Notes on Applied Statistics

Contents

Introduction to Statistics

- Basic Concepts and Definitions
- Data and Variables

The word Statistics is derived from Latin word status meaning “**state**” . Early uses of statistics involved compilation of data and graphs describing various aspects of the state or country. The word statistics has two basic meanings. We sometimes use this word when referring to actual numbers derived from data and the other refers to as a method of analysis

Overarching Definition

- **Statistics**:=deals with{ collection(data), presentation(data), analysis(data), interpretation(data)}
  - **Collection** - gathering of information
  - **Organization** or **presentation** - summarizing data
  - **Analysis** - describing the data by statistical methods or procedures
  - **Interpretation** - making conclusions on the analysed data

- **Inferential Statistics** - make inferences about the sample
  - Generalizes from samples to population
  - Performs estimations and hypothesis tests
  - Determines the relationships among variables
  - Making predictions
  - Uses probability

- **Descriptive Statistics** - describe a situation
  - Collection of data
  - Organization of data
  - Summarization of data
  - Presentation of data

Why study statistics?

- You have to be a data literate professional
- Statistics is **basic to research**: designing experiments, collecting-, organizing-, analysing-, and summarizing data to make reliable predictions or forecasts.
- You can also use the knowledge gained from studying statistics to become better consumers and citizens

Basic Concepts and Definitions

- **Variable** - characteristic or attribute that can assume different values
- **Data** - the value of the variable
- **Random data** - values of the variables are determined by chance
- **Data set** - collection of data values
- **Data value (datum)** - each value in the data set
- **Population** - all subjects that are being studied
  - **Parameter -**numerical summary or any measurement coming from a **population**
- **Sample** - group of subject selected from a **population**
  - **Statistic** - measure of the **sample**

Data and Variables

- **Quantitative** - categorical
  - Data
  - Variable
- **Qualitative** - numerical
  - Data
  - Variable
- **Dependent variable** - a variable that is **influenced by** another variable
- **Independent variable** - one that affects or **influences** another variable

Levels of Measurement

- **Nominal** - names
- **Ordinal** - categories in scale e.g. ranking
- **Interval** - comparison between the numerical differences are meaningful but not the ratio of the measurements
- **Ratio** - one can compare both the differences between measurements of the variable and the ratio of the measurements meaningfully

Data Collection

- Methods of Data Collection
- Sampling Technique

Data - facts and statistics collected together for reference or analysis.

Types of Data:

- **Primary** data - collected from an original source
- **Secondary** data - collected from published or unpublished sources.

Methods of Data Collection

1. **Direct or Interview Method** - the researcher has a direct contact with the interviewee. The researcher obtains the information needed by asking questions and inquiries from the interviewee. This method gives precise and consistent information because clarifications can be made.
2. **Indirect or Questionnaire Method** - make use of a written questionnaire. The researcher cannot expect that all distributed questionnaires will be retrieved because some respondents simply ignore the questionnaires. In addition, clarification cannot be made if the respondent does not understand the question.
3. **Registration Method** - collecting data is governed by laws.
4. **Experimental Method** - usually used to find out the cause and effect relationships

Sampling Techniques

- Determining the sample size - typically we use samples of the population and not the population itself because of the tremendous cost of using the entire population. To determine the sample size from a given population size, the **Slovin's Formula** is used: $n = \frac{N}{1+Ne^2}$; where n= sample size, N = population size, e = margin of error.

- We want unbiased samples: each interesting features should have an unbiased

- **Simple Random Sampling**

- **Subjects are selected by random numbers**
  - **Can be done by labelling each possible subject and randomly picking the subject**

- **Systematic Sampling**
  - **Number each subject of the population then selecting every k-th subject**
  - **Lecturer: we can select samples from a moving population**

- **Stratified Sampling**
  - Done by dividing the population into groups called strata according to some characteristic that is important to the study, then sampling from each group

- **Cluster Sampling**
  - The population is divided into groups called clusters by some means. Then the researcher randomly selects some of these clusters and uses all members of the selected cluster as the subject of the samples.
  - There's no guarantee that every clusters is represented

Observational and Experimental Studies

- There are different ways to classify statistical studies:

- **Observational** - researcher merely observes what is happening or what has happened in the past and tries to draw conclusions based on these observations.
- **Experimental** - the researcher manipulates one of the variables and tries to determine how the manipulation influences other variables.
-

Data Presentation

- The Frequency Distribution Table
- Graphical Presentation of Data

The Frequency Distribution Table

- **Raw data** - information obtained by observing values of a variable
- **Qualitative Data** - obtained by conserving values of a qualitative variable
- **Quantitative Data** - obtained by conserving values of a quantitative variable
  - **Discrete Data**
  - **Continuous Data**

Methods of Data Presentation

- Data can be classified as grouped or ungrouped

1. **Ungrouped data** - not organized or if arranged could only be from highest to lowest or lowest to highest
2. **Grouped data** - are data that are organized and arranged into different classes or categories

Textual Method (ungrouped)

- Ungrouped data can be presented in paragraph form.
- Involves enumerating the important characteristics, giving emphasis on significant figures and identifying important features of the data

Tabular Method (grouped)

- **Table heading** - consists of the table number and the title
- **Column header** - describes data in each column
- **Row classifier** - shows the classes or categories
- **Body** - main part of the table
- **Source note** - placed below when the data written are not original

Frequency Distribution - most commonly used by tabular method

Frequency Distribution

- Organization of raw data in table form, using classes and frequencies
- For qualitative data lists all categories and the number of elements that belong to each of the categories.
- For quantitative data, the data is grouped according to some numerical of quantitative characteristics
  - **Class limits** - endpoints of a class interval
  - **Upper class limit** - represents the largest data value that can be included in the class.
  - **Lower class limit** - represents the lowest data value that can be included in the class.
  - **Class boundaries** - used to separate the classes so that there are no gaps in the frequency distribution. The gaps are due to the limits; for example, there is a gap between 30 and 31
  - **Lower boundary** - Lower limit - 0.5
  - **Upper boundary** - Upper limit +0.5
  - **Class width** - the difference between the boundaries for any class i.e. i=upper boundary - lower boundary or i=(upper limit-lower limit)+1
  - **Class mark** - the midpoint of the class

Frequency Distribution Table (FDT) - a statistical table showing the frequency or number of observations contained in each of the defined classes or categories.

Relative frequency - a category is obtained by dividing the frequency (f) for a category by the sum of all the frequencies (n). They are commonly expressed as percentages.

To **construct a frequency distribution**, follow these rules:

1) There should be between 5 and 20 classes

    a) Although there are no hard and fast rule for the number of classes contained in a frequency distribution, it is of utmost importance to have enough classes to present a clear description of the collected data

2) It is preferable but not absolutely necessary that the class width be an odd number

    a) Ensures the midpoint of each class has the same place value as the data

3) The classes must be mutually exclusive - since they have no overlapping class limits

4) The classes must be continuous

a) There must be no gaps in a frequency distribution

b) The only exception occurs when the class with a zero frequency is the first of last class since a class with a zero frequency on either end can be omitted

5) The classes must be exhaustive.

a) There should be enough classes to accommodate all the data.

6) The classes must be equal in width

a) Avoids the distorted view of the data

Constructing an FDT

1) Determine the classes

a) Find the highest and lowest values

b) Find the range $R = highest\ value - lowest\ value$

c) Select the number of classes desired (k)

d) Find the width by dividing the range by the number of classes and rounding up. $i = \frac{R}{k}$

e) Select a starting point (usually the lowest value or any convenient number less than the lowest value); add the width to get the lower limits

f) Find the upper class limits

g) Find the boundaries

2) Tally the data

3) Find the frequencies from the tallies

Cumulative Frequency

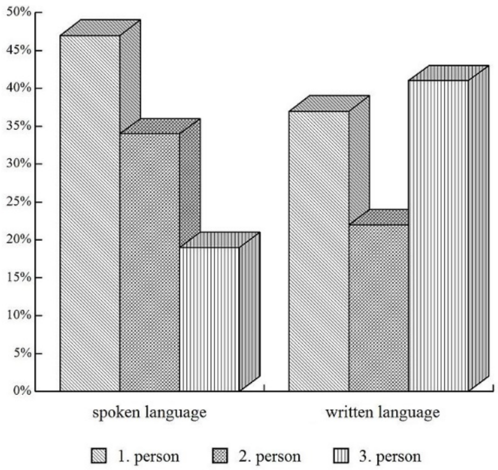
- Less than cumulative frequency (<cf) - total number of observations less than the upper boundary of a class interval
- greater than cumulative frequency (>cf) – total number of observations greater than the lower boundary of a class interval

Graphical Presentation of Data

- After you have organized the data into a frequency distribution, you can present them in a graphical form.

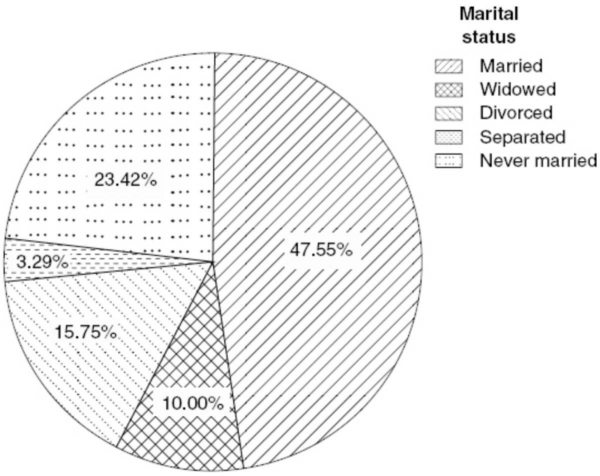
The Bar Graph

- Qualitative data

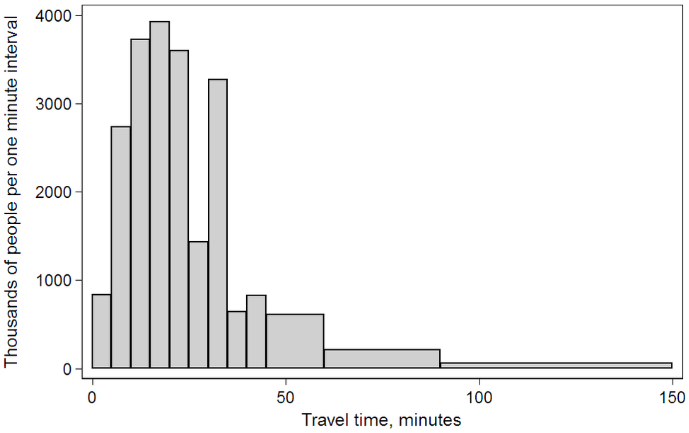


Pie Chart

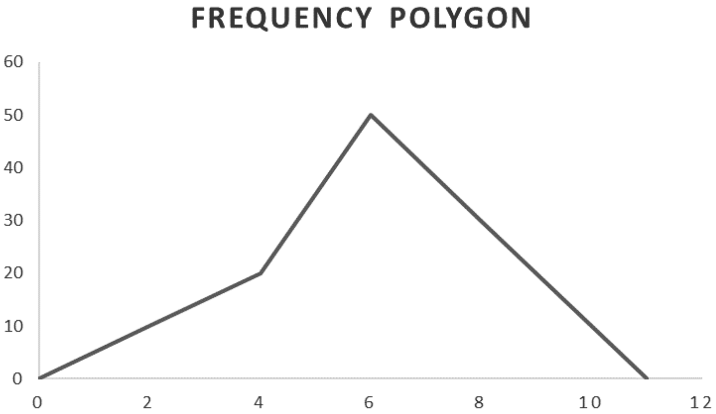
- Qualitative data



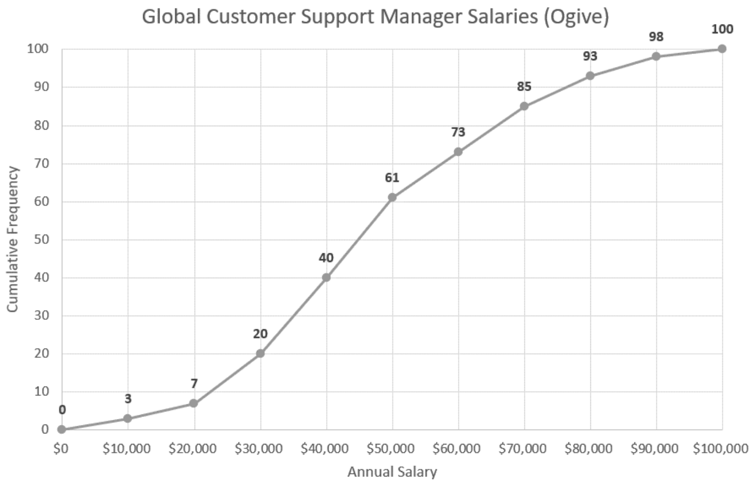
Histogram



Frequency Polygon



Ogive


Global Customer Support Manager Salaries (Ogive)
Cumulative Frequency
Annual Salary
0
10
20
30
40
50
60
70
80
90
100
$0
$10,000
$20,000
$30,000
$40,000
$50,000
$60,000
$70,000
$80,000
$90,000
$100,000
0
3
7
20
40
61
73
85
93
98
100

Measures of Central Tendency

- Measures of Central Tendency (MCT) of Ungrouped Data
- Measures of Central Tendency (MCT) for Grouped Data

Measures of Central Tendency Ungrouped

- MCT gives a single value that acts as a representative or average of the values of all the outcomes of your data set; describes the center of the distribution and represents the entire distribution to identify the single value that is the best representative for the entire set of data

Mean

- The most commonly used to measure central tendency; requires scores (or data types) that are numerical values measured on **an interval or ratio scale**.
  - Obtained by $\overline{x} = \frac{Total\ SUM}{number\ of\ dataset} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\overline{x} sample\ mean$, $\mu population\ mean$
  - For Discrete quantitative data, tabulated in a frequency table, then if possible observations are $\{x_1, x_2, \ldots, x_k\}$ and these occur with frequencies $\{f_1, f_2, \ldots, f_k\}$ respectively, so that $\sum f_n = n$, then the mean is: $\overline{x} = \frac{1}{n}\sum x_i f_i$.
  - For continuous data, the sample mean should be calculated from the original data if this is known.
    - If it is tabulated in a frequency table, and the original data is not known, the sample mean can be **estimated** by assuming that all observations in a given interval occurred at the midpoint of that interval i.e. the **class mark.** So, if the class marks of the interval are $\{m_1, m_2, \ldots m_k\}$ and the corresponding frequencies are $\{f_1, f_2, \ldots, f_k\}$ then the sample mean can be approximated by:

$$\overline{x} = \frac{1}{n}\sum m_i f_i$$

- Weighted Mean
  - If k quantities $\{x_1, x_2, \ldots, x_k\}$ have weights $\{w_1, w_2, \ldots, w_k\}$ respectively, where the weights represents measures of relative importance, then the weighted mean is:

$$\overline{x_k} = \frac{\sum_{i=1}^{k} w_i x_i}{\sum w_i}$$

- Combined mean
  - If k finite groups having $\{n_1, n_2, \ldots n_k$ measurements respectively, have means $\{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_k}\}$, the combined mean is:

$$\overline{x_c} = \frac{\sum_{i=1}^{k} n_i \; \overline{x_i}}{\sum n_i}$$

Median

- Data type: ranked
  - Obtained by
    - If even $\widetilde{x} = \frac{x_n + x_{n+1}}{2}$
    - If odd $\widetilde{x} = x_{\frac{n+1}{2}}$

Mode

- Most frequent value in a data set

Measures of Central Tendency of Grouped Data

Mean

$$\overline{x} = \frac{1}{n} \sum f x_m$$

where
f = *class frequency;*
x_m = class mark;
n=total number of observations

Median

$$\widetilde{x} = L + \left(\frac{\left(\frac{n}{2} - S_b \right)}{f_m} \right) i$$

where:
f_m = frequency of the median class;
x_i = class mark;
n = total number of observations
L = Lower boundary of the median class
S_b =< cf of the class before Median class

Mode

$$\hat{x} = L + \left(\frac{\Delta_1}{\Delta_1 + \Delta_2} \right) i$$

where:
Δ_1 = difference in the frequencies of the modal class and the next lower class
Δ_2 = differenc in the frequencies of the modal class and the next higher class
i = size of class interval
L = lower boundary of the modal class

Measures of Dispersion, Position, and Shapes

- Measures of Dispersion
- Measures of Position
- Measures of Shape

Measures of Dispersion

- Measures of variability or dispersion are measures of average distance of each observation from the center of the distribution;
- They summarize and describe the extent to which scores in a distribution differ from each other
- Tell us how spread out the scores are
    - A small dispersion would indicate that the data are:
        - Clustered closely around the mean
        - More homogeneous
        - Less variable
        - More consistent
        - More uniformly distributed

Classifications

Measures of absolute dispersion

- Are expressed in the units of the general observations; they cannot be used to compere variations of two data sets when the average of these sets differ a lot in value or when the observations differ in units of measurements

Absolute Dispersion:

- Range: difference between the highest and the lowest values; **simplest but most unreliable measure of dispersion** $Range = HV - LV$ where HV is the highest value, and LV is the lowest value

- Variance: the average of the squared deviation of each score from the **mean.**

  Population variance (raw data):

  $$\sigma^2 = \frac{1}{N}\sum f\left(x-\mu\right)^2$$

  Population Variance (grouped data):

  $$\sigma^2 = \frac{1}{N}\sum f\left(x_m-\mu\right)^2$$

  Sample variance (raw data):

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

  Sample Variance (grouped data):

  $$s^2 = \frac{1}{n-1}\sum f\left(x_m-\overline{x}\right)^2$$

- Standard Deviation: the square root of the variance

Population Standard Deviation (raw data)

$$\sigma = \sqrt{\frac{1}{n}\sum \left(x-\mu\right)^2}$$

Sample Standard Deviation (raw data)

$$s = \sqrt{\frac{1}{n-1}\sum \left(x-\overline{x}\right)^2}$$

Or

$$s = \sqrt{\frac{n\sum fx^2-\left(\sum fx\right)^2}{n\left(n-1\right)}}$$

Population Standard Deviation (grouped data)

$$\sigma = \sqrt{\frac{1}{N}\sum f\left(x_m-\mu\right)^2}$$

Sample Standard Deviation (grouped data)

$$s = \sqrt{\frac{1}{n-1}\sum f\left(x_m - \overline{x}\right)^2}$$

Measures of relative dispersion

- Are unit-less measures and are **used when one wishes to compare the scatter of the distribution with another distribution**

Relative Dispersion

- The Coefficient of Variation

Is the ration of the standard deviation to the man and is usually expressed in percentage; used to compare variability of two or more sets of data even when they are expressed in different units of measurements: $cv = \frac{1}{\overline{x}}$

1) Chebyshev's Theorem: at least the fraction of $1 - \frac{1}{k^2}$ of measurements of any set of data must lie within k standard deviations of the mean

Example:

- If the IQ's of a random sample of 1080 students at a large university have a mean score of 120 and standard deviation of 8,
  - Determine the interval containing at least 810 of the IQ's in the sample

    Solution: note that $810/1080 = 0.75$. Hence we want to determine the interval to which 75% of the IQ scores lie. By Chebyshev's theorem, 75% lie within $\mu \pm 2\sigma$, i.e. between

    $$\mu - 2\sigma = 120 - 2\left(8\right) = 104; \mu + 2\sigma = 120 + 2\left(8\right) = 136$$

    Therefore, at least 810 of the respondents score are $104 - 136$.

- In what range can we be sure that no more than 120 of scores fall?

- Solution: Results from a tells at least 810 scored $104 - 136$ .Hence the remaining not more than 120 must have scored outside this interval, i.e. less than 104 or greater than 136

Measures of Position

- Used for locating a position of non-central piece of data relative to the entire set of data

z-score

- measures how many standard deviation an observation is above or below the mean

- Population: $z = (x - \mu) / \sigma$
- Sample: $z = (x - x\ ) / s$

Fractiles or quantiles

- specific fraction or percentage of the observations in a given set must fall

1) Percentiles

   a) Ungrouped:

      (i) Arrange the data from lowest to highest
      (ii) Substitute into the formula $c = \frac{np}{100}$ where n=total number of values; p=percentile rank
      (iii) Either:
         (1) If c is not a whole number, round up to the next whole number. Starting at the lowest value, count over to the number that corresponds to the rounded up value
         (2) If c is a whole number, use the value halfway between the c-th and $(C + 1)$ st values when coming up from the lowest value

- Grouped:
  - The kth percentile on the class interval with at least $\frac{kn}{100}$ cumulative frequency. The k-th percentile is given by

$$P_k = L + \frac{\left(\frac{kn}{100} - S_b\right) i}{f_p}$$

Where:

- f_p = frequency of the percentile class
- n = total number of observations
- i = size of class interval
- L = lower boundary of the percentile class
- S_b =<cf of the class before percentile class

1) Quartiles

    a) Definition $Q_1 = P_{25}; Q_2 = P_{50}$

1) Deciles

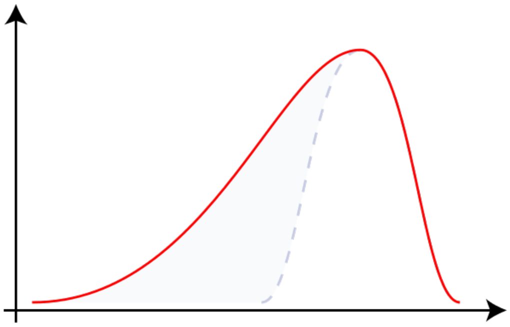
    a) Definition $D_1 = P_{10}; D_2 = P_{20}$

Measures of Shapes

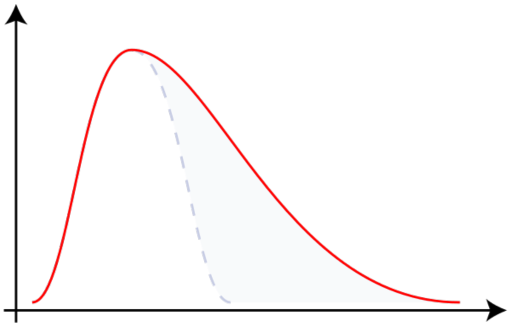
Skewness

- Refers to the degree of symmetry and asymmetry of a distribution;
- The normal distribution is bell-shaped and symmetric through the mean; it has the property of mean=median=mode

1) **Skewed to the left** (negatively skewed)- mean is less than the median; the bulk of the distribution is on the right
2) **Skewed to the right** (positively skewed) - mean is greater than median; the bulk of the distribution is on the left



Negative Skew



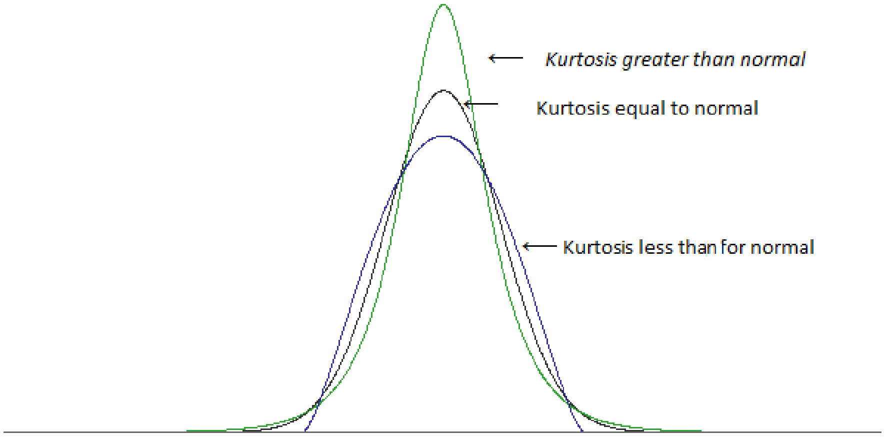
Positive Skew

The extend of skewness can be obtained by getting the coefficient of skewness

$$SK = \frac{3\left(\overline{x} - \widetilde{x}\right)}{s} | s = standard\ dev$$

- If $SK = 0$, the distribution is normal,
- If $SK < 0$, the distribution is to the left,
- If $SK > 0$, the distribution is to the right

Kurtosis



- Mesokurtic - is a normal distribution
- Leptokurtic - more peaked than the normal distribution
- Platykurtic - flatter than the normal distribution

Ungrouped

$$Ku = \frac{\sum (x - \overline{x})^4}{ns^4}$$

Grouped

$$Ku = \frac{\sum f\,(x_m - \overline{x})^4}{ns^4}$$

- If $Ku = 3$, mesokurtic
- If $Ku < 3$, leptokurtic
- If $Ku > 3$, platykurtic

Probability Theory

- Basic Concepts
- Counting Principles
- Probability of an Event

Probability Theory

Basic concepts

- Experiment - process of observing a phenomenon that has variation in its outcomes.
- Outcome - a result from a single trial of an experiment
- Sample space - the set of all possible outcomes of an experiment [usually represented by S .
  - Event - a collection of some outcome from an experiment; a subset of the sample space
    - Simple event - an event that contains one element
    - Compound event - can be expressed as a union of simple events
  - Null space (or empty space) - subset of the sample space that contains no elements; denoted by $\varnothing$

Operations with Events

- Union of two events A and B - contains both the elements of A and B; denoted by $A \cup B$
- Complement of an event A - contains the set of all elements in S that are not in A; denoted by A' or A^C
- Intersection of two events A and B - contains the common elements of A and B; denoted by $A \cap B$

Note: two events A and B are **mutually exclusive** if $A \cap B \ = \varnothing$, that is A and B has no common elements

Counting principles

- Multiplication rule - if an operation can be performed in n_1 ways, and for each of these, a second operation can be performed in n_2 ways, then the two operations can be performed in $n_1 \times n_2$ ways

  - Generalized Multiplication rule - if an operation can be performed in n_1 ways and for each of these, a second operation can be performed in n_2 ways, and so on. Then the sequence of k operations can be perfomed in $n_1 \times \ldots \times n_k$ ways

- Permutation - an **ordered** arrangement of all part of a set of objects.
- Theorem: the number of permutations in n distinct objects is $n! = n \times (n-1) \times \ldots \times 1$
  - For arranging a set of distinct objects

1) Theorem: the number of permutations of n objects can be taken r at a time is: $nP_r = \frac{n!}{(n-1)!}$

   a) e.g. arranging 10 books, 7 at a time

1) Theorem: the number of permutation of n distinct objects arranged in a cricle is $(n-1)!$

   a) Arranging objects in a circle

1) Theorem: the number of distinct permutations of n things which n_1 are not of one kind, of n_2 a second kind, ... , n_k of the k -th kind is:

$$\frac{n!}{n_1! \times n_2! \times \ldots \times n_k!}$$

- Assuming n are not (pairwise) distinct, and re-arranging a set by categories i.e. $n_1, \ldots, n_k$

1) Theorem: the number of partitioning a set of n objects into r cells with n_1 elements on the first cell, n_2 elements on the second and so on is:

$$\frac{n!}{n_1! \times n_2! \times \ldots \times n_r!}$$

$$where$$

$$n_1 + n_2 + n_r = n$$

- Involves partitioning a set where n denotes the number of objects and n_i denotes the number of objects per cell (or alloted container)

- Combination - an arrangement of objects **without regard to order**
- Theorem: the number of combinations of n distinct objects taken r at a time is:

$$nCr = \frac{n!}{(n-r)!r!}$$

Probability of an Event

- Probability - pertains to the likelihood of occurrence of an event; there are three approaches to probability
    - Subjective probability - chance of occurrence is given by a particular person based on his/her educated guess, opinion, intuition, or beliefs

- Empirical probability - probability is assigned based on the prior knowledge of the events that happened on the past, or based on research experiment

- Classical probability - applied when all possible outcomes are equally likely to happen

Probability of an Event

- In classical probability, the probability that an event E will occur is:

$$P\left(E\right) = \frac{n\left(E\right)}{n\left(S\right)} = \frac{number\ of\ outcomes\ in\ E}{number\ of\ outcomes\ in\ the\ sample\ Space}$$

Properties:

1) $0 \leq P\left(E\right) \leq 1, P\left(\varnothing\right) = 0,\ P\left(S\right) = 1$

2) If $S\{x_1,..,x_k\}$, then $P\left(\{x_1\right) = \frac{1}{k}$ and

$$\sum P\left(\{x_1\right) = P\left(\{x_1\right) + \ldots + P\left(\{x_i\right) = 1$$

- Rule of Complements
- Theorem: If E and E' are complementary events, then

  $P\left(E\right) + P\left(E^{'}\right) = 1$ or $P\left(E^{'}\right) = 1 - P\left(E\right)$

- Addition rule of Probability
- Theorem: If A and B are two events, then

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

1) Corollary: If A and B are **mutually exclusive**, then

$$P\left(A \cup B\right) = P\left(A\right) + P\left(B\right)$$

1) Corollary: if $A_1, \ldots A_n$ are mutually exclusive events, then

   $P\left(A_1 \cup \ldots \cup A_n\right) = P\left(A_1\right) + \ldots + P\left(A_n\right)$

- Conditional Probability - the probability of event B occurring when it is known that some event A, denoted as $P\left(A|B\right)$
  - Provides a way to reason with about the outcome of an experiment, based on partial information

- Let A and B be events, the probability of B, given A, denoted by $P\left(B|A\right)$ is given by

  $P\left(B|A\right) = \frac{P(A \cap B)}{P(A)}$ if $P\left(A\right) > 0$

- Theorem [Multiplication Rule]: If A and B are two events, then

  $$P\left(A\cap B\right)=P\left(A\right)P\left(B|A\right)$$

- Independent Events - two events A and B are independent if the fact that A occurs does not affect the probability of B occurring, otherwise, they are dependent

  - Theorem: two events A and B are independent if and only if:

    $$P\left(B|A\right)=P\left(B\right)\left(A|B\right)=P\left(A\right)$$

    otherwise, A and B are dependent

- Theorem: If A and B are independent events, then

  $$P\left(A\cap B\right)=P\left(A\right)P\left(B\right)$$

The Normal Distribution

- Definitions and Properties
- Area under the normal curve
- Applications

Definitions and Properties

The Normal Distribution has some notable properties that which some statistical analysis are predicated upon; we call these group of analyses as parametric statistics, the other group that which holds no assumptions with regards to the probabilistic distribution are called nonparametric statistics.

Here are the notable features of the normal distribution that would be enough to begin our exploration with different hypotheses testing:

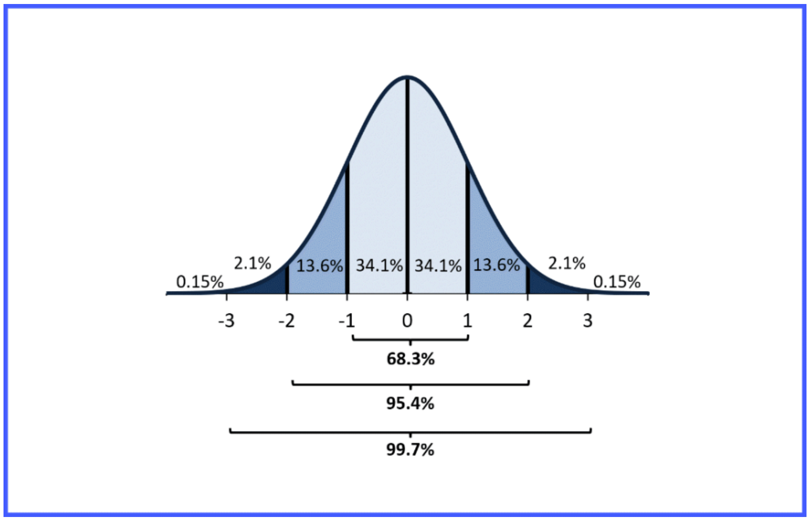
- The mean, median, and mode are located at the center of the distribution and are equal to each other;
    - the distribution is unimodal
- the curve is symmetric about the mean
- the curve never touches the x-axis
- the total area under a curve is equal to 1.00 or 100%

The simplest case of the normal distribution, which we shall cover in this notebook will be the case known as the standard normal distribution which is notable for the defining property of the distribution that has μ =0 and σ =1 and described by the probability density function

$$\Phi\left(x\right)=\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\ \pi}}e^{-\frac{t^2}{2}}dt=\frac{1}{\sqrt{2\ \pi}}\int_{-\ \infty}^{\infty}e^{-\frac{t^2}{2}}dt$$

Empirical rule

- the area under the part of a normal curve that is separated by a degree of 1 standard deviation from the mean



Standard Normal distribution

- a normal distribution with $\mu = 0; \sigma = 1$

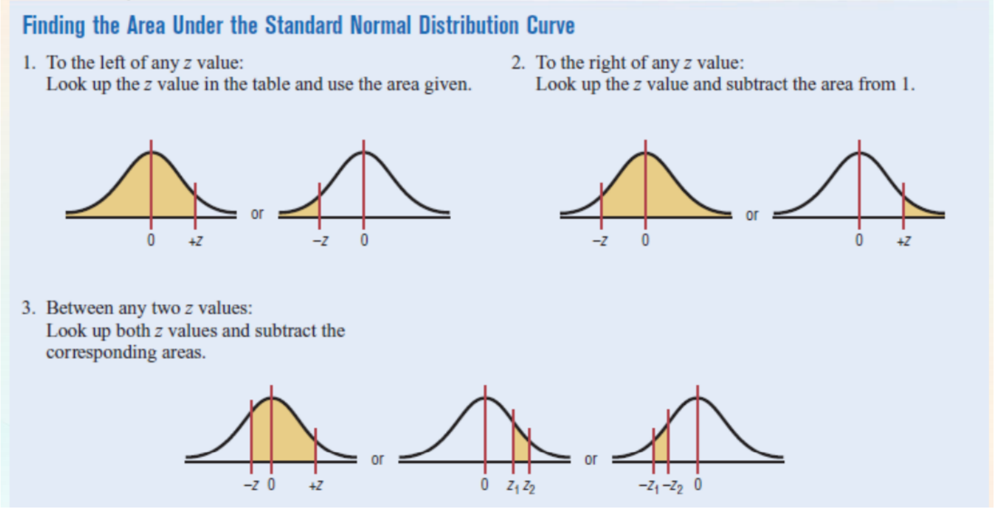
all normally distributed variables can be transformed into the standard normally distributed variable by using the formula for the standard score that is given by: $Z = \frac{X-\mu}{\sigma}$.

Area under the normal curve

- the area under the normal distribution is used for solving problems such as finding the percentage of adult women whose height is between 5ft and 4in and 5ft 7in.

Finding Areas under the standard normal distribution curve

1. draw the normal distribution curve and shade the area
2. find the appropriate figure in the procedure table and follow the directions given:



Finding the Area Under the Standard Normal Distribution Curve

1. To the left of any z value:
   Look up the z value in the table and use the area given.

2. To the right of any z value:
   Look up the z value and subtract the area from 1.

3. Between any two z values:
   Look up both z values and subtract the corresponding areas.

The z-table

- z-table gives the area under the normal distribution curve to the left of any z-value given in two decimal places.

Applications

- the area under the standard normal distribution curve can be also thought of as a probability i.e. if it were possible to select any z-value at random, the probability of choosing one, say, between z_1 and z_2 would be the same as the area under the curve between z_1 and z_2
  - for probabilities, a special notation is used: $P\left(z_1 < z < z_2\right)$

z-table and probabilities

- the z-table we are using gives the area to the left of z, hence that areas translates to $P\left(Z < z\right)$. We have the following guidelines on using the z-table:
  - $P\left(Z > z\right) = 1 - P\left(Z < z\right)$
  - $P\left(z_1 < z < z_2\right) = P\left(Z < z_2\right) - P\left(Z < z_1\right)$

Single Sample Hypothesis test for the mean

- Basic Concepts in Hypothesis Testing
- Hypothesis Test for the Mean (z-test)
- Hypothesis Test for the Mean (t-test)

Basic Concepts in Hypothesis Testing

-
  - **Hypothesis testing** is a decision-making procedure for evaluating claims about a population
    - ∗ Procedure
      - † define the population
      - † select a sample from the population (collect data)
      - † state the particular hypothesis that will be investigated
        - ▷ null hypothesis (H_0) - asserts that there is *no significant difference between a parameter and a specific value*
        - ▷ alternative hypothesis (H_a) - asserts that *there is a diffirence between parameters*
      - † give the significance level - default $\alpha = 0.05$
      - † choose the appropriate **test statistic** and establish the **rejection region**
      - † compute for the value of the test statistic from the sample
      - † make a decision. Reject H_0 if the test statistic has a value in the critical region, otherwise do not reject.
      - † reach a conclusion
  - **Statistical hypothesis** is a conjecture about a population parameter; this may be true or not
  - **Statistical test** - uses the data obtained from a sample to make decisions about whether the null hypothesis should be rejected
    - ∗ **test value** - pertains to the numerical value obtained from the statistical test
  - **Hypothesis test** - the mean is computed for the data obtained from the sample and is compared with the population mean. A decision is made to reject or not reject the null hypothesis on the basis of the **test value**.
    - ∗ if the difference is significant, the null hypothesis is rejected. Otherwise, the null hypothesis is not rejected (failed to be rejected).

- Type 1 Error (Alpha) – Happens when our significance level is too large

- Type 2 Error (Beta) – Happens when our significance level is too small

- **Decisions** - confusion matrix (Type 1 and Type 2 errors)
- **Errors**
  - **type 1** **(** α **)**: if you reject H_0 when it is true (false positive)
  - **type 2** **(** β **)** if you did not reject H_0 when it is false (false negative)

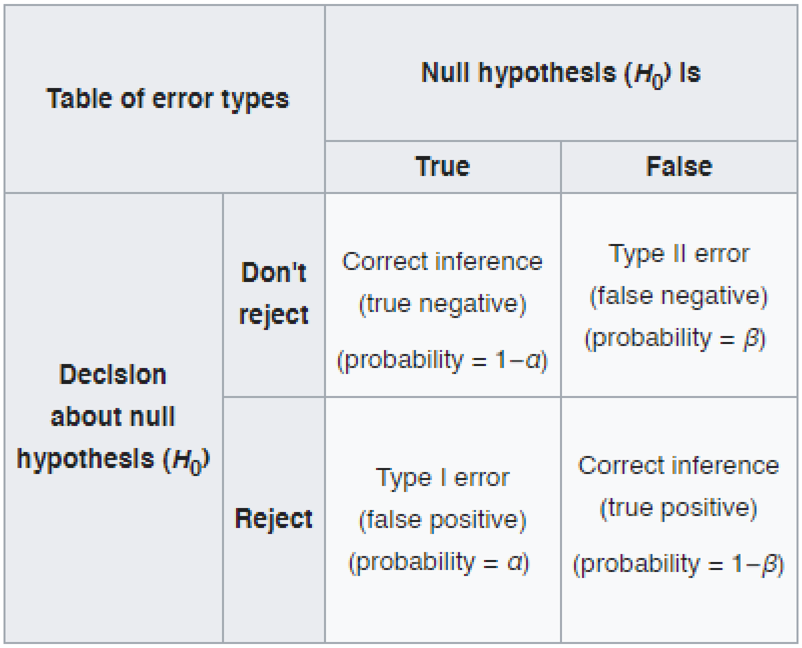
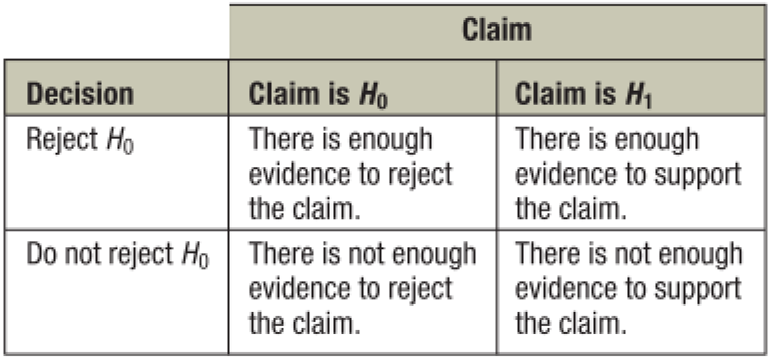


Table of error types		**Null hypothesis (H_0) is**	
		True	**False**
Decision about null hypothesis (H_0)	**Don't reject**	Correct inference (true negative) (probability = $1-\alpha$)	Type II error (false negative) (probability = β)
	Reject	Type I error (false positive) (probability = α)	Correct inference (true positive) (probability = $1-\beta$)



	Claim	
Decision	Claim is H_0	Claim is H_1
Reject H_0	There is enough evidence to reject the claim.	There is enough evidence to support the claim.
Do not reject H_0	There is not enough evidence to reject the claim.	There is not enough evidence to support the claim.

- **Level of significance** - the maximum probability of committing a **type 1 error**. This probability is symbolized as $P\left(type\ I\ error\right) = \alpha$.
  - Statisticians generally agree on using three arbitrary significance levels: 0.10, 0.05, 0.01.

- **Critical value** [after a significance level is chosen, a critical value is selected from a table for the appropriate test]
- determines critical and non-critical regions.
- can be on the *right-side* of the mean or on the *left side* of the mean for a *one-tailed tes*t; the location depends on the inequality sign of the H_a

- **Critical or rejection region** - the range of values of the **test value** that indicates that there is a **significant difference** and the H_0 should be rejected
- **non-critical regions** - the range of values of the **test value** that indicates that the difference was **probably due to chance** and that the H_0 should not be rejected

- **One-tailed** - H_0 should be rejected when the **test value** is in the **critical region** on the one side of the mean.
  - Right-tailed test
  - Left-tailed test

- **Two-tailed test** - H_0 should be rejected when the **test value** is in either of the two critical regions

Hypothesis Test for the Mean (z-test)

z-test is a statistical test for the mean of a population. It can be used when $n \geq 30$, or when the population is normally distributed and s is known.

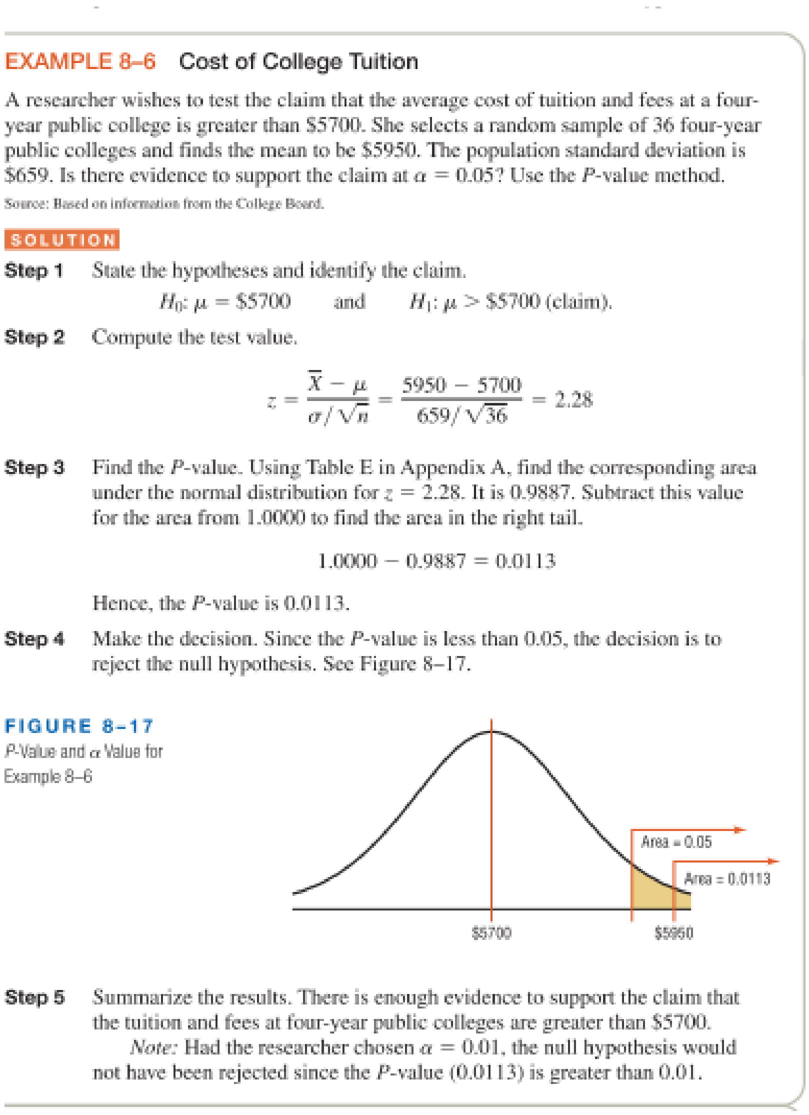
$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

$$where:$$

- $\overline{X}$ $=$ sample mean
- $\mu =$ hypothesized population mean
- $\sigma =$ population standard deviation
- $n =$ sample size

Procedure:

- state the H_0 and H_a hypotheses
- choose a level of significance
- choose the appropriate test statistic and establish the **rejection region**
- compute for the value of the test statistic from the sample
- make a decision
- make a conclusion



EXAMPLE 8–6 Cost of College Tuition

A researcher wishes to test the claim that the average cost of tuition and fees at a four-year public college is greater than \$5700. She selects a random sample of 36 four-year public colleges and finds the mean to be \$5950. The population standard deviation is \$659. Is there evidence to support the claim at $\alpha = 0.05$? Use the *P*-value method.

Source: Based on information from the College Board.

SOLUTION

Step 1 State the hypotheses and identify the claim.

$$H_0\colon \mu = \$5700 \quad \text{and} \quad H_1\colon \mu > \$5700 \text{ (claim)}.$$

Step 2 Compute the test value.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{5950 - 5700}{659/\sqrt{36}} = 2.28$$

Step 3 Find the *P*-value. Using Table E in Appendix A, find the corresponding area under the normal distribution for $z = 2.28$. It is 0.9887. Subtract this value for the area from 1.0000 to find the area in the right tail.

$$1.0000 - 0.9887 = 0.0113$$

Hence, the *P*-value is 0.0113.

Step 4 Make the decision. Since the *P*-value is less than 0.05, the decision is to reject the null hypothesis. See Figure 8–17.

FIGURE 8–17
P-Value and α Value for Example 8–6

Step 5 Summarize the results. There is enough evidence to support the claim that the tuition and fees at four-year public colleges are greater than \$5700.

Note: Had the researcher chosen $\alpha = 0.01$, the null hypothesis would not have been rejected since the *P*-value (0.0113) is greater than 0.01.

Hypothesis Test for the Mean (t-test)

when the **population standard deviation is unknown**, the z-**test** is not normally used for testing hypotheses involving means. A different test, called the t-**test** is used. **The distribution of the variable should be approximately normal.**

T-distribution
the t-distribution is *similar* to the standard normal distribution in the following ways:

- it is bell-shaped
- it is symmetric about the mean
- the mean, median, and mode are equal to 0, and are located at the center of the distribution
- the curve never touches the x-axis

The t-distribution *differs* to the standard normal distribution in the following ways:

- the variance is greater than 1
- the t-distribution is a family of curves based on the *degrees of freedom*, from which is a number related to the sample size
- as the sample increases, the t-distribution approaches the normal distribution

T-test

-
    - a statistical test for the mean of the population and is used when the population is normally or approximately normally distributed, σ is unkown, or when the sample size is small i.e. $n < 30$
    - test statistic for t-test is:

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

the degrees of freedom (df) = n-1

Procedure:

- state the H_0 and H_a hypotheses
- choose a level of significance
- choose the appropriate test statistic and establish the **rejection region**
- compute for the value of the test statistic from the sample
- make a decision
- make a conclusion

EXAMPLE 8–12 Hospital Infections

A medical investigation claims that the average number of infections per week at a hospital in southwestern Pennsylvania is 16.3. A random sample of 10 weeks had a mean number of 17.7 infections. The sample standard deviation is 1.8. Is there enough evidence to reject the investigator's claim at $\alpha = 0.05$? Assume the variable is normally distributed.

Source: Based on information obtained from Pennsylvania Health Care Cost Containment Council.

SOLUTION

Step 1 H_0: $\mu = 16.3$ (claim) and H_1: $\mu \neq 16.3$.

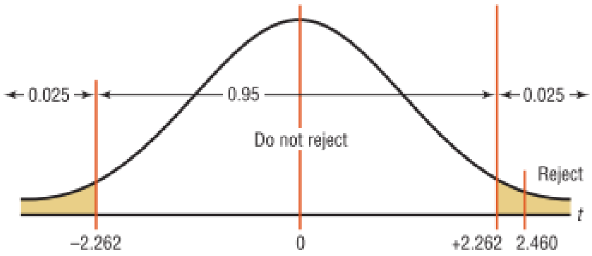
Step 2 The critical values are $+2.262$ and -2.262 for $\alpha = 0.05$ and d.f. $= 9$.

Step 3 The test value is

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{17.7 - 16.3}{1.8/\sqrt{10}} = 2.460$$

Step 4 Reject the null hypothesis since $2.460 > 2.262$. See Figure 8–20.

FIGURE 8–20
Summary of the *t* Test of Example 8–12



Step 5 There is enough evidence to reject the claim that the average number of infections is 16.3.

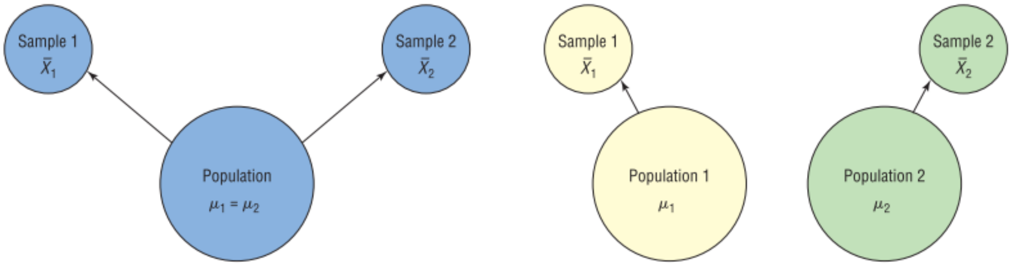
Hypothesis test for difference in Means

- Hypothesis Test for Difference of Means (z-test)
- Hypothesis Test for the Difference of Means (t-test)

Independent samples - samples are independent samples when they are not related

Hypothesis Test for Difference of Means (z-test)

FIGURE 9–2 Hypothesis-Testing Situations in the Comparison of Means



(a) Difference is not significant. The means of the populations are the same. Do not reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is not significant.

(b) Difference is significant. The means of the populations are different. Reject H_0: $\mu_1 = \mu_2$ since $\bar{X}_1 - \bar{X}_2$ is significant.

Testing difference between two means

- Researchers would wish to compare two sample means using experimental and control groups
  - For example, the average lifetimes of two different brands of bus tires might be compared to see whether there is any difference in tread wear.

Assumptions

- The sample must be independent of each other.
  - There can be no relationship between the subjects in each sample
- The standard deviation of both populations must be known, and it the sample sizes are less than 30, the populations must be normally or approximately distributed.

Test statistic

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

One-tailed tests

- Right-tailed
  - $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
  - $H_a : \mu_1 > \mu_2$ or $\mu_1 - \mu_2 > 0$

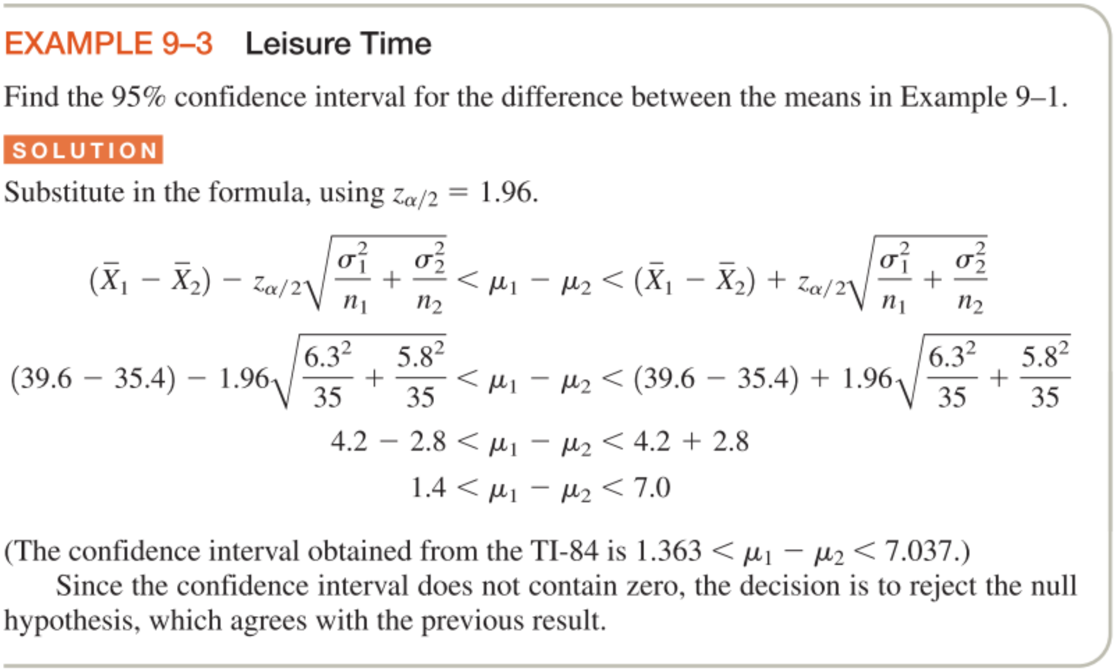
- Left-tailed
  - $H_0 : \mu_1 = \mu_2$ or $\mu_1 - \mu_2 = 0$
  - $H_a : \mu_1 < \mu_2$ or $\mu_1 - \mu_2 < 0$

Procedure:

1. state the H_0 and H_a hypotheses
2. choose a level of significance
3. choose the appropriate test statistic and establish the **rejection region**
4. compute for the value of the test statistic from the sample
5. make a decision
6. make a conclusion

Formula for the z-confidence Confidence Interval for Difference Between Two Means

$$\left(\overline{X_1} - \overline{X_2}\right) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < \left(\overline{X_1} - \overline{X_2}\right) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$



EXAMPLE 9–3 Leisure Time

Find the 95% confidence interval for the difference between the means in Example 9–1.

SOLUTION

Substitute in the formula, using $z_{\alpha/2} = 1.96$.

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$(39.6 - 35.4) - 1.96\sqrt{\frac{6.3^2}{35} + \frac{5.8^2}{35}} < \mu_1 - \mu_2 < (39.6 - 35.4) + 1.96\sqrt{\frac{6.3^2}{35} + \frac{5.8^2}{35}}$$

$$4.2 - 2.8 < \mu_1 - \mu_2 < 4.2 + 2.8$$

$$1.4 < \mu_1 - \mu_2 < 7.0$$

(The confidence interval obtained from the TI-84 is $1.363 < \mu_1 - \mu_2 < 7.037$.)

Since the confidence interval does not contain zero, the decision is to reject the null hypothesis, which agrees with the previous result.

Hypothesis Test for the Difference of Means (t-test)

Testing difference between two means

- In many situations, the conditions for z-test cannot be met in such cases t-test is used
  - Population standard deviations are not known
- t-test is used to test difference between means when the two samples are independent and when the samples are taken from two normally or approximately normally distributed populations

Assumptions

Test statistic for unequal variances - used when population variances are assumed to be unequal, we use the following test statistic.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where the degrees of freedom are equal to whichever is smaller; $n_1 - 1$ or $n_2 - 1$. Critical Values depend on the **t-distribution**

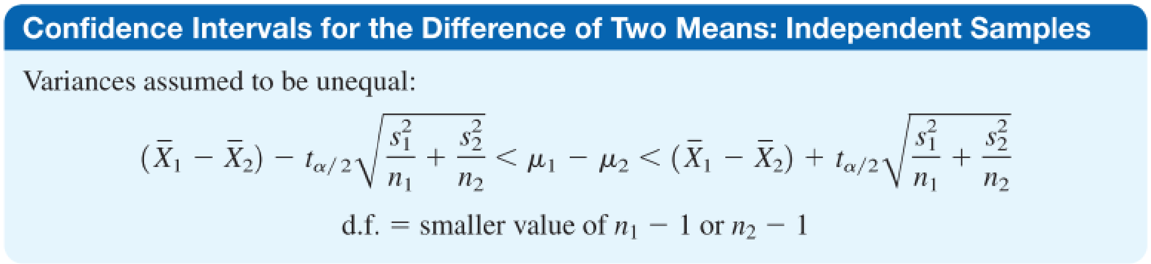
t-test for Equal Variances - used when variances are assumed to be equal.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Where the degrees of freedom are equal to $df = n_1 + n_2 - 2$. Critical Values depend on the t-distribution.

Procedure:

1. state the H_0 and H_a hypotheses
2. choose a level of significance
3. choose the appropriate test statistic and establish the **rejection region**
4. compute for the value of the test statistic from the sample
5. make a decision
6. make a conclusion



Confidence Intervals for the Difference of Two Means: Independent Samples

Variances assumed to be unequal:

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

d.f. = smaller value of $n_1 - 1$ or $n_2 - 1$

EXAMPLE 9–5 Find the 95% confidence interval for the data in Example 9–4.

SOLUTION

Substitute in the formula.

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$(123 - 116) - 2.365\sqrt{\frac{8^2}{10} + \frac{5^2}{8}} < \mu_1 - \mu_2 < (123 - 116) + 2.365\sqrt{\frac{8^2}{10} + \frac{5^2}{8}}$$

$$7 - 7.3 < \mu_1 - \mu_2 < 7 + 7.3$$

$$-0.3 < \mu_1 - \mu_2 < 14.3$$

Since 0 is contained in the interval, there is not enough evidence to support the claim that the mean weights are different.

- Bluman, A. G. (2009). *Elementary statistics: A step by step approach.* New York, NY: McGraw-Hill Higher Education.
- Walpole, R. E. (1982). *Introduction to statistics* (No. 04; QA276. 12, W35 1982.).