Statistical Methods

By

Prof. Dr/ Monir Hussein Bahgat

Professor of Internal Medicine, Mansoura University.

Trainer of SPSS and statistical analysis skills, Mansoura University.

Quality manager, Specialized Medical Hospital, Mansoura University.

Second Year

2019/2020

Acknowledgments

This two-year curriculum was developed through a participatory and collaborative approach between the Academic faculty staff affiliated to Egyptian Universities as Alexandria University, Ain Shams University, Cairo University, Mansoura University, Al-Azhar University, Tanta University, Beni Souef University, Port Said University, Suez Canal University and MTI University and the Ministry of Health and Population(General Directorate of Technical Health Education (THE). The design of this course draws on rich discussions through workshops. The outcome of the workshop was course specification with Indented learning outcomes and the course contents, which served as a guide to the initial design.

We would like to thank **Prof.Sabah Al- Sharkawi** the General Coordinator of General Directorate of Technical Health Education, **Dr. Azza Dosoky** the Head of Central Administration of HR Development, **Dr. Seada Farghly** the General Director of THE and all share persons working at General Administration of the THE for their time and critical feedback during the development of this course.

Special thanks to the **Minister of Health and Population Dr. Hala Zayed and Former Minister of Health Dr. Ahmed Emad Edin Rady** for their decision to recognize and professionalize health education by issuing a decree to develop and strengthen the technical health education curriculum for pre-service training within the technical health

Contents

Course Description

This course is prepared for undergraduate students aiming at introducing the student both to statistical reasoning and to the most commonly used statistical techniques. It should also support the students who master this material to be able to select, implement, and interpret the most common types of analyses as they undertake research in their own disciplines.

Core Knowledge

On successful completion of this course, the student will be able to :

1- **Define statistical methods.**
2- **Identify the two types of statistical methods; descriptive statistics and inferential statistics.**
3- **Recognize the commonly used terms in statistical methods.**
4- **Select the appropriate statistical method according to the goal and type of data.**
5- **Describe the different methods of descriptive statistics including measures of central tendency and measures of dispersion.**
6- **Determine the proper use of inferential statistics including its two main types; hypothesis testing and estimation statistics.**

Core General Skills

On successful completion of this course, the student will be able to:

1. Able to search for a computer software or website to help perform statistical analysis in easier way.
2. Work hardly to improve the skills in applying statistical methods.
3. Demonstrate caution and proficiency in applying statistical methods.

Core Professional Skills

On successful completion of this course, the student will be able to:

1. Sample a population in a representative manner (particularly simple random sampling).
2. Analyze the raw data in a proper way.
3. Label values correctly.
4. Make hypothesis testing.
5. Solve statistical problems.
6. Criticize a research paper.

Teaching and learning methods

1- Lectures using power point presentations.
2- Positive interaction with the lecturer by asking questions or answering them.
3- Practical sessions to solve statistical exercises.
4- Hand-outs to simplify the scientific material.
5- External readings of specialized books.
6- Training to answer model question exercises.

Teaching and learning methods for students with disabilities

1- Remedial lectures.
2- Remedial practicing lessons.
3- Simplified sketches for different statistical methods.
4- Training to answer model question exercises.

Core Skills

on successful completion of this course, the student will be able to :

1. Solve exercises on the commonly used statistical methods.
2. Apply the proper statistical method to a given dataset.
3. Interpret the results of statistical methods.
4. Report the results of statistical methods.
5. Compare the different types of statistical methods.
6. Hypothesize in a correct way what to be tested.

توصيف مقرر دراسى

بيانات المقرر		
الرمز الكودى :	اسم المقرر: طرق إحصائية Statistical Methods	الفرقة /المستوى :الثانية / شعبة تسجيل طبي
التخصص :	عدد الوحدات الدراسية: ٣ نظري ١٢ عملى	

تقويم الطلاب	
الأساليب المستخدمة	1- Practice test. 2- Mid-term test evaluation. 3- Final written test.
التوقيت	Final theoretical written exam: 3-hours.
توزيع الدرجات	1- Mid-tem test = 20 marks 2- Written test 100 marks 3- Practical test = 80 marks 4- Total marks = 200

قائمة الكتب الدراسية والمراجع	
Handouts for the lectures and practical sections	**مذكرات**
Statistical methods: 1st edition.	**كتب ملزمة**
Freud RJ, Wilson WJ and Mohr DL (2010). Statistical Methods (Third Edition). Amsterdam, Boston, Heidelberg, London New York, Oxford, Paris, San Diego San Francisco, Singapore, Sydney, Tokyo. Elsevier Inc.	**كتب مقترحة**
The Journal of Modern Applied Statistical Methods: https://digitalcommons.wayne.edu/jmasm/	**دوريات علمية أو نشرات ..... الخ**

Course Overview

ID	Theory	Practice
1st week	Introduction: Define and classify statistical methods. Outline importance of studying statistical methods.	Plot a flow chart for statistical methods
2nd week	Essential terminology to understand common statistical terms.	The student is asked to create explanatory example for each statistical term.
3rd week	Proper selection of the suitable statistical method for a particular goal and a particular type of variable.	The student is offered different scenarios and then is asked to nominate the appropriate statistical method to each scenario.
4th week	Revision for part I	Revision for part I
5th week	Descriptive statistics: Measures of central Tendency.	Problem solving for mean, median and mode
6th week	Descriptive statistics: Measures of dispersion.	Problem solving for ranges and standard deviation.
7th week	Descriptive statistics: Measures of frequencies.	Problem solving for rate, ratio, proportion, and percentage.
8th week	Revision for part II	Revision for part II
9th week	Introduction to hypothesis testing	The student is asked to create explanatory examples
10th week	5-step procedure for hypothesis testing	The student is asked to perform the procedure working on an explanatory example.
11th week	Hypothesis testing for a proportion	The student is asked to perform the procedure working on an explanatory example.
12th week	Estimation statistics: Confidence Interval	The student is asked to perform the procedure working on an explanatory example.
13th week	Revision for part III	Revision for part III

Preface

Statistical methods are mathematical formulas, models, and techniques that are used in statistical analysis of raw research data.

The application of statistical methods extracts information from research data and provides different ways to assess the strength of research outputs.

This book is designed for undergraduate students aiming at introducing the student both to statistical reasoning and to the most commonly used statistical techniques.

The goal is that students who master this material will be able to select, implement, and interpret the most common types of analyses as they undertake research in their own disciplines.

Moreover, they should be able to read research articles and in most cases understand the descriptions of the statistical results and how the authors used them to reach their conclusions.

They should understand the pitfalls of collecting statistical data, and the roles played by the various mathematical assumptions.

Statistics can be studied at several levels:

On one hand, students can learn by repetition how to plug numbers into formulas, or more often now, into a computer program, and draw a number with a neat circle around it as the answer. This limited approach is mind distressing, and rarely leads to the kind of understanding that allows students to critically select methods and interpret results.

On the other hand, there are numerous textbooks that provide introductions to the elegant mathematical backgrounds of the methods. Although this is a much deeper understanding than the first approach, its prerequisite mathematical understanding closes it to practitioners from many other disciplines.

This book seems to take a middle way by presenting enough of the formulas to motivate the techniques, and by illustrating their numerical application in a small example.

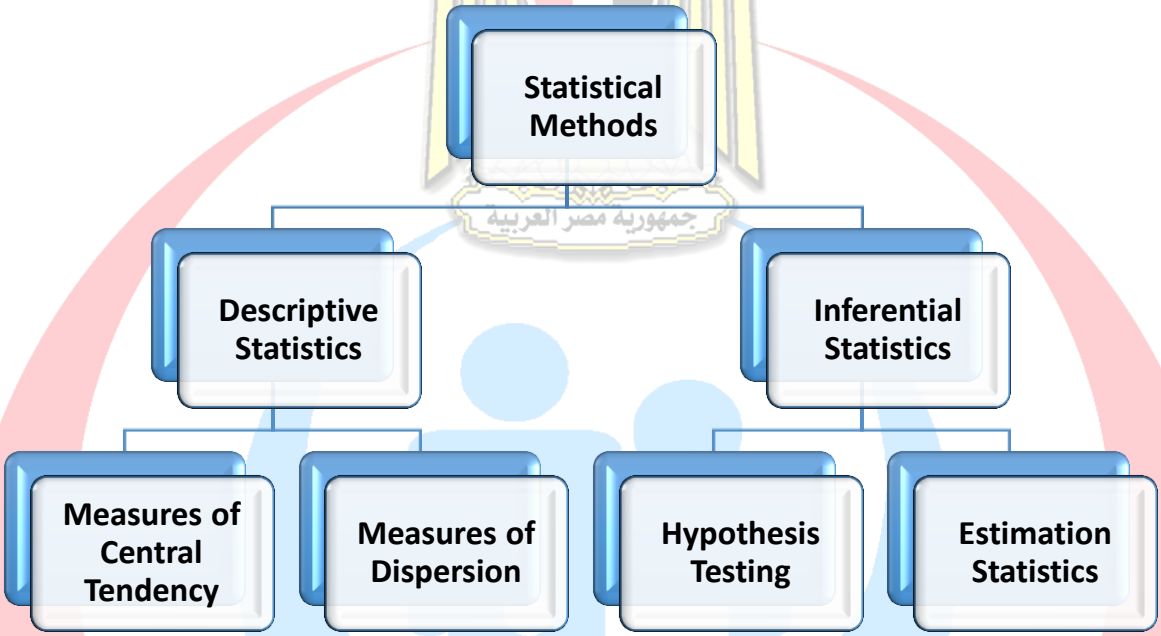
However, the focus of the discussion is on the selection of the technique, the interpretation of the results, and a critique of the validity of the analysis.

The student is advised to focus on these skills.

This book is divided into three parts:

- Part I (Introduction to statistical methods).
- Part II (Methods of descriptive statistics).
- Part III (Methods of inferential statistics).

Flow Chart of Statistical Methods:



Chapter 1
Introduction to Statistical Methods

Introduction to statistical methods

Part I (Introduction to statistical methods):

I.1. Essential terminology:

- **Population and sample:**

Definition:

A population is a data set representing the entire entity of interest while a sample is a data set consisting of a portion of a population. Normally a sample is obtained in such a way as to be representative of the population.

Explanatory example:

You are interested in studying prevalence of systemic hypertension among Mansoura University students. The study population will be 'all' students in the university. As this will consume time and resources, you'll carefully select a representative sample to do the study so that the results obtained from this sample can be generalized to the population from which this sample was selected.

- **Variables:**

Definition:

Variable is a characteristic of some event, object, or person that can take on more than one value.

Explanatory example:

When a study involves the two types of sex (coded as 1 if male and 2 if female), sex in this study is considered a variable. In another study performed on males only, sex is constant and is not considered as a variable.

Variable types:

Variables are classified into qualitative (categorical) and quantitative.

Qualitative variable is a variable that expresses a qualitative attribute (has no measurement unit) while quantitative variable has a measurement unit.

Qualitative variables are of two types: Nominal (categories have no meaningful order) and Ordinal (categories have meaningful order). Dichotomous variable is a sub-type of categorical variable with only two categories.

Quantitative variables are of two types: Continuous (data are continuous points on the scale and therefore include fractions) and Discrete (data are discrete points on the scale and therefore always come in integer form).

Quantitative variables are of two measurement types: Interval variable (a measurement where the difference between two values is meaningful) and a ratio variable (has all the properties of an interval variable, and also has a clear

definition of zero. When the variable equals 0.0, there is none of that variable). When working with ratio variables, but not interval variables, you can look at the ratio of two measurements.

Explanatory examples:

Sex (coded as 1 if male and 2 if female) is a nominal variable because it has no measurement unit with no meaningful order of the codes

Education level (coded as 1 if illiterate or just read & write, 2 if educated to a level below university and 3 if educated at a university level or more) is an ordinal variable because it has no measurement unit with meaningful order of the codes as 3 means higher education level than 2 and 2 means higher education level than 1.

Age (measured in years) is a continuous variable because it has a measurement unit (years) and fractions are accepted (patient's age might be 23.5 years).

Number of children for each female in a study is a discrete variable because it has a measurement unit (children number; e.g., 2 children, 3 children, and so on) and fractions are not accepted (a female can't have 3.5 children).

Temperature (expressed in F or C) is an example of interval variable. The difference between a temperature of 100 degrees and 90 degrees is the same difference as between 90 degrees and 80 degrees. A temperature of 0.0 on either of those scales does not mean 'no heat'.

Weight is a ratio variable. Therefore, a weight of 60 kg is twice a weight of 30 kg, while a temperature of 40 degrees C is not twice as hot as 20 degrees C, because temperature C is not a ratio variable like weight.

- **Data set:**

Definition:

A set of data is a collection of observed values representing one or more characteristics of some objects or units.

Explanatory example:

You asked a 11 patients about their age (in years), sex (coded as 1 if male and 2 if female) and education level (coded as 1 if illiterate or just read & write, 2 if educated to a level below university and 3 if educated at a university level or more). You tabulated the results as follows:

Table (1): How data set looks like

ID	Age	Sex	Education level
1	21	1	1
2	24	1	2
3	23	2	2
4	25	2	3
5	26	1	1
6	30	2	2
7	31	1	3
8	32	1	3
9	20	2	1
10	21	2	1
11	24	2	2

This data set is a collection of observed values representing age (continuous variable), sex (nominal variable) and education level (ordinal variable) of 11 patients.

- **Descriptive and inferential statistics:**

Definition:

Descriptive statistics intend to describe a data set with summary charts and tables, but do not attempt to draw conclusions about the population from which the sample was taken. You are simply summarizing the data you have like telling someone the key points of a book (executive summary) as opposed to just handing them a thick book (raw data).

Conversely, with inferential statistics, you are testing a hypothesis and drawing conclusions about a population, based on your sample.

Descriptive statistics aim to summarize a "sample" of a population.
Inferential statistics aim to draw conclusions about that "population".

Explanatory example:

Descriptive statistics:

Let's say you've tested 50 university students for hepatitis C antibody. You have a bunch of data plugged into your spreadsheet and now it is time to share the results with someone. You could hand over the spreadsheet and say "here's what

I learned" (not very informative), or you could summarize the data with some charts and graphs that describe the data and communicate some conclusions (e.g. 10% of university students have positive anti-HCV antibody). This would sure be easier for someone to interpret than a big spreadsheet. There are hundreds of ways to visualize data, including data tables, pie charts, line charts, etc. That is the idea of descriptive statistics. Note that the analysis is limited to your data and that you are not extrapolating any conclusions about the full population (the whole university students).

Inferential statistics:

Let's continue with HCV example. Let's say you wanted to know the prevalence in Mansoura University students. Well, there are >110,000 students and it would be impossible to test every single student for anti-HCV. Instead, you would try to test a representative sample of students and then extrapolate your sample results to the entire population. While this process is not perfect and it is very difficult to avoid errors, it allows researchers to make well-reasoned inferences about the population in question. This is the idea behind inferential statistics. Getting a representative sample is really important. There are many methods of sampling strategies, including random sampling. A true random sample means that everyone in the target population has an equal chance of being selected for the sample. Another key component of proper sampling is the sample size. Obviously, the larger the sample size, the better, but there are trade-offs in time and money when it comes to obtaining a large sample.

There are online calculators as well as software (free and commercial) that help determine appropriate sample sizes.

Examples of online calculators include:

https://www.surveysystem.com/sscalc.htm

https://www.calculator.net/sample-size-calculator.html

http://clincalc.com/stats/samplesize.aspx

Examples of software include:

https://www.cdc.gov/epiinfo/index.html (free)

http://www.gpower.hhu.de/en.html (free)

https://www.ncss.com/software/pass/ (commercial)

When it comes to inferential statistics, there are generally two forms: estimation statistics and hypothesis testing.

Estimation Statistics:

"Estimation statistics" is a way of saying that you are estimating population values based on your sample data. Let's think back to our sample HCV data. First, let's assume that we had a true random sample of 50 students from this university and that our full target population is >110,000 students. Let's say that 10% of students in our sample had positive anti-HCV. Can we safely extrapolate that 10% of all university students also will have positive anti-HCV? Is that the true value of the university? Well, we can't say with 100% confidence, but using inferential

statistical techniques such as the “confidence interval” we can provide a range of students that test positive for anti-HCV with some level of confidence.

Hypothesis Testing

Hypothesis testing is simply another way of drawing conclusions about a population parameter (“parameter” is simply a number, such as a mean, that includes the full population and not just a sample).

With hypothesis testing, one uses a test such as T-Test, Chi-Square, or ANOVA to test whether a hypothesis about the mean is true or not. Again, the point is that this is an inferential statistic method to reach conclusions about a population, based on a sample set of data.

Practical exercises

1. Serum albumin measured in g/dl is
    a. A nominal variable.
    b. An ordinal variable.
    c. An interval variable.
    d. A ratio variable.
2. Eye color is:
    a. A nominal variable.
    b. An ordinal variable.
    c. An interval variable.
    d. A ratio variable.
3. You examined the quality of chest x-rays whether poor, fair, good or excellent. This is:
    a. A nominal variable.
    b. An ordinal variable.
    c. An interval variable.
    d. A ratio variable.
4. A measurable characteristic of a population is:
    a. A parameter
    b. A statistic
    c. A sample
    d. An experiment

5. A measurable characteristic of a sample is:
    a. A parameter
    b. A statistic
    c. A population
    d. An experiment
6. A subset of a population is:
    a. a parameter
    b. a statistic
    c. a sample
    d. an experiment
7. the body weight in kilograms is
    a. A categorical (nominal) variable.
    b. An ordinal variable.
    c. A continuous variable.
    d. A discrete variable.
8. A true random sample means that:
    a. Everyone in the target population has an equal chance of being selected for the sample.
    b. Everyone in the target population has a 75% chance of being selected for the sample.
    c. Everyone in the target population has a 25% chance of being selected for the sample.
    d. Not everyone in the target population has an equal chance of being selected for the sample.

Answers:

1. D.
2. A.
3. B.
4. A.
5. B.
6. C.
7. C.
8. A.

Chapter 2

Proper selection of Statistical Methods

I.2. Proper selection of the statistical method:

Statistical methods are mathematical formulas, models, and techniques that are used in statistical analysis of raw research data. The application of statistical methods extracts information from research data and provides different ways to assess the strength of research outputs.

To properly select which statistical method to use, we need to know our aim (goal) and the type of data (or variable) to be tested. The goal may be just descriptive or it might be performing a comparison or looking for a relationship. Variables may be nominal (including the dichotomous variable), ordinal, or quantitative. For quantitative variables, we need to know if this variable is normally distributed (Gaussian distribution) or not (non-Gaussian distribution). This can be done by many statistical methods including Kolmogorov-Smirnov test (for large data sets) and Shapiro-Wilk test (for small data sets <50) where data will be considered normally distributed if the test result is insignificant (p value > 0.050) and will be considered non-normally distributed (skewed) if the test result is significant (p value ≤ 0.050).

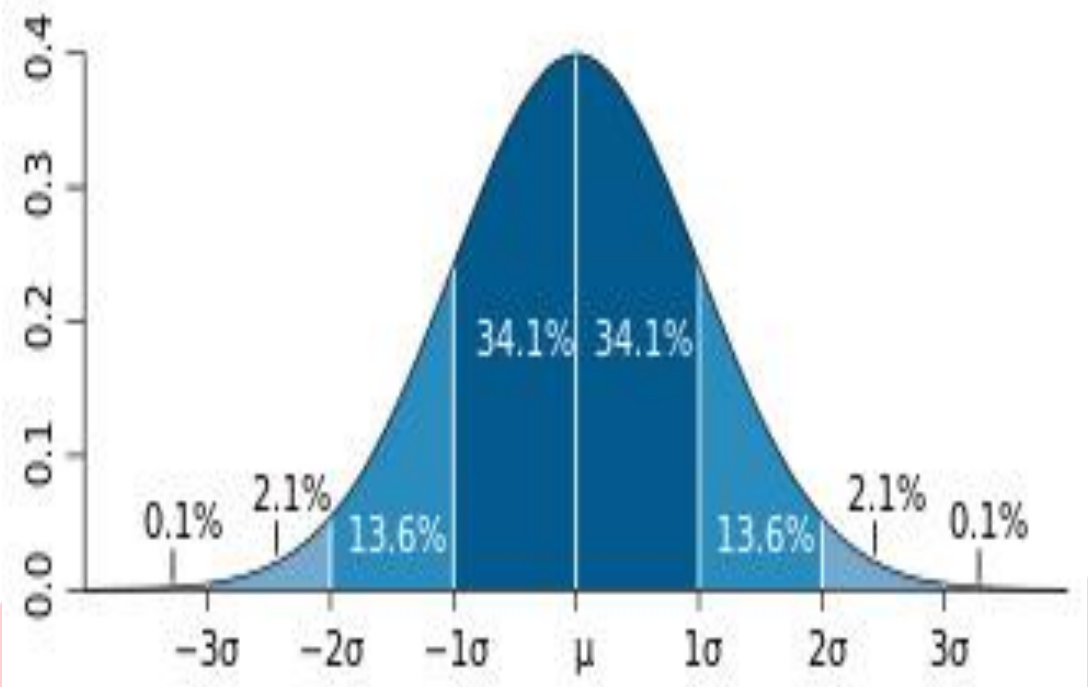


Figure 2.1: Normal (Gaussian) distribution:

μ (pronounced mu) = population mean, σ (pronounced sigma) = population standard deviation

The normal distribution is bell shaped and symmetric about the mean.

The interval (μ ± 1σ) contains approximately 68% of the observations.

The interval (μ ± 2σ) contains approximately 95% of the observations.

The interval (μ ± 3σ) contains virtually all of the observations (99.7%).

The following table simplify the choice of the proper statistical method:

Table (1): Choosing a proper statistical method for analysis

Aim (goal)	Type of data / variable		
	Quantitative (normal distribution)	- Quantitative (skewed distribution) - or ordinal	Nominal
Describe a group	Mean ± SD	Median (IQR)	Frequency (%)
Compare one-group to a hypothetical value	One-sample t-test	Wilcoxon test	One-sample Chi-Square test
Compare paired data (two sets)	Paired-samples t-test	Wilcoxon test	McNemar test
Compare repeated measures (>two sets)	Repeated-measures ANOVA	Friedman test	Cochrane Q test
Compare two-groups	Independent-samples t-test	Mann-Whitney test	Chi-Square test
Compare > two groups	One-Way ANOVA test	Kruskal Wallis H test	Chi-Square test
Relationships between variables:			
Association	Pearson's correlation	Spearman's correlation	Contingency coefficient
Prediction	Linear regression	e.g., ordinal regression	Logistic regression

SD=Standard deviation, IQR=Interquartile range, ANOVA=Analysis of variance.

Part II (Methods of descriptive statistics):

The two most important aspects of describing a data set are the location (central tendency) and the dispersion (spread) of the data.

In other words, we need to find a number that indicates where the observations are on the measurement scale and another to indicate how widely the observations vary.

Practical exercises

1. You collected the age in years for a sample of 40 patients. To test for normality, you will use:
    a. Kolmogorov-Smirnov test.
    b. Shapiro-Wilk test.
    c. Mann-Whitney test.
    d. ANOVA test.

2. You tested the normality of age in years of a sample of 300 patients using Kolmogorov-Smirnov test and you found that p value = 0.475. this means that the age distribution is:
    a. Normal.
    b. Non-Gaussian.
    c. Positively-skewed.
    d. Negatively-skewed.

3. The mean age ± standard deviation of a sample of 100 patients equals 50 ± 5. Considering that age is normally distributed, you are expected that nearly 68 patients will have their age:
    a. Exactly 50 years.
    b. Between 45 and 50 years.
    c. Between 45 and 55 years.
    d. Between 40 and 60 years.

4. To describe sex variable of a group of patients, you will use.
    a. Mean and standard deviation.
    b. Median and interquartile range.
    c. Proportions.
    d. None of the above.

5. You measured the sleeping hours for 50 patients first after taking a placebo and then after taking a sleeping pill that is claimed to be better than placebo. Sleeping hours were normally distributed after both placebo and sleeping pill. To test this hypothesis, you will use:
    a. One-sample t-test.
    b. Paired-samples t-test.
    c. Independent-samples t-test.
    d. ANOVA test.

6. You measured the sleeping hours for a group of 50 patients after taking a placebo and for another group of 50 patients after taking a sleeping pill that is claimed to be better than placebo. The two groups are of similar age and sex (matched groups). Sleeping hours were normally distributed in both groups. To test this hypothesis, you will use:
    a. One-sample t-test.
    b. Paired-samples t-test.
    c. Independent-samples t-test.
    d. ANOVA test.

7. You measured the sleeping hours after taking one of three sleeping pills to compare their effect. Each pill was tested in a separate group of 50 patients; and the three groups were of similar age and sex (matched groups). Sleeping hours were normally distributed in each of the three groups. To test this hypothesis, you will use:
    a. One-sample t-test.
    b. Paired-samples t-test.
    c. Independent-samples t-test.
    d. ANOVA test.

8. You want to compare the proportion of fall out of bed in a medical department (50 patients) against a surgical department (50 patients). You found that 25 patients fell in medical department (50%) and 10 patients fell in surgical department (20%). To compare the fall rate between the two department, you will use:
    a. Paired-samples t-test.
    b. Independent-samples t-test.
    c. ANOVA test.
    d. Chi-square test.

9. You want to test for a possible linear association between the heart rate (measured in beats / minute) of a 100 healthy women and their age (measured in years). You assumed that as age goes up, heart rate will go down (negative correlation). You found that both heart rate and age were normally distributed. To test this hypothesis, you will use:
    a. Pearson's correlation.
    b. Spearman's correlation.
    c. Contingency coefficient.
    d. Chi-square test.

10. You want to test if it is possible to predict the heart rate of a healthy woman (measured in beats / minute) by knowing her age (measured in years). To test this hypothesis, you will use:
    a. Pearson's correlation.
    b. Spearman's correlation.
    c. Linear regression.
    d. Logistic regression.

11. You want to test if it is possible to predict that a patient admitted to hospital will fall out of bed (i.e., fall or no fall). The predictors (independent variables) are age of the patient (measured in years) and the admission department (medical versus surgical). To test this hypothesis, you will use:
    a. Pearson's correlation.
    b. Spearman's correlation.
    c. Linear regression.
    d. Logistic regression.

Answers:

1. B.
2. A.
3. C.
4. C.
5. B.
6. C.
7. D.
8. D.
9. A.
10. C.
11. D.

Chapter 3
Measures of central tendency (measures of location)

Measures of central tendency (measures of location)

Three different measures are available to describe data location including:

- Mean
- Median
- Mode

Mean:

The most frequently used measure of location is the arithmetic mean, usually referred to simply as the mean or the average.

Definition:

The mean is the sum of all the observed values divided by the number of values.

Formula:

$$mean = (\sum_{i=1}^{n} yi)/n$$

n = Sample size

yi = observed values of the variable Y (where i = 1, 2, . . , *n*).

Σ = the sum of.

Explanatory example:

For the example in table (1) for the 15 patients, their mean age will be 26.07 years:

Mean = (21+24+23+25+26+30+31+32+20+21+24+26+28+29+31)/15

[Try to calculate it yourself].

Median:

This is another useful measure of location.

Definition:

The median of a set of observed values is defined to be the middle value when the measurements are arranged from lowest to highest; that is, 50% of the measurements lie above it and 50% fall below it.

The precise definition of the median depends on whether the number of observations is odd or even as follows:

- If n is odd, the median is the middle observation; hence, exactly (n - 1)/2 values are greater than and (n - 1)/2 values are less than the median, respectively.
- If n is even, there are two middle values and the median is the mean of the two middle values and n/2 values are greater than and n/2 values are less than the median, respectively.

Explanatory example:

a. 'n' is odd number of observations:

To calculate the median for the following dataset: 2, 1, 3, 5, 3, 6, 4.

1. Arrange from lowest to highest: 1, 2, 3, 3, 4, 5, 6
2. As the n is odd (7), median is the middle value (4th value) which is 3

b. 'n' is even number of observations:

To calculate the median for the following dataset: 2, 1, 3, 5, 3, 6, 4, 2.

1. Arrange from lowest to highest: 1, 2, 2, 3, 3, 4, 5, 6
2. As the n is even (8), median is the mean of the two middle values (mean of 4th and 5th values) which are 3 and 3. So, median = (3+3)/2 = 6/2 =3

Comparing mean and median:

Which one should be used for descriptive statistical method?

The choice of the measure to be used may depend on its ultimate interpretation and use.

For symmetric or nearly symmetric distributions (bell-shaped), the mean and median will be the same or nearly the same, while for skewed distributions the value of the mean will tend to be “pulled” toward the long tail.

This is illustrated in the following figure:

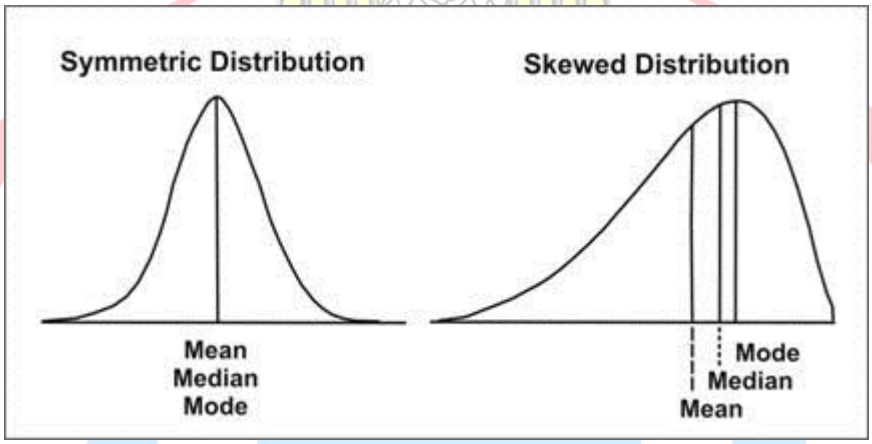


Figure 3.1: Normal and skewed distribution:

This phenomenon can be explained by the fact that the mean can be interpreted as the center of gravity of the distribution. That is, if the observations are viewed as weights placed on a plane, then the mean is the position at which the weights on each side balance. It is a well-known fact of physics that weights placed further from the center of gravity exert a larger degree of influence (also called leverage); hence the mean must shift toward those weights in order to achieve balance. However, the median assigns equal weights to all observations regardless of their actual values; hence, the extreme values have no special leverage.

Explanatory example:

For the following two datasets X and Y we'll calculate mean and median for each dataset:

X: 1, 2, 3, 3, 4, 5

Y: 1, 1, 1, 2, 5, 8

Mean values:

For X = (1 + 2 + 3 + 3 + 4 + 5) / 6 = 18 / 6 = 3

For Y = (1 + 1 + 1 + 2 + 5 + 8) / 6 = 18 / 6 = 3

So, the mean value for these two different datasets is equal.

Median values:

For X = (3 + 3) / 2 = 3

For Y = (1 + 2) / 2 = 1.5

So the median value for these two datasets is different.

The reason for this difference is due to different distribution which is symmetric for variable X (note that mean and median for this dataset is the same) and skewed to the right for variable Y.

Pearl:

"Use the mean as the single measure of location unless the distribution of the variable is skewed."

Table (1)

Normal distribution	Skewed distribution
Mean	Median

The mean is calculated using the value of each observation, so all the information available from the data is utilized. This is not so for the median. For the median, we only need to know where the "middle" of the data is. Therefore, the mean is the more useful measure and, in most cases, the mean will give a better measure of the location of the data. However, as we have seen, the value of the mean is heavily influenced by extreme values and tends to become a distorted measure of location for a highly skewed distribution. In this case, the median may be more appropriate.

Mode:

Definition:

The mode is the most frequently occurring value.

This measure may not be unique in that two (or more) values may occur with the same greatest frequency.

Also, the mode may not be defined if all values occur only once, which usually happens with continuous numeric variables.

Explanatory example:

For the following dataset:

1, 1, 2, 1, 3, 2, 4, 3, 1, 3, 4, 2, 3, 1, 1, 1, 1, 3, 4, 5, 6, 3, 6, 2, 5, 1, 1

The most frequently occurring value is ‘1’, which occurs 10 times. So the mode is ‘1’.

Practical exercises

1. Which one of the following is a measure of central tendency?
    a. Range.
    b. Interquartile range.
    c. Standard deviation.
    d. Median.

2. For a dataset of 25, 18, 5, 12, 24, and 16, the mean is
    a. 12
    b. 16.7
    c. 20.5
    d. 18

3. For a dataset of 25, 18, 5, 12, 24, and 16, the median is
    a. 16
    b. 17
    c. 18
    d. 20

4. If 5, 3, 4 and 2 of your sample cases have age of 20, 23, 30, and 25 years, respectively, the age **mode** is:
    a. 20 years.
    b. 23 years.

c. 25 years.

d. 30 years.

5. The median is a better measure of central tendency than the mean if:
   a. The variable is discrete.
   b. The distribution is skewed.
   c. The variable is continuous.
   d. The distribution is symmetric.

6. The median can be defined as:
   a. The 25th percentile.
   b. The 50th percentile.
   c. The 75th percentile.
   d. The average.

7. For a dataset of 1, 4, 6, 3, 2, and 5:
   a. The mean is more than the median.
   b. The mean less than the median.
   c. The mean equals the median.
   d. The mean does not equal the median.

Answers:

1. D.
2. B.
3. B.
4. A.
5. B.
6. B.
7. C.

Chapter 4
Measures of dispersion (spread)

Measures of dispersion (spread):

Although location is generally considered to be the most important single characteristic of a distribution, the variability or dispersion of the values is also very important.

The mean or median of a variable provides an inadequate description of the distribution of that variable since the list of values would include a wide range.

Therefore, description is best presented by using a measure of central tendency as well as a measure of dispersion.

For normally distributed data, using mean and standard deviation is the most appropriate while for non-normally distributed data, using median and interquartile range is more appropriate.

Measures of dispersion include:

- Range.
- Variance.
- Standard deviation.

Range:

Definition:

The difference between the largest and smallest observed values.

Although it is the simplest and most obvious measure of variability, the range has one very serious drawback: It completely ignores any information from all the other values in the data.

Explanatory example:

For a dataset of 2,4,5,4,6,3,7,5,1,9,5

The minimum value is 1

The maximum value is 9

So, the range equals 9-1 = 8

Interquartile Range (IQR):

Definition:

The interquartile range is the length of the interval between the 25th and 75th percentiles and describes the range of the middle half of the distribution.

Formula:

IQR = Upper quartile (Q3) – Lower quartile (Q1)

Steps to calculate:

- Organize dataset in ascending order.
- Divide data into two halves:
  - Odd number: Midpoint will be the exact middle number.
  - Even number: Midpoint will be between the two middlemost numbers.
- Find the median of:
  - Upper half of data (=Q3).
  - Lower half of data (=Q1).
- Subtract Q3 – Q1 to determine the IQR.

Explanatory example:

- For a dataset of 7,6,3,4,8,2,1,5
  - Data in ascending order: 1,2,3,4,5,6,7,8
  - Data in two halves: 1,2,3,4 and 5,6,7,8
  - Q3 = (6 + 7) / 2 = 13 / 2 = 6.5
  - Q1 = (2 + 3) / 2 = 5 / 2 = 2.5
  - IQR = 6.5 – 2.5 = 4
- For a dataset of 9,7,6,3,4,8,2,1,5
  - Data in ascending order: 1,2,3,4,5,6,7,8,9
  - Data in two halves: 1,2,3,4 and 6,7,8,9 (5=exact middle number)
  - Q3 = (7 + 8) / 2 = 15 / 2 = 7.5
  - Q1 = (2 + 3) / 2 = 5 / 2 = 2.5
  - IQR = 7.5 – 2.5 = 5

Variance:

Definition:

The variance is mathematically defined as the average of the squared differences from the mean.

Formula:

The sample variance, denoted by s2 (pronounced sigma square), of a set of n observed values having a mean $\bar{y}$ is the sum of the squared deviations divided by n - 1:

$$Variance = \sum_{i=1}^{n} (yi - \bar{y})2/(n-1)$$

n = Sample size.

yi = observed values of the variable Y (where i = 1, 2, . . , *n*).

$\bar{y}$ = mean value.

Σ = the sum of.

(yi - $\bar{y}$)2 = Squared deviations.

Explanatory example:

For a dataset of 1, 2, 3, 4, 5

To calculate the variance:

- Calculate the mean:
  - Mean = (1 + 2 + 3 + 4 + 5) / 5 = 15 / 5 = 3

- Calculate the sum of squared deviations:
  - Sum of squared deviations:
    $= (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2$
    $= (-2)^2 + (-1)^2 + (0)^2 + (+1)^2 + (+2)^2$
    $= 4 + 1 + 0 + 1 + 4$
    $= 10$

- Calculate variance:
  - Variance:
    = Sum of squared deviations / (n - 1)
    = 10 / (5 - 1)
    = 10 / 4
    = 2.5

Notes:

1. Note that the variance is actually an average or mean of the squared deviations and is often referred to as a mean square.

2. Note also that we have divided the sum by (n - 1) rather than n:

- One of the uses of the sample variance is to estimate the population variance.

- Dividing by n tends to underestimate the population variance; therefore by dividing by (n - 1) we get, on average, a more accurate estimate.

- Recall that we have already noted that the sum of deviations $(y_i - \bar{y})$ = 0; hence, if we know the values of any (n - 1) of these values, the last one must have that value that causes the sum of all deviations to be zero.

- Thus there are only (n - 1) "free" deviations. Therefore, the quantity (n - 1) is called the degrees of freedom.

Standard deviation:

Definition:

The standard deviation of a set of observed values is defined to be the positive square root of the variance.

Formula:

This measure is denoted by s and does have a very useful interpretation as a measure of dispersion:

$$s = \sqrt{variance}$$

Explanatory example:

For the above dataset with variance calculated to be 2.5

s = $\sqrt{2.5}$

s = 1.58

Usefulness of the Mean and Standard Deviation:

Although the mean and standard deviation (or variance) are only two descriptive measures, together the two actually provide a great deal of information about the distribution of an observed set of values.

This is illustrated by the empirical rule:

If the shape of the distribution is nearly bell-shaped, the following statements hold:

- The interval ($\bar{y} \pm s$) contains approximately 68% of the observations.
- The interval ($\bar{y} \pm 2s$) contains approximately 95% of the observations.
- The interval ($\bar{y} \pm 3s$) contains virtually all of the observations.

Note that for each of these intervals the mean is used to describe the location and the standard deviation is used to describe the dispersion of a given portion of the data.

Explanatory example:

For a normally distributed dataset, n = 50, mean ($\bar{y}$) = 20, s = 2

According to empirical rule:

- ($\bar{y} \pm s$), which is 20 ± 2, defines the interval 18 to 22 and should include (0.68)*(50) = 34 observations,
- ($\bar{y} \pm 2s$), which is 20 ± 4, defines the interval from 16 to 24 and should include (0.95)*(50) = 48 observations,
- ($\bar{y} \pm 3s$) which is 20 ± 6, defines the interval from 14 to 26 and should include all 50 observations.

Calculating the standard deviation (s) from the range:

The empirical rule provides us with a quick method of estimating the standard deviation of a bell-shaped distribution.

Since at least 95% of the observations fall within 2 standard deviations of the mean in either direction, the range of the data covers about 4 standard deviations.

Thus, we can roughly estimate the standard deviation by taking the range divided by 4.

Explanatory example:

For a dataset, the minimum is 16 and the maximum is example 28, therefore, the range of this dataset is 28 - 16 = 12. Divided by 4 we get 3. The estimated standard deviation had a crude value of 3.

Practical exercises

1. Which one of the following is a measure of dispersion?
    a. Mean.
    b. Mode.
    c. Standard deviation.
    d. Median.
2. For a dataset of 10, 14, 16, 22, 12, and 19, the range is
    a. 10.
    b. 11.
    c. 12.
    d. 13.
3. For a dataset of 1, 2, 3, 4, and 5, the variance is:
    a. 2
    b. 2.5
    c. 3
    d. 3.5
4. If the variance of a dataset is 4, the standard deviation is
    a. 2
    b. 4
    c. 8
    d. 16

5. If the variance of a dataset of 100 observations is 4, the standard error of the mean (SEM) is
    a. 0.2
    b. 0.4
    c. 0.6
    d. 0.8
6. The standard error of the mean (SEM) is calculated by dividing the standard deviation by:
    a. Mean.
    b. Median.
    c. Sample size.
    d. Square root of sample size.
7. For a dataset of 0, 1, 2, 3, 4, 5, 6, and 7, interquartile range is:
    a. 2
    b. 4
    c. 6
    d. 8
8. In a secondary school, you found that the smallest height among students was 152 cm and the tallest student has a height of 176 cm. the estimated standard deviation would roughly be equal to:
    a. 2
    b. 4
    c. 6

d. 8

Answers:

1. C.
2. C.
3. B.
4. A.
5. A.
6. D.
7. B.
8. C (Range = 176 – 152 = 24, standard deviation = Range / 4 = 24 / 4 = 6).

Chapter 5
Measures of frequency:

These measures show how often a value occurs

Measures of frequency include:

- Frequency.
- Ratio.
- Rate.
- Proportion.
- Percentage.

Absolute and relative frequency:

Definition:

Absolute frequency: The number of times a certain value occurs in the data.

Relative frequency: The number of times a certain value occurs in the data (absolute frequency) relative to the total number of values for that variable.

The relative frequency may be expressed in ratios, rates, proportions, and percentages.

Ratios:

Definition:

Ratios compare the frequency of one value for a variable with another value for the same variable.

Explanatory examples:

Example (1):

The sex frequency of a certain disease was studied in 40 participants. They were found to be 30 females and 10 males. Therefore, the ratio of female to male is 30:10 (= 3:1).

Example (2):

In 30 participants, a certain adverse effect to an experimental drug occurred in 2 participants only, therefore, the ratio of an experimental drug's adverse effect to no adverse effects is 2:28 (= 1:14).

Rate:

Definition:

Rate is the measurement of one value for a variable in relation to the entire sample of values within a given period.

Explanatory example:

In a total of 30 participants, there are 2 who show adverse effects after taking an experimental drug; therefore, the rate of adverse effects is 2/30 participants (= 1 / 15).

Proportion:

Definition:

Proportion is the fraction of a total sample that has some value.

Explanatory example:

In a total of 30 participants, with two participants having adverse drug effects, the proportion of adverse effects is 2/30 = 0.067

Percentage:

Definition:

Percentage is another way of expressing a proportion as fraction of 100. The total percentage of an entire dataset should always add up to 100%.

Explanatory example:

Example (1):

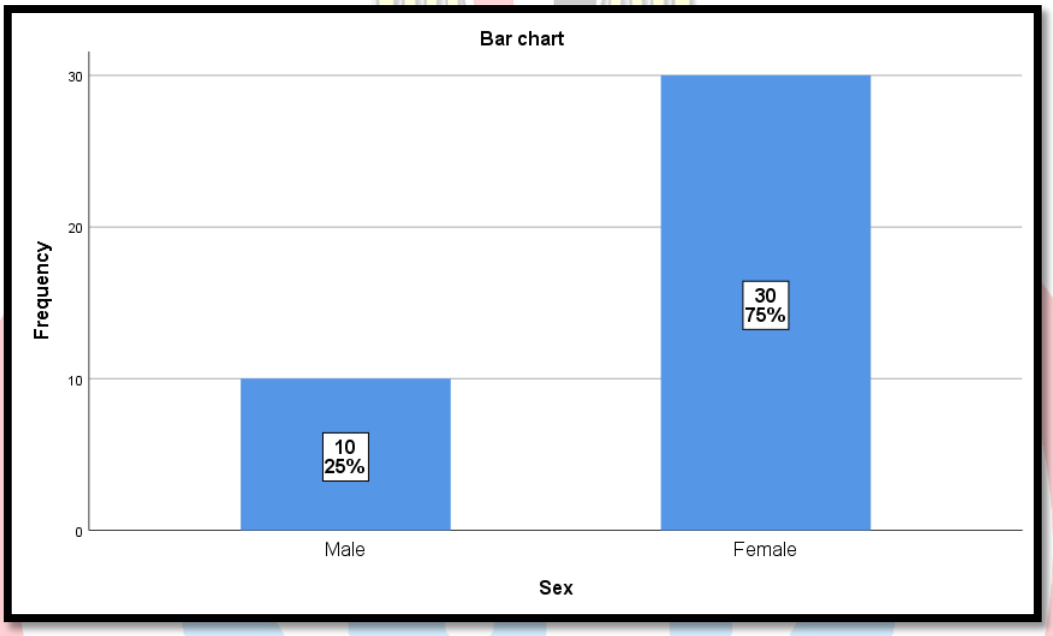
The sex frequency of a certain disease was studied in 40 participants. They were found to be 30 females and 10 males. Therefore, the percentage of female is (30 / 40)*100 = (3/4)*100 = (0.75)*100 = 75%. On the other hand, the percentage of male is (10 / 40)*100 = (1/4)*100 = (0.25)*100 = 25%. Note that 75% + 25% = 100%.

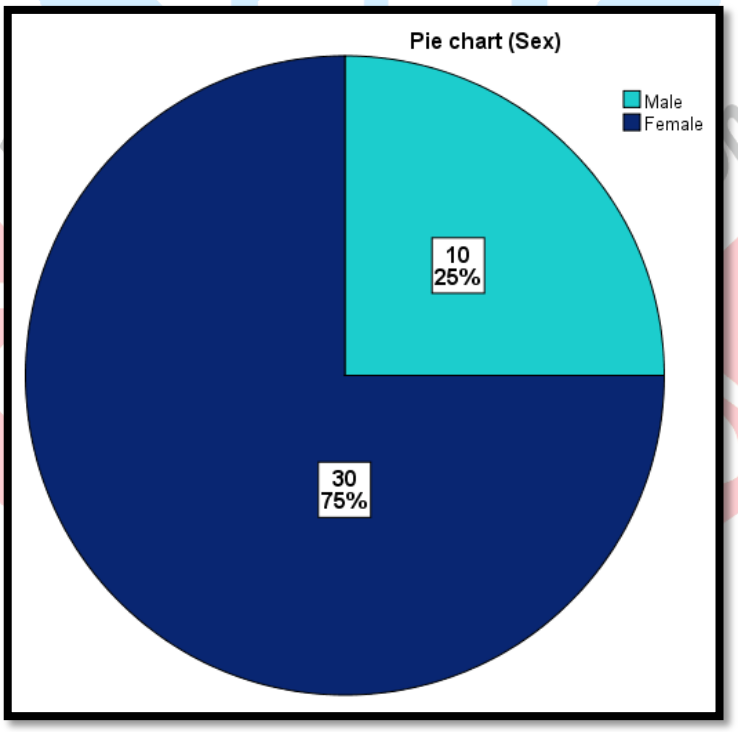
Example (2):

In total of 30 participants, where 2 experience adverse drug effects, (2/30)*100 = (0.067)*100 = 6.7% of participants experience adverse effects.

Graphical presentation:

The above measures of frequency are often expressed visually in the form of tables, histograms (for quantitative variables), or pie or bar graphs (for qualitative variables) to make the information more easily interpretable.





Practical exercises

1. Which of the following is a measure of frequency:
    a. Rate.
    b. Mean.
    c. Median.
    d. Mode.
2. The number of times a particular value occurs in the data is called:
    a. Rate.
    b. Absolute frequency.
    c. Relative frequency.
    d. Proportion.
3. The fraction of a total sample that has some value is called:
    a. Rate.
    b. Ratio.
    c. Proportion.
    d. Percentage.
4. In a study on 50 cardiac patients receiving nitrite, severe headache occurred in 10 patients. The proportion of severe headache is:
    a. 10:50
    b. 10:40
    c. 20%
    d. 0.2

Answers:

1. A.
2. B.
3. C.
4. D.

Chapter 6
Hypothesis testing

To reach conclusions about a population, based on a sample dataset using hypothesis testing, one uses a statistical test such as T-Test, Chi-Square, or ANOVA test to assess whether a hypothesis about the mean is true or not.

A hypothesis usually results from assumption concerning observed behavior, natural phenomena, or established theory. If the hypothesis is stated in terms of population parameters such as the mean and variance, the hypothesis is called a statistical hypothesis. Data from a sample (which may be an experiment) are used to test the validity of the hypothesis. A procedure that enables us to agree or disagree with the statistical hypothesis using data from a sample is called a test of the hypothesis.

Explanatory example:

A test of the effect of a diet pill on weight loss would be based on observed weight losses of a sample of healthy adults. If the test concludes the pill is effective, the manufacturer can safely advertise to that effect.

The Hypotheses:

Statistical hypothesis testing starts by making a set of two statements about the parameter(s) in question.

These are usually in the form of simple mathematical relationships involving the parameters.

The two statements are exclusive and comprehensive, which means that one or the other statement must be true, but they cannot both be true.

The first statement is called the null hypothesis and is denoted by H_0, and the second is called the alternative hypothesis and is denoted by H_1.

Definitions:

The null hypothesis:

The null hypothesis is a statement about the values of one or more parameters. This hypothesis represents the status quo and is usually not rejected unless the sample results strongly imply that it is false.

The alternative hypothesis:

The alternative hypothesis is a statement that contradicts the null hypothesis. This hypothesis is accepted if the null hypothesis is rejected. The alternative hypothesis is often called the research hypothesis.

Explanatory example:

Suppose that you want to conduct a study to test whether smoking causes lung cancer. The current status at that time is that there is no evidence yet to say that and this actually what motivated you to conduct this study. So, the null hypothesis will be that smoking in NOT causing lung cancer (statement of "no effect") while the alternative hypothesis will be that smoking causes lung cancer. You will not be able to reject the null hypothesis (and accept the alternative hypothesis) until you finished your research and found that the null hypothesis is false.

The rejection region:

The rejection region (also called the critical region) is the range of values of a sample statistic that will lead to rejection of the null hypothesis.

Possible Errors in Hypothesis Testing:

The results of a hypothesis test may be subject to two distinctly different errors, which are called type I and type II errors.

Definition:

Type I error:

A type I error occurs when we incorrectly reject H_0, when H_0 is actually true and our sample-based inference procedure rejects it.

Type II error:

A type II error occurs when we incorrectly fail to reject H_0, that is, when H_0 is actually false, and our inference procedure fails to detect this fact.

Table (1): Results of a hypothesis test:

Decision	H_0 in the population	
	True	False
H_o is rejected	Type I error	Correct decision
H_o is not rejected	Correct decision	Type II error

Probabilities of making errors:

Definition:

***a*-** Denotes the probability of making a type I error.

***b*-** Denotes the probability of making a type II error.

Importance:

The ability to provide these probabilities is a key element in statistical inference, because they measure the reliability of our decisions.

Five-Step Procedure for Hypothesis Testing:

- **Step 1:**

Specify H_0, H_1, and an acceptable level of *a*.

This value is based on the seriousness or cost of making a type I error in the problem being considered.

The significance level of a hypothesis test is the maximum acceptable probability of rejecting a true null hypothesis.

Why Do We Focus on the Type I Error?

In general, the null hypothesis is usually constructed to be that of the status quo; that is, it is the hypothesis requiring no action to be taken, no money to be spent, or in general nothing changed. This is the reason for denoting this as the null or nothing hypothesis. Since it is usually expensive to incorrectly reject the status quo than it is to do the reverse, this characterization of the null hypothesis does indeed cause the type I error to be of greater concern.

On the other hand, the alternative hypothesis describes conditions for which something will be done. It is the action or research hypothesis. In an experimental or research setting, the alternative hypothesis is that an established (status quo) hypothesis is to be replaced with a new one. Thus, the research hypothesis is the one we actually want to support, which is accomplished by rejecting the null hypothesis with a sufficiently low level of α such that it is unlikely that the new hypothesis will be erroneously pronounced as true. The significance level represents a standard of evidence. The smaller the value of α, the stronger the evidence needed to establish H_1.

Historically and traditionally, α has been chosen to have values of 0.10, 0.05, or 0.01, with 0.05 being most frequently used.

- **Step 2:**

Define a sample-based test statistic & rejection region for the specified H0.

The test statistic is a sample statistic whose sampling distribution can be specified for both the null and alternative hypothesis case.

After specifying the appropriate significance level of α, the sampling distribution of this statistic is used to define the rejection region.

The rejection region comprises the values of the test statistic for which

- The probability when the null hypothesis is true is less than or equal to the specified α and
- The probabilities when H1 is true are greater than they are under H_0.

- **Step 3:**

Collect the sample data and calculate the test statistic.

- **Step 4:**

Make a decision to either reject or fail to reject H_0.

This decision will normally result in a recommendation for action.

- **Step 5:**

Interpret the results in the language of the problem in such a way that the results be usable by the practitioner.

Since H_1 is of primary interest, this conclusion should be stated in terms of whether there was or was not evidence for the alternative hypothesis.

The p value:

The p value is the probability of committing a type I error if the actual sample value of the statistic is used as the boundary of the rejection region.

It is therefore the smallest level of significance for which we would reject the null hypothesis with that sample.

Consequently, the p value is often called the “attained” or the “observed” significance level.

It is also interpreted as an indicator of the weight of evidence against the null hypothesis.

The Probability of a Type II Error:

In presenting the procedures for hypothesis and significance tests we have concentrated exclusively on the control over α, the probability of making the type I error.

However, just because that error is the more serious one, we cannot completely ignore the type II error. There are many reasons for ascertaining the probability of that error, for example:

- The probability of making a type II error may be so large that the test may not be useful.
- Because of the trade-off between α and ß, we may find that we may need to increase α in order to have a reasonable value for ß.

Power of a test:

Definition:

The power of a test is the probability of correctly rejecting the null hypothesis when it is false. The power of a test is (1 - ß) and depends on the true value of the parameter μ.

As a practical matter we are usually more interested in the probability of not making a type II error, that is, the probability of correctly rejecting the null hypothesis when it is false.

Obviously high power is a desirable property of a test. If a choice of tests is available, the test with the largest power should be chosen. In certain cases, theory leads us to a test that has the largest possible power for any specified alternative hypothesis, sample size, and level of significance. Such a test is considered to be the best possible test for the hypothesis and is called a "uniformly most powerful" test.

One-tailed and two-tailed hypothesis tests:

Two-tailed test:

If you are using a significance level of 0.05, a two-tailed test allots half of your alpha to testing the statistical significance in one direction and half of your alpha to testing statistical significance in the other direction.

This means that .025 is in each tail of the distribution of your test statistic.

When using a two-tailed test, regardless of the direction of the relationship you hypothesize, you are testing for the possibility of the relationship in both directions.

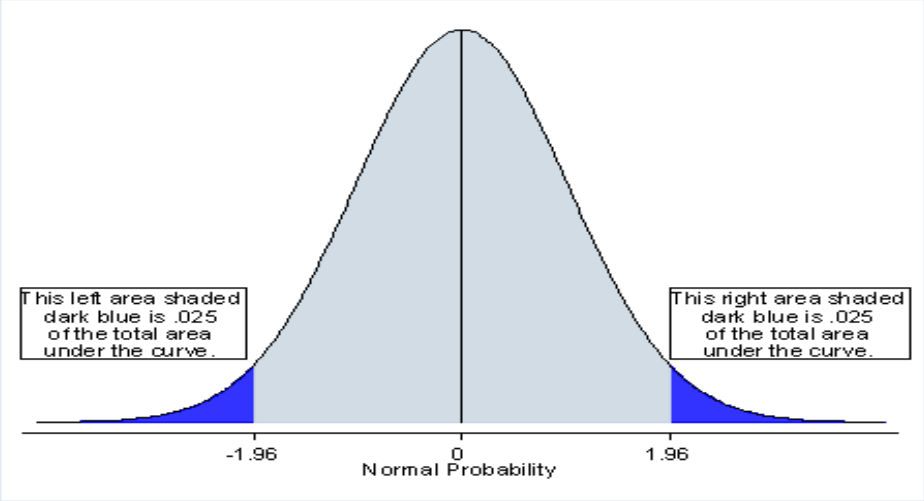


Figure 6.1: Two-tailed test

One-tailed test:

If you are using a significance level of .05, a one-tailed test allots all of your alpha to testing the statistical significance in the one direction of interest.

This means that .05 is in one tail of the distribution of your test statistic.

When using a one-tailed test, you are testing for the possibility of the relationship in one direction and completely disregarding the possibility of a relationship in the other direction.

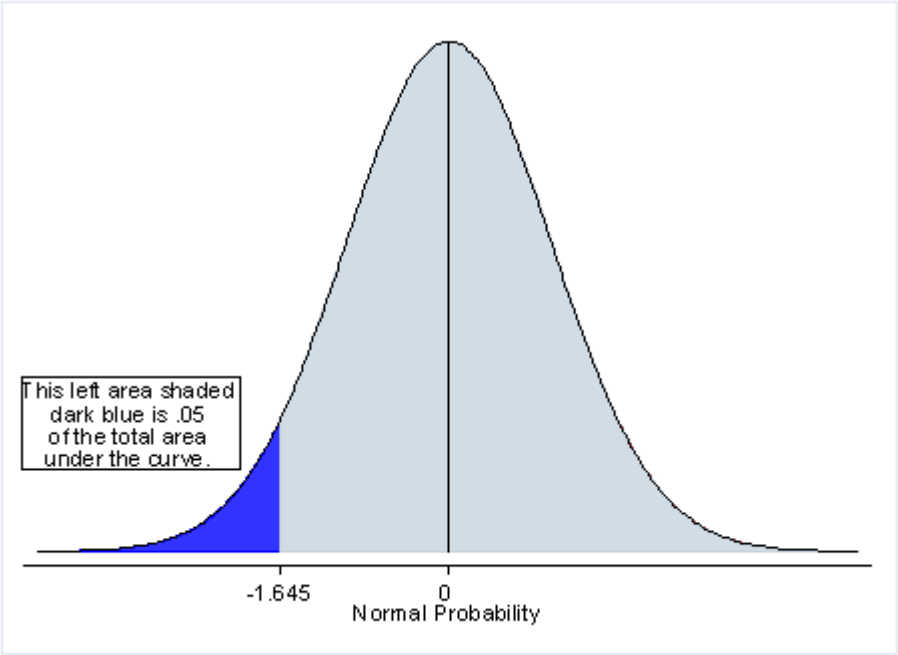


Figure 6.2: One-tailed test:

Explanatory example:

We may wish to compare the mean of a sample to a given value 'x' using a t-test.

Our null hypothesis is that the mean is equal to x.

A two-tailed test will test both if the mean is significantly greater than x and if the mean significantly less than x.

The mean is considered significantly different from x if the test statistic is in the top 2.5% or bottom 2.5% of its probability distribution, resulting in a p-value less than 0.05.

A one-tailed test will test either if the mean is significantly greater than x or if the mean is significantly less than x, but not both.

Then, depending on the chosen tail, the mean is significantly greater than or less than x if the test statistic is in the top 5% of its probability distribution or bottom 5% of its probability distribution, resulting in a p-value less than 0.05.

The one-tailed test provides more power to detect an effect in one direction by not testing the effect in the other direction.

Hypothesis testing for a proportion:

Definition:

Hypothesis testing for a proportion is used to determine if a sampled proportion is significantly different from a specified population proportion.

Explanatory example:

If you expect the proportion of male births to be 50 percent, but the actual proportion of male births is 53 percent in a sample of 1000 births. Your aim will be to test if this is significantly different from the hypothesized population parameter.

Steps to perform a hypothesis testing for a proportion:

- **Step 1:**

Formulate your research question.

- **Step 2:**

Check to see if assumptions are met.

- **Step 3:**

State the null hypothesis and the alternative hypothesis.

- **Step 4:**

Set an appropriate significance (alpha) level. By definition, the α-level is the probability of rejecting the null hypothesis when the null hypothesis is true.

- **Step 5:**

Calculating the test statistic (z):

$$z = \frac{\mathrm{P1} - \mathrm{P0}}{\sqrt{\frac{P0 * [1 - P0]}{n}}}$$

P1 = Sample proportion.

P0 = Hypothesized population proportion.

n = sample size.

- **Step 6:**

Convert the test statistic to a p value.

- **Step 7:**

Decide between null and alternative hypotheses.

- **Step 8:**

State a conclusion about the research question.

Explanatory example:

- **Step 1**

Research question:

Is the proportion of babies born male different from 50 percent?

- **Step 2:**

To test this claim, 1000 deliveries were surveyed using 'simple random sampling'. In this sample, 530 gave birth to male boys.

- **Step 3:**

H_0: p = 0.5

H_1: p > 0.5

This is a one-sided (right-tailed) test because we want to know if the proportion is greater than 0.5

- **Step 4:**

Alpha-level (α) is set at 0.05

- **Step 5:**

$$z = \frac{0.53 - 0.50}{\sqrt{\frac{0.50 * [1 - 0.50]}{1000}}}$$

$$z = \frac{0.03}{\sqrt{\frac{0.25}{1000}}}$$

$$z = \frac{0.03}{\sqrt{0.00025}}$$

$$z = \frac{0.03}{0.0158}$$

$$z = 1.894$$

- **Step 6:**

Convert the test statistic to a p value.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952

Figure 6.3:

Our z value = 1.894

According to the table (Appendix 2), P (z<1.894) = 0.9678

Therefore, P (z ≥ 1.894) = 1.0000 – 0.9678 = 0.0322

- **Step 7:**

As p value < 0.050, our decision is to reject the null hypothesis.

- **Step 8:**

Conclusion about the research question:

There is a statistical evidence to state that the proportion of babies born male is different from 50%.

Practical exercises

1. A type I error occurs when we:
    a. Reject H_0, when H_0 is actually true.
    b. Do not reject H_0, when H_0 is actually true.
    c. Reject H_0, when H_0 is actually false.
    d. Do not reject H_0, when H_0 is actually false.

2. A type II error occurs when we:
    a. Reject H_0, when H_0 is actually true.
    b. Do not reject H_0, when H_0 is actually true.
    c. Reject H_0, when H_0 is actually false.
    d. Do not reject H_0, when H_0 is actually false.

3. The probability of making a type I error is denoted by:
    a. Alpha.
    b. Beta.
    c. One minus alpha.
    d. One minus beta.

4. The most frequently used significance (alpha) level is:
    a. 0.100
    b. 0.050
    c. 0.010
    d. 0.001

5. The probability of making a type II error is denoted by:
    a. Alpha.
    b. Beta.
    c. One minus alpha.
    d. One minus beta.

6. The power of a test is:
    a. Alpha.
    b. Beta.
    c. One minus alpha.
    d. One minus beta.

7. You want to test if the mean serum bilirubin is higher in a group of viral hepatitis patients as compared to a control group. This is:
    a. One-tailed test.
    b. Two-tailed test.

8. You want to test if the mean age is higher or lower in a group of viral hepatitis patients as compared to a control group. This is:
    a. One-tailed test.
    b. Two-tailed test.

Answers:

1. A.
2. D.
3. A.
4. B.
5. B.
6. D.
7. A.
8. B.

Chapter 7
Estimation statistics

Estimation statistics

"Estimation statistics" is a way of saying that you are estimating population values based on your sample data.

This is performed by using inferential statistical methods such as the "Confidence Interval".

Confidence Interval (CI):

Definition:

A confidence interval consists of a range of values together with a percentage that specifies how confident we are that the parameter lies in the interval.

A confidence interval is an indicator of your measurement's *precision*.

It is also an indicator of how *stable* your estimate is (a measure of how close your measurement will be to the original estimate if you repeat your experiment).

Formula:

CI = mean ± margin of error

Margin of error:

Definition:

The maximum error of estimation (= the margin of error) is an indicator of the precision of an estimate and is defined as one-half the width of a confidence interval.

The 6-Steps to calculate CI:

- **Step 1:**

Write down the phenomenon you'd like to test.

- **Step 2:**

Select a sample from your chosen population.

- **Step 3:**

Calculate your sample mean and sample standard deviation.

- **Step 4:**

Choose your desired confidence level. The probability used to construct the interval is called the level of confidence or confidence coefficient.

- **Step 5:**

Calculate your margin of error.

- **Step 6:**

State your confidence interval.

Explanatory example:

- **Step 1:**

Say that the mean body weight of a male student in Mansoura University is 70 kg and the standard deviation is 10 kg (70 ± 10 kg). You'll be testing how accurately you will be able to predict the weight of male students in Mansoura University within a given confidence interval.

- **Step 2:**

You have randomly selected 400 male students.

- **Step 3:**

The sample mean was 70 kg and sample standard deviation was 10 kg.

- **Step 4:**

The most commonly used confidence levels are 90%, 95% and 99%.

Let's say you've chosen 95%.

- **Step 5:**

Calculating the margin of error:

Margin of error = Critical value * Standard error of the mean (SEM)

= (Za/2) * σ/√(n)

Critical value = (Za/2) where a = confidence level

As confidence level chosen is 95% = 0.95

So, a/2 = 0.95 / 2 = 0.475

Check out the z-table (see appendix) to find the corresponding value that goes with 0.475.

You will see that the closest value is 1.96, at the intersection of row 1.9 and the column of .06.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817

Figure 7.1: Z-table

Standard error of the mean (SEM) = σ/√(n) where σ = Standard deviation and n = sample size.

So, SEM $= 10/\sqrt{(400)}$

$= 10/20$

$= 0.5$

Therefore,

Margin of error = 1.96 * 0.5

= 0.98

- **Step 6:**

The 95% Confidence interval = 70 ± 0.98

So, the Confidence limits (lower and upper boundary values of the interval) are

Lower bound = 70 - 0.98

= 69.02

Upper bound = 70 + 0.98

= 70.98

This means that we are 95% confident that the population mean lies in the interval 69.02–70.98

Practical exercises

1. The margin of error is calculated by multiplying the critical value by:
    a. Mean.
    b. Standard deviation.
    c. Standard error of the mean (SEM).
    d. Range.

2. If the confidence level chosen is 95%, the critical value is the value in z-table that goes with:
    a. 0.05
    b. 0.95
    c. 0.475
    d. 0.095

3. Upper and lower boundaries of interval of confidence are classified as
    a. error biased limits
    b. marginal limits
    c. estimate limits
    d. confidence limits

4. If point estimate is 8 and margin of error is 5 then confidence interval is:
    a. 3 to 13
    b. 4 to 14

c. 5 to 15
d. 6 to 16

5. To develop interval estimate of any parameter of population, value which is added or subtracted from point estimate is classified as
   a. margin of efficiency
   b. margin of consistency
   c. margin of biasedness
   d. margin of error
6. for a chosen confidence level of 95%, the critical value is:
   a. 1.63
   b. 1.74
   c. 1.85
   d. 1.96
7. In a study involving 100 participants, the mean height was 160 cm and the standard deviation was 10 cm. Considering 95% confidence interval (with a critical value of 1.96), the margin of error is:
   a. 0.98
   b. 1.96
   c. 3.92
   d. 5.88
8. In a study involving 100 participants, the mean height was 160 cm and the standard deviation was 10 cm. Considering 95% confidence interval (with a critical value of 1.96), the confidence limits are:
   a. 159.02 and 160.98
   b. 158.04 and 161.96
   c. 156.08 and 163.92

d. 154.12 and 165.88

Answers:

1. C.
2. C.
3. D.
4. A.
5. D.
6. D.
7. B.
8. B.

For questions 7 and 8:

SEM = 10 / √100

= 10 /10

= 1

So, margin of error = Critical value * SEM

= 1.96 * 1

= 1.96

So, confidence limits = Mean ± Margin of error

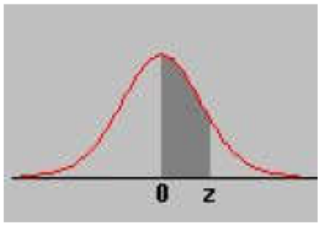
= 160 ± 1.96

= 160 – 1.96 and 160 + 1.96

= 158.04 and 161.96

Appendix A

Standard Normal (Z) Table
Area between 0 and z



	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015

1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Appendix B

Standard Normal Probabilities

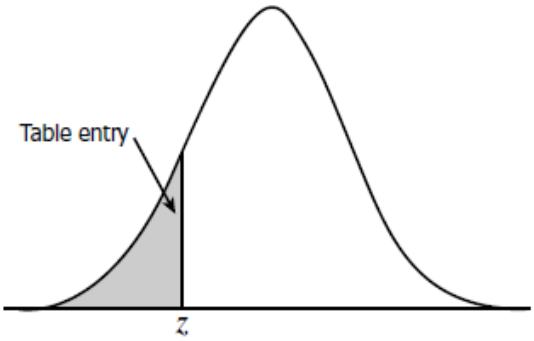


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367

−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

Standard Normal Probabilities

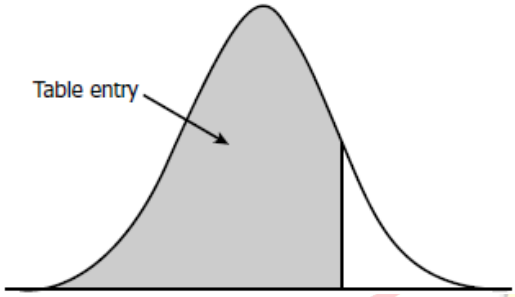


Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974

2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Appendix C

Summary Table for Statistical Techniques

	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
1	Estimating a mean	One population mean μ	Sample mean $\bar{x}$	Numerical	What is the average weight of adults? What is the average cholesterol level of adult females?	1-sample t-interval $\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$	Stat > Basic statistics > 1-sample t	Data approximately normal OR Have a large sample size ($n \geq 30$)
2	Test about a mean	One population mean μ	Sample mean $\bar{x}$	Numerical	Is the average GPA of juniors at Penn State higher than 3.0? Is the average Winter temperature in State College less than 42°F?	$H_o: \mu = \mu_o$ $H_a: \mu \neq \mu_o$ or $H_a: \mu > \mu_o$ or $H_a: \mu < \mu_o$ The one sample t test: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	Stat > Basic statistics > 1-sample t	Data approximately normal OR Have a large sample size ($n \geq 30$)
3	Estimating a proportion	One population proportion p	Sample proportion $\hat{p}$	Categorical (binary)	What is the proportion of males in the world? What is the proportion of students that smoke?	1-proportion Z-interval $\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	Stat > Basic statistics > 1-sample proportion	Have at least 5 in each category
4	Test about a proportion	One population proportion p	Sample proportion $\hat{p}$	Categorical (binary)	Is the proportion of females different from 0.5? Is the proportion of students who fail Stat200 less than 0.1?	$H_o: p = p_o$ $H_a: p \neq p_o$ or $H_a: p > p_o$ or $H_a: p < p_o$ The one proportion Z-test: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	Stat > Basic statistics > 1-sample proportion	$n p_o \geq 5$ and $n (1-p_o) \geq 5$

	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
5	Estimating the difference of two means	Difference in two population means $\mu_1-\mu_2$	Difference in two sample means $\bar{x}_1-\bar{x}_2$	Numerical	How different are the mean GPAs of males and females? How many fewer colds do vitamin C takers get, on average, than non vitamin C takers?	2-sample t-interval $(\bar{x}_1-\bar{x}_2)\pm t^* \times s.e.(\bar{x}_1-\bar{x}_2)$	Stat > Basic statistics > 2-sample t	Independent samples from the two populations Data in each sample are about normal or large samples
6	Test to compare two means	Difference in two population means $\mu_1-\mu_2$	Difference in two sample means $\bar{x}_1-\bar{x}_2$	Numerical	Do the mean pulse rates of exercisers and non-exercisers differ? Is the mean EDS score for dropouts greater than the mean EDS score for graduates?	$H_o: \mu_1=\mu_2$ $H_a: \mu_1\neq\mu_2$ or $H_a: \mu_1>\mu_2$ or $H_a: \mu_1<\mu_2$ The two sample t test: $t=\frac{(\bar{x}_1-\bar{x}_2)-0}{s.e.(\bar{x}_1-\bar{x}_2)}$ See text, page 445, for the s.e. of the difference	Stat > Basic statistics > 2-sample t	Independent samples from the two populations Data in each sample are about normal or large samples
7	Estimating a mean with paired data	Mean of paired difference μ_D	Sample mean of difference $\bar{d}$	Numerical	What is the difference in pulse rates, on the average, before and after exercise?	paired t-interval $\bar{d}\pm t_{\alpha/2}\frac{s_d}{\sqrt{n}}$	Stat > Basic statistics > Paired t	Differences approximately normal OR Have a large number of pairs ($n\geq 30$)
8	Test about a mean with paired data	Mean of paired difference μ_D	Sample mean of difference $\bar{d}$	Numerical	Is the difference in IQ of pairs of twins zero? Are the pulse rates of people higher after exercise?	$H_o: \mu_D=0$ $H_a: \mu_D\neq 0$ or $H_a: \mu_D>0$ or $H_a: \mu_D<0$ $t=\frac{\bar{d}-0}{s_d/\sqrt{n}}$	Stat > Basic statistics > Paired t	Differences approximately normal OR Have a large number of pairs ($n\geq 30$)

	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
9	Estimating the difference of two proportions	Difference in two population proportions $p_1 - p_2$	Difference in two sample proportions $\hat{p}_1 - \hat{p}_2$	Categorical (binary)	How different are the percentages of male and female smokers? How different are the percentages of upper- and lower-class binge drinkers?	two-proportions Z-interval $(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2} \times s.e.(\hat{p}_1 - \hat{p}_2)$	Stat > Basic statistics > 2 proportions	Independent samples from the two populations. Have at least 5 in each category for both populations.
10	Test to compare two proportions	Difference in two population proportions $p_1 - p_2$	Difference in two sample proportions $\hat{p}_1 - \hat{p}_2$	Categorical (binary)	Is the percentage of males with lung cancer higher than the percentage of females with lung cancer? Are the percentages of upper- and lower-class binge drinkers different?	$H_0: p_1 = p_2$ $H_a: p_1 \neq p_2$ or $H_a: p_1 > p_2$ or $H_a: p_1 < p_2$ The two proportion z test: $z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ $\hat{p} = \frac{x_1 + x_2}{n_1 + n_2}$	Stat > Basic statistics > 2 proportions	Independent samples from the two populations. Have at least 5 in each category for both populations.
11	Relationship in a 2-way table	Relationship between two categorical variables OR Difference in two or more population proportions	The observed counts in a two-way table	Categorical	Is there a relationship between smoking and lung cancer? Do the proportions of students in each class who smoke differ?	H_0: The two variables are not related H_a: The two variables are related The chi-square statistic: $\chi^2 = \sum_{\substack{all \\ cells}} \frac{(Observed - Expected)^2}{Expected}$	Stat >Tables > CrossTabulation > Chi-Square analysis	All expected counts should be greater than 1 At least 80% of the cells should have an expected count greater than 5

	Inference	Parameter	Statistic	Type of Data	Examples	Analysis	Minitab Command	Conditions
12	Test about a slope	Slope of the population regression line β_1	Sample estimate of the slope b_1	Numerical	Is there a linear relationship between height and weight of a person?	$H_o: \beta_1 = 0$ $H_a: \beta_1 \neq 0$ or $H_a: \beta_1 > 0$ or $H_a: \beta_1 < 0$ The t test with n-2 degrees of freedom: $t = \frac{b_1 - 0}{s.e.(b_1)}$	Stat > Regression > Regression	The form of the equation that links the two variables must be correct The error terms are normally distributed The errors terms have equal Variances The error terms are independent of each other
13	Test to compare several means	Population means of the *t* populations $\mu_1, \mu_2, \ldots, \mu_t$	Sample mean of the t populations $x_1, x_2, \ldots, x_t$	Numerical	Is there a difference between the mean GPA of Freshman, Sophomore, Junior and Senior classes?	$H_o: \mu_1 = \mu_2 = \cdots = \mu_t$ H_a: not all the means are equal The F test for one-way ANOVA: $F = \frac{MSTR}{MSE}$	Stat >ANOVA > Oneway	Each population is normally Distributed Independent samples from the *t* populations Equal population standard deviations
14	Test of Strength & Direction of Linear Relationship of 2 Quantitative Variables	Population Correlation ρ "rho"	Sample correlation r	Numerical	Is there a linear relationship between height and weight?	$H_o: \rho = 0$ $H_a: \rho \neq 0$ $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$	Stat > Basic Statistics > Correlation	2 variables are continuous Related pairs No significant outliers Normality of both variables Linear relationship between the variables
15	Test to Compare Two Population Variances	Two population variances σ_1^2, σ_2^2	Sample variances of 2 populations s_1^2, s_2^2	Numerical	Are the variances of length of lumber produced by Company A different from those produced by Company B	$H_o: \sigma_1^2 = \sigma_2^2$ $H_a: \sigma_1^2 \neq \sigma_2^2$ $F = s_1^2 / s_2^2$	Stat > Basic statistics > 2 variances	Each population is normally Distributed Independent samples from the 2 populations

Appendix D

	Formula	Note
Continuous outcome		
One sample	$n \geq \frac{\sigma^2}{\delta^2}(Z_\alpha + Z_\beta)^2$	δ is the detected difference σ is the population standard deviation
Two independent samples with common standard deviation	$n \geq \frac{4\sigma^2}{\delta^2}(Z_\alpha + Z_\beta)^2$	δ is the relevant difference in means σ is the population standard deviation
Two paired samples	$n \geq \frac{\sigma_d^2}{\delta_d^2}(Z_\alpha + Z_\beta)^2$	σ_d is the standard deviation of the mean difference δ_d is the mean difference
Categorical outcome		
One sample	$n \geq \frac{(Z_\alpha \sqrt{p_0(1-p_0)} + Z_\beta \sqrt{p_1(1-p_1)})^2}{(p_1 - p_0)^2}$	p_0 is the success rate under null hypothesis p_1 is the success rate under alternative hypothesis
Two independent groups	$n \geq \frac{p_C(1-p_C) + p_E(1-p_E)}{\delta^2}(Z_\alpha + Z_\beta)^2$	p_C is the success rate in control group p_E is the success rate in treatment group δ is $p_E - p_C$
Two paired sample	$n \geq \frac{(Z_\alpha + Z_\beta)^2 f}{d^2}$	f is the proportion of discordant pairs d is the proportion difference

Where α is type I error level and β is the type II error level.

Reference

1- Kaur P, Stoltzfus J, Yellapu V. (2018). Descriptive statistics. Int J Acad Med 2018;4:60-3

2- Freud RJ, Wilson WJ and Mohr DL (2010). Statistical Methods (Third Edition). AMSTERDAM, BOSTON, HEIDELBERG, LONDON NEW YORK, OXFORD, PARIS, SAN DIEGO SAN FRANCISCO, SINGAPORE, SYDNEY, TOKYO. Elsevier Inc.

3- Yan, F., Robert, M., & Li, Y. (2017). Statistical methods and common problems in medical or biomedical science research. International Journal of Physiology, Pathophysiology and Pharmacology, 9(5), 157–163.

4- Hollander, M. (2014). Nonparametric statistical methods (Third edition.). Hoboken, New Jersey: John Wiley & Sons, Inc.

5- SPINA, D. (2011). Statistical methods in research. Methods Mol Biol. 2011; 746: 443-472.

6- Agresti, A. (1984). Analysis of ordinal categorical data. New York: Wiley.

7- Bancroft, T. A. (1968). Topics in intermediate statistical methods. Ames, IA: Iowa State University Press.

8- Bancroft, T. A. (1968). Topics in intermediate statistical methods. Ames, IA: Iowa State University Press.

9- https://onlinecourses.science.psu.edu/stat500/sites/onlinecourses.science.psu.edu.stat500/files/lesson14/summary_table/index.pdf. Accessed online on 07-10-2018.

10- Bhaskar, S. B., & Manjuladevi, M. (2016). Methodology for research II. Indian Journal of Anaesthesia, 60(9), 646-651. http://doi.org/10.4103/0019-5049.190620

11- Ali Z, Bhaskar SB. (2016). Basic statistical tools in research and data analysis. Indian J Anaesth. 2016;60:662-9.

Book Coordinator ; Mostafa Fathallah

General Directorate of Technical Education for Health

https://www.facebook.com/mostafa.fathallah.3