Research Methods and Statistics

Muhammad Adeel Javaid

Preface

I wrote this book because there is a large gap between the elementary statistics course that most people take and the more advanced research methods courses taken by graduate and upper-division students so they can carry out research projects. These advanced courses include difficult topics such as regression, forecasting, structural equations, survival analysis, and categorical data, often analyzed using sophisticated likelihood-based and even Bayesian methods. However, these advanced courses typically devote little time to helping students understand the fundamental assumptions and machinery behind these methods. Instead, they teach the material like witchcraft: Do this, do that, and voilà—Statistics! Students thus have little idea what they are doing and why they are doing it. Like trained parrots, they learn how to recite statistical jargon mindlessly. The goal of this book is to make statistics less like witchcraft, to treat students like intelligent humans and not like trained parrots—thus the title, *Research Methods and Statistics*.

This book will cause students and researchers to think differently about things, not only about math and statistics, but also about research, the scientific method, and life in general. It will teach them how to do good modeling—and hence good statistics—from a standpoint of *deep* knowledge rather than *rote* knowledge. It will also provide them with tools to think critically about the claims they see in the popular press, and to design their own studies to avoid common errors.

This book is not a "cookbook." Cookbooks tell you all about the *what* but nothing about the *why*. With computers, software and the Internet readily available, it is easier than ever for students to lose track of the *why* and focus on the *what* instead. This book takes exactly the opposite approach. It will empower students and researchers to use advanced statistical methods with confidence.

Contents

Research your idea. See if there's a demand. A lot of people have great ideas, but they don't know if there's a need for it. You also have to research your competition.

1

THE NATURE OF DATA

Anything that can be counted or measured is called a variable. Knowledge of the different types of variables, and the way they are measured, play a crucial part in choice of coding and data collection. The measurement of variables can be categorized as categorical (nominal or ordinal scales) or continuous (interval or ratio scales).

Categorical measures can be used to identify change in a variable, however, should you wish to measure the magnitude of the change you should use a continuous measure.

A *nominal scale* allows for the classification of objects, individual and responses based on a common characteristic or shared property. A variable measured on the nominal scale may have one, two or more sub-categories depending on the degree of variation in the coding. Any number attached to a nominal classification is merely a label, and no ordering is implied: social worker, nurse, electrician, physicist, politician, teacher, plumber, etc.

An *ordinal scale* not only categorizes objects, individuals and responses into sub-categories on the basis of a common characteristic it also ranks them in descending order of magnitude. Any number attached to an ordinal classification is ordered, but the intervals between may not be constant: GCSE, A-level, diploma, degree, postgraduate diploma, higher degree, and doctorate.

The *interval scale* has the properties of the ordinal scale and, in addition, has a commencement and termination point, and uses a scale of equally spaced intervals in relation to the range of the variable. The number of intervals between the commencement and termination points is arbitrary and varies from one scale to another. In measuring an attitude using the Likert scale, the intervals may mean the same up and down the scale of 1 to 5 but multiplication is not meaningful: a rating of ‘4’ is not twice as ‘favourable’ as a rating of ‘2’.

In addition to having all the properties of the nominal, ordinal and interval scales, the *ratio scale* has a zero point. The ratio scale is an absolute measure allowing multiplication to be meaningful. The numerical values are ‘real numbers’ with which you can conduct mathematical procedures: a man aged 30 years is half the age of a woman of 60 years.

Categorical			Continuous
Unitary	***Dichotomous***	***Polytomous***	***Interval or Ratio Scale***
Name Occupation Location Site	[1] . . . Yes [0] . . . No [1] . . . Good [0] . . . Bad [1] . . . Female [0] . . . Male [1] . . . Right [0] . . . Wrong [1] . . . Extrovert [0] . . . Introvert [1] . . . Psychotic [0] . . . Neurotic [1] . . . Assertive [0] . . . Passive [1] . . . Present [0] . . . Absent	Attitudes (Likert Scale): [5] . . . strongly agree [4] . . . agree [3] . . . uncertain [2] . . . disagree [1] . . . strongly disagree Age: [4] . . . Old [3] . . . Middle-aged [2] . . . Young [1] . . . Child Income: [3] . . . High [2] . . . Medium [1] . . . Low Socio-Economic Status: [5] . . . A [4] . . . B [3] . . . C1 [2] . . . C2 [1] . . . D [0] . . . E	Income (£000s per annum) Age (in years) Reaction Time (in seconds) Absence (in days) Distance (in kilometres) Length (metres) Attitude (Thurstone & Cheve)

Qualitative	Quantitative
Sex (Male/Female) Age (Old/Young) Attitude (Favourable/Unfavourable) Attitude (Likert scale) Achieved Educational Level (High/Low) Style (Autocratic/Participative) Location (Urban/Rural) Performance (Good/Bad)	Age (in years) Attitude (Guttman scale) Attitude (Thurstone & Cheve scale) Performance (errors or faults per minute) Achieved Educational Level (number of years post-secondary school education)

Table I
A Two-Way Classification of Variables

Methods of Data Collection

The major approaches to gathering data about a phenomenon are from primary sources: directly from subjects by means of experiment or observation, from informants by means of interview, or from respondents by questionnaire and survey instruments. Data may also be obtained from secondary sources: information that is readily available but not necessarily directly related to the phenomenon under study. Examples of secondary sources include published academic articles, government statistics, an organization's archival records to collect data on activities, personnel records to obtain data on age, sex, qualification, length of service, and absence records of workers, etc. Data collected and analyzed from published articles, research papers and journals *may* be a primary source if the material is directly relevant to your study. For instance, primary sources for a study conducted using the Job Descriptive Index may be Hulin and Smith (1964-68) and Jackson (1986-90), whereas a study using an idiosyncratic study population, technique and assumptions, such as those published by Herzberg, et alia (1954-59), would be a secondary source.

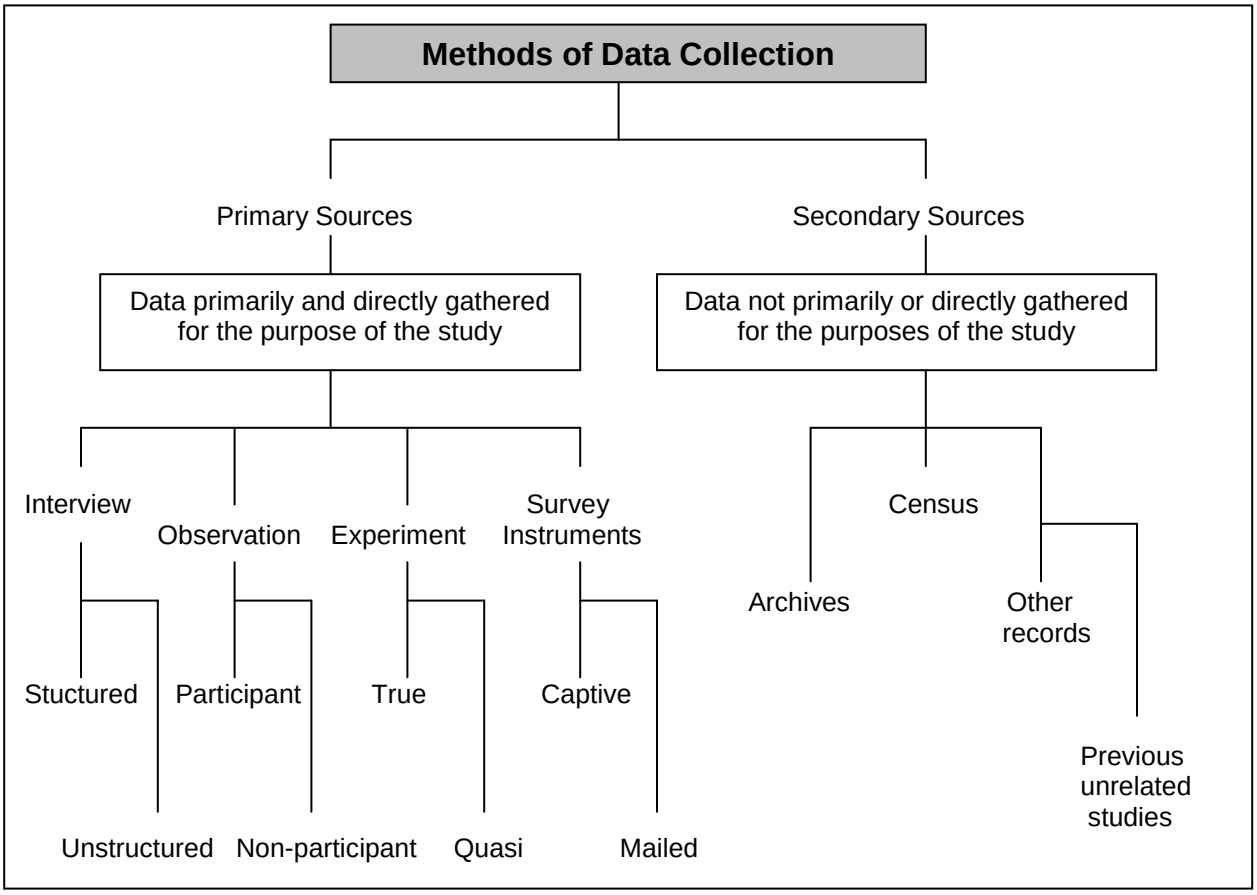


Figure 1
A Classification of Methods of Data Collection
Procedures for Coding Data

A coding frame is simply a set of instructions for transforming data into codes and for identifying the location of all the variable measured by the test or instrument. Primary data gathered from subjects and informants is amenable to control during the data collection phase. The implication is that highly structured data, usually derived from tests, questionnaires and interviews, is produced directly by means of a calibrated instrument or is readily produced from raw scores according to established rules and conventions. Generally, measures such as physical characteristics such as height and weight are measured on the ratio scale. Whereas psychological attributes such as measures of attitude and standard dimensions of personality are often based on questions to which there is no appropriate response. However, the sum of the responses is interpreted according to a set of rules and provides a numerical score on the interval scale but is often treated as though the measures relate the ratio scale. Norms are available for standard tests of physical and psychological attributes to establish the meaning of individual scores in terms of those derived from the general population. A questionnaire aimed at determining scores as a measure of a psychological attribute are said to be pre-coded; that is, the data reflects the coder's prior structuring of the population. The advantages of pre-coding are that it reduces time, cost and coding error in data handling. Ideally, the pre-coding should be sufficiently robust and discriminating as to allow data processing by computer.

A coding frame should include information for the variable to be measured:

- the source data (e.g. Question 7 or 'achieved educational level');
- a list of the codes (e.g. number of years post-secondary school education);
- column location of the variable on the coded matrix.

Once we have a coding frame, data relating to an individual or case can be read off just as we would read the data in a sales catalogue or a coded matrix from a holiday brochure:

Destination *Paphos*	**Number of days**	**Class of hotel**	**Number of rooms**	**Full/half board**	**Car hire Included**	**Price £**
Athenaeum	10	***	45	F	N	450
Polis	7	**	23	H	N	299
Roman Court	5	*****	225	F	Y	999
Aphrodite	7	**	30	F	Y	520

The following statements may or may not be true for your work organization. For each item below, please answer by ticking (/) the appropriate box below the categories at the top of the page. If you cannot decide then tick the box which is closest to your answer.

		Definitely True	More True than False	More False than True	Definitely False
1.	I feel that I am my own boss in most matters	☐	☐	☐	☐
2.	The company is generally quick to use improved work methods	☐	☐	☐	☐
3.	The company has a real interest in the welfare and happiness of those who work here	☐	☐	☐	☐
4.	The company tries to improve working conditions	☐	☐	☐	☐
5.	A person can makes his or her own decisions around here without checking with anybody else	☐	☐	☐	☐
6.	Everyone has a specific job to do	☐	☐	☐	☐
7.	The company has clear-cut reasonable goals and objectives	☐	☐	☐	☐
8.	Most people here make their own rules on the job	☐	☐	☐	☐
9.	Work activities are sensibly organized	☐	☐	☐	☐
10.	I feel that those above me are very receptive to my ideas and suggestions	☐	☐	☐	☐

Case #	Age	Sex	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
101	18	1	1	3	1	2	2	2	3	1	1	3
102	27	1	5	3	5	1	2	1	2	1	5	3
103	54	1	2	3	1	1	2	1	2	1	4	3
104	23	1	2	4	1	1	2	1	2	1	4	3
105	19	0	3	3	4	1	2	2	3	2	2	4
106	34	1	4	2	3	3	3	1	2	2	3	5
107	51	1	4	3	3	1	3	3	3	1	3	3
108	25	0	5	5	1	3	3	3	2	3	3	2
109	25	0	5	3	2	3	2	3	3	3	4	3
110	31	0	2	3	2	1	2	3	2	1	4	4
111	21	1	1	4	5	1	2	2	3	2	4	3
112	44	1	1	4	3	2	2	2	3	3	4	2

Table 1

Survey Instrument and Coding Frame Matrix for Research Data

If you choose to use an unstructured interview schedule or questionnaire, containing open-ended questions, it is necessary to transform the responses from these questions into a mode that allows for efficient analysis. It may be possible to conduct a pilot study to establish the kind of responses expected from the sample of respondents, thus deriving a code list that may allow for some degree of pre-coding. For small samples, say less than thirty people, the analysis may take the form of listening to the voice tape, or scanning the written responses, to gain an overall impression of the categories of response. Considering key statements relating to the phenomenon under study (see Table 2 below) can then reduce the data from the responses. This technique can be used during the piloting the interview or questionnaire, and even if we decide not to pre-code the responses, we have a useful set of anticipated responses that can be used for training interviewers and be extended, if need be, by adding further categories from the survey.

Key Responses to Statement: *What do you particularly like about your job?*	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1. Convenience	X		X	X		X		X
2. Responsibility		X			X		X	
3. Making decisions		X			X		X	
4. Friendships						X		X
5. Use of abilities and talents		X			X		X	
6. Working conditions	X		X					X
7. Opportunities for promotion								
8. Pay			X					X
9. Interesting work		X			X		X	
10. Job security	X		X					X

Table 2
Response Matrix Representing Categories of Interest

The design of a coding frame is also determined by the approach we take in respect of the data: what the data signifies, and useful ways of understanding the data once collected. After Swift (1996), three approaches can be identified:

1. Representational Approach
   The response of the informant is said to express the surface meaning of what is “out there” requiring the researcher to apply codes to reduce the data, whilst at the same time, reflecting this meaning as faithfully as possible. At this stage of the process, the data must be treated independently from any views the researcher may hold about underlying variables and meanings.

2. Anchored-in Approach

The researcher may view the responses as having additional and implicit meanings that come from the fact that the responses are dependent on the data-gathering context. For example, in investigating worker involvement, we might want to conduct this with a framework comprising of types of formal and informal worker/manager interactions. As a consequence, the words given by informants can be interpreted to produce codes on more than one dimension relating to the context: (a) nature of the contact: formal versus informal, intermittent versus continuous contact, etc. (b) initiator of contact: worker versus manager. The coding frame using this approach takes into account "facts" as being anchored to the situation, rather than treating the data as though they are context-free.

3. Hypothesis-Guided Approach

Although similar to the second approach, we may view the data as having multiple meanings according the paradigm or theoretical perspective from which they are approached (e.g. phenomenological or hermeneutic approach to investigating a human or social phenomenon). The hypothesis-guided approach recognizes that the data do not have just one meaning which refers to some reality approachable by analysis for the surface meaning of the words: words have multiple meanings, and "out there" is a multiverse rather than a universe. In the hypothesis-guided approach, the researcher might use the data, and other materials, to create or investigate variables that are defined in terms of the theoretical perspective and construct propositions. For example, a data set might contain data on illness and minor complaints that informants had experienced over a period of say, one year. Taking the hypothesis-guided approach, the illness data might be used as an indicator of occupational stress or of a reaction to transformational change. Hence, the coding frame is based on the researcher'' views and hypotheses rather than on the surface meanings of the responses.[1]

[1] In the case of anchored and hypothesis-guided coding frames, there may well be categories that are not represented in small samples of data. Indeed, this is very likely for the hypothesis-guided coded frame; that is, you may want the frame to be able to reflect important concept, if only by showing that they did *not* occur among the responses.

II

ANALYSIS OF INDIVIDUAL OBSERVATIONS

In the analysis of individual observations, or ungrouped data, consideration will be given to all levels of measurement to determine which descriptive measures can be used, and under what conditions each is appropriate.

One of the most widely used descriptive measures is the ‘average’. One speaks of the ‘average age’, average response time’, or ‘average score’ often without being very specific as to precisely what this means. The use of the average is an attempt to find a single figure to describe or represent a set of data. Since there are several kinds of 'average', or measures of central tendency, used in statistics, the use of precise terminology is important: each ‘average’ must be clearly defined and labelled to avoid confusion and ambiguity. At least three kinds of common uses of the ‘average’ can be described:

1. An average provides a summary of the data. It represents an attempt to find one figure that tells more about the characteristics of the distribution of data than any other. For example, in a survey of several hundred undergraduates the average intelligence quotient was 105: this one figure summarizes the characteristic of intelligence.

2. The average provides a common denominator for comparing sets of data. For example, the average score on the Job Descriptive Index for British managers was found to be 144, this score provides a quick and easy comparison of levels of felt job satisfaction with other occupational groups.

3. The average can provide a measure of typical size. For example, the scores derived for a range of dimensions of personality can be compared to the norms for the group the sample was taken from; thus, one can determine the extent to which the score for each dimension is above, or below, that to be expected.

The Mode

The mode can be defined as the most frequently occurring value in a set of data; it may be viewed as a single value that is most representative of all the values or observation in the distribution of the variable under study. It is the only measure of central tendency that can be appropriately used to describe nominal data. However, a mode may not exist, and even if it does, it may not be unique:

1 2 3 4 5 6 7 8 9 10	. . .	. . .	. . .	. . .	No mode
Y Y N Y N N N N Y	. . .	. . .	. . .	. . .	Unimodal (N)
1 2 2 3 4 4 4 4 5 5	. . .	. . .	. . .	. . .	Unimodal (4)
1 2 2 2 3 4 5 5 5 6	. . .	. . .	. . .	. . .	Bimodal (2, 5)
1 2 2 3 4 4 5 6 6 7	. . .	. . .	. . .	. . .	Multimodal (2, 4, 6)

With relatively few observations, the mode can be determined by assembling the set of data into an array. Large numbers of observations can be arrayed by means of Microsoft EXCEL, or other statistical software programs:

Subject	Reaction Time (in m/seconds)	Array
000123	625	460
000125	500	480
000126	480	500
000128	500	500
000129	460	500
000131	500	**500** Mode
000134	575	510
000137	530	525
000142	525	530
000144	500	575
000145	510	625

The Median

When a measurement of a set of observation is at least ordinal in nature, the observations can be ranked, or sorted, into an array whereby the values are arranged in order of magnitude with each value retaining its original identity. The median can be defined as the value of the middle item of a set of items that form an array in ascending or descending order of rank: the [N+1}/2 position. In

simple terms, the median splits the data into two equal parts, allowing us to state that half of the subjects scored below the median value and half the subjects scored above the median value. If an observed value occurs more than once, it is listed separately each time it occurs:

Subject	Reaction Time in m/secs	Reaction Time shown in array: shortest to longest	Reaction Time shown in array: longest to shortest
000123	625	460	625
000125	500	480	575
000126	480	500	530
000128	500	500	525
000129	460	500	510
000131	500	500	**500** Median
000134	575	510	500
000137	530	525	500
000142	525	530	500
000144	500	575	480
000145	510	625	460

The Arithmetic Mean

Averages much more sophisticated than the mode or median can be used at the interval and ratio level. The arithmetic mean is widely used because it is the most commonly known, easily understood and, in statistics, the most useful measure of central tendency. The arithmetic mean is usually given the notion μ and can be computed:

$$\mu = [x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + \ldots x_n] / N$$

where, x_1, x_2, x_3 . . . x_n are the values attached to the observations; and, N is the total number of observations:

Subject	000123 x_1	000125 x_2	000126 x_3	000128 x_3	000129 x_4	000131 x_5	000134 x_6	000137 x_7	000142 x_8
Reaction Time (m/secs)	625	500	480	500	460	500	575	530	525

Using the above formula, the arithmetic mean can be computed:

$$\mu = \Sigma (x)/N = 4695/8 = 586.875 \text{ m/secs.}$$

The fact that the arithmetic mean can be readily computed does not mean that it is meaningful or even useful. Furthermore, the arithmetic mean has the weakness of being unduly influenced by small, or unusually large, values in a data set. For example: five subjects are observed in an experiment and display the following reaction times: 120, 57, 155, 210 and 2750 m/secs. The arithmetic mean is 658.4 m/secs, a figure that is hardly typical of the distribution of reaction times.

Measures of Dispersion

The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data. Various measures of dispersion or variation are available such as range, mean deviation, semi-interquartile range, and the standard deviation.

The Mean Absolute Deviation

The mean absolute deviation of a data set of N numbers, x_1, x_2, x_3 . . . x_n is defined by:

$$\text{Mean Deviation} = \Sigma[x - \mu]/N$$

Where μ is the arithmetic mean of the data set and $[x-\mu]$ is the absolute value of the deviation of x from μ (note that in mathematics, the absolute value of a number is the number without the associated sign and is indicated by two vertical lines: |-4| = 4, |+4| = 4).

Compute, and make observation on, the mean absolute deviation for the two sets of data derived from sample (a) and sample (b):

(a) 2, 3, 6, 8, 11 (b) 2, 4, 7, 8, 9

(a)	Arithmetic Mean	=	[2+3+6+8+11]/5	=	6
	Mean Absolute Deviation	=	[|2-6|+|3-6|+|6-6|+|8-6|+|11-6|]/5	=	2.8
(b)	Arithmetic Mean	=	[2+4+7+8+9]/5	=	6
	Mean Absolute Deviation	=	[|2-6|+|4-6|+|7-6|+|8-6|+|9-6|]/5	=	2.4

Conclusion

The mean absolute deviation indicates that sample (b) shows less dispersion than sample (a).

Variance and Standard Deviation

The most useful measures of dispersion, and those with the most desirable mathematical properties, are *variance* and *standard deviation*. The standard deviation is particularly useful when dealing with the *Normal Distribution*, since any Normal Distribution can be completely determined when the arithmetic mean and standard deviation are known. When it is necessary to distinguish the standard deviation of a population from the standard deviation of a sample drawn from this population, we often use the symbol σ for the former and *s* for the latter. For ungrouped data the variance and standard deviation from a sample can be determined from the following derived formulae:

$$(1)\ s^2 = \Sigma[x - \mu]^2/N \qquad (2)\ s^2 = [\Sigma x^2/N] - \mu^2$$

where, s^2 = variance; μ = the arithmetic mean; $\Sigma[x - \mu]^2$ = the squared difference between the observed value and the arithmetic mean; and N = the number of observations. To compute the variance or standard deviation of a sample from a population, the value for N should be substituted by [N - 1]; however, with large samples, say >50, the difference is so small as to be of no significance. The standard deviation is the *square root of the variance*. Formula (2) above is often easier to compute for manual operations:

Find the standard deviation of each set of data: (a) 2, 3. 6, 8, 11 and (b) 2, 4, 7, 8, 11.

(a)	Arithmetic Mean	=	$[2+3+6+8+11]/5$	=	6
	Variance	=	$[(2^2+3^2+6^2+8^2+11^2)/5] - 6^2$	=	10.8
	hence, the standard deviation	=	$10.8^{0.5}$	=	<u>3.28634</u>
(b)	Arithmetic Mean	=	$[2+4+7+8+9]/5$	=	6
	Variance	=	$[(2^2+4^2+7^2+8^2+9^2)/5] - 6^2$	=	6.8
	hence, the standard deviation	=	$6.8^{0.5}$	=	<u>2.60768</u>

The above results should be compared with those of the mean absolute deviation computed previously. It will be noted that the standard deviation does indicate that the data set (b) shows less dispersion than data set (a). However, the effect is masked by the fact that extreme values affect the standard deviation much more than the mean absolute deviation. This of course is to be expected since the deviations are squared in computing the standard deviation.

For *grouped data* the arithmetic mean is computed as: $\mu = \Sigma fm/\Sigma f$ and the variance: $s^2 = \Sigma f[m - \mu]^2/\Sigma f$ where *m* is the class midpoint.

The properties of the standard deviation for approximately normal, or moderately skewed distributions, are:

1. 68.27 percent of observations (x) are included between the value of the arithmetic mean and +/- one standard deviation

2. 95.45 percent of observations (x) are included between the value of the arithmetic mean and +/- two standard deviations; and,

3. 99.73 per cent of observations (x) are included between value of the arithmetic mean and +/- three standard deviations.

For moderately skewed distribution the above percentages may hold approximately. These powerful properties are of pivotal importance to probability and hypothesis testing in applying the Normal deviate, or its derivatives. Furthermore, such measures may be said to represent the distribution for which they were calculated. For instance, in the validation of learning gain acquired by two similar training programmes the respective properties are:

Traditional Programme	Computer Based Programme
$\mu_1 = 63.500$	$\mu_2 = 75.950$
$\sigma_1 = 15.220$	$\sigma_2 = 11.400$
$n_1 = 60$	$n_2 = 40$

Whilst the traditional training programme shows a lower mean score than that of the computer-based programme, how can we compare the relative *variability* bearing in mind that one shows a lower mean than the other?

The *Pearson Coefficient of Variability* can assist in such instances, where *s* is the standard deviation and μ is the arithmetic mean of the sample:

$$\text{Variability } (\varpi) = 100\ s/\mu$$

Hence, the traditional training programme has a variability (ϖ) of 23.97 per cent whereas the computer-based training programme has a variability of 15.01 per cent. We can conclude that not only is the traditional training programme more

variable in an absolute sense, but the variability of the observed values expressed as a percentage are also greater.

Empirical Relations between Measures of Dispersion

The consequences for the normal distribution are that the mean deviation and the semi-interquartile range (or median) are equal respectively to 0.7979 and 0.6745 times the standard deviation. For moderately skewed distributions we have the empirical formulae:

Mean Deviation	=	0.7979(standard deviation)
Median (Q)	=	0.6745(standard deviation)

Example for Computation of Mean and Standard Deviation for Ungrouped Data

Scores of Ten Sampled Subjects as at 31 March 20XX

Name	Score (x)	$[x - \mu]$	$[x - \mu]^2$	x^2
Abrams	22.5	16.00	256.00	506.25
Brown	56.0	17.50	306.25	3136.00
Curtis	42.3	3.80	14.44	1831.84
Cuthbert	20.8	17.70	313.29	432.64
Dunning	33.4	5.10	26.01	1115.56
Easton	48.2	9.70	94.09	2323.24
Fallowes	50.0	11.50	132.25	2500.00
Gorman	37.4	1.10	1.21	1398.76
Graham	41.0	2.50	6.25	168.10
Harris	33.4	5.10	26.01	1115.56
Σ =	*385.0*		*1175.80*	*15998.30*

Arithmetic Mean (μ) $= \Sigma x/N = 385/10 =$ <u>38.50</u>

Variance (s^2) $= \Sigma[x - \mu]^2/N-1 = 1175.80/9 = 130.6444$

hence, $s = 130.6444^{0.5} =$ <u>11.43</u>

Note: Since the set of data is a sample drawn from a population, the number of observation (N) has been adjusted to N-1.

Standardized Variables and Standard Scores

The variable $z = [x - \mu]/s$

Which measures the deviation from the mean in units of the standard deviation is called a *standardized variable* and is a dimensionless quantity, that is, independent of the units used. If deviations from the mean are given in units of the standard deviation, they are said to be expressed in *standard scores* or *standard units*. These are of great value in comparison of distributions and are often used in aptitude and educational testing.

In a test of aptitude, a subject received a score of 84 in a numerical test for which a mean score was 76 and the standard deviation was 10. In the test for verbal reasoning for which the mean score was 82 and the standard deviation was 16, the subject received a score of 90. In which test was the subject's relative standing higher?

Computation

The standardized variable $z = [x-\mu]/s$ measures the deviation of x from the mean μ in terms of standard deviation *s*.

For the numerical test, $z = [84-76]/10 = 0.8$
For the verbal reasoning test, $z = [90-82]/16 = 0.5$

Thus, the subject had a score 0.8 of a standard deviation above the mean for the numerical test, but only 0.5 of a standard deviation above the mean in verbal reasoning. Thus, the subject's standing was higher in numerical aptitude.

Miscellaneous Averages

Weighted Arithmetic Mean

It is not always the case that values of a variable (x) are of equal importance, for example in the assessment of a learning event. To compute the arithmetic mean certain weights (w) may be attached to the variable:

$$\mu = [w_1x_1 + w_2x_2 + w_3x_3 + \ldots w_nx_n] / [w_1 + w_2 + w_3 + \ldots w_n]$$

$$= \Sigma w_j x_j [1/\Sigma x]$$

Assessment	Score (x)	Weight (w)	wx
Test of knowledge I	65	0.20	13
Practical assignment I	70	0.30	21
Test of knowledge II	60	0.30	18
Practical assignment II	60	0.20	12
Total:		1.00	64

Geometric Mean

It is possible to average observations of data on the ratio scale by means of the geometric mean (*G*). This is particularly useful in the computation of the average increase in index numbers or percentiles. The geometric mean can be defined as the n^{th} root of the product of n values of a variable (x):

$$G = [x_1 * x_2 * x_3 * \ldots x_n]^{1/n}$$

For large values of X, the geometric mean can be most easily computed by the application of logarithms_{10}

$$\text{Log}_{10}G = \Sigma[\log_{10}x]1/n$$

Period end year	Number of subjects	Percentage increase over previous period (x)	$\text{Log}_{10}x$
1950	138	-	0.000000
1960	292	211.594	2.324952
1970	385	131.849	2.120078
1980	596	154.805	2.189786
1990	938	157.383	2.196957
2000	1233	131.450	2.118760
Total			10.950533

hence,

$$\text{Log}_{10}G = 10.950533[1/5] = 2.1901066$$

$$G = 154.9196831 \text{ say, 155 percent increase over each ten year period.}$$

Alternatively, a rough and ready estimate can be derived from the following:

$$G = [\text{no. of subjects at end period /no. of subjects at beginning period}]^{1/n}$$

$$= [1233/138]^{1/5}$$

$$= 154.9589975 \text{ percent}$$

But this increase is for each period of ten years, to obtain the annual rate of increase over the total of fifty years:

$$G = [1233/138]^{1/50} - 1$$

$$= 1.044772373 - 1$$

$$= 4.47724 \text{ percent a year}$$

The nature of the geometric mean is that if any value for x is zero, then the product will be zero. It is also meaningless of any of the values are negative. This can be overcome by expressing the values of x with care, and selecting the values of x to be averaged.

Harmonic Mean

The harmonic mean (*H*) is useful for solving problems that involve variables expressed in time, such as reaction times, number of errors per hour, etc. The harmonic mean can be computed:

$$H = N/[\Sigma 1/x]$$

Subject (n)	**Minutes to complete task (x)**	**Number of tasks completed per hour**
104/64	2.5	24
104/74	2.0	30
104/77	1.5	40
104/85	6.0	10
Total:	12.0	104

The arithmetic mean would be 3 minutes to complete the task (12 minutes divided by 4 subjects), which on average is 20 tasks per hour/subject or 80 tasks

per hour for all four subjects. However, when computed separately for each subject, the total number of tasks completed per hour is 104. The total number of tasks completed per hour for all four subjects must be computed some other way:

$$H = N/[1/x_1 + 1/x_2 + 1/x_3 + 1/x_4 \ldots 1/x_n] = 4/[0.4+0.5+0.67+0.17]$$

$$= 2.31 \text{ minutes}$$

Since the four subjects participated in the experiment for one hour, representing 240 minutes of running time, the average number of tasks completed by the group of four subjects is:

240 minutes/2.31 minutes per task = 104 tasks per hour

III
Descriptive Statistics

PART A: DESCRIPTIVE STATISTICS

Data

Data are very important for scientific study, and statistics is a discipline that deals with the collection, presentation and analysis of data. In this chapter we are going to study how we can summarize and describe a set of data. When we study a set of data we need to identify the following important characteristics of the dataset.

- Primary and secondary data. When the data are collected by us it is called primary data. We always have the individual values of the data. When the dataset is collected by others, it is called secondary data. Sometimes the data is grouped into a table, and is called grouped data.

- Population and sample data. Population refers to the totality of elements in which we are interested. Suppose we want to study the salary of Hong Kong people, our population includes all those persons who work in Hong Kong. However as the population is so big that it is not practical and economical to collect salary data of all the working people, we always select randomly only a subset of the population and the data is sample.

- Discrete and continuous data. It is important to identify whether the data is continuous or discrete. For example data on the number of persons in a household is discrete, and data on salary is continuous. Different statistical techniques are used for handling discrete or continuous data.

Frequency Distribution

Statistical data obtained by means of census, sample surveys or experiments usually consist of raw, unorganized sets of numerical values. Before these data can be used as a basis for inferences about the phenomenon under investigation or as a basis for decision, they must be summarized and the pertinent information must be extracted.

Example 1

A random sample of 100 households in a town was selected and their monthly town gas consumption (in cubic metres) in last month were recorded as follows:

55	82	83	109	78	87	95	94	85	67
80	109	83	89	91	104	90	103	67	52
107	78	86	29	72	66	92	99	60	75
88	112	97	88	49	62	70	66	88	62
72	85	81	78	77	41	105	92	94	74
78	75	87	83	71	99	56	69	78	60
119	39	104	86	67	79	98	102	82	91
46	120	73	125	132	86	48	55	112	28
42	24	130	100	46	57	31	129	137	59
102	51	135	53	105	110	107	46	108	117

A useful method for summarizing a set of data is the construction of a frequency table, or a frequency distribution. That is, we divide the overall range of values into a number of classes and count the number of observations that fall into each of these classes or intervals.

The general rules for constructing a frequency distribution are:

(i) There should not be too few or too many classes.

(ii) Insofar as possible, equal class intervals are preferred. But the first and last classes can be open-ended to cater for extreme values.

In example 1, the sample size is 100 and the range for the data is 113 (137 - 24). A frequency distribution with six classes is appropriate and it is shown below.

Frequency distribution of household town gas consumption

Town gas monthly consumption (in cubic metres)	Number of households
20 - 39	5
40 - 59	15
60 - 79	25
80 - 99	30
100 - 119	18
120 - 139	7
Total	100

Class limits: are the numbers that typically serve to identify the classes in a listing of a frequency distribution. Thus, in the above frequency distribution, for the class whose frequency is 30, its lower class limit is 80 and upper class limit is 99.

As contrasted to a class limit, a **class boundary** is the precise point that separates one class from another, rather than being a value indicated in one of the classes. A class boundary is typically located midway between the upper limit of a class and the lower limit of the next higher class adjoining it. Therefore the class boundary separating the class 60-79 and the class 80-99 is halfway between 79 and 80, that is, at the point 79.5.

Class interval: is the width of a class. The class interval of a class is computed by subtracting the lower limit (boundary) of the class from the lower limit (boundary) of the next class.

Class midpoint or class mark: is the point dividing the class into equal halves on the basis of class interval. This point can be obtained by adding the lower and upper limits (boundaries) of a class and dividing by 2.

Relative frequency of a class: is the frequency of the class divided by the total frequency of the distribution.

Cumulative frequency distribution: shows the number of items of a series that are less than (or more than) certain specified values.

Measure of Central Tendency

A value that would describe the 'centre' of a distribution would be visually located near the spot where most of the data seem to be concentrated. Consequently, values that fulfil this role are called measures of central tendency.

The most common measures of the central tendency of a data set are arithmetic mean or simply as mean, median and mode.

The mean of a set of numerical data is the sum of the set divided by the number of observations, that is, their average.

The median of a distribution is the value which divides the distribution so that an equal number of values lie on either side of it, i.e., half of the items have values smaller or equal to it and half of the items have values larger or equal to it.

The mode of a set of numerical data is the value which occurs most frequently.

Example 1 (calculating mean, median and mode for individual data)

The following table shows the hourly wage rates of eight sampled construction workers.

Worker i	1	2	3	4	5	6	7	8
Hourly wage rate (x_i)	\$35	38	46	60	65	69	72	78

Mean $(\bar{x}) = \frac{\sum_{i=1}^{8} x_i}{8} \quad (= \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8}{8})$

$$= \frac{463}{8} = 57.875 (\$)$$

Location of the median: $\frac{n+1}{2} = \frac{9}{2} = 4.5th$

Median = $\frac{x_4 + x_5}{2} = \frac{60 + 65}{2} = 62.5 (\$)$

Mode: the sample size is too small, mode cannot be identified.

Example 2 (calculating mean, median and mode for grouped data)

The following table shows the daily wages of a random sample of construction workers. Calculate its mean, median and mode.

Daily Wages ($)	Number of Workers
200 - 399	5
400 - 599	15
600 - 799	25
800 - 999	30
1000 - 1199	18
1200 - 1399	7
Total	100

Solution

Daily Wages ($)	Number of Workers f_i	Class Mark x_i	$f_i x_i$
200 - 399	5	299.5	1,497.5
400 - 599	15	499.5	7,492.5
600 - 799	25	699.5	17,489.5
800 - 999	30	899.5	26,985.5
1000 - 1199	18	1,099.5	19,791.0
1200 - 1399	7	1,299.5	9,096.5
Total	100		82,350.0

Mean $(\bar{x}) = \dfrac{\sum_{i=1}^{6} f_i x_i}{\sum_{i=1}^{6} f_i} = \dfrac{82{,}350.0}{100} = 823.5$ ($

Daily Wages ($)	Number of Workers f_i	Cumulative Frequency F_i
200 - 399	5	5
400 - 599	15	20
600 - 799	25	45
800 - 999	30	75
1000 - 1199	18	93
1200 - 1399	7	100
Total	100	

As 0.5n = 0.5(100) = 50, so the median lies in the 4*th* class.

Median = $L_4 + \dfrac{0.5n - F_3}{f_4}(c_4)$ where L is the lower class boundary,

c is the class interval.

$$= 799.5 + \frac{0.5(100) - 45}{30}(200) = 832.8\,(\$)$$

Daily Wages ($)	Number of Workers f_i	Class Interval c_i	Relative Density $f_i^{'} = \dfrac{f_i}{c_i}(200)$
200 - 399	5	200	5
400 - 599	15	200	15
600 - 799	25	200	25
800 - 999	30	200	30
1000 - 1199	18	200	18
1200 - 1399	7	200	7
Total	100		

As $f_4^{'} = 30$ is the largest relative density, so mode lies in the 4*th* class.

Mode $= L_4 + \dfrac{f_4^{'} - f_3^{'}}{(f_4^{'} - f_5^{'}) + (f_4^{'} - f_3^{'})}(c_4)$

$$= 799.5 + \frac{30 - 25}{(30 - 18) + (30 - 25)}(200) = 858.3\,(\$)$$

Advantages and disadvantages of each measure

Mean

Advantages:
(i) All values in the distribution are used in its calculation, so it can be regarded as more representative than the other two measures.

(ii) Its method of calculation is simple and most people understand the meaning of its result.

(iii) Its result can easily be used in further analysis.

Disadvantages:
(i) Its result can be easily distorted by extreme values. As such, its result may be rather lower or higher than the bulk of the values and becomes unrepresentative.

(ii) In case of open end classes, mean can be calculated only if their class marks are determined. If such classes contain a large proportion of the values, then the mean may be subjected to substantial error.

Median

Advantage: Its result will not be affected by extreme values and open end classes.

Disadvantage: It has to be supplemented by other statistics because it does not reflect the distribution in the way that the mean does, that is, including all values.

Mode

Advantages:
(i) Its result will not be affected by extreme values and open end classes.

(ii) If data are not grouped, it can be determined easily.

Disadvantages:
(i) It has to be supplemented by other statistics.

(ii) It is difficult to obtain an accurate estimate of the mode if the values are classified into a frequency distribution.

How to select a suitable measure

(i) Always select the mean whenever there is no special reason for choosing the other two measures.

(ii) Select the median is the distribution consists of substantial amount of extreme large or small values.

(iii) Select the mode if integral result is preferred as in cases the data are in ordinal scales.

Measure of data variation (variability)

A measure of central tendency is almost never, by itself, sufficient to provide an adequate summary of the characteristics of a set of data. We will usually require, in addition, a measure of the amount of variation in the data.

Example 1

Consider the following measurements, in grams, for two samples of strawberry jam bottled by companies A and B:

Sample for Company A	31	32	32	33	32
Sample for Company B	28	29	32	35	36

Both samples have the same mean, 32 grams. It is obvious that company A, in comparison with company B, bottles strawberry jam with a more consistent content. We say that the variability of the observations is smaller for company A. Therefore in buying strawberry jam we would feel more confident that the bottle we select will be closer to the advertised average content if we buy from company A.

The most important measures of variability or dispersion are the ***range, mean deviation, standard deviation and variance.***
(There are some other measures like quartile deviation and percentiles. We shall not study these measures. Read our textbook if interested)

The ***range*** of a set of numbers is the difference between the largest and the smallest number in the set.

Example 2 (For individual data)

The following table shows the hourly wage rates of eight sampled construction workers.

Worker i	1	2	3	4	5	6	7	8
Hourly wage rate (x_i)	$35	38	46	60	65	69	72	78

The range is $78 - $35 = $43.

Though range is simple and can be obtained easily, its result is unstable. This is particularly true if the sample size is large. So whenever the sample size is over 10, we seldom choose to use range to indicate variability of the data.

Mean deviation is the average of the absolute deviation of the numerical data from their mean.

Worker i	1	2	3	4	5	6	7	8
Hourly wage rate (x_i)	\$35	38	46	60	65	69	72	78
$\lvert x_i - \bar{x}\rvert$ $= \lvert x_i - 57.875\rvert$	22.875	19.875	11.875	2.125	7.125	11.125	14.125	20.125

$$\text{Mean deviation } \frac{\sum_{i=1}^{8}\lvert x_i - 57.875\rvert}{8} = \frac{109.25}{8} = 13.656 (\$)$$

The mean deviation is a good measure to show the extent of variation of the data in a distribution. However, when this measurement is used in further analysis, it would give rise to some unnecessary tedious mathematical problem as a result of its absolute value term. To avoid this pitfall, we can use the standard deviation instead.

Standard deviation of a population **(**σ) is the square root of the average of the squared distances of the observations from the mean.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}, \quad \text{where } \mu \text{ is the population mean}$$

To compute the **sample** standard deviation (s) we use the above formula, replacing μ by $\bar{x}$ and N by $n-1$.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Worker i	1	2	3	4	5	6	7	8	Total
Hourly wage rate (x_i)	$35	38	46	60	65	69	72	78	463

$$s = \sqrt{\frac{(x - 35.875)^2}{7}} = 16.226(\$)$$

Variance is the square of the standard deviation.

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

Example 3 (for grouped data)

The following table shows the daily wages of a random sample of construction workers. Calculate its mean deviation, variance, and standard deviation.

Daily Wages ($)	Number of Workers
200 - 399	5
400 - 599	15
600 - 799	25
800 - 999	30
1000 - 1199	18
1200 - 1399	7
Total	100

Solution

Daily Wages ($)	Number of Workers f_i	Class Mark x_i	$f_i\lvert x_i - \bar{x}\rvert = f_i\lvert x_i - 823.5\rvert$
200 - 399	5	299.5	2,620
400 - 599	15	499.5	4,860
600 - 799	25	699.5	3,100
800 - 999	30	899.5	2,280
1000 - 1199	18	1,099.5	4,968
1200 - 1399	7	1,299.5	3,332
Total	100		21,160

$$\text{Mean deviation} = \frac{\sum_{i=1}^{6} f_i\lvert x_i - \bar{x}\rvert}{\sum_{i=1}^{6} f_i} = \frac{21{,}160}{100} = 211.60 \text{ (\$)}$$

Daily Wages ($)	Number of Workers f_i	Class Mark x_i	$f_i(x_i - \bar{x})^2$
200 - 399	5	299.5	1, 372,880
400 - 599	15	499.5	1,574,640
600 - 799	25	699.5	384,400
800 - 999	30	899.5	173,280
1000 - 1199	18	1,099.5	1,371,168
1200 - 1399	7	1,299.5	1,586,032
Total	100		6,462,400

$$\text{Variance}\,(s^2) = \frac{6462400}{99} = 65,276.77$$

Standard deviation =

$$\sqrt{65276.77} = 255.49$$

Comparison of the variation of two distributions

The values of the standard deviations cannot be used as the bases of the comparison because:

(a) units of measurements of the two distributions may be different, and

(b) average values of two distributions may be widely dissimilar.

The correct measure that should be used is the coefficient of variation (CV).

$$CV = \frac{s}{\bar{x}}100\%$$

Example 4

The following table shows the summary statistics for the daily wages of two types of workers.

Worker's Type	Daily Wages	
	Mean	Standard deviation
I	\$100	\$20
II	\$150	\$24

Compare these two daily wages distributions.

Solution

In comparison	Distribution	Reason
Average magnitude	II > I	$\bar{x}_{II} = 150 > \bar{x}_I = 100$
Variation	I > II	$CV_I = \frac{20}{100}100\% = 20\% > CV_{II} = \frac{24}{150}100\% = 16\%$

PART B: Probability

Introduction and concepts

"Perhaps it was man's unquenchable thirst for gambling that led to the early development of probability theory. In an effort to increase their winnings, gamblers called upon the mathematicians to provide optimum strategies for various games of chance." ---- from Walpole R.E. Introduction to Statistics

Probability is the basis upon which the discipline of statistics has been developed and applied in many fields associated with chance occurrences such as politics, business, weather forecasting, and scientific research. Probability may be taken as a tool with which we may solve problems involving uncertainties. In fact uncertainty is a basic element of human experiences. To cite some examples: travelling time, number of customers, rainfall, temperature, share price movement, length of our life, etc.

There are three approaches to understand probability. In the empirical approach, probability may be taken as a relative frequency. As such the probability of an aeroplane arriving its destination on time may be taken as the proportion of times the aeropline has been on time in the past, say, one thousand times.

Suppose in a trial of an experiment, there are k possible outcomes which are equally likely. The probability of the occurrence of an outcome is therefore 1/k. Thus in throwing a coin, the probability of having a head is ½. In our course, we shall adopt this approach but the empirical approach is always useful in giving us some intuition to understand the problem.

The third approach is very mathematical. A number of axioms have been set up and from these some theorems of probability have been developed. This approach is too abstract and usually used by mathematicians.

Some Basic Concepts

Sample space: is a set of all possible outcomes of an experiment.

Event: is a subset of a sample space.

To find the probability of an event we need to count the number of outcomes of the event and the number of all possible outcomes of the experiment, and then to divide the former by the latter. Hence the following counting rules may be helpful.

Some counting rules

Example 1

Three items are selected at random from a manufacturing process. Each item is inspected and classified defective (D) or non-defective (N).

Its sample space is = $\begin{Bmatrix} DDD & DDN & DND & NDD \\ DNN & NDN & NND & NNN \end{Bmatrix}$

Example 2

The event that the number of defectives in above example is greater than 1.

Its sample space is = {DDD DDN DND NDD}

The probability of the event is 4/8 or ½.

Example 3

Suppose a licence plate containing two letters following by three digits with the first digit not zero. How many different licence plates can be printed?

	1st Letter	2nd Letter	1st Digit	2nd Digit	3rd Digit
Number of Choices	A - Z (26)	A - Z (26)	1 - 9 (9)	0 - 9 (10)	0 - 9 (10)

Number of different licence plates that can be printed is

(26)(26)(9)(10)(10) = 608,400

Example 4

Find the possible permutations (the number of ways where sequence of the letters is counted) from 3 letters A, B, C.

- **The number of permutations of n distinct objects is**

$${}_nP_n = (n)(n-1)(n-2).....(2)(1) = n!$$

- **The number of permutations of n distinct objects taken r at a time is**

$${}_nP_r = (n)(n-1)...(n-r+2)(n-r+1)$$

$$= \frac{\big((n)(n-1)...(n-r+2)(n-r+1)\big)\big((n-r)(n-r-1)...(2)(1)\big)}{(n-r)(n-r-1)...(2)(1)} = \frac{n!}{(n-r)!}$$

e.g. The number of 3-letter words formed from 5 letters is

$${}_5P_3 = \frac{5!}{(5-3)!} = 60$$

- **The number of distinct permutations of n objects of which n_1 are alike of the first kind, n_2 are alike of the second kind, ..., n_k are alike of the kth kind and $n_1 + n_2 + ... + n_k = n$ is** $\dfrac{n!}{(n_1!)(n_2!)...(n_k!)}$

Find the possible permutations of the following 5 letters: A A A B C

There are five objects of which three are alike.

$$\therefore \text{The answer} = \frac{{}_5P_5}{3!} = \frac{5!}{3!}$$

Example 5

How many 7-letter words can be formed using the letters of the word 'BENZENE'?

(there are 1 B, 3 E, 2 N and 1 Z)

The number of 7-letter words that can be formed is $\dfrac{7!}{(1!)(3!)(2!)(1!)} = 420$

- **The number of combinations (number of ways where sequence is not counted) of n distinct objects taken r at a time is**

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

Find the possible combinations of 5 distinct objects taken 3 at a time.

The answer

$$= \frac{5!}{3!(5-3)!}$$

Example 6

The number of 3-person committees that can be formed from a group of 4 persons is

$${}_4C_3 = \frac{4!}{3!(4-3)!} = 4$$

Example 7

A box contains 8 eggs, 3 of which are rotten. Three eggs are picked at random. Find the probabilities of the following events.
(a) Exactly two eggs are rotten.
(b) All eggs are rotten.
(c) No egg is rotten.

Solution:

(a) The 8 eggs can be divided into 2 groups, namely, 3 rotten eggs as the first group and 5 good eggs as the second group.

Getting 2 rotten eggs in 3 randomly selected eggs can occurred if we select randomly 2 eggs from the first group and 1 egg from the second group.

The number of this outcome is $\left({}_3C_2\right)\left({}_5C_1\right) = 15$

Total number of possible outcomes of selecting 3 eggs randomly from the total 8 eggs is ${}_8C_3 = 56$.

Thus the probability of having exactly two rotten among the 3 randomly selected eggs is $\frac{\left({}_3C_2\right)\left({}_5C_1\right)}{{}_8C_3} = \frac{15}{56}$

(b) Similarly, the probability of having all 3 rotten eggs is

$$\frac{({}_3C_3)({}_5C_0)}{{}_8C_3} = \frac{1}{56}$$

(c) The probability of having no rotten egg is

$$\frac{({}_3C_0)({}_5C_3)}{{}_8C_3} = \frac{10}{56} = \frac{5}{28}$$

Rules of probability

The following rules may help us to find the probability of an event.

Addition Rule: For any events that are not **mutually exclusive**

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

where $A \cup B$ is the union of two sets A and B, it is the set of elements that belong to A or to B or to both.

$A \cap B$ is the intersection of two sets A and B, it is the set of elements that are common to A and B.

Illustrative example

180 students took examinations in English and Mathematics. Their results were as follows:

Number of students passing English = 80

Number of students passing Mathematics = 120

Number of students passing at least one subject = 144

Then we can rewrite the above results as:

Probability that a randomly selected student passed English = $\frac{80}{180} = \frac{4}{9}$

Probability that a randomly selected student passed Mathematics $= \frac{120}{180} = \frac{2}{3}$

Probability that a randomly selected student passed at least one subject $= \frac{144}{180} = \frac{4}{5}$

Find the probability that a randomly selected student passed both subject.

Solution

Let E be the event of passing English, and M be the event of passing Mathematics.

It is given that: $P(E) = \frac{4}{9}$; $P(M) = \frac{2}{3}$; $P(M \cup E) = \frac{4}{5}$

As $P(M \cup E) = P(E) + P(M) - P(M \cap E)$

$\therefore \quad P(M \cap E) = P(E) + P(M) - P(M \cup E) = \frac{4}{9} + \frac{2}{3} - \frac{4}{5} = \frac{14}{45} = 0.31$

Example 8

A card is drawn from a complete deck of playing cards. What is the probability that the card is a heart or an ace?

Solution

Let A be the event of getting a heart, and B be the event of getting an ace.

The probability that the card is a heart or an ace is $P(A \cup B)$.

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

$$= \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = \frac{4}{13}$$

For mutually exclusive events, $P(A \cup B) = P(A) + P(B)$

What is the probability of getting a total of '7' or '11' when a pair of dice are tossed?

Solution

Total number of possible outcomes = (6)(6) = 36

Possible outcomes of getting a total of '7' :{1,6; 2,5; 3,4; 4,3; 5,2; 6,1}

Possible outcomes of getting a total of '11' : {5,6; 6,5}

Let A be the event of getting a total of '7', and B be the event of getting a total of '11'.

The probability of getting a total of '7' or '11' is $P(A \cup B)$.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = P(A) + P(B)$ ...A and B are mutually exclusive

$$= \frac{6}{36} + \frac{2}{36} = \frac{2}{9}$$

If A and A' are complementary events then $P(A) = 1 - P(A')$

Example 9

A coin is tossed six times in succession. What is the probability that at least one head occurs?

Let A be the number of heads occurs in six successive tosses.

$P(A \geq 1) = 1 - P(A = 0)$

$$= 1 - \left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right)\left(\tfrac{1}{2}\right) = \frac{63}{64}$$

Conditional Probability

Let A and B be two events. The conditional probability of event A given that event B has occurred, denoted by $P(A/B)$ is defined as

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$ provided that P(B) > 0.

Similarly, the conditional probability of B given that event A has occurred is defined as $P(B/A) = \frac{P(A \cap B)}{P(A)}$, provided P(A) > 0.

Example 10

A hamburger chain found that 75% of all customers use mustard, 80% use ketchup, and 65% use both, when ordering a hamburger. What are the probabilities that:

(a) a ketchup-user uses mustard?
(b) a mustard-user uses ketchup?

Solution

Let A be the event of using mustard, and B be the event of using ketchup.

It is given that: $P(A) = 0.75$; $P(B) = 0.80$; $P(A \cap B) = 0.65$

(a) P(a ketchup-user uses mustard) $= P(A/B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{0.65}{0.80} = 0.8125$

(b) P(a mustard-user uses ketchup) $= P(B/A) = \dfrac{P(A \cap B)}{P(A)} = \dfrac{0.65}{0.75} = 0.8667$

Multiplicative Rule

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

or $= P(B)P(A/B)$

Statistically Independence: the occurrence or non-occurrence of one event has no effect on the probability of occurrence of the other event.

Two events A and B are independent if and only if $P(A \cap B) = P(A)P(B)$

Example 11

A pair of fair dice are thrown twice. What is the probability of getting totals of 7 and 11?

Solution

Let A_i be the event of getting '7' in the i-th throw and B_j be the event of getting '11' in the j-th throw.

P(Getting totals of 7 and ll) $= P(A \cap B) = P(A_1 \cap B_2) + P(B_1 \cap A_2)$

$= P(A_1)P(B_2/A_1) + P(B_1)P(A_2/B_1)$

$= P(A_1)P(B_2) + P(B_1)P(A_2)$ .... A_i, B_j are independent

$$= \left(\frac{6}{36}\right)\left(\frac{2}{36}\right) + \left(\frac{2}{36}\right)\left(\frac{6}{36}\right) = \frac{1}{54}$$

Theorem of Total Probability

If the events $B_1, B_2, ..., B_k$ constitute a partition of the sample space S such that $P(B_i \neq 0)$ for i = 1, 2, ... , k, then for any event A of S

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_k)$$

$$= P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + ... + P(B_k)P(A/B_k)$$

Example 12

Suppose 50% of the cars are manufactured in the United States and 15% of these are compact; 30% of the cars are manufactured in Europe and 40% of these are compact; and finally, 20% are manufactured in Japan and 60% of these are compact. If a car is picked at random from the lot, find the probability that it is a compact.

Let A be the event that the car is compact,

B_1 be the event that the car is manufactured is United States,

B_2 be the event that the car is manufactured in Europe, and

B_3 be the event that the car is manufactured in Japan.

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3)$$

$$= P(B_1)P(A/B_1) + P(B_2)P(A/B_2) + P(B_3)P(A/B_3)$$

$$= (0.50)(0.15) + (0.30)(0.40) + (0.20)(0.60) = 0.315$$

Baye's Theorem

If $E_1, E_2, ..., E_k$ are mutually exclusive events such that $E_1 \cup E_2 \cup ... \cup E_k$ contains all sample points of S, then for any event D of S with $P(D) \neq 0$,

$$P(E_i/D) = \frac{P(E_i \cap D)}{P(D)} = \frac{P(E_i \cap D)}{\sum_{j=1}^{k} P(E_j \cap D)}$$

$$= \frac{P(E_i)P(D/E_i)}{P(E_1)P(D/E_1) + P(E_2)P(D/E_2) + ...P(E_k)P(D/E_k)}$$

Example 13

Suppose a box contains 2 red balls and 1 white ball and a second box contains 2 red ball and 2 white balls. One of the boxes is selected by chance and a ball is drawn from it. If the drawn ball is red, what is the probability that it came from the 1st box?

Solution

Let A be the event of drawing a red ball and B be the event of choosing the 1st box.

Given: $P(B) = P(B') = \frac{1}{2}$; $P(A/B) = \frac{2}{3}$; $P(A/B') = \frac{2}{4}$

P(Coming from the 1st box/the drawn ball is red) $= P(B/A)$

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

$$= \frac{P(B)P(A/B)}{P(B)P(A/B) + P(B')P(A/B')} = \frac{(1/2)(2/3)}{(1/2)(2/3) + (1/2)(2/4)} = \frac{4}{7}$$

PART C: Probability Distributions

To cope with uncertainties of outcome, a statistical model that describes the behavior of the outcome is needed. These theoretical models which are very similar to relative frequency distributions, are called probability distributions.

Random Variables - A random variable is a variable that takes on different numerical values determined by the outcomes of a random experiment.

Example 1

An experiment of tossing a coin 3 times.

Let random variable, X be the number of heads achieved.

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

so X = {0, 1, 2, 3}

Discrete random variable - in a given interval, only a specified number of values can occur.

Continuous random variable - in a given interval, any value can occur.

Probability Distribution of a random variable - is a representation of the probabilities for all the possible outcomes.

Example 2

The probability distribution of the number of heads occurred when a coin is tossed 4 times.

x	0	1	2	3	4
P(X=x)	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

That is, $P(X = x) = \frac{{}_4C_x}{16}$, $x = 0,\ 1, 2, 4$

Example 3

Consider an experiment of tossing two fair dice.

Let random variable, X be the sum of the two dice. Then the probability distribution of X is:

x	2	3	4	5	6	7	8	9	10	11	12
P(X=x)	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

The probability function $f(x)$, of a discrete random variable X expresses the probability that X takes the value x, as a function of x. That is

$$f(x) = P(X = x)$$

where the function is evaluated at all possible values of x.

Properties of probability function $P(X = x)$:-

1. $P(X = x) \geq 0$ for any value x.

2. $\sum_x P(X = x) = 1.$

Mathematical Expectations

The <u>expected value</u>, E(X), of a discrete random variable X is defined as

$$E(X) \text{ or } \mu_x = \sum_x xP(X = x)$$

It is the mean of the probability distribution.

Let X be a random variable. The expectation of the squared discrepancy about the mean, $(X - \mu_x)^2$, is called the **variance**, denoted σ_x^2, and given by

$$Var(X) \text{ or } \sigma_x^2 = E[(X - \mu_x)^2]$$

$$= \sum_x (x - \mu_x)^2 P(X = x)$$

$$= \sum_x x^2 P(X = x) - \mu_x^2$$

Example 4

Calculate the mean and variance of the discrete probability distribution in example 2 and 3.

The Normal Distribution

Normal distribution is probability distribution of a continuous random variable. It is based on the Law of Errors which states that

1. Errors are inevitable.
2. Large errors are less likely than small errors.
3. Positive and negative errors are equally likely.

Definition :

A continuous random variable X is defined to be a normal random variable if its probability function is given by

$$f(x) = \frac{1}{\sigma\sqrt{(2\pi)}} \exp[-\frac{1}{2}(\frac{x-\mu}{\sigma})^2] \qquad \text{for } -\infty < x < +\infty$$

where μ = the mean of X, σ = the standard deviation of X,

π = 3.14154

Notation : X ~ N(μ, σ^2)

Properties of the normal distribution:-

1. It is a continuous distribution.

2. The curve is symmetric and bell-shaped about a vertical axis through the mean μ.

3. The total area under the curve and above the horizontal axis is equal to 1.

4. Area under the normal curve:

   - Approximately 68% of the values in a normally distributed population are within 1 standard deviation from the mean.

- Approximately 95.5% of the values in a normally distributed population are within 2 standard deviation from the mean.

- Approximately 99.7% of the values in a normally distributed population are within 3 standard deviation from the mean.

The standard normal curve :

The distribution of a normal random variable with $\mu = 0$ and $\sigma = 1$ is called a **standard normal distribution**. Usually a standard normal random variable is denoted by Z.

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

Remark : Usually a table of Z is set up to find the probability $P(Z \geq z)$ for $z \geq 0$.

Example 7

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

(a) $P(Z > 1.73) = 0.0418$

(b) $P(0 < Z < 1.73) = P(Z > 0) - P(Z > 1.73) = 0.5 - 0.0418 = 0.4582$

(c) $P(-2.42 < Z < 0.8) = 1 - P(Z < -2.42) - P(Z > 0.8)$
$= 1 - 0.00776 - .2119 = 0.78034$

(d) $P(1.8 < Z < 2.8) = P(Z > 1.8) - P(Z > 2.8) = 0.0359 - 0.00256 = 0.03334$

(e) the value z that has

(i) 5% of the area below it

Let the corresponding z value be z_1, then we have $P(Z < z_1) = 0.05$.

From the standard normal distribution table we have $P(Z < -1.64) = 0.05$.

So $z_1 = -1.64$

(ii) 39.44% of the area between 0 and z.

Let the corresponding z value be z_1, then we have $P(0 < Z < z_1) = 0.3944$.

From the standard normal distribution table we have $P(0 < Z < 1.25) = 0.3944$.

So $z_1 = 1.25$

Theorem :

If X is a normal random variable with mean μ and standard deviation σ, then

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

is a standard normal random variable and hence

$$P(x_1 < X < x_2) = P(\frac{x_1 - \mu}{\sigma} < Z < \frac{x_2 - \mu}{\sigma})$$

Example 8

Given X ~ N(50, 10^2), find P(45 < X < 62).

Solution: $$P(45 < X < 62) = P\left(\frac{45 - \mu}{\sigma} < \frac{X - \mu}{\sigma} < \frac{62 - \mu}{\sigma} \right)$$

$$= P\left(\frac{45 - 50}{10} < Z < \frac{62 - 50}{10} \right)$$

$$= P(-0.5 < Z < 1.2)$$

$$= 1 - P(Z < -0.5) - P(Z > 1.2)$$

= 1 - 0.3085 - .1151 = 0.5764

Example 9

The charge account at a certain department store is approximately normally distributed with an average balance of $80 and a standard deviation of $30. What is the probability that a charge account randomly selected has a balance

(a) over $125;
(b) between $65 and $95.

Let X be the balance in the charge account. X ~ N(80, 30^2)

(a) $P(X > 125) = P(\frac{X-\mu}{\sigma} > \frac{125-\mu}{\sigma})$

$$= P(Z > \frac{125-80}{30}) = P(Z > 1.5) = 0.0668$$

(b) $$P(65 < X < 95) = P\left(\frac{65-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{95-\mu}{\sigma}\right)$$

$$= P\left(\frac{65-80}{30} < Z < \frac{95-80}{30}\right)$$

$$= P(-0.5 < Z < 0.5) = 1 - P(Z < -0.5) - P(Z > 0.5)$$

$$= 1 - 0.3085 - 0.3085 = 0.3830$$

Example 10

On an examination the average grade was 74 and the standard deviation was 7. If 12% of the class are given A's, and the grades are curved to follow a normal distribution, what is the lowest possible A and the highest possible B?

Let X be the examination grade and x_1 be the lowest grade for A.

$$P(X > x_1) = 0.12 \Rightarrow P\left(Z > \frac{x_1 - 74}{7}\right) = 0.12$$

From the standard normal distribution, we get

$$P(Z > 1.17) = 0.1210, \text{ and } P(Z > 1.18) = 0.1190$$

so $P(Z > 1.175) \cong 0.12$

Thus $\frac{x_1 - 74}{7} = 1.175$

i.e. $x_1 = 74 + (7)(1.175) = 82.2 \cong 83$

The highest possible B is 82.

The Binomial Distribution

A binomial experiment possesses the following properties :

1. There are n identical observations or trials.
2. Each trial has two possible outcomes, one called “success” and the other “failure”. The outcomes are mutually exclusive and collectively exhaustive.
3. The probabilities of success p and of failure 1 – p remain the same for all trials.
4. The outcomes of trials are independent of each other.

Example 11

1. In testing 10 items as they come off an assembly line, where each test or trial may indicate a defective or a non-defective item.

2. Five cards are drawn with replacement from an ordinary deck and each trial is labelled a success or failure depending on whether the card is red or black.

Definition :

In a binomial experiment with a constant probability p of success at each trial, the probability distribution of the binomial random variable X, the number of successes in n independent trials, is called the binomial distribution.

Notation : $X \sim b(n, p)$

$P(X = x) = \binom{n}{x} p^x q^{n-x}$ $\quad x = 0, 1, \ldots, n$

$p + q = 1$

Example 12

Of a large number of mass-produced articles, one-tenth is defective. Find the probabilities that a random sample of 20 will obtain

(a) exactly two defective articles;
(b) at least two defective articles.

Let X be the number of defective articles in a random sample of 20. X ~ b(20, $\frac{1}{10}$)

(a) $$P(X=2)=\binom{20}{2}\left(\frac{1}{10}\right)^2\left(\frac{9}{10}\right)^{18}=0.28517$$

(b) $$P(X\geq 2)=1-P(X=0)-P(X=1)$$
$$=1-\binom{20}{0}\left(\frac{1}{10}\right)^0\left(\frac{9}{10}\right)^{20}-\binom{20}{1}\left(\frac{1}{10}\right)\left(\frac{9}{10}\right)^{19}=1-.12158-0.27017=0.60825$$

Example 13

A test consists of 6 questions, and to pass the test a student has to answer at least 4 questions correctly. Each question has three possible answers, of which only one is correct. If a student guesses on each question, what is the probability that the student will pass the test?

Let X be the no. of correctly answered questions among 6 questions. X ~ b(6, $\frac{1}{3}$)

$$P(X\geq 4)=\sum_{x=4}^{6}P(X=x)=\sum_{x=4}^{6}\binom{6}{x}\left(1/3\right)^x\left(2/3\right)^{6-x}$$

$$=\binom{6}{4}\left(1/3\right)^4\left(2/3\right)^2+\binom{6}{5}\left(1/3\right)^5\left(2/3\right)^1+\binom{6}{6}\left(1/3\right)^6\left(2/3\right)^0=0.10014$$

Theorem

The mean and variance of the binomial distribution with parameters of n and p are
μ = np and σ^2 = npq respectively where p + q = 1.

Example 14

A packaging machine produces 20 percent defective packages. A random sample of ten packages is selected, what are the mean and standard deviation of the binomial distribution of that process?

Let X be the no. of defective packages in a sample of 10 packages. X ~ b(10, 0.2)

Its mean is μ = np = (10)(0.2) = 2

Its standard deviation is $\sigma = \sqrt{npq} = \sqrt{(10)(0.2)(0.8)} = 1.265$

The Normal Approximation to the Binomial Distribution Theorem :

Given X is a random variable which follows the binomial distribution with parameters n and p, then

$$P(X = x) = P(\frac{(x-0.5)-np}{\sqrt{(npq)}} < Z < \frac{(x+0.5)-np}{\sqrt{(npq)}})$$

if n is large and p is not close to 0 or 1.

Remark : If both np and nq are greater than 5, the approximation will be good.

Example 15

A process yields 10% defective items. If 100 items are randomly selected from the process, what is the probability that the number of defective exceeds 13?

Let X be the no. of defective in a random sample of 100 items. X ~ b(100, 0.1)

$\mu = np = (100)(0.1) = 10$, $\sigma = \sqrt{npq} = \sqrt{(100)(0.1)(0.9)} = 3$

$P(X > 13) \cong P(X' > 13.5)$ by normal approximation

$$= P\left(\frac{X'-\mu}{\sigma} > \frac{13.5-\mu}{\sigma}\right) = P\left(Z > \frac{13.5-10}{3}\right) = P(Z > 1.167) = 0.121$$

Example 17

A multiple-choice quiz has 200 questions each with four possible answers of which only one is the correct answer. What is the probability that sheer guesswork yields from 25 to 30 correct answers for 80 of the 200 problems about which the student has no knowledge?

Let X be the no. of correct answers for 80 with sheer guesswork. X ~ b(80, 0.25)

$\mu = np = (80)(0.25) = 20$, $\sigma = \sqrt{npq} = \sqrt{(80)(0.25)(0.75)} = \sqrt{15}$

$P(25 \le X \le 30) \cong P(24.5 < X' < 30.5)$ by normal approximation

$$= P\left(\frac{24.5-20}{\sqrt{15}} < Z < \frac{30.5-20}{\sqrt{15}}\right) = P(1.16 < Z < 2.71) = 0.1230 - 0.00336 = 0.1196$$

The Poisson Distribution

Experiments yielding numerical values of a random variable X, the number of successes (observations) occurring during a given time interval (or in a specified region) are often called Poisson experiments.

A Poisson experiment has the following properties:

1. The number of successes in any interval is independent of the number of successes in other interval.
2. The probability of a single success occurring during a short interval is proportional to the length of the time interval and does not depend on the number of successes occurring outside this time interval.
3. The probability of more than one success in a very small interval is negligible.

Examples of random variables following Poisson Distribution

1. The number of customers arrived during a time period of length t.
2. The number of telephone calls per hour received by an office.
3. The number of typing errors per page.
4. The number of accidents occurred at a junction per day.

Definition :

The probability distribution of the Poisson random variable X is called the Poisson distribution.

Notation : $X \sim P_o(\lambda)$

where λ is the average number of successes occuring in the given time interval.

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

$$e = 2.718283$$

Theorem: In a Poisson Distribution mean is equal to variance, i.e., $\mu = \sigma^2$.

Example 17

The average number of radioactive particles passing through a counter during 1 millisecond in a laboratory experiment is 4. What is the probability that 6 particles enter the counter in a given millisecond?

Let X be the no. of particles entering the counter in a given millisecond. $X \sim P_o(4)$

$$P(X = 6) = \frac{e^{-4}4^{6}}{6!} = 0.1042$$

Example 19

Ships arrive in a harbour at a mean rate of two per hour. Suppose that this situation can be described by a Poisson distribution. Find the probabilities for a 30-minute period that

(a) No ships arrive;
(b) Three ships arrive.

Let X be the no. of ship arriving in a harbour for a 30-minute period. $X \sim P_o(2/2 = 1)$

(a) $$P(X = 0) = \frac{e^{-1}1^{0}}{0!} = 0.3679$$

(b) $$P(X = 3) = \frac{e^{-1}1^{3}}{3!} = 0.0613$$

Theorem :

The mean and variance of the Poisson distribution both have mean λ.

Poisson approximation to the binomial distribution

If n is large and p is near 0 or near 1.00 in the binomial distribution, then the binomial distribution can be approximated by the Poisson distribution with parameter np.

Example 20

If the prob. that an individual suffers a bad reaction from a certain injection is 0.001, determine the prob. that out of 2000 individuals, more than 2 individuals will suffer a bad reaction.

Sol^n : According to binomial :

The required probability

$$= 1 - \left\{ \binom{2000}{0} (0.001)^0 (0.999)^{2000} + \binom{2000}{1} (0.001)^1 (0.999)^{1999} + \binom{2000}{2} (0.001)^2 (0.999)^{1998} \right\}$$

Using Poisson distribution:

P(0 suffers) = $\frac{2^0 e^{-2}}{0!} = \frac{1}{e^2}$ $\because$ $\lambda = np = 2$

P(1 suffers) = $\frac{2^1 e^{-2}}{1!} = \frac{2}{e^2}$

P(2 suffer) = $\frac{2^2 e^{-2}}{2!} = \frac{2}{e^2}$

Then the required probability = $1 - \frac{5}{e^2} = 0.323$

General speaking, the Poisson distribution will provide a good approximation to binomial when

(i) n is at least 20 and p is at most 0.05; or

(ii) n is at least 100, the approximation will generally be excellent provided p< 0.1.

Example 21

Two percent of the output of a machine is defective. A lot of 300 pieces will be produced. Determine the probability that exactly four pieces will be defective.

Let X be the no. of defective pieces among 300 pieces. X ~ b(300, 0.02)

$$P(X = 4) = {}_{300}C_4 (0.02)^4 (0.98)^{296} = 0.1338$$

By Poisson Approximation:

$$\lambda = np = (300)(0.02) = 6$$

$$P(X = 4) = \frac{e^{-6} 6^4}{4!} = 0.1338$$

PART D: SAMPLING DISTRIBUTIONS AND ESTIMATION

Definition

1. A sample statistic is a characteristic of a sample.
   A population parameter is a characteristic of a population.

2. A statistic is a random variable that depends only on the observed random sample.

3. A sampling distribution is a probability distribution for a sample statistic. It indicates the extent to which a sample statistic will tend to vary because of chance variation in random sampling.

4. The standard deviation of the distribution of a sample statistic is known as the standard error of the statistic.

An illustrating example

Suppose a population consists of four elements, {0,1,2,3}. A simple random sample of two elements is to be drawn.

The population has two parameters: a mean μ of 1.5 and a variance σ^2 of 1.6667.

Obviously there are six possible samples ($C_2^4 = 6$). They are

Sample	Sample mean	error	Probability
0,1	0.5	-1.0	1/6
0,2	1.0	-0.5	1/6
0,3	1.5	0	1/6
1,2	1.5	0	1/6
1,3	2.0	0.5	1/6
2,3	2.5	1.0	1/6

From the above table, we can see that if we draw a sample and use the sample mean to estimate the population mean, the accuracy of our estimate depends on which sample we have drawn, which in turn depends on chance.

The probability distribution of sample mean is known as a sampling distribution of sample mean, as compiled in the following table:

Sample mean	0.5	1.0	1.5	2.0	2.5
Probability	1/6	1/6	2/6	1/6	1/6

The expected value of sample mean is

$$E(\bar{y}) = \sum \bar{y} * P(\bar{y}) = 0.5*1/6 + 1.0*1/6 + 1.5*2/6 + 2.0*1/6 + 2.5*1/6 = 1.5 = \bar{Y}.$$

Hence the average value of the sample mean is equal to the population mean. We call the sample mean an unbiased estimator of the population mean.

The variance of the sample mean (i.e., the average square deviation of the sample mean from the population mean) is: $V(\bar{y}) = \sum (\bar{y} - \bar{Y})^2 P(\bar{y})$ =

$$(0.5-1.5)^2 * \frac{1}{6} + (1.0-1.5)^2 * \frac{1}{6} + (1.5-1.5)^2 * \frac{2}{6} + (2.0-1.5)^2 * \frac{1}{6} + (2.5-1.5)^2 * \frac{1}{6}$$

=0.4167

Sampling Distribution of Mean

The Central Limit Theorem

> If repeated samples of size n are drawn from any infinite population with mean μ and variance σ^2, and n is large ($n \geq 30$), the distribution of $\bar{x}$, the sample mean, is approximately normal, with mean μ (i.e. $E(\bar{x}) = \bar{\mu}$) and variance σ^2/n (i.e. $V(\bar{x}) = \frac{\sigma^2}{n}$), and this approximation becomes better as n becomes larger.

Notes: As in the previous illustrating example, we can see the following modifications:

(i) If the population is finite, $V(\bar{x}) = (1 - \frac{n}{N}) \frac{\sigma^2}{n}$; where (1-n/N) is known as the finite population correction factor. When N is very big, the factor is equal to 1.

(ii) If n is small, say less than 30, the sampling distribution is not so normal. A t-distribution will be used (discussed later).

In the above example, N=4, n=2, $V(\bar{x}) = (1-\frac{n}{N})\frac{\sigma^2}{n}$ =(1-2/4)(1.6667/2) = 0.4167. If the population is big (or the sample is drawn with replacement), then $V(\bar{x}) = \frac{\sigma^2}{n}$ =1.6667/2=0.8333.

In this course we assume a big population or sampling with replacement.

Example 1

An electrical firm manufactures light bulbs that have a length of life that is approximately normal distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Let $\bar{X}$ be the average life of the 16 bulbs. $\bar{X}$ ~ N($\mu_{\bar{x}} = \mu = 800$, $\sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{40^2}{16}$)

$$P(\bar{X} < 775) = P\left(\frac{\bar{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} < \frac{775 - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(Z < \frac{775 - 800}{\frac{40}{\sqrt{16}}}\right)$$

$$= P(Z < -2.5) = 0.00621$$

Example 2

The mean IQ scores of all students attending a college is 110 with a standard deviation of 10.

(a) If the IQ scores are normally distribution, what is the probability that the score of any one student is greater than 112?
(b) What is the probability that the mean score in a random sample of 36 students is greater than 112?
(c) What is the probability that the mean score in a random sample of 100 students is greater than 112?

Solution

(a) Let X be the student's IQ score. X ~ N(110, 10^2)

$$P(X > 112) = P\left(Z > \frac{112 - 110}{10}\right) = P(Z > 0.2) = 0.4207$$

(b) Let $\overline{X}_1$ be the mean score of a sample of 36 students.

$$\overline{X}_1 \sim \text{N}\left(\mu_{\bar{x}} = \mu = 110,\ \sigma^2_{\bar{x}} = \frac{\sigma^2}{n} = \frac{10^2}{36}\right)$$

$$P(\overline{X}_1 > 112) = P\left(Z > \frac{112 - 110}{\frac{10}{\sqrt{36}}}\right) = P(Z > 1.2) = 0.1151$$

(c) Let $\overline{X}_2$ be the mean score of a sample of 100 students.

$$\overline{X}_2 \sim \text{N}\left(\mu = 110,\ \frac{\sigma^2}{n} = \frac{10^2}{100}\right)$$

$$P(\overline{X}_2 > 112) = P\left(Z > \frac{112 - 110}{\frac{10}{\sqrt{100}}}\right) = P(Z > 2) = 0.0.228$$

Estimation

Estimation is the process of using statistics from sample data to estimate the parameters of the population. A statistic is a random variable which depends on which sample is drawn from a population.

The followings are some examples

	Estimator	Population parameter
1.	$\bar{x}$	μ
2.	s^2	σ^2
3.	P	P

There are two important properties for an estimator, namely, unbiasedness and efficiency.

Unbiased estimator: An estimator, for example, $\bar{x}$, is unbiased if and only if $E(\bar{x}) = \mu$.

Efficiency: The efficiency of an estimator, for example, $\bar{x}$, is given by $V(\bar{x})$. The smaller the $V(\bar{x})$, the more accurate will be the $\bar{x}$ as an estimator.

There are two types of estimate

1. A point estimate is a single-value estimate of a population parameter, for example, $\hat{\mu} = \bar{x}; \hat{P} = p$.

2. An interval estimate of a population parameter gives an interval that may contain the true value of the parameter with a certain probability (i.e. confidence); for example, $\Pr(a < \mu < b) = 0.99.$

For a point estimate, both the accuracy and reliability of the estimation are unknown. For an interval estimate, the width of the interval gives the accuracy and the probability gives the reliability of the estimation.

Examples 3

(a) The mean and standard deviation for the quality point averages of a random sample of 36 college seniors are calculated to be 2.6 and 0.3, respectively. Find a 95% confidence interval for the mean of the entire senior class.

(b) How large a sample is required in (a) if we want to be 95% confident of μ is off by less than 0.05?

Solution

Let μ be the mean of the entire senior class.

Given: n = 36, $\bar{x} = 2.6$, s = 0.3, (1 - α) = 0.95 $\Rightarrow \alpha = 0.05$

(a) A 95% confidence interval estimate for the μ is

$$\bar{x} - z_{0.025}\frac{\hat{\sigma}}{\sqrt{n}} < \mu < \bar{x} + z_{0.025}\frac{\hat{\sigma}}{\sqrt{n}} \Rightarrow 2.6 - 1.96\left(\frac{0.3}{\sqrt{36}}\right) < \mu < 2.6 + 1.96\left(\frac{0.3}{\sqrt{36}}\right)$$

$$\Rightarrow 2.502 < \mu < 2.698$$

(b) Let n_1 be the required sample size.

To be 95% confident that μ if off by less than 0.05 would imply

$$z_{0.025}\left(\frac{\hat{\sigma}}{\sqrt{n_1}}\right) \le 0.05 \Rightarrow 1.96\left(\frac{0.3}{\sqrt{n_1}}\right) \le 0.05$$

$$\therefore \qquad n_1 \ge \left[\frac{(1.96)(0.3)}{0.05}\right]^2 = 138.30 \cong 139$$

A summary table for constructing (1 – α)% confidence interval for mean and proportion

Estimating	Conditions	Formula
Mean	Large samples (n ≥ 30) <u>OR</u> σ is known	$\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
Mean *	Small samples and σ unknown	$\bar{X} - t_{\nu,\alpha/2}\frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{\nu,\alpha/2}\frac{s}{\sqrt{n}}$ v=n-1
Proportion	Large sample	$\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$
Difference of means	Large sample OR σ_1 and σ_2 are known	$(\bar{X}_1 - \bar{X}_2) \pm Z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\hat{\sigma}_2}{n_2}}$
Difference of means	Small sample & σ_1 and σ_2 are unknown, assume $\sigma_1 = \sigma_2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{\nu,\alpha/2}(s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $\nu = n_1 + n_2 - 2$, pooled estimate of sample standard deviation:

		$s_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$
Difference of means	Small sample & σ_1 and σ_2 are unknown, assume $\sigma_1 \neq \sigma_2$	$(\bar{X}_1 - \bar{X}_2) \pm t_{v,\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\upsilon = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}$
Difference of means	Paired observations	$\bar{d} \pm t_{v,\alpha/2}\frac{s_d}{\sqrt{n}}$; $d = \bar{x}_1 - \bar{x}_2$ and v=n-1
Difference of proportions	Large samples	$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$

Example 4

The contents of seven similar containers of sulfuric acid are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2 and 9.6 liters. Find a 95% confidence interval for the mean of all such containers, assuming an approximate normal distribution.

Solution

Let μ be the mean of all such containers.

Given: n = 7 $\quad \sum x = 70 \quad \sum x^2 = 700.48$

$$\bar{x} = \frac{\sum x}{n} = \frac{70}{7} = 10 \qquad s^2 = \frac{\sum x^2 - (\sum x)^2/n}{n-1} = \frac{700.48 - 70^2/7}{6} = 0.08$$

$\therefore$ s = $\sqrt{0.08} = 0.2828$; $\quad$ (1 - α) = 0.95 $\Rightarrow \alpha = 0.05$, $\quad t_{6,0.025} = 2.447$

A 95% confidence interval estimate for the μ is

$$\bar{x} - t_{6,0.025}\frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{6,0.025}\frac{s}{\sqrt{n}} \Rightarrow 10 - 2.447\left(\frac{0.2828}{\sqrt{7}}\right) < \mu < 10 + 2.447\left(\frac{0.2828}{\sqrt{7}}\right)$$

$$\Rightarrow 9.738 < \mu < 10.262$$

Example 5

In a random sample of n = 500 families owning television sets in the city of Hamilton, Canada, it was found that x = 340 owned color sets. Find a 95% confidence interval for the actual proportion of families in this city with colour sets.

Let P be the actual proportion of families in this city with colour sets.

Given: n = 500, $\hat{p} = {}^{x}\!/_{n} = {}^{340}\!/_{500} = 0.68$, $(1-\alpha) = 0.95 \Rightarrow \alpha = 0.05$

A 95% confidence interval for P is

$$\hat{p} - z_{0.025}\left(\sqrt{{}^{\hat{p}\hat{q}}\!/_{n}}\right) < P < \hat{p} + z_{0.025}\left(\sqrt{{}^{\hat{p}\hat{q}}\!/_{n}}\right)$$

$$\Rightarrow 0.68 - 1.96\sqrt{\frac{(.68)(.32)}{500}} < P < 0.68 + 1.96\sqrt{\frac{(.68)(.32)}{500}} \Rightarrow 0.64 < P < 0.72$$

Examples 6

A standardized chemistry test was given to 50 girls and 75 boys. The girls made an average grade of 76 with a standard deviation of 6, while the boys made an average grade of 82 with a standard deviation of 8. Find a 96% confidence interval for the difference μ_1 and μ_2, where μ_1 is the mean score of all boys and μ_2 is the mean score of all girls who might take this test.

Given: $n_1 = 75$, $n_2 = 50$, $\bar{x}_1 = 82$, $s_1 = 8$, $\bar{x}_2 = 76$, $s_2 = 6$,

$$(1-\alpha) = .96 \Rightarrow \alpha = 0.04$$

A 96% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) - z_{0.02}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{0.02}\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}$$

$$\Rightarrow (82-76) - 2.05\sqrt{\frac{8^2}{75} + \frac{6^2}{50}} < \mu_1 - \mu_2 < (82-76) + 2.05\sqrt{\frac{8^2}{75} + \frac{6^2}{50}},$$

$n_1 > 30$ & $n_2 > 30$, so $\hat{\sigma}_1 = s_1$ & $\hat{\sigma}_2 = s_2$

$\Rightarrow 3.43 < \mu_1 - \mu_2 < 8.57$

Example 7

In a batch chemical process, two catalysts are being compared for their effect on the output of the process reaction. A sample of 12 batches is prepared using catalyst 1 and a sample of 10 batches was obtained using catalyst 2. The 12 batches for which catalyst 1 was used gave an average yield of 85 with a sample standard deviation of 4, while the average for the second sample gave an average of 81 and a sample standard deviation of 5. Find a 90% confidence interval for the difference between the population means, assuming the populations are approximately normally distributed with equal variances.

Solution

Let μ_1 and μ_2 be the mean population yield using catalyst 1 and catalyst 2, respectively.

Given: $n_1 = 12$, $n_2 = 10$, $\bar{x}_1 = 85$, $s_1 = 4$, $\bar{x}_2 = 81$, $s_2 = 5$,

$$(1-\alpha) = .90 \Rightarrow \alpha = 0.10, \quad \nu = n_1 + n_2 - 2 = 12 + 10 - 2 = 20, \quad t_{20,0.05} = 1.725$$

pooled estimate of sample standard deviation $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$

$$= \sqrt{\frac{(12-1)4^2 + (10-1)5^2}{12+10-2}} = 4.478$$

A 90% confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) - t_{20,0.05}(s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{20,0.05}(s_p)\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\Rightarrow (85-81) - (1.725)(4.478)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < (85-81) + (1.725)(4.478)\sqrt{\frac{1}{12} + \frac{1}{10}}$$

$\Rightarrow 0.69 < \mu_1 - \mu_2 < 7.31$

Example 8

The weight of 10 adults selected randomly before and after a certain new diet was introduced was recorded as follows:

Adult	Before (x_1)	After (x_2)	Difference
1	76	81	-5
2	60	52	8
3	85	87	-2
4	58	70	-12
5	91	86	5
6	75	77	-2
7	82	90	-8
8	64	63	1
9	79	85	-6
10	88	83	5

Find a 98% confidence interval for the mean difference in weight.

Solution

$\bar{d} = \frac{\sum d_i}{n}$ = -1.6 $\qquad s_d^2 = \frac{\sum (d_i - (-1.6))^2}{n-1} = 40.7$

For v = n-1 = 9; $t_{0.01} = 2.821$.

A 98% confidence interval is $-1.6 \pm (2,821)\left(\frac{6.38}{\sqrt{10}}\right)$

That is $-7.29 < \mu_d < 4.09$

Example 9

A certain change in a manufacturing procedure for component parts is being considered. Samples are taken using both the existing and the new procedure in order to determine if the new procedure results in an improvement. If 75 of 1500 items from the existing procedure were found to be defective and 80 of 2000 items from the new procedure were found to be defective, find a 90% confidence interval for the true difference in the fraction of defectives between the existing and the new process.

Solution

Let P_1 and P_2 be the true fraction of defectives of the existing and the new processes, respectively.

Given: $n_1 = 1500$, $x_1 = 75$, $\hat{p}_1 = {}^{75}\!/_{1500} = 0.05$

$n_2 = 2000$, $x_2 = 80$, $\hat{p}_2 = {}^{80}\!/_{2000} = 0.04$

$(1-\alpha) = .90 \Rightarrow \alpha = 0.10$

A 90% confidence interval for P_1 - P_2 is:

$$(\hat{p}_1 - \hat{p}_2) - z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}} < P_1 - P_2 < (\hat{p}_1 - \hat{p}_2) + z_{0.05}\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$$

$$\Rightarrow (0.01) - 1.64\sqrt{\frac{(.05)(.95)}{1500} + \frac{(.04)(.96)}{2000}} < P_1 - P_2 < (0.01) + 1.64\sqrt{\frac{(.05)(.95)}{1500} + \frac{(.04)(.96)}{2000}}$$

$$\Rightarrow -0.001697 < P_1 - P_2 < 0.021697$$

PART E: Introduction to Test of Hypothesis

Statistical Hypothesis

Consider the following example:

A manufacturer of sports equipment has developed a new synthetic fishing line that he claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that μ = 8 kilograms against the alternative that $\mu \neq$ 8 kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

When a random sample is drawn from a population (the 50 lines randomly selected), the sample information can be used assess the validity of some conjecture, or hypothesis. Here μ = 8 kilograms is known as the null hypotheis and $\mu \neq$ 8 kilograms is the alternative hypothesis. They are complementary to each other, and we need to decide which one to accept on the basis of the sample result of 50 lines.

Now let us make a 95% confidence interval about the mean breaking strength of the population as below:

$$P\left(7.8-1.96*\frac{0.5}{\sqrt{50}}<\mu<7.8+1.96*\frac{0.5}{\sqrt{50}}\right)=0.95;$$

i.e., $P(7.6614<\mu<7.9386)=0.95$.

As there is a probability of 0.95 that the mean breaking strength is between 7.66 kg and 7.94 kg, it is highly unlikely that the null hypothesis μ = 8 kg is true and hence should be rejected.

There are four possible situations for the above decision making exercise:

	H_0 is correct	H_0 is wrong
Accept H_0	Correct decision	Type 2 error
Reject H_0	Type 1 error	Correct decision

We still have a probability of 1-0.95, or 0.05 to reject a true H_0. We call this probability 'level of significance' or α, which is the probability of committing a type 1 error.

The rationale of hypothesis testing is simply outlined as above. There are however some formal concepts and procedures to conduct the test. The details are put down below.

Some Hypothesis Testing Terminology

1. Null hypothesis, H_0

   A hypothesis that is held to be true until very strong evidence to the contrary is obtained.

   $$H_0 : \mu = \mu_0$$

2. Alternative hypothesis, H_1

   It is a hypothesis that is complement to the null hypothesis. Hence it will be accepted if the null hypothesis is rejected.

   $H_1 : \mu \neq \mu_0$ (two-tail test)
   $H_1 : \mu > \mu_0$ (One-tail test)
   $H_1 : \mu < \mu_0$ (One-tail test)

   In the one-tail test we have some expectation about the direction of the error when the null hypothesis is wrong, while in the two-tail test we don't have such expectation.

3. Test statistics

   is the value, based on the sample, used to determine whether the null hypothesis should be rejected or accepted.

4. Critical region

   is a region in which if the test statistic falls the null hypothesis will be rejected.

5. Types of error

   (a) Type I error: Reject H_0 when H_0 is true

   (b) Type II error: Accept H_0 when H_a is true

6. The significance level, α

   is the probability of committing a type 1 error, i.e., P(Type I error) = α.

   The probability of committing a type 2 error is β; i.e., P(Type II error) = β.

Basic Steps in Testing Hypothesis

1. Formulate the null hypothesis.
2. Formulate the alternative hypothesis.
3. Specify the level of significance to be used.
4. Select the appropriate test statistic and establish the critical region.
5. Compute the value of the test statistic.
6. Conclusion: Reject H_0 if the statistic has a value in the critical region, otherwise accept H_0.

Tests concerning means

H_0	Conditions	Test statistic
$\mu = \mu_0$	Large samples ($n \geq 30$) OR σ is known	$z = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$
$\mu = \mu_0$	Small samples and σ unknown	$t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ **with** $\upsilon = n-1$
$\mu_1 - \mu_2 = d_0$	Large samples OR σ_1 and σ_2 are known	$z = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$
$\mu_1 - \mu_2 = d_0$	Small sample & σ_1 and σ_2 are unknown, assume $\sigma_1 = \sigma_2$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{s_p\sqrt{\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ **with** $\upsilon = n_1 + n_2 - 2$ **and** $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}$ **if** σ_1 **=** σ_2 **but unknown**
$\mu_1 - \mu_2 = d_0$	Small sample & σ_1 and σ_2 are unknown, assume $\sigma_1 \neq \sigma_2$	$t = \dfrac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ **with** $\upsilon = \dfrac{(s_1^2/n_1 + s_2^2/n_2)^2}{\dfrac{(s_1^2/n_1)^2}{n_1 - 1} + \dfrac{(s_2^2/n_2)^2}{n_2 - 1}}$
$\mu_1 - \mu_2 = d_0$	Paired observations	$t = \dfrac{\bar{d} - d_0}{s_d/\sqrt{n}}$ **with** $\upsilon = n-1$
$p = p_0$	Large sample	$z = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$
$p_1 - p_2 = 0$	Large samples	$z = \dfrac{(\hat{p}_1 - \hat{p}_2)}{\sqrt{\hat{p}(1-\hat{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$ $\hat{p} = \dfrac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$

Example 1

A manufacturer of sports equipment has developed a new synthetic fishing line that he claims has a mean breaking strength of 8 kilograms with a standard deviation of 0.5 kilogram. Test the hypothesis that μ = 8 kilograms against the alternative that $\mu \neq$ 8 kilograms if a random sample of 50 lines is tested and found to have a mean breaking strength of 7.8 kilograms. Use a 0.01 level of significance.

Null hypothesis: $\mu = 8$ kilograms

Alternative hypothesis: $\mu \neq 8$ kilograms

Level of significance: 0.01

Critical region: Z > $z_{0.005} = 2.58$ or Z < $-z_{0.005} = -2.58$

Computation:

n = 50 $\quad \bar{x} = 7.8 \quad \sigma = 0.5$

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{7.8 - 8}{0.5 / \sqrt{50}} = -2.828$$

Conclusion: As the sample z (= -2.828) falls inside the critical region, so reject the null hypothesis at 0.01 level of significance and conclude that μ is significantly smaller than 8 kilograms.

Example 2

The average length of time for students to register for fall classes at a certain college has been 50 minutes with a standard deviation of 10 minutes. A new registration procedure using modern computing machines is being tried. If a random sample of 12 students had an average registration time of 42 minutes with a standard deviation of 11.9 minutes under the new system, test the hypothesis that the population mean is now less than 50, using a level of significance of (1) 0.05, and (2) 0.01. Assume the population of times to be normal.

Let μ be the population mean time for students to register in the new registration procedure.

(1) Null hypothesis: $\mu = 50$ minutes

Alternative hypothesis: $\mu < 50$ minutes

Level of significance: 0.05

Critical region: (n = 12 < 30; and the new σ is unknown, so t-test should be used) degree of freedom (ν) = n -1 = 12 -1 =11

$$\therefore \text{ t} < t_{11,0.05} = -1.796$$

Computation:

n = 12 $\bar{x} = 42$ s = 11.9

$$\therefore\ t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{42 - 50}{11.9/\sqrt{12}} = -2.329$$

Conclusion: As sample t (= -2.329) falls inside the critical region, so reject the null hypothesis at 0.05 level of significance and conclude that μ is significantly smaller than 50 minutes.

(2) Identical with those of (1) except the critical region would be replaced by:

$$t < t_{11,0.01} = -2.718$$

and the corresponding conclusion would be changed as follows:

As sample t (= -2.329) falls outside the critical region, so reject the alternative hypothesis at 0.01 level of significance and conclude that μ is not highly significantly smaller than 50 minutes.

Example 3

An experiment was performed to compare the abrasive wear of two different laminated materials. Twelve pieces of material 1 were tested, by exposing each piece to a machine measuring wear. Ten pieces of material 2 were similarly tested. In each case, the depth of wear was observed. The samples of material 1 gave an average (coded) wear of 85 units with a standard deviation of 4, while the samples of material 2 gave an average of 81 and a standard deviation of 5. Test the hypothesis that the two types of material exhibit the same mean abrasive wear at the 0.10 level of significance. Assume the populations to be approximately normal with equal variances.

Let μ_1 and μ_2 be the mean abrasive wear of material 1 and 2 respectively.

Null hypothesis: $\mu_1 = \mu_2$, i.e. $\mu_1 - \mu_2 = 0$

Alternative hypothesis: $\mu_1 \neq \mu_2$, i.e. $\mu_1 - \mu_2 \neq 0$

Level of significance: 0.10

Critical region: (As both n_1 and n_2 are smaller than 30 and their standard deviations are unknown, so t-test has to be used.)

$$\nu = n_1 + n_2 - 2 = 12 + 10 - 2 = 20\ ,\ \therefore\ t > t_{20,0.05} = 1.725 \quad \text{or}$$
$$t < -t_{20,0.05} = -1.725$$

Computation:

$$n_1 = 12 \qquad \bar{x}_1 = 85 \qquad s_1 = 4$$

$$n_2 = 10 \qquad \bar{x}_2 = 81 \qquad s_2 = 5$$

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(11)4^2 + (9)5^2}{12 + 10 - 2} = 20.05$$

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{(85 - 81) - 0}{\sqrt{20.05}\sqrt{\frac{1}{12} + \frac{1}{10}}} = 2.086$$

Conclusion: As the sample t (=2.086) falls inside the critical region, so reject the null hypothesis at 0.10 level of significance and conclude that the mean abrasive wear of material 1 is significantly higher than that of the material 2.

Example 4

Five samples of a ferrous-type substance are to be used to determine if there is a difference between a laboratory chemical analysis and an X-ray fluorescence analysis of the iron content. Each sample was split into two sub-samples and the two types of analysis were applied. Following are the coded data showing the iron content analysis:

	Sample				
Analysis	1	2	3	4	5
x-ray	2.0	2.0	2.3	2.1	2.4
Chemical	2.2	1.9	2.5	2.3	2.4

Assuming the populations normal, test at the 0.05 level of significance whether the two methods of analysis give, on the average, the same result.

Let μ_1 and μ_2 be the mean iron content determined by the laboratory chemical analysis and X-ray fluorescence analysis respectively; and

μ_D be the mean of the population of differences of paired measurements.

Null hypothesis: $\mu_1 = \mu_2$ or $\mu_D = 0$

Alternative hypothesis: $\mu_1 \neq \mu_2$ or $\mu_D \neq 0$

Level of significance: 0.05

Critical region: (As n = 5 < 30, so t-test should be used.)

$$\therefore\ t > t_{4,0.025} = 2.776 \quad \text{or} \quad t < -t_{4,0.025} = -2.776$$

Computation:

	Sample				
Analysis	1	2	3	4	5
x-ray	2.0	2.0	2.3	2.1	2.4
Chemical	2.2	1.9	2.5	2.3	2.4
d_i	-0.2	0.1	-0.2	-0.2	0
d_i^2	0.04	0.01	0.04	0.04	0

$$\sum_{i=1}^{5} d_i = -0.5 \qquad \sum_{i=1}^{5} d_i^2 = 0.13$$

$$\bar{d} = \frac{\sum d}{5} = \frac{-0.5}{5} = -0.1$$

$$s_d^2 = \frac{n\sum d^2 - \left(\sum d\right)^2}{n(n-1)} = \frac{(5)(0.13) - (-0.5)^2}{(5)(4)} = 0.02$$

$$t = \frac{\bar{d} - \mu_D}{s_d / \sqrt{n}} = \frac{(-0.1) - 0}{\sqrt{0.02} / \sqrt{5}} = -1.5811$$

Conclusion: As the sample t (=-1.5811) falls outside the critical region, so reject the alternative hypothesis at 0.05 level of significance and conclude that there is no significant difference in the mean iron content determined by the above two analyses.

Tests Concerning Proportions

Example 5

A manufacturing company has submitted a claim that 90% of items produced by a certain process are non-defective. An improvement in the process is being considered that they feel will lower the proportion of defective below the current 10%. In an experiment 100 items are produced with the new process and 5 are defective. Is this evidence sufficient to conclude that the method has been improved? Use a 0.05 level of significance.

Let P be the proportion of defective product in the new production process.

Null hypothesis: P = 0.1

Alternative hypothesis: P < 0.1

Level of significance: 0.05

Critical region: $Z < -z_{0.05} = -1.64$

Computation:

n = 100 x = 5 $\hat{p} = \frac{x}{n} = \frac{5}{100} = 0.05$

$$z = \frac{\hat{p} - P}{\sqrt{\frac{P(1-P)}{n}}} = \frac{0.05 - 0.1}{\sqrt{\frac{(0.1)(0.9)}{100}}} = -1.667$$

Conclusion: As the sample z (=-1.667) falls inside the critical region, so reject the null hypothesis at 0.05 level of significance and conclude that P is significantly smaller than 0.1. That is, the production method has been improved in lowering the proportion of defective below the current 10%.

Example 6

A vote is to be taken among the residents of a town and the surrounding country to determine whether a proposed chemical plant should be constructed. The construction site is within the town limits and for this reason many voters in the country feel that the proposal will pass because of the large proportion of town voters who favor the construction. To determine if there is a significant difference in the proportion of town voters and county voters favoring the proposal, a poll is taken. If 120 of 200 town voters favor the proposal and 240 of 500 county residents favor it, would you agree that the proportion of town voters favoring the proposal is higher than the proportion of county voters? Use a 0.025 level of significance.

Let P_1 and P_2 be the proportions of town voters and country voters, respectively, favouring the proposal.

Null hypothesis: $P_1 = P_2$ or $P_1 - P_2 = 0$

Alternative hypothesis: $P_1 > P_2$ or $P_1 - P_2 > 0$

Level of significance: 0.025

Critical region: $Z > z_{0.025} = 1.96$

Computation:

$$n_1 = 200 \qquad x_1 = 120 \qquad \hat{p}_1 = \frac{x_1}{n_1} = \frac{120}{200} = 0.6$$

$$n_2 = 500 \qquad x_2 = 240 \qquad \hat{p}_2 = \frac{x_2}{n_2} = \frac{240}{500} = 0.48$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{(200)(0.6) + (500)(0.48)}{200 + 500} = 0.514$$

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (P_1 - P_2)}{\sqrt{\hat{p}(1-\hat{p})\left[\frac{1}{n_1} + \frac{1}{n_2}\right]}} = \frac{(0.6 - 0.48) - 0}{\sqrt{(0.514)(0.486)\left[\frac{1}{200} + \frac{1}{500}\right]}} = 2.870$$

Conclusion: As sample z (=2.870) falls inside the critical region, so reject the null hypothesis at 0.025 level of significance and conclude that the proportion of town voters favouring the proposal is significantly larger than that of the country voters.

PART F: Chi-square Tests

There are two types of chi-square tests: goodness-of-fit test and tests for independence.

Goodness-of-fit Test

A test to determine if a population has a specified theoretical distribution. The test is based on how good a fit we have between the frequency of occurrence of observations in an observed sample and the expected frequencies obtained from the hypothesized distribution.

Theorem: A goodness-of-fit test between observed and expected frequencies is based on the quantity

$$\chi^2_{\text{test}} = \sum \frac{(O_i - E_i)^2}{E_i}$$

where χ^2_{test} is a value of the random variable whose sampling distribution is approximated very closed by the Chi-square distribution,

O_i is the observed frequency of cell i, and E_i is the expected frequency of cell i.

The number of degrees of freedom in a Chi-square goodness-of-fit test is equal to the number of cells minus the number of quantities obtained from the observed data that are used in the calculations of the expected frequencies.

For a level of significance equal to $\alpha, \chi^2_{\text{test}} > \chi^2_{\alpha}$ constitutes the critical region.
The decision criterion described here should not be used unless each of the expected frequencies is at least equal to 5.

Example 1

Consider the tossing of a die 120 times.

	Faces					
	1	2	3	4	5	6
Observed Expected	20	22	17	18	19	24

By comparing the observed frequencies with the expected frequencies, one has to decide whether the die is fair die or not.

Null hypothesis: the die is a fair die, i.e. $P(X = i) = 1/6$ for i = 1, 2, 3, 4, 5, and 6

Alternative hypothesis: the die is not a fair die

Level of significance: 0.05

Critical region: $\nu = n - 1 = 6 - 1 = 5$; $\therefore\ \chi^2 > \chi^2_{5,0.05} = 11.07$

Computation:

Expected value = $nP(X = i) = 120(\frac{1}{6}) = 20$

i	1	2	3	4	5	6
Observed (O_i)	20	22	17	18	19	24
Expected (E_i)	20	20	20	20	20	20
$O_i - E_i$	0	2	-3	-2	-1	4

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{0^2}{20} + \frac{2^2}{20} + \frac{(-3)^2}{20} + \frac{(-2)^2}{20} + \frac{(-1)^2}{20} + \frac{4^2}{20} = 1.7$$

Conclusion: As the sample χ^2 (= 1.7) falls outside the critical region, so reject the alternative hypothesis and conclude that the die is a fair die.

Example 2

The following distribution of battery lives may be approximated by the normal distribution.

Class boundaries	O_i	z-value	p-value	E_i

1.45 - 1.95	2
1.95 - 2.45	1
2.45 - 2.95	4
2.95 - 3.45	15
3.45 - 3.95	10
3.95 - 4.45	5
4.45 - 4.95	3

Chi-squared test can be applied to test whether the above frequency distribution can be approximated by a normal distribution or not.

Null hypothesis: the distribution can be approximated by a normal distribution

Alternative hypothesis: the distribution cannot be approximated by a normal distribution

Level of significance: 0.05

Critical region: $\chi^2 > \chi^2_{\nu,0.05}$ where $\nu = n-3$, and n is the number of cells.

Computation:

For finding the expected values, the mean and standard deviation of the frequency distribution have to be found first.

Class boundaries	(f_i) O_i	Class mark (x_i)
1.45 - 1.95	2	1.7
1.95 - 2.45	1	2.2
2.45 - 2.95	4	2.7
2.95 - 3.45	15	3.2
3.45 - 3.95	10	3.7
3.95 - 4.45	5	4.2
4.45 - 4.95	3	4.7

n = 40 $\quad \sum fx = 136.5 \quad \sum fx^2 = 484.75$

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{136.5}{40} = 3.4125$$

$$s = \sqrt{\frac{\sum fx^2 - \left(\sum fx\right)^2 / n}{n-1}} = \sqrt{\frac{484.75 - \left(136.5^2 / 40\right)}{39}} = 0.6969$$

$$z - value = \left[\frac{L_i - 3.4125}{0.6969} < Z < \frac{U_i - 3.4125}{0.6969} \right];$$

where L_i and U_i are the Lower and Upper Boundaries of the *i*th class.

$$p - value = \left[\frac{L_i - 3.4125}{0.6969} < Z < \frac{U_i - 3.4125}{0.6969} \right]$$

$$E_i = (40)P\left[\frac{L_i - 3.4125}{0.6969} < Z < \frac{U_i - 3.4125}{0.6969} \right]$$

Class boundaries	O_i	z-value	p-value	E_i
1.45 - 1.95	2	Z<-2.10	.0179	0.716
1.95 - 2.45	1	-2.10<Z<-1.38	.0659	2.636
2.45 - 2.95	4	-1.38<Z<-0.66	.1708	6.832
2.95 - 3.45	15	-0.66<Z<0.05	.2653	10.612
3.45 - 3.95	10	0.05<Z<0.77	.2595	10.38
3.95 - 4.45	5	0.77<Z<1.49	.1525	6.1
4.45 - 4.95	3	1.49<Z	.0681	2.724

In order to satisfy the rule that the expected value in each cell is larger than or equal to 5, we have to combine the first three classes in to one cell and the last two classes into another cell. As such, the number of cells (n) is 4.

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(7-10.184)^2}{10.184} + \frac{(15-10.612)^2}{10.612} + \frac{(10-10.38)^2}{10.38} + \frac{(8-8.824)^2}{8.824}$$
$$= 2.901$$

Conclusion: Since $\chi^2_{1,0.05} = 3.841$, so the sample χ^2 (= 2.901) falls outside the critical region. As such, reject the alternative hypothesis and conclude that the distribution of battery lives can be approximated by a normal distribution.

Test for Independence

The Chi-square test procedure can also be used to test the hypothesis of independence of two variables/attributes. The observed frequencies of two variables are entered in a two-way classification table, or contingency table.

Remark: The expected frequency of the cell in the i^{th} row and j^{th} column in the contingency table

$$E_{ij} = \frac{(\text{total of row i}) * (\text{total of column j})}{\text{grand total}}$$

The degrees of freedom for the contingency table is equal to (r – 1) (c –1) where r is the number of rows and c is the number of columns in the table.

Example 3

Suppose that we wish to study the relationship between grade point average and appearance.

	Grade Point Average				
Appearance	1	2	3	4	Totals
attractive	14 ( )	11 ( )	10 ( )	5 ( )	40
ordinary	10 ( )	16 ( )	16 ( )	14 ( )	56
unattractive	3 ( )	4 ( )	7 ( )	10 ( )	24
Totals	27	31	33	29	120

Null hypothesis: There is no relationship between grade point average and appearance. That is, the two characteristics are independent.

Alternative hypothesis: There is a relationship between grade point average and appearance. That is, the two characteristics are not independent.

Level of significance: 0.05

Critical region: $\chi^2 > \chi^2_{\nu,0.05}$, where ν = (r -1)(c - 1)

Computation:

	Grade Point Average

Appearance	1	2	3	4	Totals
attractive	14 (*9*)	11 (*10.33*)	10 (*11*)	5 (*9.67*)	40
ordinary	10 (*12.6*)	16 (14.47)	16 (*15.4*)	14 (*13.53*)	56
unattractive	3 (*5.4*)	4 (*6.2*)	7 (*6.6*)	10 (*5.8*)	24
Totals	27	31	33	29	120

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(14-9)^2}{9} + \frac{(11-10.33)^2}{10.33} + \frac{(10-11)^2}{11} + \frac{(5-9.67)^2}{9.67}$$

$$+ \frac{(10-12.6)^2}{12.6} + \frac{(16-14.47)^2}{14.47} + \frac{(16-15.4)^2}{15.4} + \frac{(14-13.53)^2}{13.53}$$

$$+ \frac{(3-5.4)^2}{5.4} + \frac{(4-6.2)^2}{6.2} + \frac{(7-6.6)^2}{6.6} + \frac{(10-5.8)^2}{5.8} = 10.818$$

Conclusion: Since $\chi^2_{6,0.05} = 12.596$, so sample χ^2 (=10.818) falls outside the critical region. So reject the alternative hypothesis and conclude that there is no evidence to support there is relationship between grade point average and appearance.

Test for Homogeneity

To test the hypothesis that several population proportions are equal.

Remark: The approach for the test of homogeneity is the same as for the test of independence of variables/attributes.

Example 4

A study of the purchase decisions for 3 stock portfolio managers, A, B, and C was conducted to compare the rates of stock purchases that resulted in profits over a time period that was less than or equal to one year. One hundred randomly selected purchases obtained for each of the managers showed the results given in the table. Do these data provide evidence of differences among the rates of successful purchases for the three portfolio managers? Test with $\alpha = 0.05$.

Result	Manager		
	A	B	C
Purchases show profit	63	71	55
Purchase do no show profit	37	29	45
Total	100	100	100

Null hypothesis: the rates of stock purchases that resulted in profit were the same for the three stock portfolio managers

Alternative: their rates were not all the same

Level of significance: 0.05

Critical region: $\nu = (2-1)(3-1) = 2$; $\therefore\ \chi^2 > \chi^2_{2,0.05} = 5.991$

Computation:

Result	Manager			
	A	B	C	Total
Purchases show profit	63 (*63*)	71 (*63*)	55 (*63*)	189
Purchase do no show profit	37 (*37*)	29 (*37*)	45 (*37*)	111
Total	100	100	100	300

$$\chi^2 = \frac{(63-63)^2}{63} + \frac{(71-63)^2}{63} + \frac{(55-63)^2}{63} + \frac{(37-37)^2}{37} + \frac{(29-37)^2}{37} + \frac{(45-37)^2}{37} = 5.49$$

Conclusion: As the sample χ^2 (= 5.49) falls outside the critical region so reject the alternative hypothesis and conclude that there is no sufficient evidence to support the rates of purchases resulted in profit of the three portfolio managers were different.

IV
Standard Probability Distributions

Suppose an event *E* can happen in *h* ways out of a total of *n* possibly equally likely ways. Then the probability of occurrence of the event is denoted by

$$p = h/n$$

and, the probability of non-occurrence of the event is denoted by

$$q = 1 - p$$

The probability scale runs from impossibility on the one hand to absolute certainty on the other. The two poles of this continuum are given the numerical values of 0 and 1 respectively. Therefore, the probability of any event occurring must lie somewhere between these two extremes. The probability of my death is absolutely certain indeed, the probability of my death next year can be determined by a study of death rates in the past:

Age Group	Deaths per 1000 per Year	Probability of Dying in Year+1
0 - 4	3.4	0.0034
5 - 9	0.3	0.0003
10 -14	0.3	0.0003
15 -19	0.9	0.0009
20 - 24	0.9	0.0009
25 - 34	0.9	0.0009
35 - 44	2.0	0.0020
45 - 54	6.6	0.0066
55 - 64	18.7	0.0187
65 - 74	47.6	0.0476
75 - 84	110.3	0.1103
> 85	234.4	0.2344

However, the above definition of probability is circular because we are essentially defining probability in term of itself: "equally likely" appears to be synonymous with "equally probable". For this reason a statistical definition of probability is advocated: the estimated probability, or empirical probability, of an event is the *relative frequency* of occurrence of the event when the number of observations is very large. The probability itself is the *limit* of the relative frequency as the number of observations increase indefinitely.

Population

In collecting data concerning characteristics of a group of individuals, such as attitudes, aptitudes and personality dimensions of university students, it is often impossible or impractical to observe the entire group, especially if it is large. Instead of examining the entire group, called the *population* or universe, one examines a small part of the group called a sample. A population may be finite or infinite. A finite population is defined as an entity with a boundary around it, this boundary may be one of sex, time, location or other attribute determined by the researcher; for example, the population of students may be defined as: all students registered with the University of Newcastle upon Tyne as at 31 March 2001. Whereas an infinite population is one consisting of all possible outcomes of an event (successive tosses of a coin or guessing the colour of a card is infinite).

Sampling Theory and Sampling

Sampling theory is the study of relationships existing between a population and samples drawn from the population. It is of value in estimating of unknown population parameters (such as population arithmetic mean, variance, etc.) from knowledge of corresponding sample statistics. Sampling theory is also useful in determining whether observed differences between two samples are actually due to chance variation or whether they are really significant. The answers involve use of so-called *tests of significance and hypotheses*. In general, a study of inferences made concerning a population by use of samples drawn from it, together with indications of the accuracy of such inferences using probability theory, is called statistical inference.

A sample is a set of individuals or objects selected from a population. The purpose of sampling is to infer a characteristic or characteristics of a given population from a subset of measurements obtaining from the sample. In a quantitative research design, the sampling methods and the sample design are crucial to obtain unbiased and consistent estimates of the characteristic population inferred from the sample; and, enable one to statistically validate any conclusions based on the observations and findings. In order that conclusions of sampling theory and statistical inference be valid, samples must be chosen so as to be representative of the population. A study of methods of sampling and the related problems, which arise, is called the *design of the experiment*. One

way in which a representative sample may be obtained is by a process of random sampling, according to which each member of the population has an equal chance of being included in the sample. One technique for obtaining a random sample is to assign a number to each member of the population, write each number on a separate piece of paper, place them in a container and then draw numbers from the container, being careful to mix thoroughly before each drawing. Alternatively, this can be replaced by using a table of random numbers specially constructed for such purposes.[2]

How well a sample represents a given population depends on the sample frame, the sample size and the specific sample design:

- Sample Frame: the set of subjects who have a chance of being selected from the study population, given the sampling approach chosen
- Sample Design: the specific procedures to be used for selecting the subjects in the sample
- Sample Size: the planning of, and reasons for choosing, the number of subjects in the sample.

Sample Frame

A sample frame is a complete list of individuals or units in the population to be studied. An initial step in sampling is to provide a *clear and accurate definition* of the population exposed to your study. This study population may comprise of a group of individuals who go somewhere or do something that enables them to be sampled. Sampling is carried out in two stages; the first involves sampling something other than the individuals to be finally selected, the second entails creating a list of sampling units (individuals) from which a final selection is made. Whilst the sample frame should be representative of the study population, it may be necessary to control mediating or intervening variables. For instance, suppose I wish to ascertain the attitude of workers, employed by the Kenkei Electronics Company, toward some object or subject. A number of workers in the study population will have been recently employed by the company, and may retain beliefs, opinions and attitudes from previous employers. To eliminate these from the study, the sampling frame may be constructed to include workers with at least one year's service with the Kenkei Electronics Company.

[2] Random numbers can be generated from scientific calculators, or be extracted from statistical tables.

Sample Design

The sample design influences the *precision* of the subset of measurements obtained from the sample. Sampling strategies can be categorized as random probability, non-random probability and mixed sampling. A random probability sample design provides for an equal and independent chance of selection, i.e. each individual has a known probability of selection set by the sampling procedure. Non-random probability sample designs are used when the population size is either unknown or cannot be discretely identified. A mixed sampling design has characteristics of both random and non-random probability sampling. Systematic sampling is the most common strategy used, based on a known population with the sample size determined *a priori.* The advantages of systematic sampling over simple random sampling is that it is more efficient, in terms of information per respondent cost, and easier to perform thus reducing sampling error. Systematic sampling will generally be an adequate form of random sampling to the extent that the placing of any sampling subject or unit is independent of the placing of other sampling individuals or units, i.e. there is no systematic bias introduced into the listing of sampling units. Should such a risk of sampling bias be known, it may be avoided by choosing a suitable sampling interval; or, after a predetermined number of units have been drawn, a fresh random start can be made.[3]

Sample Size

How big should my sample be? It all depends on what you want to do with your findings and the type of relationship you want to establish in your study. The sample size is crucially important in a correlational research design, i.e. tests of hypotheses and significance, or establishing an association or relationship between two of more variables. However, there is no relationship between the sample size and the size of the study population, sample size is determined by the variability of the factor, element or characteristic prevalent in the study population. Other things being equal, precision increases steadily up to sample sizes of 50-200; after that, there is only a modest gain in precision to increasing sample size. Whilst increasing sample size reduces errors attributable to sampling, managing large amounts of data may increase non-sampling errors (e.g. field-work problems, interviewer-induced bias, clerical errors in transcribing data, etc.). In determining the size of a sample, consideration should be given to:

- the degree of accuracy required in the estimation of the variables in the chosen study population
- the level of confidence demanded from the sample to test the significance of the findings or hypotheses

[3] See Jakobowitz, K., (1999), *How to Develop and Write a Business Research Proposal*, pp.58-74, ISBN 0-9538664-0-8

- the extent to which the variability of the factor in the chosen study population is known, or can be estimated.[4]

You have developed a survey instrument and, after a pilot-study of ten subjects, have determined a sample estimate of mean to be 52 with a standard deviation of 15. Having decided *a priori* that the sample of the population needs to be sufficient to allow for 95 percent confidence in your findings and conclusions, it is necessary to compute the required sample size:

Computation

The standard deviation of the study population is unknown, hence the normal distribution cannot be used but a derivative, Student's *t*-distribution, can be applied:

$$n_s = \left[t_{0.025}s/L\right]^2$$

where:

n_s = sample size

t = confidence limit value derived from statistical tables for v = n – 2 degrees of freedom, and $t_{\alpha/2}$ ($t_{-0.025} - t_{+0.025}$)

s = sample standard deviation obtained from the pilot-study

L = level of tolerance or error (say, 5 points on the instrument scale of 100)

Hence:

$$n_s = [2.306(15)/5]^2 = \underline{47.858724} \text{ (say, 48)}$$

Note The normal distribution provides a sample size of $[1.96(15)/5]^2 = 35$; and, using an estimate of the population standard deviation [(n/n-1)/s] provides a sample size of $[1.96(16.667)/5]^2 = 43$

[4] It is rare indeed to have a known population variance and so be able to determine the population standard deviation. Hence, deviates of the normal distribution cannot be applied, and a derivative such as Student's t-test should always be used for an unknown standard deviation of the study population.

Sampling Error

One purpose of using random probability sampling methods is to apply a variety of statistical techniques to estimate the confidence one can have that the characteristics of a sample accurately reflect the study population. Sampling error is a random product of sampling. However, when random sampling methods are used it is possible to compute how much the sample-based estimate of the characteristic will vary from the study population by chance because of sampling. The larger the sample size and the less variability of the characteristic being measured, the more accurate a sample-based estimate will be. Sampling error can be defined as the variation around the true population value that results from random sample differences drawn from the population.

The *standard error of mean* is the most commonly used statistic to describe sampling error[5]:

$$SE = [s^2/n]^{0.5}$$

Where: s^2 is the variance derived from the sample
n is the sample size
and $[sum]^{0.5}$ is the square root of the product

Alternatively, the standard error of mean is more easily computed from a proportion statement, since the variance of a proportion is expressed as p[1-p]: the standard error of mean of a proportion is computed from: $[p(1-p)/n]^{0.5}$.

In a survey of job satisfaction, a pilot study of 21 subjects provided a sample mean of 136.81 with a sample standard deviation of 27.42. Estimate the true population mean from the pilot-study.

Computation

The standard error of mean = $[s^2/n]^{0.5} = [(27.42)^2/21]^{0.5} = 5.98$

Conclusion

We can be 67 percent certain that the true population mean lies $\pm$ one standard error from the sample mean, i.e. 136.81 $\pm$ 5.98, or between 130.83 and 142.79 on the survey instrument scale.

Also, we can be 95 percent certain that the true population mean lies $\pm$ two standard errors from the sample mean, i.e. 136.81 $\pm$ 11.96, or between 124.85 and 148.77 on the survey instrument scale.

Source: Jackson (1989)

[5] For a normal, or approximately asymmetrical distribution, the 95 percent confidence limits of the population true mean can be computed: sample mean $\pm$ $1.96s/(n^{0.5})$.

In a study to estimate a characteristic in a population of 1000 subjects, a sample of 50 is chosen at random. In responding to a question there is only two dichotomous values: 0 (No) and 1 (Yes). In the sample 15 respondents say 'yes' and 35 say 'no'. Estimate the standard error.

The mean of the sample response is: $\Sigma x/n = (15/50) = 0.3$ (in other words, a proportion of thirty percent of the sample claim the characteristic).

The standard error of a proportion = $[p(1-p)/n]^{0.5} = [0.3(1-0.3)/50]^{0.5} = 0.0648$

Thus, we can be 67 percent certain (i.e. $\pm$ one standard deviation from the mean) that the value of the true population characteristic lies between one standard error of the sample mean: 0.3 $\pm$ 0.0648, or between 23.52 and 36.48 percent of the study population. Also, we can be 95 percent confident that the true population mean lies $\pm$ two standard errors of the sample mean: 0.03 $\pm$ 0.1296, or between 17.04 and 42.96 percent of the study population.

Summary of Sampling Distribution Formulae

1. Calculation of Confidence Limits for an Unknown Percentage Mean

μ lies between the sample mean percentage $\pm z[p(100-p)/n]^{0.5}$

where,
- n = sample size
- p = sample estimate of the mean
- z_1 = 1.96 for 95 percent confidence limits
- z_2 = 2.58 for 99 percent confidence limits

Out of a random sample of 1000 subjects, 500 say they have a positive attitude towards the object or subject of the study. What conclusions can we draw about the percentage of the total population who may hold this view?

Our sample estimate (p) is, in effect, a random selection from a normal distribution (n>30) with mean = π and a standard error (= standard deviation) of $[\pi (100 - \pi)/n]^{0.5}$. Accordingly we can state with 95 percent confidence that the true population mean lies between:

$$\pi \pm [1.96[\pi(100-\pi)/n]]^{0.5} = 50 \text{ percent} \pm [1.96[50(100-50)/1000] = 3.1 \text{ percent}$$

Conclusion We can state with 95 percent confidence that the true proportion of people who hold this view is within the range of 46.9 to 53.1 percent

2. Calculating the Sample Size for Established Accuracy from a Large Pilot-Study

$$n_s = 4\pi(100-\pi)/L^2$$

where n_s = required sample size
π = sample proportion estimate of the mean
L = specified percentage of accuracy

In order to evaluate the success of an attitudinal change programme, a company interviewed 400 workers exposed to the change programme. Some 120 workers displayed evidence of the required positive attitudinal change. How accurately does this survey reflect the percentage of workers who have undergone a positive attitudinal change? How many more workers must be surveyed in order to establish this percentage with $\pm$ 2 percent accuracy with 95 percent confidence?

In the survey, $n_s = 400$ and $\pi = (120/400) = 30$ percent and provides the computed accuracy at the 95 percent confidence limits:

$$\mu \text{ lies between } \pi \pm 1.96\left[\pi(100-\pi)/n\right]^{0.5}$$
$$= 30 \text{ percent} \pm 1.96\left[30(100-30)/400\right]^{0.5}$$
$$= 30 \text{ percent} \pm 4.6 \text{ percent}$$

To compute the sample size necessary to reduce the standard error of $\pm$ 4.6 percent to one of $\pm$ 2 percent, with L = 2 and $\pi = 30$:

$$n_s = 4\pi(100-\pi)/L^2$$
$$= 4(30)(100-30)/2^2$$
$$= 2100$$

Conclusion In order to be establish that 30 percent of workers have undergone a positive attitudinal change, within $\pm$ 2 percent accuracy, we need a sample of 2100 workers to be 95 percent confident.

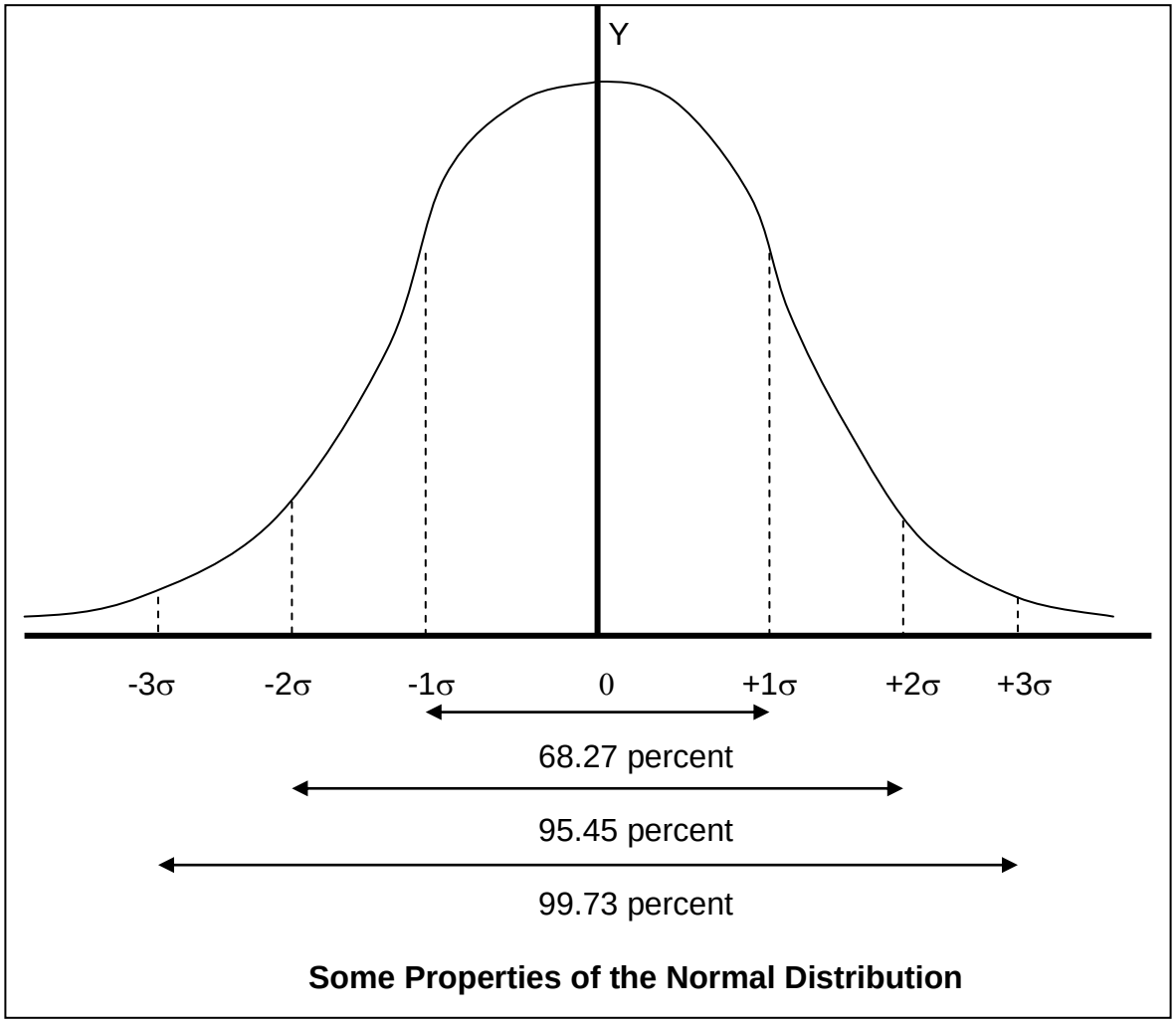
V
The Normal Distribution

The most important continuous probability distribution in the field of statistics is the normal distribution:

- empirical data is often normally, or approximately normally, distributed
- the assumption of normality allows for the application of powerful statistical analyses
- the distribution of many sample statistics tends to normality as the sample size increases (>30)
- many population distributions can be readily transformed to normality

The properties of the normal distribution are:

- observations tend to cluster at the mean: mean = mode = median
- the distribution of observations is symmetrical about the vertical axis through the mean
- the total area under the curve is equal to unity, i.e. 1.0000
- the normal curve continues to decrease in height as one proceeds in either direction away from the mean, but never reaches the horizontal axis, i.e. there is a presumption of negative and positive infinity
- the area under the curve between two ordinates, $X = a$ and $X = b$ where $a < b$, represents the probability that X lies between a and b and can be expressed by the probability of $a < X < b$
- when the variable X is expressed in standard units, $z = (X - \mu)/\sigma$



Some Properties of the Normal Distribution

How to Find the Area of the Standardized Normal Distribution

Statistical tables by White, Yeats and Skipworth (1991) and Murdoch and Barnes (1985) of the areas of the standardized normal distribution give the probability that a random variable will be greater than the mean (U), i.e. the area in the tail. Hence, for the z = 1.2 we are given a value of 0.11507, but this represents the probability of a random sample being greater than the mean, i.e. the probability of being in the tail of the distribution. To find the probability of a random sample falling in the acceptance region (z = 0 to z = 1.2), we must subtract this value from 0.5:

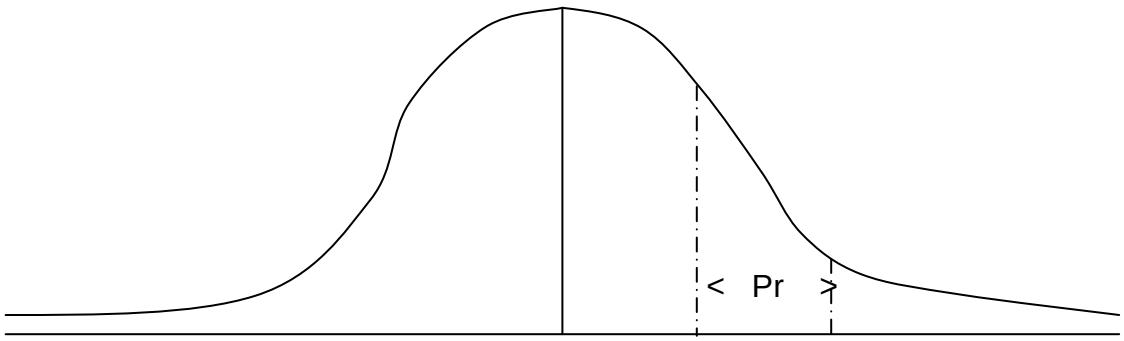
Probability (z = 0 to z = 1.2)	=	0.50000 – 0.11507
	=	0.38493 (say, 38.5 percent)

The intelligence quotient of a sample of 5000 British adults provided a mean of 100 with a standard deviation of 15. Assuming that intelligence quotient is normally distributed, how many adults would we expect to have an IQ between 120 and 130?

Computation

The intelligence quotient recorded as 120 and 130 can actually have a value from 119.5 and 130.5 assuming that the IQ is recorded to the nearest unit of measurement. Hence:

119.5 in standard units	=	$(x-\mu)/\sigma$	=	(119.5-100)/15
			=	+1.3000
150.5 in standard units	=	$(x-\mu)/\sigma$	=	(130.5-100)/15
			=	+2.0333



Intelligence Quotient:	100	120	130
Standard Score (z):	0	+1.30	+2.03

The required proportion of standard units = (area between z = 1.3000) + (z = 2.0333). Using White, Yeats and Skipworth (1991), *Tables for Statisticians*, pp.14:

Area between z = 0 to z = 1.3000 . . .	. . .	=	0.09680
Area between z = 0 to z = 2.0333 . . .	. . .	=	0.02118
Subtract:			0.07562

Conclusion

The number of British adults in the sample having an IQ of between 120 and 130 = 5000 (0.07562) = 378.1 (say, 378)

VI
Statistical Decision Theory: Tests of Hypothesis and Significance

Very often in practice we are called upon to make decisions about populations on the basis of sample information. Such decisions are called *statistical decisions*. For example, we may wish to decide on the basis of sample data whether one psychological procedure is better than another, whether the findings from survey data are representative of the population, whether the conclusions reached as to an experiment are valid, etc.

What is a Hypothesis?

A hypothesis refers to conjectures that can be used to explain observations; a hunch, or educated guess, which is advanced for the purpose of being tested. A hypothesis provides us with a guiding idea to determine what was relevant and what was irrelevant. This mode of accounting for problems possesses three steps:

1. the proposal of a hypothesis to account for a phenomenon

2. the deduction from the hypothesis that certain phenomena should be observed in given circumstances

3. the checking of this deduction by observation.

Example

We may reason that a deprived family background causes low reading attainment in children. We may then produce empirical evidence that low family income and overcrowding are associated with poor reading attainment. If no such evidence was forthcoming, then the hypothesis must be decisively rejected. But if the predicted relationship was found, could we conclude that the hypothesis was correct, i.e. poor family background does cause low reading attainment? The answer must be

'no'. It might equally be the case that a low level of school resources is also to blame. The main point to note is that the scientific process never leads to *certainty* in explanation, only to the rejection of existing hypotheses and the construction of new ones.

How do we Formulate Hypotheses?

Generally, a hypothesis is derived from theory or the literature on the problem. But regardless of source, a hypothesis must meet one criterion: *it must be stated in such a way that it can be confirmed or refuted.* In other words, it must be amenable to test.

Example "*Is a student's authoritarian behaviour directly related to his or her attitudes concerning punishment received from his or her parents*?"

Although the statement of purpose is quite detailed, it conveys little information unless the conceptual propositions are detailed: what do we mean by *authoritarian behaviour*, attitudes *concerning punishment* and *received*? Although most individuals may know the meanings, they lack scientific precision.

Once the conceptual propositions have been established (i.e., the meanings in scientific terms) we then need an operational proposition that defines the concepts in such a way that they can be observed and measured. This may be derived from a score achieved on a particular scale of authoritarianism; indirectly, we may study the relationship between childhood aggression and exposure to violent television programmes, but we still need to define both the variables under study – *aggression* and *television violence* – in operational terms. The former might be simply a tally of aggressive acts such as hitting, fighting, damaging property, etc. Or it might be based on the analysis of projective test material (Thematic Apperception Test). A panel of judges may be used to develop an operational definition of aggression by watching a child in a free-play situation and then rate the child's aggressiveness on a five-point scale. Alternatively, we could observe children as they play with a selection of toys we had previously classified as aggressive (guns, tanks, knives, etc,) and toys classified as non-aggressive (cars, dolls, puzzles, etc.).

Defining violence may be a little more difficult to agree on. What constitutes television violence? The problem here is both cultural and the difference in precision between what the general public will accept in defining a term and what researchers will accept. To operationalize the concept of television violence we could use a checklist of items, such as "Was there physical contact of an aggressive nature?" "Has an illegal act taken place?" etc. Perhaps you can establish a criterion that a violent television programme must have five or more items checked 'yes' for it to be considered violent.

From the General to the Operational

Problem or General Hypothesis: You expect some children to read better than others because they come from homes, in which there are positive values and attitudes to education.

Research Hypothesis: Reading ability in nine-year-old children is related to parental attitudes towards education.

Operational Hypothesis: There is a significant relationship between reading ability for nine-year-old children living in Carlisle as measured by standardized reading test (state test) and parental attitudes to education as measured by the attitudinal scale derived from test (state test).

Criteria for Judging Hypotheses

1. Hypotheses should be clearly stated. General terms such as *personality, self-esteem, moral fibre,* etc. should be avoided: the statement demands concise, scientific definition of the concepts and terms:
   "Personality as measured by the Eysenck's Personality Inventory . . ."

2. Hypotheses predict an outcome, an obvious necessity is that instrument exists to provide valid and reliable measures of the variables involved.
3. Hypotheses should state differences or relationships between variables. A satisfactory hypothesis is one in which the expected relationship is made explicit.

4. Hypotheses should be limited in scope. Hypotheses of global significance are not required. Those that are specific and relatively simple to test are preferable.

5. Hypotheses should not be inconsistent with known facts. All hypotheses should be grounded in past knowledge. The hypothesis should not lead the cynical reader to say: "Whatever led you to expect that?" or "You made this one up after you collected the data."

Unconfirmed Hypotheses

Does an unconfirmed hypothesis invalidate prior knowledge or the literature? Well, either the hypothesis is false, or some of the previous information is erroneous, or other information has been overlooked, or some information may have been misinterpreted b the researcher, or the experimental design might have been incorrect. A new hypothesis may need to be formulated and tested using a different study – scientific progress in developing alternative paradigms! Even if the hypothesis is refuted, knowledge is advanced

Statistical Hypotheses

In attempting to reach decisions, it is useful to make assumptions about the population involved in the study. Such assumptions, which may or may not be true, are called *statistical hypotheses* and in general are statement about the probability distributions of the population. In many instances we formulate a statistical hypothesis for the sole purpose of rejecting or nullifying it. For example, if we want to decide whether one psychological procedure is better than another, we formulate the hypothesis that there is *no significant difference* between the procedures (i.e. any observed differences are merely due to sampling error from the *same* population). Such hypotheses are called *null hypotheses* and are denoted by the symbol H_0. Any hypothesis that differs from a given hypothesis is called an *alternative hypothesis.* A hypothesis alternative to the null hypothesis is denoted H_1.

Tests of Hypotheses and Significance

If on the supposition that a particular hypothesis is true we find that results observed in a random sample differ markedly from those expected under the hypothesis. On the basis of chance using sampling theory, we would say that the observed differences are *significant* and we would be inclined to reject the hypothesis (or at least not accept it on the basis of the evidence obtained). Procedures which enable us to decide whether to accept or reject hypotheses or to determine whether observed samples differ significantly from the expected results are called *tests of hypotheses, tests of significance* or *rules of decision*.

Type I and Type II Errors

If we reject a hypothesis when it should be accepted, we say a *Type I error* has been made. On the other hand, if we accept a hypothesis when it should be rejected, we say that a *Type II error* has been made. In either case a wrong decision has been made or an error of judgement has occurred. In order for any tests of hypotheses to be sound, they must be designed so as to minimize these errors of judgement. The probability of making a Type I error is determined *before* the experiment. A Type I error can be limited by properly choosing a level of significance (α): a level of significance of 0.05 implies a 5 percent chance of making the wrong decision. The probability of making a Type II error cannot be determined before the experiment. It is possible to avoid risking a Type II error altogether by simply not making them, which amounts to never accepting hypotheses. However, for any given sample size, an attempt to reduce one type of error is often accompanied by an increase in the other type of error. The only way to reduce both types of error is to increase the sample size, which may or may not be possible.

One-Tailed and Two-Tailed Tests

Often we display interest in extreme values of the statistic *X* or its corresponding standard *z* score on both sides of the mean, i.e. in both "tails" of the distribution. For this reason such tests are called *two-tailed tests.* In such tests the area covered by critical regions in both tails is equal to the level of significance, i.e. $\alpha/2$. However, we may be interested only in extreme values to one side of the mean, for example when we are testing the hypothesis that one procedure is better than another. Such tests are called *one-tailed tests* and the critical region to one side of the distribution has an area equal to the level of significance. The critical values of *z* for levels of significance are found by use of tables of the normal distribution:

Level of significance α	0.10	0.05	0.01	0.005	0.002
Critical values of *z* for one-tailed test	-1.28 *or* +1.28	-1.64 *or* +1.64	-2.33 *or* +2.33	-2.58 *or* +2.58	-2.88 *or* +2.88
Critical values of *z* for two-tailed test	-1.645 *and* +1.645	-1.96 *and* +1.96	-2.58 *and* +2.58	-2.81 *and* +2.81	-3.08 *and* +3.08

Tests Involving the Normal Distribution

Suppose that under a given hypothesis the sampling distribution of a statistic *X* is normally distributed with mean μ and a standard deviation σ, then the distribution of the standardized score (*z*) is given by $(X - \mu)/\sigma$.

In an experiment on extra-sensory perception a subject in one room is asked to state the colour (red or blue) of a card chosen from a deck of 50 well-shuffled cards by an individual in another room. It is unknown to the subject how many red or blue cards are in the deck. If the subject identified 35 cards correctly, determine whether the results are significant at the 0.05 level of significance.

Step 1

A statistical or Null Hypothesis is set up

This is the initial assumption is almost invariably an assumption about the value of a population parameter: the probability (p) of the subject choosing the colour of the card correctly is 0.5, hence the subject is merely guessing and the experimental results are due to chance:

$$H_0 \quad : \quad p = 0.5$$

Step 2

An Alternative Hypothesis is defined that is accepted if the test permits us to reject the null hypothesis

The subject is not guessing and the experimental results are indicative of the subject having powers of extra-sensory perception:

$$H_1 \quad : \quad p > 0.5$$

Step 3

An appropriate level of significance (α) is established

By convention, a level of significance of 0.05 (or 5 percent) is sufficient in most situations. However, should the consequences of wrongly rejecting the null hypothesis (H_0) be serious enough to warrant it, a 0.01 level of significance can be applied. We choose a one-tailed test, since we are not interested in the ability to obtain extremely low scores but rather the ability to obtain high scores:

$$\alpha \quad = \quad 0.05$$

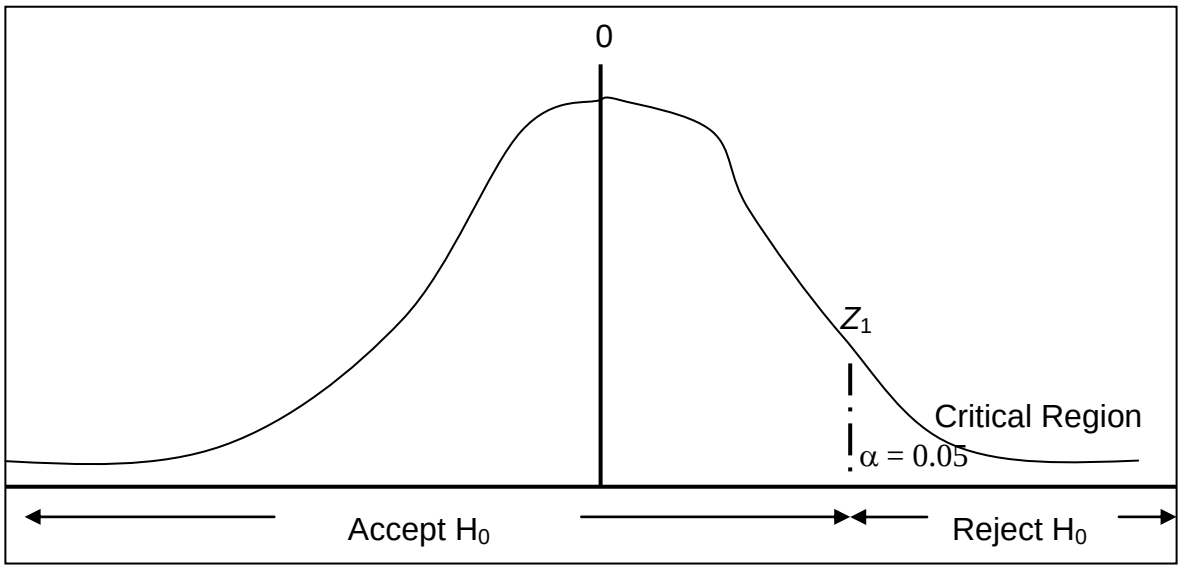
Step 4

The appropriate sampling distribution is defined and the critical values established to identify the accept/reject regions

If the null hypothesis (H_0) is true, the mean and standard deviation of the number of cards identified correctly is given by:

μ = (number of cards in the pack * probability) = Np = 50(0.5) = 25

$\sigma = [Np(1-p)]^{0.5} = [50(0.5)(1-0.5)]^{0.5} = (12.5)^{0.5} = 3.54$



Step 5

The decision rule is established

For a one-tailed test at a level of significance of 0.05 we must choose z_1 so that the area of the critical region of high scores is 0.05. Then the area between 0 and z_1 = 0.4500, and z_1 = 1.645 (read from statistical tables). Thus our decision rule or test of significance is:

(i) If the *z* score observed is greater than 1.645, the results are significant at the 0.05 level and the subject has powers of extra-sensory perception: accept H_0

(ii) If the *z* score is less than 1.645 the results are due to chance, i.e. not significant at the 0.05 level, and reject H_0.

Step 6

The position of the Sample result is computed

If the null hypothesis is valid, the sampling distribution will have a mean of $\mu = Np = 50(0.5) = 25$, and a standard deviation of $\sigma = [(Np(1-p)]^{0.5} = 50(0.5)(0.5) = 3.54$

However, we need to apply a continuity correction, since 32 on a continuous scale is between 31.5 and 32.5, hence

$$z = (X-\mu)/\sigma = (31.5-25)/3.54 = 1.84$$

Step 7

Conclusion

Since 32 in standard score = 1.84, which is greater than 1.645, decision (i) holds, i.e. we accept H_0 : we conclude that the subject has powers of extra-sensory perception. It does not follow from this that the null hypothesis is true, it merely means that there is insufficient evidence to reject it (this is equivalent to a "not proven" verdict).

Tests Involving Differences in Means and Proportions

Let m_1 and m_2 be the sample means obtained from a large sample of sizes n_1 and n_2 drawn from respective populations having means μ_1 and μ_2 and standard deviations σ_1 and σ_2. Consider the null hypothesis that there is *no difference* between the population means, i.e. $\mu_1 = \mu_2$ or the two samples are drawn from populations having the same mean.

An aptitude test was given to two groups of people consisting of 40 and 50 subjects respectively. In the first group the mean score was 74 with a standard deviation of 8, while in the second group the mean score was 78 with a standard deviation of 7. Is there a significant difference between the performance of the two groups at the 0.05 level of significance?

Suppose the two groups come from two populations having the respective means of μ_1 and μ_2. Then we have to decide between the hypotheses:

H_0 : $\mu_1 = \mu_2$ and the difference is merely due to chance

H_1 : $\mu_1 < \mu_2$, or $\mu_1 > \mu_2$ and there is a significant difference between the two groups

Under the null hypothesis it is assumed that both groups come from the same population. Since the population standard deviation is not known, an *estimate* of the population standard deviation is computed by pooling the two sample standard deviations:

$\sigma_{est} = [(\sigma_1^2/n_1) + (\sigma_2^2/n_2)]^{0.5} = [(8^2/40) + (7^2/50)]^{0.5} = 1.606$

Then $z = (m_1 - m_2)/\sigma_{est} = (74 - 78)/1.606 = -2.49$

For a two-tailed test the results are significant at the 0.05 level if *z* lies outside the range –1.96 to +1.96. Hence we conclude that at a 0.05 level there is a significant difference in performance of the two groups and that the second group is *probably* better.

Two groups, A and B, consist of 100 subjects each who have a phobia towards spiders. A psychoanalytical technique is administered to group A but not group B (which is called the control group); otherwise the two groups are treated identically. It is found that in groups A and B, 75 and 65 subjects, respectively, manage the phobia effectively. Test the hypothesis that the psychoanalytical technique is effective using a 0.05 level of significance.

Let p_1 and p_2 denote respectively the population proportions effectively managing their phobia (a) using psychoanalysis, (b) not using psychoanalysis. We must decide between two hypotheses:

H_0 : $p_1 = p_2$ and observed differences are due to chance i.e. psychoanalysis is ineffective

H_1 : $p_1 > p_2$ and psychoanalysis is effective

Under the hypothesis H_0 the average proportion of success is given by:

$\mu = (p_1 + p_2/n_1 + n_2) = (75+65)/(100+100) = 70$

and an estimate of the population standard deviation is given by:

$\sigma_{est} = [p(1-p)(1/n_1 + 1/n_2)]^{0.5} = [(0.70)(0.30)(1/100+1/100)]^{0.5}$

$= 0.0648$

Then $z = (p_1 - p_2)/\sigma_{est} = (0.75 - 0.65)/0.0648$

$= +1.54$

On the basis of a one-tailed test at a 0.05 level of significance, we would reject the null hypothesis only if the *z* score were greater than 1.645. Since the *z* score is 1.54, we must conclude that the results are due to chance at this level of significance, and that the psychoanalytical technique is no more effective in managing arachnophobia than a placebo.

It should be noted that our conclusions depend on how much we are willing to risk being wrong. If the results are actually due to chance and we conclude that the psychoanalytical treatment is effective (Type I error), we might proceed to treat large numbers of people only to find then that it is ineffective. However, in concluding that the psychoanalytical technique does not help when it actually does (Type II error), may be a dangerous conclusion especially if people's well being is at stake.

Associated Normal Approximations

The Binomial Distribution

The binomial distribution is a discrete probability of an event occurring in any single trial (called the probability of success). Probabilities associated with the binomial distribution are readily available from the formula:

$$p(X) = \left[n!/x!(n-x)!\right] p^{x}(1-p)^{n-x}$$

When n is large the computation becomes extremely tedious to operate. However, if n > 50 and p > 0.1 and np > 5 the normal distribution can be used to approximate the binomial:

It has been decided to administer a multiple-choice questionnaire to the first year candidates studying for a degree in psychology. The examination will comprise of 100 questions, and the candidates are asked to choose the correct answer from four possible answers for each question. However, some academic colleagues, who claim that a pass score of 40 could be easily achieved by guessing, greet this departure from the traditional examination with skepticism. Test this claim at the 0.05 level of significance.

In this case n = 100 questions and the probability of successfully answering each question is 1 in 4 = 0.25. By substitution:

μ	=	np	=	100(0.25)	=	25
σ	=	$[np(1-p)]^{0.5}$	=	$[(100)(0.25)(0.75)]^{0.5}$	=	4.33

Then $z = (x - \mu)/\sigma = (40 – 25)/4.33 = 3.364$

From tables of the normal distribution we can determine the probability of a candidate deriving a standard score:

$z = +3.364 = 0.00039$ (or, less than 0.04 percent).

We can conclude that the multiple-choice questionnaire is reliable if we are willing to accept one candidate in 2500 to obtain 40 or more correct answers (and obtain a pass grade) simply by guessing.

VII
Small Sampling Theory

For samples of size n>30, called large samples, the sampling distributions of many statistics are approximately normal. In such cases we can use the statistic *z* to formulate decision rules or tests of hypotheses and significance. In each case the results hold good for infinite populations. In a test of means in sampling without replacement from finite populations, the *z* score is given by:

$$z = \frac{m - \mu}{\sigma/(n-1)^{0.5}}$$

where
- z — the statistic for the normal distribution
- m — sample mean
- μ — population mean
- σ — population standard deviation
- n — sample size
- and, $\sigma/(n-1)^{0.5}$ is the standard error

However, for sample of size n<30, called small samples, this particular approximation is not good and becomes worse with decreasing sample size. Hence, exact sampling with appropriate modifications must be made. If we consider sample of size *n* drawn from a normal, or assumed normal, distribution with a population mean of μ, using the sample mean *m* and sample standard deviation *s*, the sampling distribution for the *t* statistic can be expressed:

$$t = \frac{m - \mu}{s/(n-1)^{0.5}}$$

$$= [(m - \mu)/s]\, n\text{-}1^{0.5}$$

where
- t — the statistic for Student's t-distribution
- m — sample mean
- μ — population mean
- s — sample standard deviation
- n — sample size

Tests of hypotheses and significance are easily extended to problems involving small samples, the only difference is that the *z* statistic is replaced by the more suitable *t* statistic.

Degrees of Freedom

In order to compute the statistic or population parameter such as an estimate of the population standard deviation, it is necessary to use observations obtained from a sample. The *number of degrees of freedom* of a statistic generally denoted by ν is defined as the number of independent observations in the sample (i.e. the sample size = n) minus the number of population parameters (k) which must be estimated from sample observations. In other words $\nu = n - k$.

Example From a sample of ten observations we have derived the sample mean and sample standard deviation, *m* and *s* respectively, in order to compute an estimate of the population mean μ. However, since we must estimate the population parameter μ, k = 1 and so $\nu = (n - 1) = (10 - 1) = 9$. The degrees of freedom (ν) is used to derive the value of t from the tables of the t distribution: the confidence limits for α = 0.4 to 0.0005 is read horizontally at the top of the table, the degrees of freedom ν = 1 to >120 is read vertically. Hence, for a one-tailed test at the 0.05 level of significance for ten observations we find $t_{0.05}$ for (10-1) = 9 degrees of freedom to be 1.833; and, for a two-tailed test, we find $t_{0.025}$ for (10-1) = 9 degrees of freedom to be 2.262

Confidence Limits

In a similar manner to the normal distribution, we can define 95 percent or other confidence limits by using the table of the *t* distribution. Likewise, we can estimate within specified limits of confidence the population mean μ. For example, for a 95 percent confidence interval the values are $\pm t_{0.975}$ (i.e. we have 0.025 of the area in each tail of the *t* distribution):

$$\mu_{est} = \text{sample mean} \pm t_{calc}\, s/(n-1)^{0.5}$$

A survey sample of 10 subjects gives a mean of 30 with a standard deviation of 4. Estimate the population mean with 95 percent confidence.

1. We are dealing with a small sample of n<30 and the population standard deviation σ is unknown. Hence, the sample mean of 30 is a random selection from a *t* distribution with a population standard deviation = standard error = $s/n^{0.5}$. Thus, we can state that the sample mean is no more than $t_{0.025}$ standard errors from the population mean:

$$\mu \text{ lies between } 30 \pm t_{0.025}\, s/n^{0.5}$$

2. From tables of the *t* distribution the value for $t_{0.025}$ based on (n-1) = 8 degrees of freedom = $\pm$ 2.306. Hence, substituting in the formula:

$$\mu \text{ lies between } 30 \pm 2.306(4/9^{0.5}) = 30 \pm 3.07$$

3. We can conclude that we are 95 percent confident that the population mean score lies between 26.93 and 33.07

One- or Two-Tailed Tests

To test the hypothesis that the observed sample mean (m) is equal to the population mean (μ), or that the observed mean (m_1) from one sample is equal to the observed mean of another sample (m_2):

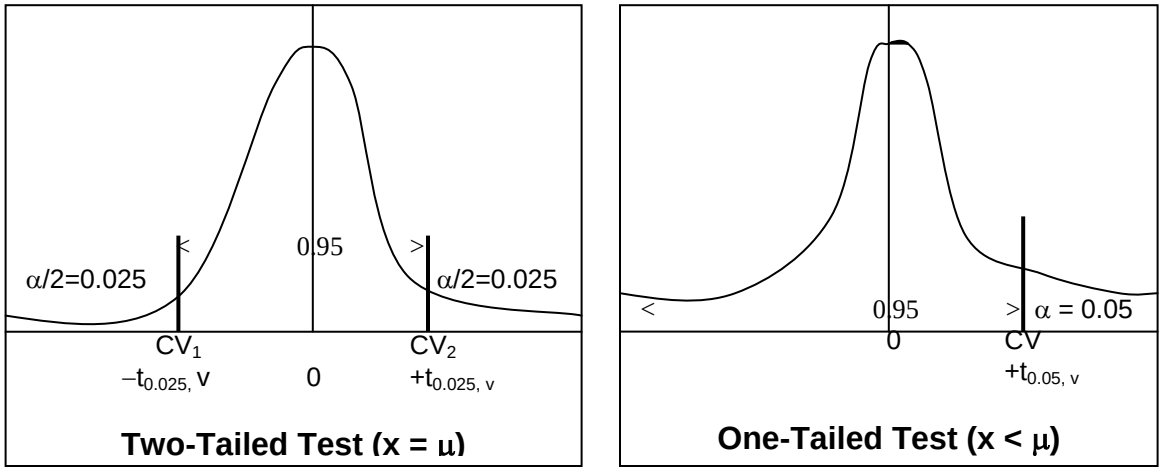
	H_0	:	m	=	μ		
or,	H_0	:	m_1	=	m_2	. . . . .	two-tailed test

To test the hypothesis that the observed sample mean (m) is greater than, or less than, the population mean (μ):

	H_0	:	m	>	μ		
or,	H_0	:	m	<	μ		
or,	H_0	:	m_1	>	m_2		
or,	H_0	:	m_1	<	m_2	. . . . .	one-tailed test

Similarly, to test the hypothesis that an observed sample mean is *not equal to* the population mean, or that an observed sample mean is not equal to the observed mean of another sample, we use a two-tailed test. In a two-tailed test, the area representing the level of significance must be distributed between the two tails.

Thus, for a level of significance of $\alpha = 0.05$ the area in each tail is equal to $\alpha/2 = 0.025$:



Tests of Hypotheses and Significance

Comparative Test of Means

To test the hypothesis H_0 that $m = \mu$ or $m_1 = m_2$ from a population with unknown standard deviation:

Step 1

Establish the Null Hypothesis

The mean scores achieved in the post-test by subjects following the traditional learning programme are equal to the mean scores achieved in the post-test by subjects exposed to the computer-aided learning programme,

i.e. $H_0 : m_1 = m_2$

Alternative Hypotheses

The mean scores achieved in the post-test by subjects following the traditional learning programme are not equal to the mean marks achieved in the post-test by subjects exposed to the computer-aided leaning programme.

H_1 : $m_1 > m_2$

or, H_1 : $m_1 < m_2$

i.e. H_1 : m_1 is not equal to m_2

Step 2

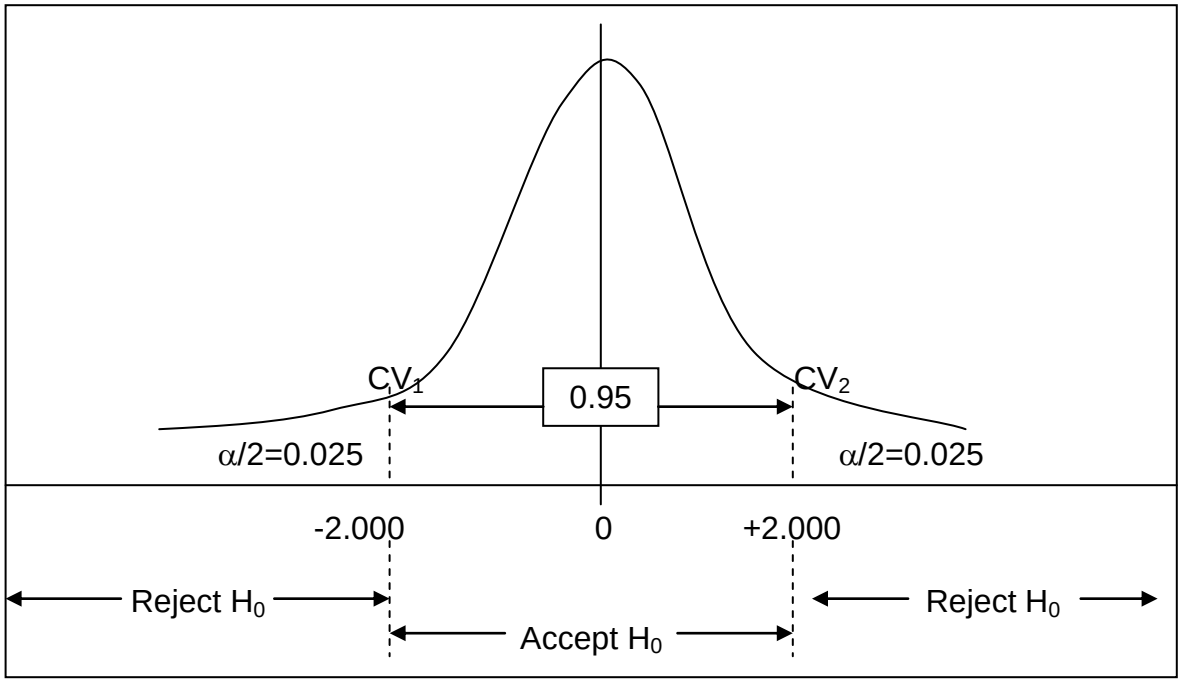
Decision Rule

For a level of significance $\alpha = 0.05$ with $v = (n_1 + n_2 - 2) = (40+20-2) = 58$ degrees of freedom:

If the t calculation is less than –2.000, or greater than +2.000, then reject the null hypothesis and accept the alternatives hypotheses, otherwise accept H_0.

Step 3

Graphically describe the distribution showing the critical values and the region of acceptance/rejection.



Step 4

Collect Sample Data and Compute the Result of the Test

Subject	Pre-Test Score	Post-Test Score Traditional Programme	Post-Test Score Computer-Aided Programme
111011	23	64	
111219	40	95	
111232	27		93
111234	41	90	
111235	15	41	
111241	37	61	
111243	37	58	
111247	47		84
111248	31		68
111250	36	44	
111251	34	82	
111252	49		71
111254	10	37	
111255	35	73	
111257	47		93
111258	29		64
111260	49	50	
111261	29		68
111266	19	77	
111269	43	84	
111270	28	63	
111271	21	84	
111272	25	64	
111275	34		71
111277	51		87
111278	44	78	
111279	21	79	
111280	31	42	
111281	17	58	
111282	14	52	
111283	19	35	
111284	18	60	
111285	26	49	
111286	31	82	
111287	34		85
111290	20		63
111291	14		73
111292	27	63	
111293	32	53	

111294	16		76
111295	15	54	
111296	31		83
111297	17	70	
111298	48	74	
111299	18		54
111300	32	67	
111301	11	57	
111302	43		94
111303	30	91	
111310	42	60	
111311	49		84
111312	33	48	
111315	17	56	
111318	29	65	
111321	24		63
111322	24	66	
111323	14	45	
111324	16		70
111325	23	69	
111326	31		75

From which we derive:

Traditional Learning Programme: $n_1 = 40$, $m_1 = 63.50$, $s_1 = 15.41228$

Computer-Aided Learning Programme: $n_2 = 20$, $m_2 = 75.95$, $s_2 = 11.40395$

Since the population standard deviation is unknown, a calculated estimate can be obtained by pooling the two sample standard deviations whilst accommodating for two unequal sample sizes:

$$\sigma_{\varepsilon\sigma\tau} = [(n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2 - 2)]^{0.5}$$

$$= [(40)(15.41228^2) + (20)(11.40395^2)]/(40+20-2)$$

$$= 17.8462$$

Hence:

$$t_{calc} = (m_1 - m_2)/[\sigma(1/n_1 + 1/n_2)]^{0.5}$$

$$= (63.50 - 75.95)/[17.8462(0.025+0.05)]^{0.5}$$

$$= -2.5474$$

Step 5

Conclusion

There is a significant difference between the two sample means.

Since t_{calc} = -2.5474 which is less than the critical value of –2.000 we must reject the null hypothesis at the 0.05 level of significance, and accept the alternative hypothesis. The mean score obtained by the sample of subjects exposed to the computer-aided learning package is not equal to the mean score by the sample of subjects following the traditional learning programme.

Indeed, one can claim that the mean score of subjects exposed to the traditional learning programme is less than the mean score of subjects following the computer-aided learning programme.

Testing a Specified Population Mean from a Small Sample

Using the standard procedure for significance testing, for example can derive the appropriate test of a specified population mean:

The population mean score of a proprietary test of occupational aptitude is 100 with an unknown standard deviation. A sample of eight experienced and competent subjects are given the test and provide a mean score of 120 with a standard deviation of 5. How likely is it that the population mean score is an adequate statistic for the test of aptitude?

1. Null Hypothesis

$$H_0 : \mu = 100$$

Alternative Hypothesis

$$H_1 : \mu > 100$$

2. Level of Significance

$\alpha = 0.05$ is considered to be adequate

3. Critical Value

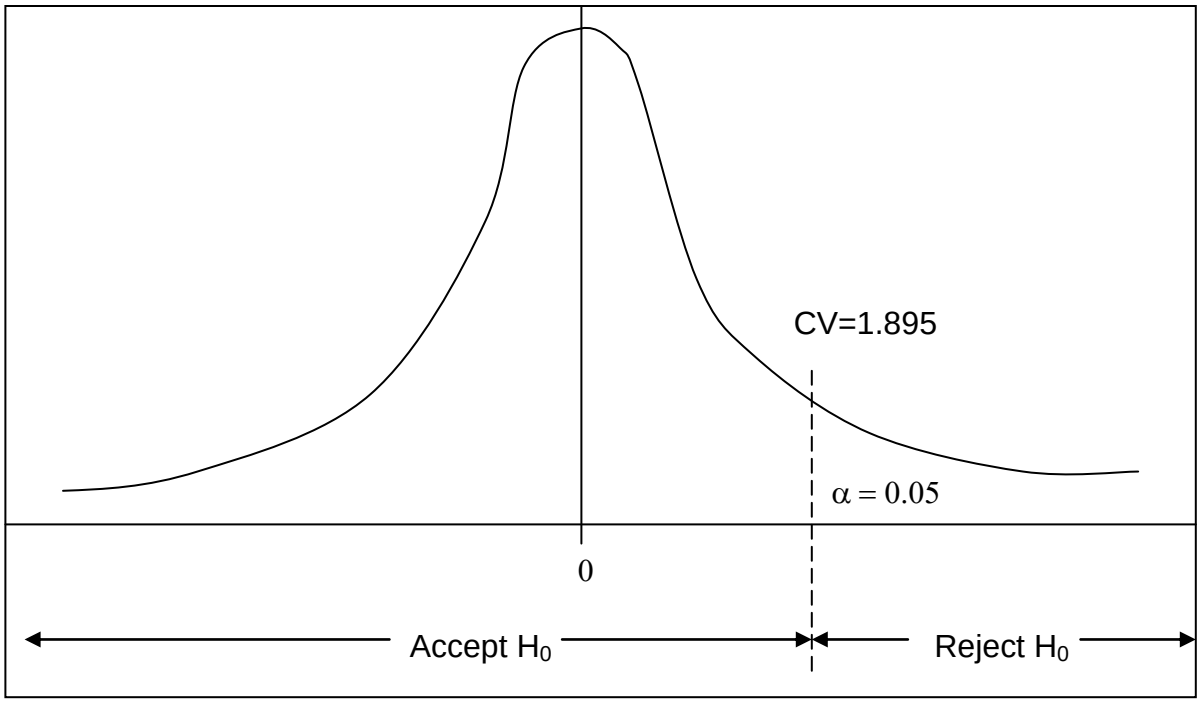
Since the sample size n<30 and the population standard deviation is unknown, then t is the appropriate sampling distribution, with a mean of 100 and a standard error of $s/(n^{0.5}) = 5/(8^{0.5}) = 1.7678$.

The reject region will be $t_{0.05}$ standard errors above the population mean; from tables for $t_{0.05}$ with (n-1) = (8-1) = 7 degrees of freedom the critical value is established as +1.895

4. Decision Rule

If the t_{calc} for the sample mean is less than +1.895 we must accept the null hypothesis. However, if the t_{calc} for the sample mean is greater than +1.895 then the null hypothesis must be rejected and the alternative hypothesis accepted.

Displayed graphically:



5. Computation of *t*

To find the position of the observed sample mean in the t distribution:

t = (m - μ)/standard error
= (110 – 100)/1.7678
= +5.657

6. Conclusion

Since the t score of +5.657 is greater than the critical value of +1.895 we must reject the null hypothesis at the 0.05 level of significance, and accept the alternative hypothesis. The population mean for the proprietary test of aptitude is not 100; or, a very rare event has occurred. Further sampling would be advised to confirm this conclusion.

Comparing Two Sample Means from Small Samples

A sample of six subjects in an experiment have an average performance of 96 on the measurement scale with a standard deviation of 4. Another sample of five subjects in the same experiment have an average performance of 92 with a standard deviation of 3.5 on the measurement scale. Do these samples indicate that the overall mean in the two sample differ?

$m_1 = 96$ $s_1 = 4.0$ $n_1 = 6$
$m_2 = 92$ $s_2 = 3.5$ $n_2 = 5$

1. Null Hypothesis: There is no significant difference between the sample mean scores achieved by the two experiments:

$$H_0 : m_1 = m_2$$

Alternative Hypothesis: The mean scores achieved by the two samples are not equal:

$$H_1 : m_1 > m_2; \text{ or, } m_1 < m_2$$

2. Decision Rule: For $v = (n_1 + n_2 - 2) = (6+5-2) = 9$ degrees of freedom at the 0.05 level of significance ($\alpha/2$ for a two-tailed test): if the calculation value for t in the distribution is less than-2.262 or greater than +2.262 then reject the null hypothesis and accept the alternative hypothesis, otherwise accept.

2. Compute the Result of the Test:

$$t = (m_1 - m_2)/SE$$

where the standard error $= \sigma(1/n_1+1/n_2)^{0.5}$

Since the population standard deviation (σ) is unknown, this can be estimated from pooling the two sample standard deviations (s_1 and s_2):

$$\sigma = [(n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2 - 2)]^{0.5} = [(6)(4^2) + (5)(3.5^2)] / (6+5-2)]^{0.5}$$

$$= 4.18$$

Hence, $SE = 4.18(1/6+1/5)^{0.5} = 2.5311$

and, $t = (m_1 - m_2) / SE = (96-92) / 2.5311$

$$= \underline{1.58}$$

4. Conclusion

Since the calculated value of *t* is greater than the critical value of –2.262, and is less than the critical value of +2.262, we must accept the null hypothesis: there is no significant difference between the two sample means at the 0.05 level of significance.

Computing the Sample Size from a Pilot Study

We have seen that to be 95 percent certain that the sample mean is not more than a specified limit from the population, or true, mean we can compute:

$$\text{Population Mean} = \text{Sample Mean} \pm t_{0.025}s/(n^{0.5})$$

transposing this formula gives:

$$\text{Sample Size } (n_{0.95}) = [t_{0.025}s/L]^2$$

where L is the specified limit of acceptable error, for example:

We have piloted a survey questionnaire using nine respondents and have a sample mean of 50 with a standard deviation of 8. We wish to be 95 percent confident of our estimate of the sample size, whilst accepting an error of 3 units on the measurement sale

For n = 9, s = 8, and m = 50 the 95 percent confident limits of the sample size, accepting an error of 3 units on the measurement scale, can be computed:

$$N = [t_{0.025}s/L]^2$$

$$= [(2.306)(8)/3]^2$$

$$= 37.814 \text{ (say, 38)}$$

Comparing the Variability of Two Samples (Analysis of Variance)

While the *z* and *t* tests are useful for examining single differences, such as sample means, there can be problems when a set of differences is to be examined. A significance test considers how likely it is that a given result is due to *sampling error* rather than representing a real difference. By convention, we reject the null hypothesis of 'no difference' if the probability of this is as low as

0.05 (i.e. this means that if we do twenty experiments or tests we are very likely to be making at least one Type I error).

The F-distribution is used to compare the variability of two samples rather than the sample mean values, for example:

We have designed an instrument to measure an attribute and piloted this on two samples of the population under study. In a series of ten trials we compare the results obtained with the measurement instrument with those obtained from a well validated and reliable measurement instrument derived from the literature. The results indicate that whilst the sample exposed to our instrument provides a standard deviation of 4.099, the sample exposed to the existing instrument gives a standard deviation of 2.011. Does the pilot study indicate that our measurement instrument is less consistent than the existing measurement instrument?

$n_1 = 10 \qquad s_1^2 = 4.099$
$n_2 = 10 \qquad s_2^2 = 2.011$

1. Null Hypothesis

There is no significant variability between the two measurement instruments

$$H_0 : \sigma_1^2 = \sigma_2^2$$

Alternative hypothesis

$$H_1 : \sigma_1^2 > \sigma_2^2$$

2. Level of Significance

A level of significance of $\alpha = 0.05$ (i.e. at the five percent level)

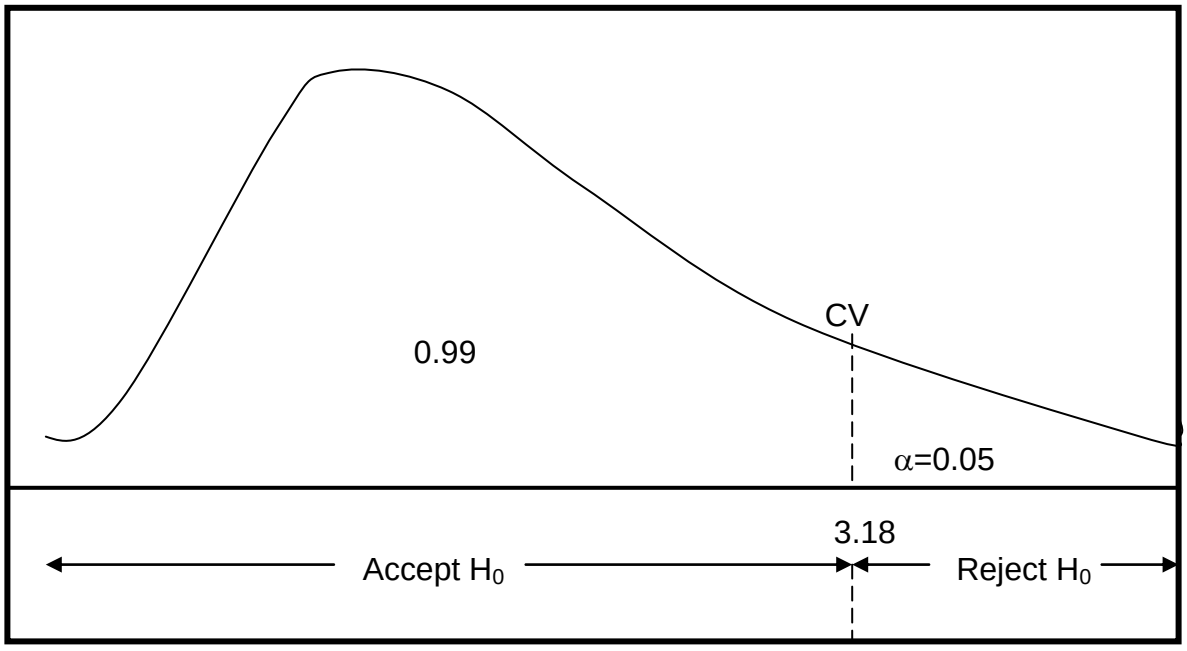
3. Decision Rule

The critical value for *F* is derived from table of the *F*-distribution, the degrees of freedom being:

s_1^2 (the numerator) is read horizontally $= v_1 = n_1 - 1 = 9$

s_2^2 (the denominator) is read vertically $= v_2 = n_2 - 1 = 9$

Hence, for $F_{0.05, 9, 9} = 3.18$



4. Compute the Result of the Test

$$F = s_2^2/s_1^2$$

$$= 16.802/4.044$$

$$= 4.153$$

5. Conclusion

Since the calculated *F* value of 4.153 is greater than the critical value of 3.18, we must reject the null hypothesis and accept the alternative hypothesis[6]. There is a significant difference in the variability of the two tests at the 0.05 level of significance: the scores derived from our measurement instrument are significantly more variable than those derived from the existing measurement instrument. However, we would accept the null hypothesis at the 0.01 level of significance where the *F* calculated value of 4.135 is less than $F_{cv} = 5.35$

[6] Occasionally, we may have no prior knowledge of the variability of two observations and a two-tailed test may be called for. In this instance, the lower critical value for *F* is the reciprocal of the upper critical value: in the example above, for $v_1 = 9$, $v_2 = 9$ at 0.05 level of significance, the upper critical value 4.03 and the reciprocal is 0.2421.

The Chi-Square (χ^2) Test of Significance

As we have seen, the results obtained in samples do not always agree precisely with theoretical results expected according to the rules of probability. With problems of categorical data, the previous methods of applying the *z* and *t* sampling distributions are unsatisfactory. χ^2 is a measure of the discrepancy existing between observed and expected frequencies. Suppose that in a particular sample a set of events E_1, E_2, E_3 . . E_n are observed to occur with frequencies o_1 o_2 o_3 . . . o_n called *observed frequencies*, and that according to probability rules they are expected to occur with frequencies e_1 e_2 e_3 . . . e_n called *expected frequencies.* Often we wish to know whether *observed* frequencies differ significantly from *expected frequencies.*

Event	E_1	E_2	E_3	E_4	E_5	. . .	E_n
Observed Frequency	o_1	o_2	o_3	o_4	o_5	. . .	o_n
Expected Frequency	e_1	e_2	e_3	e_4	e_5	. . .	e_n

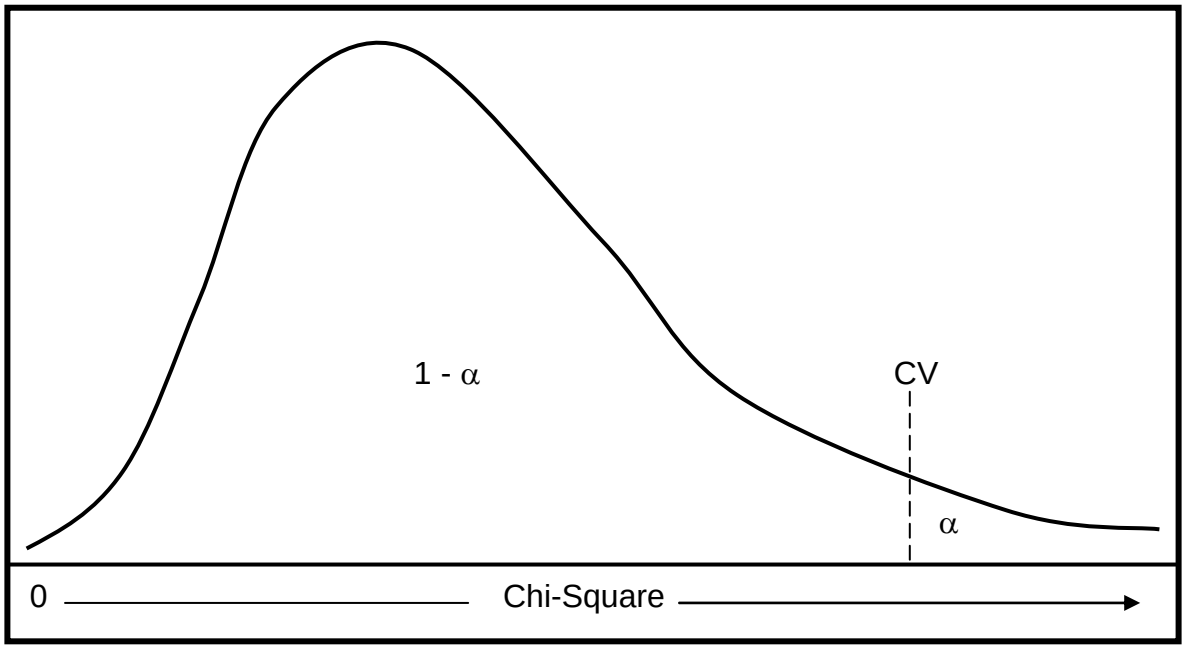
If $\chi^2 = 0$, observed and theoretical frequencies agree exactly, while if $\chi^2 > 0$ the do not agree exactly. The larger the value of χ^2, the greater is the discrepancy between observed and expected frequencies.

The χ^2 distribution can be defined as:

$$\chi^2 = \sum (\text{observed value} - \text{expected value})^2 / \text{expected value}$$

$$= (o_1 - e_1)^2/e_1 + (o_2 - e_2)^2/e_2 + (o_3 - e_3)^2/e_3 + \ldots (o_n - e_n)^2/e$$

The chi-square distribution is used to test if a series of observed values differs significantly from that which is expected. The chi-square distribution is typically skewed:



Degrees of Freedom

The number of degrees of freedom of the χ^2 statistic, denoted *v*, is defined as the number of *n* independent observations in the sample. However, since we must estimate the population mean (μ) from samples of the population, by convention $v = n - 1$. The degrees of freedom (*v*) for the χ^2 distribution is determined:

(Number of Columns (*k*) in the Table– 1)(Number of Rows (*h*) in the Table– 1)

Thus, for a table of six columns and four rows, $v = (k-1)(h-1) = (6-1)(4-1) = 15$

Confidence Intervals for χ^2

As with the normal and *t* distributions, we can set a level of significance and define confidence intervals with tables of the χ^2 distribution, for example:

α / ***v***	0.05	0.025	0.01	0.005
1	3.84146	5.02389	6.63490	7.87944
5	11.0705	12.8325	15.0863	16.7496
10	18.3070	20.4832	23.2093	25.1882
15	24.9958	27.4884	30.5779	32.8013
20	31.4104	34.1696	37.5662	39.9968

Contingency Tables

Corresponding to each observed frequency in a (*h*)(*k*) contingency table, there is an expected or theoretical frequency, which is computed according to the rules of probability. These frequencies which occupy the cells of a contingency table are called *cell frequencies*. The total frequency in each row of column is called the *marginal frequency.* To investigate the agreement between observed and expected frequencies, we compute the statistic

$$\chi^2 = \sum (o_n - e_n)^2/e_n$$

The observed frequencies of a phenomenon are displayed:

Column (*k*) / Row (*h*)	**I**	**II**	**Total**
A	a_1	a_2	N_A
B	b_1	b_2	N_B
Total	N_1	N_2	N

Applying the cell frequency coding above, the expected frequencies are computed as follows:

Column (*k*) / Row (*h*)	**I**	**II**	**Total**
A	N_1N_a/N	N_2N_a/N	N_A
B	N_1N_b/N	N_2N_b/N	N_B
Total	N_1	N_2	N

For example:

The table below shows the results of an experiment to investigate the effect of a hypnotherapy technique on a group of subjects who complained that they did not sleep well. Some subjects were exposed to hypnotherapy while others were given a placebo treatment (although they all thought they were getting hypnotherapy). They were later asked whether the hypnotherapy helped them sleep well or not. The results of the responses are shown in the table below:

	Slept well	**Did not sleep well**
Exposed to the hypnotherapy technique	**44**	**10**
Given a placebo treatment	**81**	**35**

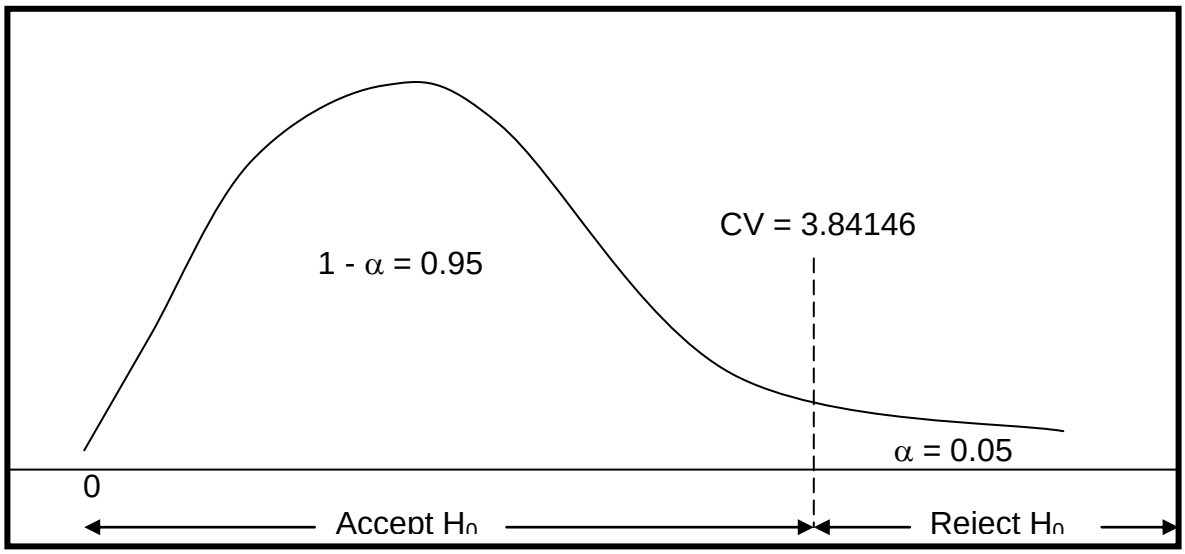
Assuming that all subjects told the truth, test the hypothesis that there is no difference between treatment by hypnotherapy and the placebo at a level of significance of 0.05

1. Null Hypothesis There is no significant difference between the two treatments, i.e. the application of the hypnotherapy technique to subjects is independent of the subjects who claimed to have slept well.

   Alternative Hypothesis There is a significant difference between the subjects exposed to the hypnotherapy technique and claimed that they slept well, and those subjects exposed to the placebo treatment.

2. Level of Significance A level of significance of $\alpha = 0.05$ is considered to be acceptable

3. Decision Rule For *k* columns and *h* rows, the degrees of freedom = $v = (k\text{-}1)(h\text{-}1) = (2\text{-}1)(2\text{-}1) = 1$. Hence, for a one-tailed test, for $v = 1$, and $\alpha = 0.05$ the critical value of $\chi^2 = 3.84146$



If the calculated value of χ^2 is less than 3.84146 we must accept the null hypothesis, i.e. that there is no difference between the hypnotherapy treatment and the placebo. Otherwise reject.

4. Computation of χ^2

The table of frequencies expected under H_0 is calculated from the observations:

	Slept well	Did not sleep well	Total
Exposed to hypnotherapy	(125)(54)/170 = **39.71**	(45)(54)/170 = **14.29**	54
Exposed to placebo	(125)(116)/170 = **85.29**	(45)(116)/170 = **30.71**	116
Total:	125	45	170

Hence, $\chi^2 = \Sigma (o_j - e_j)^2/e_j =$ $(44-39.71)^2/39.71 + (81-85.29)^2/85.29 + (10 - 14.29)^2/14.29 + (35 - 30.71)^2/30.71$

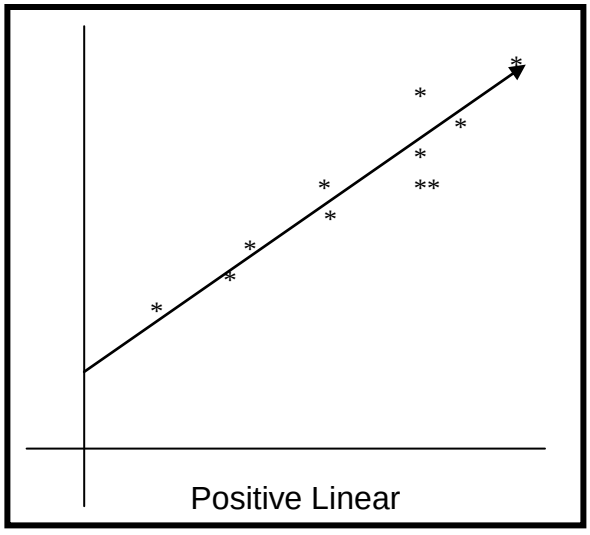
$= 2.56643$

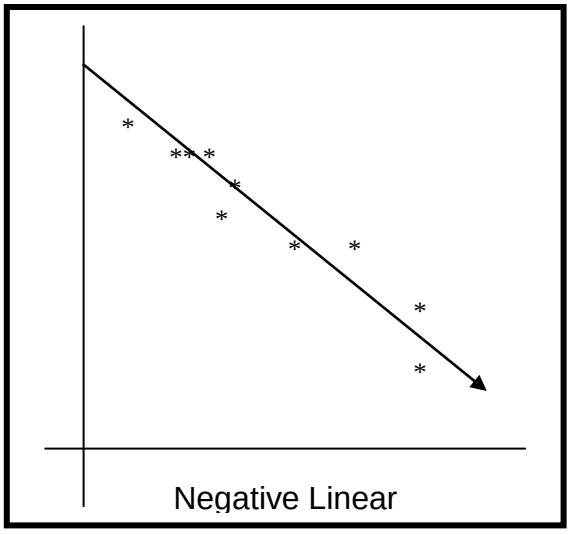
4. Conclusion

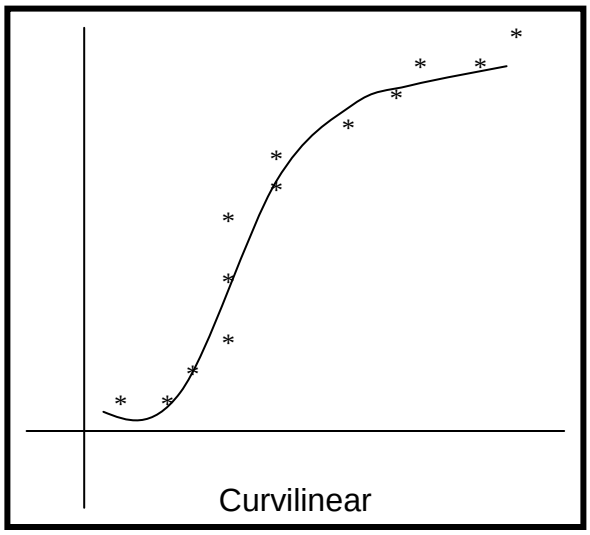
Since the calculated χ^2 value of 2.56643 is less than the critical value of 3.84146, the null hypothesis cannot be rejected at the 0.05 level of significance: there is no significant difference between the two treatments.

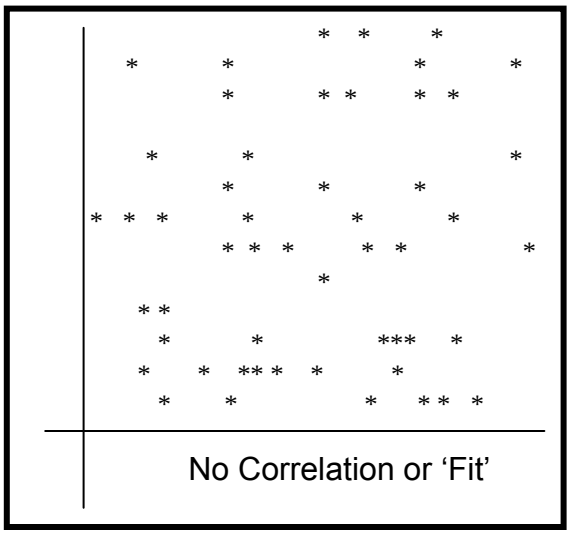
VIII
Linear Regression and Correlation

If *X* and *Y* denote two variables under consideration, a scatter diagram shows the location points *X,Y* on a rectangular co-ordinate system. If all the points in this scatter diagram seem to lie near a perceived line, the correlation or fit is termed *linear.*


Positive Linear


Negative Linear


Curvilinear


No Correlation or 'Fit'

Linear Regression

Very often in practice a linear relationship is found to exist between two variable say, *X* and *Y*. It is frequently desirable to express this relationship in mathematical form by determining the regression equation that connects these two variables. The *least square line* is one with the best goodness of fit in that deviation or error is minimum. The least square line can be expressed:

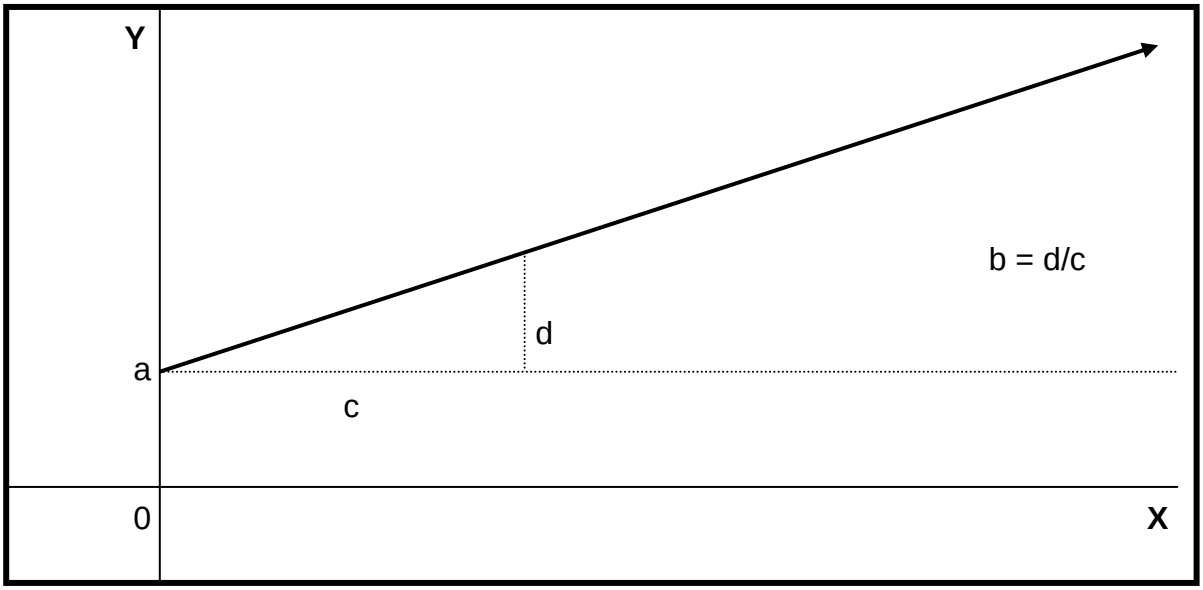
$$Y = a + bX$$

Where *a* is the intersect of the line with the vertical axis and can be computed:

$$a = \frac{\Sigma Y \Sigma X^2 - \Sigma X \Sigma XY}{N\Sigma X^2 - (\Sigma X)^2}$$

and *b*, the gradient or slope of the line:

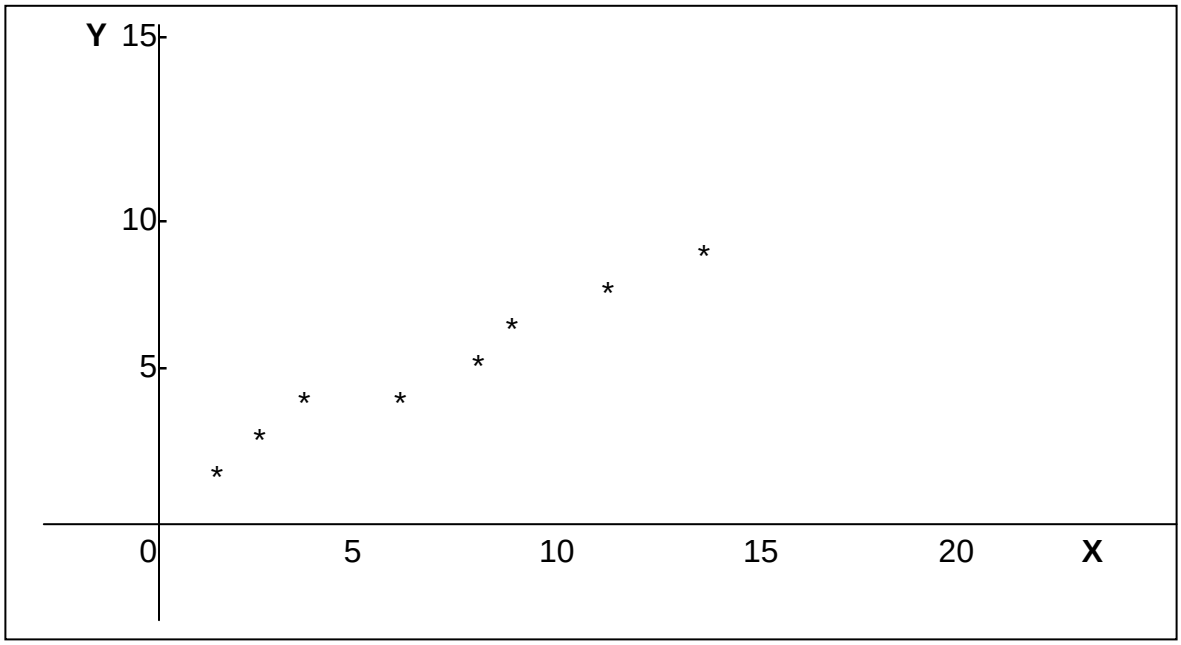
$$b = \frac{N\Sigma XY - \Sigma X \Sigma Y}{N\Sigma X^2 - (\Sigma X)^2}$$



Compute the regression equation for the following data, using X as the independent variable:

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

Step 1 Construct a scatter diagram to identify the model and get an estimate of the intersect *a*



The scatter diagram indicates that the estimate for the value *a* is positive, and lies between 0 and 5

Step 2 Compute the values for *a* and *b*

X	**Y**	**X^2**	**XY**	**Y**
1	1	1	1	1
3	2	9	6	4
4	4	16	16	16
6	4	36	24	16
8	5	64	40	25
9	7	81	63	49
11	8	121	88	64
14	9	196	126	81
Total	56	40	524	364	256

Subsequently, the standard equations become:

$a = [(40)(524) – (56)(364)] / [(8)(524) – (56)(56)] = 0.545$

$b = [(8)(364) – (56)(40)] / [(8)(524) – (56)(56)] = 0.636$

Hence, the regression equation can be expressed:

$$Y = 0.545 + 0.636X$$

An alternative method, particularly useful in manually processing complex numbers, uses the following equations:

$$\Sigma Y = aN + b\Sigma X$$

$$\Sigma XY = a\Sigma X + b\Sigma X^2$$

hence,

$$40 = 8a + 56b \quad . . . \quad . . . \quad . . . \quad (1)$$
$$364 = 56a + 524b \quad . . . \quad . . . \quad . . . \quad (2)$$

by multiplying equation (1) by 7, we can solve the equations simultaneously and by substitution:

$$280 = 56a + 392b \quad . . . \quad . . . \quad . . . \quad (1)$$
$$364 = 56a + 524b \quad . . . \quad . . . \quad . . . \quad (2)$$

subtract equation (1) from equation (2):

$$84 = 132b \quad . . . \quad . . . \quad . . . \quad (3)$$

hence, $b = 84/132 = 0.636$

substitute the value for *b* into equation (1):

$$40 = 8a + 56(0.636)$$

$$= 8a + 35.6364$$

$$a = (40 – 35.6364) / 8$$

$$= 0.545$$

<u>The regression equation allows us to compute the value of *Y* for any value of *X*.</u>

Linear Correlation

The statistic *r* is called the *coefficient of correlation* and is defined by the equation for Pearson's Product-Moment formula:

$$r = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{[[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]]^{0.5}}$$

The equation automatically gives the sign of *r* as well as clearly showing the symmetry between the variable *X* and *Y*. It should be noted that in the instance of linear regression, the quantum *r* is the same, regardless of whether *X* or *Y* is considered as the independent variable. Thus, *r* gives a very good measure of the relationship between two or more variables and lies between –1 and +1. The close the value of *r* is to 0, the less there is a linear relationship between the variables.

In the example provided to compute linear regression, we have:

$N = 8$	$\Sigma X = 56$	$\Sigma X^2 = 524$	$\Sigma XY = 364$	$\Sigma Y^2 = 256$	$\Sigma Y = 40$

Substituting in the equation:

$$r = (8)(364) - (56)(40) / [[(8)(524) - (56)(56)][(8)(256) - (40)(40)]]^{0.5}$$

$$= 672 / [(1056)(448)]^{0.5}$$

$$= 0.977$$

Coefficient of Determination

The ratio of the explained variation to the unexplained variation is termed the *coefficient of determination*. If there is no explained variation (i.e. $r = 0$), this ratio is zero. If there is no unexplained variation (i.e. $r = 1$), this ratio is one. Since the ratio is always positive, we denote the coefficient of determination as:

$$R = r^2 = \text{(Explained variation)/(Unexplained variation)}$$

$$= \frac{\Sigma(Y_{estimate} - Y_{mean})^2}{\Sigma(Y - Y_{mean})^2}$$

. . . since *r* is a dimensionless quantity.

For practical purposes, the coefficient of determination is computed as the square of the coefficient of correlation (i.e. r^2). Using the example above, the coefficient of correlation has been computed to be $r = 0.9777$, hence R = 0.95589. Subsequently, we can claim that say, 95.6 per cent of the variability of *Y* is explained by the variability of *X*; and, 4.4 percent of the variability of *Y* is unexplained.

However, it should be noted that both the coefficient of correlation and the coefficient of determination do not measure a cause-and-effect relationship, but merely the strength of association between two or more variables. For instance, research in the 1930's showed a strong positive correlation between the incidence of prostitution and the output of steel in Pennsylvania. It would be nonsense to claim that the former 'causes' the latter.

Sampling Theory of Correlation

The *n* pairs of values (*X, Y*) of two variables can be thought of as a sample from a population of all possible such pairs. Hence, this bivariate population can be assumed to have a bivariate normal distribution. The population coefficient of correlation may be denoted π. Tests of significance concerning π require the sampling distribution of *r*:

1. Test of Hypothesis: There is no correlation between the two variables in the population, therefore any explained variability in the sample is chance:

$$\pi = 0$$

$$t = \frac{r(n-2)^{0.5}}{(1-r^2)} \quad \text{for } v = n - 2 \text{ degrees of freedom}$$

2. Test of Hypothesis: There is a correlation between the two variables in the population, therefore any unexplained variability in the sample is chance:

$$\pi < 0, \text{ or } \pi > 0$$

$$z = 0.5\log_e [(1 + r)/(1 - r)] = 1.1513\log_{10}[(1 + r)/(1 - r)]$$

The former equation (1) is more generally applicable. Hence, for $r = 0.9777$:

$$t = \frac{0.9777(8-2)^{0.5}}{(1-0.9559)} = 54.306$$

For a two-sided test, the critical value for *t* at the 0.05 level of significance and for $v = (8 - 2) = 6$ degrees of freedom, = $\pm$2.447. Since the calculated *t* value of 54.306 is greater than the critical value of $\pm$2.447, we must reject the null hypothesis and conclude that the coefficient of correlation is greater than 0. Hence, our observation is significant at the 5 percent level (see footnote 7: for $v = 6$, *r* must be greater than 0.7067 at the 5 percent level of significance)

Point Biserial Correlation

Point biserial correlation is a derivative of the Pearson product-moment correlation *r*, similarly, the tables used for the product-moment correlation can determine the significance. Point biserial correlation (r_{pbis}) is used when one of the variables is in the form of a continuous score and the other is a categorical dichotomy. For example, the reaction times in an experiment (times is continuous data) and the sex of the subject (the dichotomy male/female is categorical data); levels of job satisfaction (assumed continuous data albeit on a Likert scale) and the categories of manager/worker, etc.

Point biserial correlation is particularly useful to determine whether a particular survey (or test) item discriminates between high and low scores on the questionnaire as a whole. The continuous data, in terms of the overall test score, of each individual is designated as the *X* variable and the categorical data (for instance, yes/no, right/wrong, agree/disagree, male/female) the *Y* variable. An item that has a positive correlation with the total survey score is discriminating between candidates *in the same way* as the overall survey. An item that has a low positive or negative correlation is not discriminating between the low and high scoring individuals in the survey (or test) in the same way as the overall survey; hence, it would be advisable to exclude such an item from the questionnaire (or test).

We have conducted a survey to assess the levels of felt job satisfaction in a department of ten employees. Recognizing that some of the employees are managers and others workers, we wish to correlate the scores with the category of the respondents: manager or worker. The individual total scores and the workers' perceived level in the hierarchy are given below:

Respondent (random number)	Overall JDI Score	Perceived Level (manager = 1, worker = 0)
285	60	0
271	56	0
952	51	0
067	58	0
502	49	1
176	48	0
331	55	1
642	45	1
858	47	1
441	55	0

The point biserial correlation is defined as:

$$r_{pbis} = \left[\frac{X_0 - X_1}{\sigma_x} \right] [(p)(1-p)]^{0.5}$$

where

X_0	=	mean score of X for respondents scoring 0 on Y	=	54.67
X_1	=	mean score of X for respondents scoring 1 on Y	=	49.00
σ_x	=	standard deviation of all scores X	=	4.820
p	=	proportion of respondents scoring 0 on Y	=	0.6
(1-p)	=	proportion of respondents scoring 1 on Y	=	0.4

Thus,

$r_{pbis} = [(54.67 – 49.00)/4.82)] * [(0.6)(0.4)]^{0.5}$

$= (5.67/4.82) * (0.24)^{0.5}$

$= 0.5763$

Conclusion: There is a moderately high positive correlation between the perceived occupational status of the respondents and the overall score in the level of job satisfaction. However, the correlation is significant at the 0.10 level of significance ($r > 0.5494$), but not at the 0.05 level ($r > 0.6319$).[7]

[7] See White, J., Yeats, A. and Skipworth, (1991), *Tables for Statisticians*, Table 15, p.29, Critical values of the product-moment correlation.