic (T.S.): $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

$$Z = \frac{0.5333 - 0.70}{\sqrt{\frac{(0.7)(0.3)}{45}}} = -2.44$$

Rejection Region of H_o (R.R.):

Critical values:

$Z_{\alpha/2} = Z_{0.05} = -1.645$

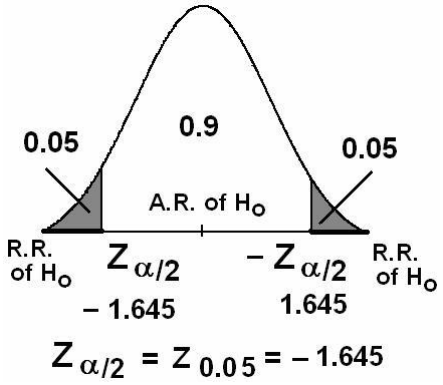
$-Z_{\alpha/2} = -Z_{0.05} = 1.645$

We reject H_o if:

$Z < Z_{\alpha/2}=Z_{0.05}= -1.645$

or

$Z > -Z_{\alpha/2}= -Z_{0.05}= 1.645$



Decision:

Since $Z= -2.44 \in$Rejection Region of H_o (R.R), we reject H_o: p=0.7 and accept H_A: $p \neq 0.7$ at α=0.1. Therefore, we do not agree with the claim stating that 70% of the patients in this population are females.

Example:

In a study on the fear of dental care in a certain city, a survey showed that 60 out of 200 adults said that they would hesitate to take a dental appointment due to fear. Test whether the proportion of adults in this city who hesitate to take dental

appointment is less than 0.25. Use a level of significance of 0.025.

Solution:

p = Proportion of adults in the city who hesitate to take a dental appointment.

n= 200 (large)

X= no. of adults who hesitate in the sample = 60

$\hat{p} = \ proportion\ of\ adults\ who\ hesitate\ in\ \ the\ sample$

$$\hat{p} = \frac{X}{n} = \frac{60}{200} = 0.3$$

p_o=0.25

α=0.025

Hypotheses:

Ho: $p \geq 0.25$ (p_o= 0.25)

H_A: $p < 0.25$ (research hypothesis)

Level of significance:

α=0.025

Test Statistic (T.S.):

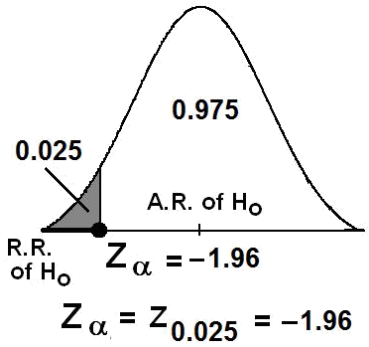
$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{0.3 - 0.25}{\sqrt{\frac{(0.25)(0.75)}{200}}} = 1.633$$

Rejection Region of Ho (R.R.):

Critical value: $Z\alpha = Z_{0.025} = -1.96$

Critical Region:

We reject Ho if: $Z < Z\alpha = Z_{0.025} = -1.96$

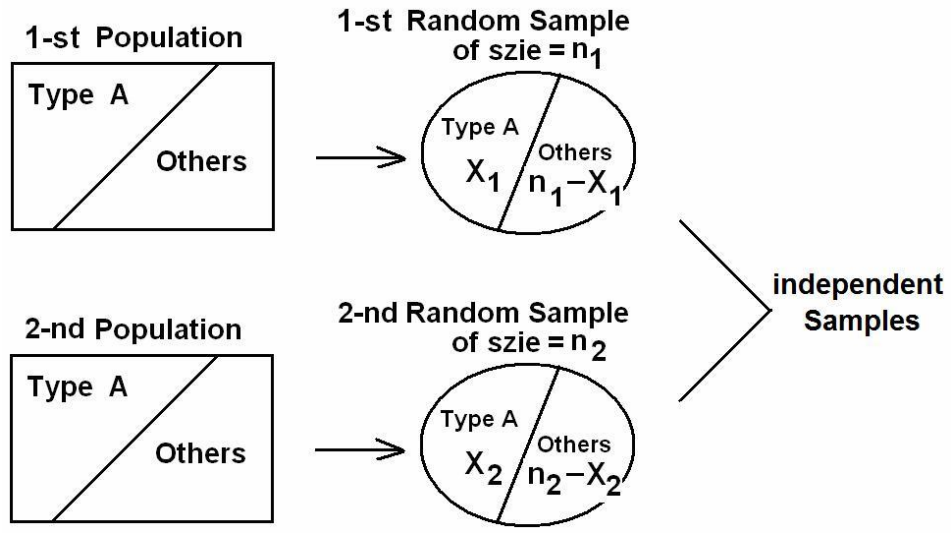


Decision:

Since Z=1.633 ∈ Acceptance Region of H_o (A.R.), we accept (do not reject) H_o: $p \geq 0.25$ and we reject H_A: $p < 0.25$ at α=0.025. Therefore, we do not agree with claim stating that the proportion of adults in this city who hesitate to take dental appointment is less than 0.25.

Hypothesis Testing: The Difference Between Two Population Proportions $(P_1 - P_2)$:

In this section, we are interested in testing some hypotheses about the difference between two population proportions $(P_1 - P_2)$.



Suppose that we have two populations:

- p_1= population proportion of the 1-st population.
- p_2= population proportion of the 2-nd population.
- We are interested in comparing p_1 and p_2, or

equivalently, making inferences about $p_1 - p_2$.

- We <u>independently</u> select a random sample of size n_1 from

the 1-st population and another random sample of size n_2 from the 2nd population:

- Let X_1 = no. of elements of type A in the 1-st sample.
- Let X_2 = no. of elements of type A in the 2-nd sample.

$$\widehat{p_1} = \frac{X_1}{n_1} = the\ sample\ proportion\ of\ the\ 1st\ sample$$

$$\widehat{p_2} = \frac{X_2}{n_2} = \ the\ sample\ proportion\ of\ the\ 2nd\ sample$$

- The sampling distribution of $\widehat{p_1} - \widehat{p_2}$ is used to make inferences about $p_1 - p_2$.
- For large n_1 and n_2, we have

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1) \quad (Approximately)$$

- $q = 1 - p$

Hypotheses:

We choose one of the following situations:

(i) H_o: $p_1 = p_2$ against H_A: $p_1 \neq p_2$

(ii) H_o: $p_1 \geq p_2$ against H_A: $p_1 < p_2$

(iii) H_o: $p_1 \leq p_2$ against H_A: $p_1 > p_2$

or equivalently,

(i) H_o: $p_1 - p_2 = 0$ against H_A: $p_1 - p_2 \neq 0$

(ii) H_o: $p_1 - p_2 \geq 0$ against H_A: $p_1 - p_2 < 0$

(iii) H_o: $p_1 - p_2 \leq 0$ against H_A: $p_1 - p_2 > 0$

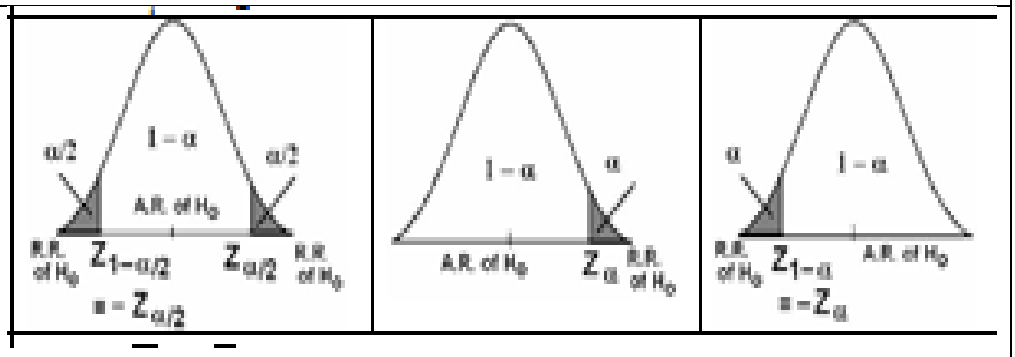
Note, under the assumption of the equality of the two population proportions (H_o: $p_1 = p_2 = p$), the pooled estimate of the common proportion p is:

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} \qquad (\bar{q} = 1 - \bar{p})$$

The test statistic (T.S.) is

$$Z = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \sim N(0,1)$$

Testing Procedure:

Hypothesis	$H_0: p_1 - p_2 = 0$ $H_A: p_1 - p_2 \neq 0$	$H_0: p_1 - p_2 \leq 0$ $H_A: p_1 - p_2 > 0$	$H_0: p_1 - p_2 \geq 0$ $H_A: p_1 - p_2 < 0$
Test Statistic (T.S.)	$Z = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} \sim N(0,1)$		
R.R. & A.R. of H_0			
Decision:	Reject H_0 (and accept H_A) at the significance level α if Z∈R.R:		
Critical Values	$Z > Z_{\alpha/2}$ Or $Z < -Z_{\alpha/2}$ Two-Sided Test	$Z > Z_\alpha$ One-Sided Test	$Z < -Z_\alpha$ One-Sided Test

Example:

In a study about the obesity (overweight), a researcher was interested in comparing the proportion of obesity between males and females. The researcher has obtained a random sample of 150 males and another independent random sample of 200 females. The following results were obtained from this study.

	n	Number of obese people
Males	150	21
Females	200	48

Can we conclude from these data that there is a difference between the proportion of obese males and proportion of obese females? Use $\alpha = 0.05$.

Solution:

p_1 = population proportion of obese males

p_2 = population proportion of obese females

$\widehat{p_1} = sample\ proportion\ of\ obese\ males$

$\widehat{p_2} = sample\ proportion\ of\ obese\ females$

Males

$n_1 = 150$

$X_1 = 21$

$\widehat{p_1} = \frac{X_1}{n_1} = \frac{21}{150} = 0.14$

Females

$n_2 = 200$

$X_2 = 48$

$\widehat{p_2} = \frac{X_2}{n_2} = \frac{48}{200} = 0.24$

The pooled estimate of the common proportion *p is:*

$$\bar{p} = \frac{X_1 + X_2}{n_1 + n_2} = \frac{21 + 48}{150 + 200} = 0.197$$

Hypotheses:

H_o: $p_1 = p_2$

H_A: $p_1 \neq p_2$

or

$H_0: p_1 - p_2 = 0$

$H_A: p_1 - p_2 \neq 0$

Level of significance: α=0.05

Test Statistic (T.S.):

$$Z = \frac{(\widehat{p_1} - \widehat{p_2})}{\sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}}} = \frac{(0.14 - 0.24)}{\sqrt{\frac{0.197(0.803)}{150} + \frac{0.197(0.803)}{200}}}$$

$$= -2.328$$

Rejection Region (R.R.) of H_o:

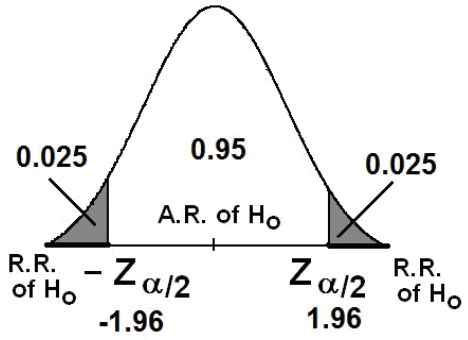
Critical values:

$Z_{\alpha/2} = Z_{0.025} = -1.96$

$Z_{1-\alpha/2} = Z_{0975} = 1.96$

Critical region:

Reject H_o if: $Z < -1.96$ or $Z > 1.96$



Decision:

Since Z= -2.328 ∈R.R., we reject H_o: $p_1 = p_2$ and accept H_A: $p_1 \neq p_2$ at α=0.05. Therefore, we conclude that there is a difference between the proportion of obese males and the proportion of obese females. Additionally, since, $\widehat{p_1} = 0.14 < \widehat{p_2} = 0.24$, we may conclude that the proportion of obesity for females is larger than that for males.