EC124
Statistical
Techniques B

Lecture Notes &
Course Outline

UNIVERSITY OF WARWICK
DEPARTMENT OF ECONOMICS

EC124 STATISTICAL TECHNIQUES B 2015/16

Aims of the course

The course aims to provide students with important skills, which are of both academic and vocational value, being an essential part of the intellectual training of an economist and also useful for a career. In particular the course aims to equip students with the following competence.

(1) Understanding the nature of uncertainty and methods of dealing with it.

(2) An awareness of the empirical approach to economics.

(3) Experience in the analysis and use of empirical data in economics.

Teaching method and course organisation and textbook

There will be two hours of lecture each week and these lectures will be supplemented by a weekly small group tutorial classes which will go through weekly exercise sheets. On the EC124 website there is available a series videos which go through step-by-step some questions which are linked to the more theoretical material taught in the lectures and these are covering the topics according to the lecture handouts. In addition, there are also a series of videos which go through the exam papers for a number of the past years as well as example questions for each of the main topics (see Revision Videos section). There is also series of on-line multiple choice type questions (split according to tutorial work), in which you should test your knowledge (see Problem Sets section).

The recommended textbook is:

Peck, R. (2013), *Statistics: Learning from Data*, Cengage

There are plenty of textbooks around for a 1st year statistics module for business or economics students. In general, these books are aimed at students taking a slightly more basic statistics module than this one, however, they are generally well written and have plenty of example exercises in them. Two worthy of a mention are:

Newbold, P. and Carlson, W. L., Thorne, B (2012), *Statistics for Business and Economics* (8th edition), Pearson.

Anderson, D. R., Sweeney, D. J., Williams, T. A., Freeman, J. and Shoesmith, E., (2010), *Statistics for Business and Economics*, (2nd edition). Thomson, South-Western.

Tutorial classes

There are weekly tutorial classes. Exercise sheets are all handed out in lecture 1. Exercise sheet questions will be covered in the tutorial class on the Tuesday, but you must look over these questions prior to walking into your class.

Assessment

Assessment will be by examination and coursework. The examination component will comprise two short tests of 50 minutes and a one a half hour exam in June. This module is worth 40% of your overall mark towards the 1st year composite module Quantitative Techniques (EC120). A further 40% will come from the module Mathematical Techniques A (EC121) or B (EC123) and the remaining 20% will come from the module Computer and Data Analysis (EC125). The 40% for this module is accounted for by the June exam (worth 30%) and two tests worth 5% each (10% in total). The June exam is a 1.5 hour exam (plus 15 minutes reading time).

Calculators

The University policy only allows non-programmable calculators in examinations of the form given to you at the start of the academic year.

Examination Syllabus

The list of topic below is definitive of the examination syllabus. The scheduling however is tentative and it may be necessary to cover different topics in different weeks. Students should regard this list as a guide to their reading.

Timetable for tests

Test 1 Thursday (6-7pm) 25th February.

Test 2 Thursday (6-7pm) 5th May.

These tests are multiple choice question tests and will be marked electronically (a sample of the answer sheet which you will have to use to do the test is attached). The tests will be 5 marks for a correct answer and minus 1 for an incorrect answer. You will need to take with you to the tests:

(i) Statistical tables,
(ii) Calculator,
(iii)2 sides of A4, with any notes you want.[1]

[1] In the June exam for this module you will be provided with the statistical tables and you will be provided with a formula sheet (you cannot take in your own sheet of notes).

STATISTICAL TECHNIQUES B: Lecture material

Topics	Peck (Ch)	Newbold (Ch)	Anderson (Ch.)
Basic descriptive statistics: measures of central tendency, measures of dispersion, covariance and correlation	1, 2 3	1, 2	1, 2, 3
Probability: An introduction to the idea of probability using Venn diagrams and basic set theory. Introduce the basic concept of probability and some rules of probability, conditional and marginal probability and Bayes Theorem.	5	3	4
Discrete and continuous distributions: Single variable density functions. Methods for calculating means, variances.	6	4, 5	5, 6
Special distributions: Binomial, Poisson, Uniform, Normal, Chi-squared, F- and t- distributions.	6	4, 5	6
Joint probability density functions: Bivariate distributions, marginal and conditional distributions, independence.	8	4, 5	5, 6
Central Limit Theorem and Point Estimation.	12	6	7
Hypothesis Testing: for single sample means and the difference in sample means.	10, 12	9	9, 10
Hypothesis Testing (cont'd): for sample variances and the power of a test.	11, 13	10	9, 10, 11
Confidence Intervals: for single sample means and the difference in sample means as well as for sample variances.	12	7, 8	8
Non-parametric tests: Sign Test, Mann-Whitney Test, Goodness-of-Fit Tests, Contingency Tables.	15	14	12,19

7151578969

ECONOMICS
EC124 Statistical Techniques B
ANSWER SHEET

Student ID Number							

Please use a blue or black pen and mark answers by putting a cross in a relevant box

Question 1	☐ A	☐ B	☐ C	☐ D	☐ E
Question 2	☐ A	☐ B	☐ C	☐ D	☐ E
Question 3	☐ A	☐ B	☐ C	☐ D	☐ E
Question 4	☐ A	☐ B	☐ C	☐ D	☐ E
Question 5	☐ A	☐ B	☐ C	☐ D	☐ E
Question 6	☐ A	☐ B	☐ C	☐ D	☐ E
Question 7	☐ A	☐ B	☐ C	☐ D	☐ E
Question 8	☐ A	☐ B	☐ C	☐ D	☐ E
Question 9	☐ A	☐ B	☐ C	☐ D	☐ E
Question 10	☐ A	☐ B	☐ C	☐ D	☐ E
Question 11	☐ A	☐ B	☐ C	☐ D	☐ E
Question 12	☐ A	☐ B	☐ C	☐ D	☐ E
Question 13	☐ A	☐ B	☐ C	☐ D	☐ E
Question 14	☐ A	☐ B	☐ C	☐ D	☐ E
Question 15	☐ A	☐ B	☐ C	☐ D	☐ E
Question 16	☐ A	☐ B	☐ C	☐ D	☐ E
Question 17	☐ A	☐ B	☐ C	☐ D	☐ E
Question 18	☐ A	☐ B	☐ C	☐ D	☐ E

EC124: STATISTICAL TECHNIQUES B

Descriptive Statistics

1. Introduction

Statistics are numerical facts or figures. Therefore statistics as a science essentially deal with numbers. It is generally taken to include the systematic collection, classification and analysis and synthesis of the important features of a large body of numerical information. The objective being to make sense of the data by summarising it in such a way that a readily understood picture emerges while little of importance is lost, and the information is presented without misleading or misrepresenting it. The actual methods adopted will depend on the nature of the numerical information and how it is to be used.

Statistics is the science of uncertainty, we deal with what *could be*, what *might be*, or what *probably is*. We will therefore be introducing procedures for attacking problems where the conclusion will necessarily be uncertain. This will involve introducing the concept of probability.

The subject of statistics can usefully be divided into two parts; descriptive statistics and inferential statistics. Descriptive statistics are used to summarise information which would otherwise be too complex to take in, by use of numerical and graphical techniques. Inferential statistics concerns the relationship between a sample of data and the population from which it is drawn. It asks what inferences can be drawn about the population from the sample, where we assume the population parameter of interest is a constant (even though it is unknown). Clearly it is not possible to learn anything about the population precisely from the sample statistic, as some uncertainty remains and learning about uncertainty will form an integral part of this module.

Population - refers to all elements of interest who have the characteristic in which you are interested (size=N). For example, suppose we are interested in the outcome of a vote to join the EMS. In this case the population is all individuals in the UK eligible to vote in a referendum – the issue here being who will be 18 years and older at the time of the referendum. Alternatively, suppose one is interested in the average weekly salary of females in full-time employment in the UK. The population is all females who are in paid

employment. However, there is an issue of what we mean by full-time employment, as there is a relatively large black-economy which is undetectable. In addition, how do we deal with over-time payment?

Sample – a subset (hopefully random, and therefore representative, using computer generated random numbers) of the population (size=n). One takes a sample because it is prohibitively expensive (either in time or financial terms) to interview the whole population, in order to make a statement about the population.

1.1 Types of economic data

Time Series – Macroeconomists and financial economists are interested in things such as Gross Domestic Product (GDP), Unemployment, Inflation, Interest Rates, Exchange Rates and Stock Prices etc and data such as these is often collected at specific points in time and are observed at different frequencies. Common frequencies are: annually, (once a year), quarterly (four times a year), monthly (twelve times a year), weekly (52 times a year). Time series data are often presented in chronological order.

Cross-sectional – Researchers often work with data that is characterised by individual units, such as: people, companies or countries. With cross-sectional data, the ordering of the data typically does not matter (unlike for time series data).

1.2 Graphical representation of data

One important way of representing/summarizing data is to use charts and tables. There are many different types of plots (bar charts, pie charts, scatter plots etc). Since most economic data is either time series or cross-sectional, we will briefly illustrate plots for both of these.

Time series graphs – As many data sets may contain a great many observations on some variable X, it is difficult to gauge much information from the raw data. A simple way of presenting the data might be as a time series graph, which plots the variable of interest (on the vertical axis) against time (on the horizontal axis). Figure 1a, is a time series graph of annualised GDP growth for the UK using quarterly observations over the period from 1990-2013 and captures the main features of the series, which are the 2 recessions (early 1990s - associated with the collapse in the housing market and 2008 from the

financial crisis) and the relative prosperity of the UK economy since the early 1990s (for which Gordon Brown wanted to take the credit). Figure 1b is a time series plot of annualised inflation for the UK using quarterly observations over the period from 1990-2013 and captures the main features of the series, which are the relatively high inflation rates in the early 1900s associated with the defending the £ and the relative stability and low inflation since the mid-1990s and the Independence of the Central Bank followed by the more recent turmoil associated with the financial crisis.

Cross-sectional – Plots of the nature of Figures 1a or 1b are often inappropriate for cross-sectional data (see Figure 1c). Again it is the case that the data sets may contain a great many observations on some variable X, and it is difficult to gauge much information from simply staring at the raw numbers. The first step in an attempt to understand data may involve creating a FREQUENCY TABLE. This is done by counting the number of times, that is, the frequency (f_i), the each value of X, denoted x_i (where x_1 denote the first observation, x_2 denote the second observation and so on) occurs and then present the results in tabular form.

Table 1: Frequency table for the discrete random variable X

Value	Frequency
x_1	f_1
x_2	f_2
x_3	f_3
…	…
x_k	f_k
Total	$\sum_{i=1}^{k} f_i = n$

A graph of these frequencies against x_i is called a frequency density function (or a bar chart/histogram) - these graphs give a pictorial representation of the distribution of the variable X.

However, our instinct is to think in terms of the proportion of observations in each class. It seems more desirable that these proportions, or relative frequencies ($p_i=f_i/n$), be shown. The relative frequencies are constructed by dividing the frequencies (f_i) by the total number of observations in the sample (n). In addition, we often want to consider the proportion of observations that are either in that one or one of the earlier classes. These

proportions are called cumulative relative frequencies, and are obtained by adding the relative frequency for the class to the cumulative relative frequency of the previous class. The interpretation of these quantities is straightforward. F_j is the proportion of observations less than of equal to x_j.

A plot of the relative frequencies (or frequency probabilities) against the values of x_i is called the probability density (mass) function (PDF). A plot of the cumulative relative frequencies against x_i is called the cumulative distribution function (CDF).

Table 2: Relative freq and cumulative relative freq for X

Value	Freq.	Rel. Freq	Cumulative Relative Frequency
x_1	f_1	$p_1=f_1/n$	$F_1=f_1/n$
x_2	f_2	$p_2=f_2/n$	$F_2=(f_1+f_2)/n$
x_3	f_3	$p_3=f_3/n$	$F_3=(f_1+f_2+f_3)/n$
…	…	…	…
x_k	f_k	$p_k=f_k/n$	$F_k=\sum_{i=1}^{k} f_i / n = 1$
Total	n	1	

Consider the following example, in which we have information on the age left education (X) for a sample of 106 working males and 154 working females.

Table 3a: Males

15	15	15	15	15	15	15	15	15	16	16	16	16	16	16	16	16
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	17
17	17	17	17	17	17	17	18	18	18	18	18	18	18	18	18	18
19	19	19	19	19	20	21	21	21	21	21	21	21	22	22	22	22
23	23	23	23	23	23	23	23	24	24	24	24	24	25	25	25	25
26	26	27	27													

Table 3b: Females

15	15	15	15	15	15	15	15	15	15	16	16	16	16	16	16	16
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16	16
16	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	18
18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18	18
18	18	19	19	19	19	19	19	20	20	20	20	20	20	21	21	21
21	21	21	21	21	21	21	21	21	21	22	22	22	22	22	22	22
22	22	22	23	23	23	23	23	24	24	24	24	24	25	25	25	26
26																

From this collection of 106 numbers for males and 154 for females it is difficult to gauge much information. We present this data more concisely as a FREQUENCY TABLE, by counting the number of times, that is, the frequency, an individual left school at age=15 at age=16, etc and then present the results in tabular form (compare this table with Table 3)

Table 4: Relative freq and cumulative relative freq for leaving school

	Males			Females		
Age	f_i	p_i	F_i	f_i	p_i	F_i
15	9	0.085	0.085	10	0.065	0.065
16	41	0.387	0.472	59	0.383	0.448
17	8	0.076	0.547	15	0.097	0.546
18	10	0.094	0.642	20	0.130	0.675
19	5	0.047	0.689	6	0.039	0.714
20	1	0.009	0.698	6	0.039	0.753
21	7	0.066	0.764	13	0.084	0.838
22	4	0.038	0.802	10	0.065	0.903
23	8	0.076	0.877	5	0.033	0.935
24	5	0.047	0.925	5	0.033	0.968
25	4	0.038	0.962	3	0.020	0.987
26	2	0.019	0.981	2	0.013	1.000
27	2	0.019	1.000	0	0.000	1.000
Total	106	1.000		154	1.000	

Diagrammatically this can be represented as Figures 2a and 2b. These graphs give a pictorial representation of the distribution of the number of jobs males and females had. Unfortunately, the two distributions are not directly comparable as the number of males and females used in the sample are different. To compare two samples with different numbers of observations, we might think of reporting relative and cumulative relative frequencies (compare with Table 4).

From the table we observe that, 76.4% of the males had left school by the age of 21, whereas 83.8% of females were in this position. Figure 3 plots the relative frequency for males and females together. We can see from this figure that the most likely occurrence is for both males and females to leave school at 16 (these are the modal classes). However, there is some chance that males leave school as late as 27 (0.019).

2. Measures of central tendency

2.1 Arithmetic (simple) mean

Population: $E(X) = \mu = \frac{\sum_{i=1}^{N} x_i}{N}$

Sample[1]: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

Note: the sample mean is your estimate of the population mean. For rules on summation see Appendix 1.

2.2 Median

Middle value of an ordered set of observations, for an odd number of observations this is:

$x_1, x_2, \ldots, x_n : x_{0.5(n+1)}$

For an even number of observations it is $\frac{x_{0.5n} + x_{0.5n+1}}{2}$.

2.3 Mode

Value which occurs most frequently from the set of observations the largest.

[1] For grouped data this would be calculated as $\bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{n}$, where there are k categories on the variable x and f_i represents the frequency of times that x_i occurs.

3. Measures of spread (dispersion)

3.1 Variance

The variance expresses how spread out are a set of numbers and is constructed as the average squared deviation around the mean.

Population: $V(X) = E(X - E(X))^2 = \sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$

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

where n-1 are the degrees of freedom i.e. the number of observations of x_i you can freely choose.

3.2 Standard deviation

Population: $\sigma = \sqrt{\sigma^2}$

Sample: $s = \sqrt{s^2}$

3.2.1 Tchebychev's rule

For any population with mean μ and standard deviation σ, at least $100(1-1/m^2)\%$ of the population lie within m standard deviations around the mean, for $m>1$.

3.3 Coefficient of variation

This is a scaled standard deviation, to ensure the spread is comparable across variables with very different means.

Population: $\frac{\sigma}{\mu}$

Sample: $\frac{s}{\bar{x}}$

[2] For grouped data this would be calculated as $s^2 = \frac{\sum_{i=1}^{k} f_i(x_i - \bar{x})^2}{n-1}$, where there are k categories on the variable x and f_i represents the frequency of times that x_i occurs and $\bar{x}$ is calculated as in footnote 1.

3.4 Mean absolute deviation

Population: $MAD = \frac{\sum_{i=1}^{N}|x_i - \mu|}{N}$

Sample: $MAD = \frac{\sum_{i=1}^{n}|x_i - \bar{x}|}{n}$

3.5 Range

This is the difference between the largest and smallest observations. Assuming the data has been sorted by size, from smallest to largest: $x_n - x_1$

NOTE: This measure can clearly be severely affected by extreme points and does not account for how the points are distributed.

3.6 Interquartile range

Assuming the data has been sorted by size, from smallest to largest, this measures the difference between the 25th and 75th percentile points: $x_{0.75(n+1)} - x_{0.25(n+1)}$

NOTE: This measure is robust to extreme points as it discards the highest and lowest points from the construction of the statistic.

4. Other measures

4.1 Skewness (using frequencies)

In much applied work, only measures of central tendency and dispersion (typically mean and standard deviation) are calculated. However, in some cases this parsimony in reporting of summary statistics may be misleading. Skewness gives a numerical measure of how asymmetric is a distribution. This is calculated as:

Sample:[3] $\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^3}{ns^3}$

A zero value implies a symmetric distribution (see Figure 5), a positive number implies a distribution skewed to the right (positively skewed) (see Figure 6) and a negative number implies a distribution skewed to the left (negatively skewed) (see Figure 7). The figures also show the relationship between the three measures of central tendency when the distribution is asymmetric.

4.2 Kurtosis (using frequencies)

Kurtosis gives a measure of how many observations lie in the tails of the distribution.

Sample[4]: $\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^4}{ns^4}$

A value of three implies the distribution has the same proportion of observations in the tails as a normal distribution. A value less than three implies the distribution is platykurtic – meaning the distribution is flat-topped. A value greater than three implies the distribution is leptokurtic – meaning the distribution is more peaked.

[3] The grouped version of this skewness statistic can be calculated in a similar way to that outlined for the sample variance in footnote 2.
[4] The grouped version of this kurtosis statistic can be calculated in a similar way to that outlined for the sample variance in footnote 2.

5. Continuous grouped data

The analysis of section 4 to data has assumed that the data was discrete in nature and had only a finite number of possible outcomes. Alternatively, the data might be continuous in nature, in which case we may be only have data on the number of occurrences in a given range, for example, Table 5, presents data from a sample of n=833 born in 1958, who fall into various interval ranges for birth weight.

Table 5: Frequency table for a continuous random variable *X* (Birth weights)

Birth weights	freq
1.00-2.00	21
2.00-2.50	41
2.50-2.75	50
2.75-3.00	97
3.00-3.25	167
3.25-3.50	147
3.50-3.75	157
3.75-4.00	74
4.00-4.50	56
4.50-5.00	0
5.00-6.00	23
	n= 833

A histogram plots the standardized frequencies against x_i, such that the areas contained by the bars being drawn proportional to the frequencies of the classes they represent. (A bar chart has vertical bars of equal width, whose height is proportional to the frequencies). We adopt a standardized width of say 0.25kg for the data in Table 5 above. However, the first interval width has 21 babies born in the 1kg width between 1.0-2.0kg. Assuming the baby's births are equal spread over the interval we would expect 5.25 babies to have weights in the interval 1.0-1.25 and 5.25 babies to be in the interval 1.25-1.5 etc. For the next group we have 41 babies born in the 0.5kg width between 2.0-2.5 and we would expect 20.5 babies to be born between 2.0-2.25kg and 20.5 between 2.25-2.5kg. In a histogram it is the frequency in the standardized width one reports on the vertical axis (see Figure 8).

In order to analyse this data it is necessary to establish a representative point for each interval range, for instance we know that 21 outcomes were between 1.0 and 2.0. However, we do not know where in this range the f_2 occurrences lie. In order to make progress, some approximation is needed. Since the exact location is unknown, one

obvious solution is to assume that the occurrences are evenly distributed across the class interval, in which case we can assign to each occurrence in the class interval the mid-point of that interval. The only complicated issue is the fact that the top class (and possibly the bottom class) are often not closed and therefore the mid-point is unbounded. In this case we determine a likely (reasonable) upper (lower) bound. If we do this we can proceed to calculate the mean and variance as we did in the case for when the data was discrete see section 4.1 when we had data on the number of different job held by both males and females.

Table 6: Frequency table for the continuous random variable *X* using mid-points

Birth weights	*f*	*x*	*fx*	*fx*^{2}
1.00-2.00	21	1.500	31.500	47.25
2.00-2.50	41	2.250	92.250	207.5625
2.50-2.75	50	2.625	131.250	344.5313
2.75-3.00	97	2.875	278.875	801.7656
3.00-3.25	167	3.125	521.875	1630.859
3.25-3.50	147	3.375	496.125	1674.422
3.50-3.75	157	3.625	569.125	2063.078
3.75-4.00	74	3.875	286.75	1111.156
4.00-4.50	56	4.125	231.000	952.875
4.50-5.00	0	4.750	0.000	0.000
5.00-6.00	23	5.500	126.500	695.75
	n= 833	$\sum f_i x_i$ =	2765.25	$\sum f_i x_i^2$ = 9529.25
		$\bar{x}$ =	3.320	*s*= 0.648

where $m_j=(x_{j-1}+x_j)/2$.

The mean and standard deviation can then be easily calculated as previously and the modal group, which occurs most frequently is 3.0-3.25kg. The median group is that which includes the 50th percentile, the 416th case based on a sample size of 833. For the median sometimes people actual calculate the median point in the interval as:

$F_{50} = L + \left[\frac{0.5n - f_L}{f}\right]c$. In the example above $F_{50} = 3.25 + \left[\frac{416-376}{147}\right]0.25 = 3.32$

where n is the number of observations, L is the lower boundary in that class in which the $0.5n^{th}$ case falls, and c is width of that class, f_L is the cumulative frequency up to the lower point L and f is the number of observations in the median class. In general, we can calculate any percentile point as

$$F_j = L + \left[\frac{\frac{jn}{100} - f_L}{f}\right]c$$

such that the 25th percentile is

$$F_{25} = L + \left[\frac{0.25n - f_L}{f}\right]c$$

6. Measures of linear association

In Economics and business we are interested in the relation between 2 or more random variables, for example:

Advertising expenditure and sales revenue

Personal consumption and disposable income

Investment and interest rates

Earnings and schooling

while there are many ways in which these pairs of random variables might be related – a linear relationship is often a useful first approximation.

6.1 Covariance

The association might be STRONG, when a scatter plot, of *Y* against *X*, will be tightly clustered around a straight line, or weak with a scatter plot more widely dispersed about a line. A plot of the data is a necessary preliminary to data analysis, but more sophisticated techniques than a graphical inspection are often required.

The sample covariance between two random variables *X* and *Y* is defined as:

$$s_{XY} = \text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}.\bar{y}}{n-1}$$

the degrees of freedom are only *n*-1 as we actually only need to know the mean of *x* or *y*.

The covariance measures the average cross product or *x* relative to its mean with *y* relative to its mean. Consequently, if high (low) values of *x* - relative to its mean - are associated with high (low) values of *y* - relative to its mean – then we get a high positive covariance (see Figure 9). Conversely if high (low) values of *x* are associated with low (high) values of *y* we get a negative covariance (see Figure 10). A zero covariance occurs when there is no predominant association between the *x* and *y* values (see Figure 11). NOTE: The covariance is a linear association between *x* and *y* values and would be approximately zero for a quadratic association (see Figure 12).

6.2 Correlation

The covariance measure is not scale free and multiplying the *x* variable by 100 multiplies the covariance by 100.

A scale free measure is a correlation:

$$\rho_{XY} = corr(X,Y) \equiv \rho(X,Y) = \frac{\text{cov}(X,Y)}{\sqrt{V(X)V(Y)}}$$

as

$$corr(aX,Y) = \frac{\text{cov}(aX,Y)}{\sqrt{V(aX)V(Y)}} = \frac{a\,\text{cov}(X,Y)}{\sqrt{a^2V(X)V(Y)}} = \frac{a\,\text{cov}(aX,Y)}{a\sqrt{V(X)V(Y)}} = corr(X,Y)$$

ρ is a population parameter of association between the random variables *X* and *Y*. In practice for a sample of *n* pairs of observations $(x_1, y_1)\ldots(x_n, y_n)$ the estimated sample correlation, r, is calculates as:

$r_{XY} = \frac{s_{XY}}{s_X s_Y}$ where,

$s_X^2 = \sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)$, $s_Y^2 = \sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1)$, $\bar{x} = \sum_{i=1}^{n} x_i/n$ and $\bar{y} = \sum_{i=1}^{n} y_i/n$.

6.2.1 Properties of correlation

1. $-1 \le \rho(X,Y) \le 1$
2. $\rho(X,Y) = -1 \Rightarrow$ perfect negative association
3. $\rho(X,Y) = 1 \Rightarrow$ perfect positive linear association
4. $\rho(X,Y) = 0 \Rightarrow$ no linear association
5. As $|\rho(X,Y)|$ increases $\Rightarrow$ stronger association.

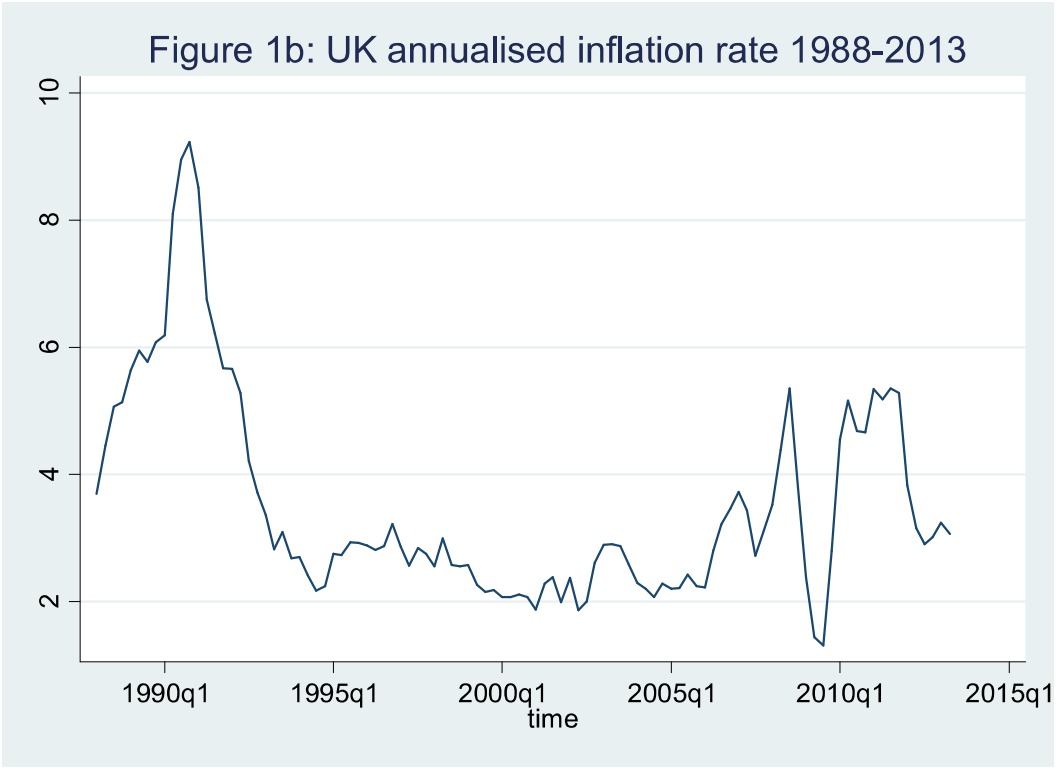


Figure 1b: UK annualised inflation rate 1988-2013
10
8
6
4
2
1990q1
1995q1
2000q1
2005q1
2010q1
2015q1
time

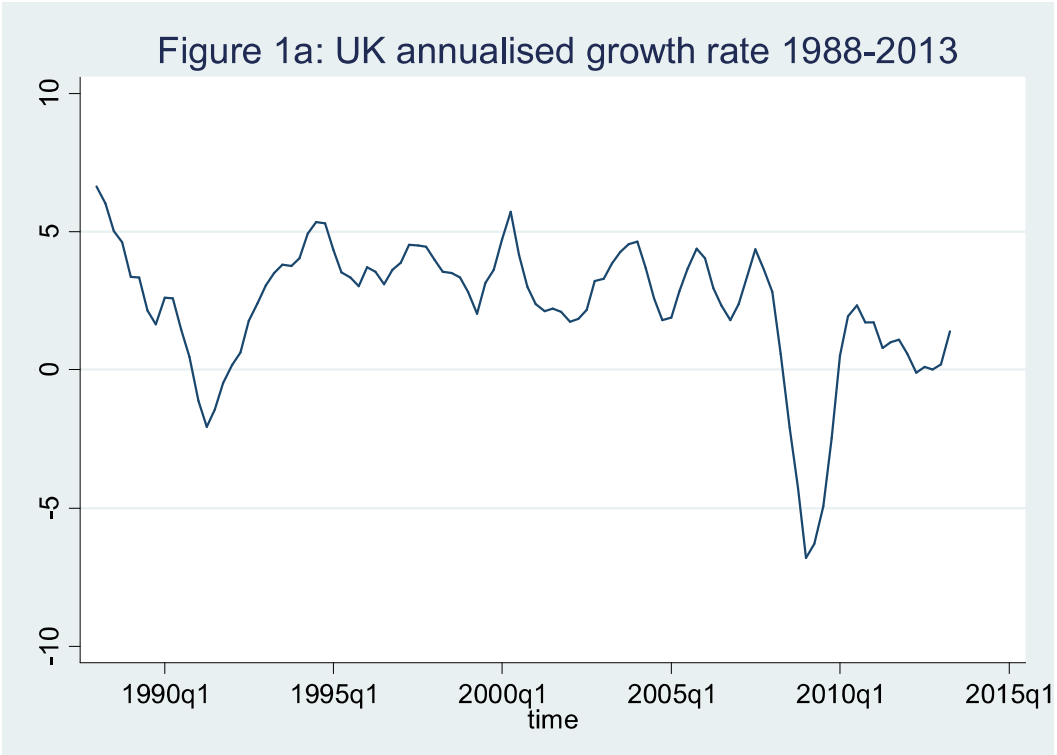


Figure 1a: UK annualised growth rate 1988-2013
10
5
0
-5
-10
1990q1
1995q1
2000q1
2005q1
2010q1
2015q1
time

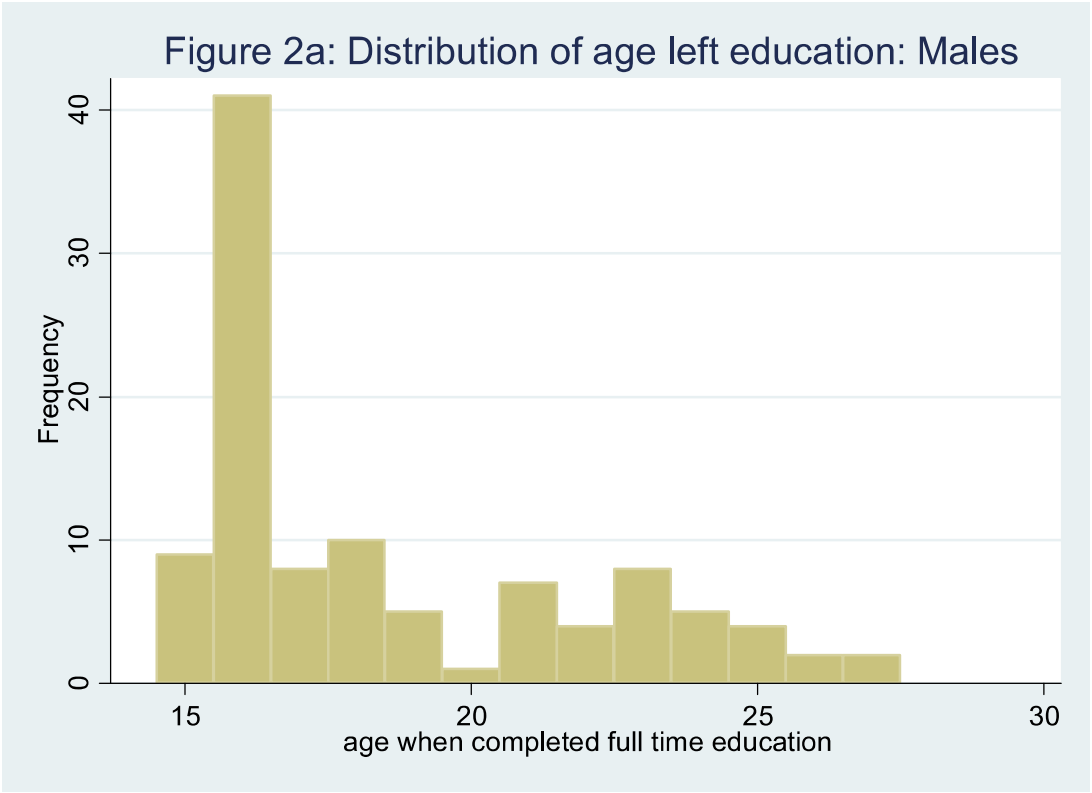


Figure 2a: Distribution of age left education: Males
Frequency
0
10
20
30
40
15
20
25
30
age when completed full time education

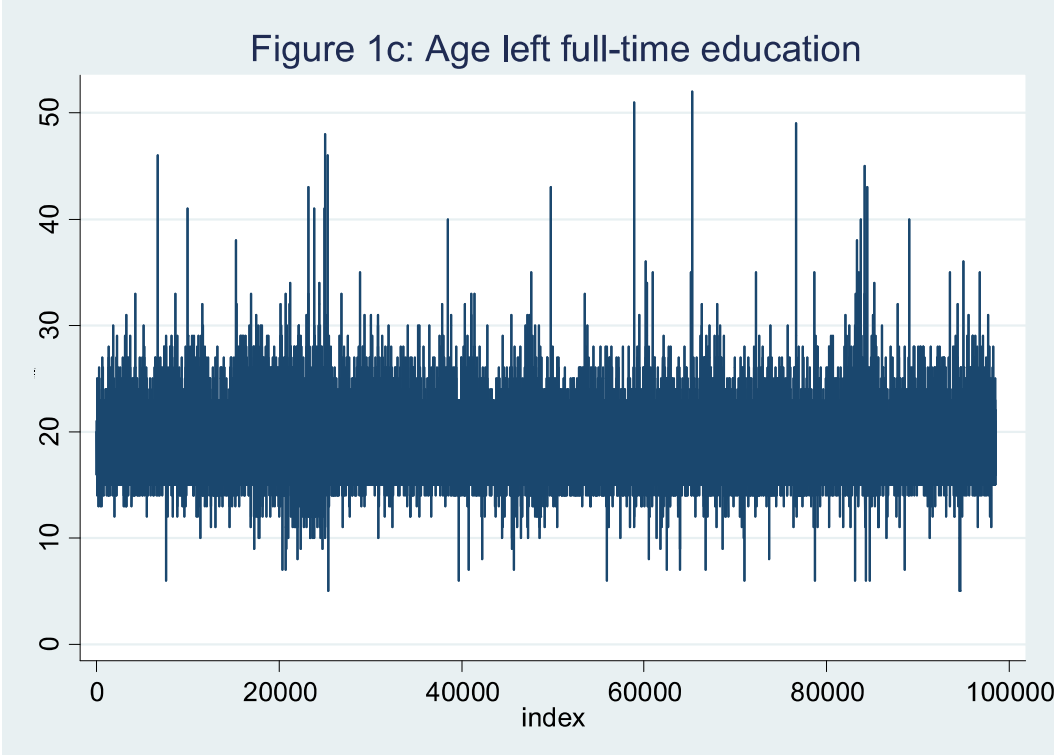


Figure 1c: Age left full-time education
0
10
20
30
40
50
0
20000
40000
60000
80000
100000
index

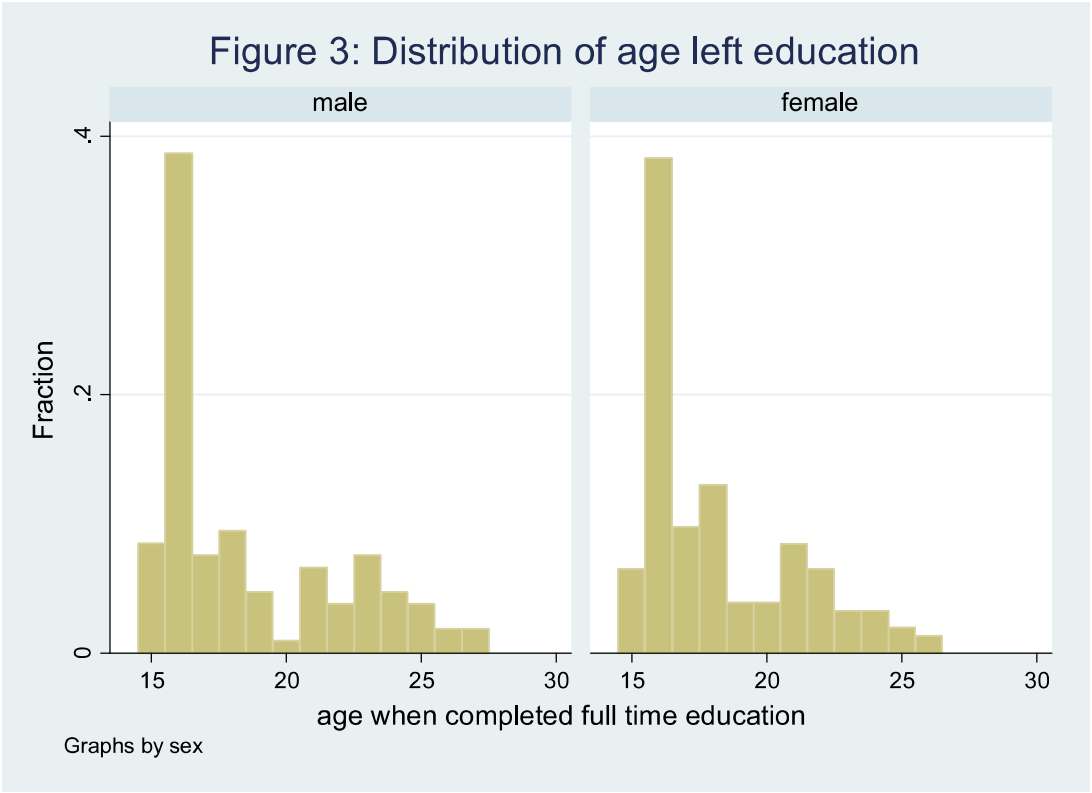


Figure 3: Distribution of age left education
male
female
Fraction
0
.2
.4
15
20
25
30
age when completed full time education
Graphs by sex

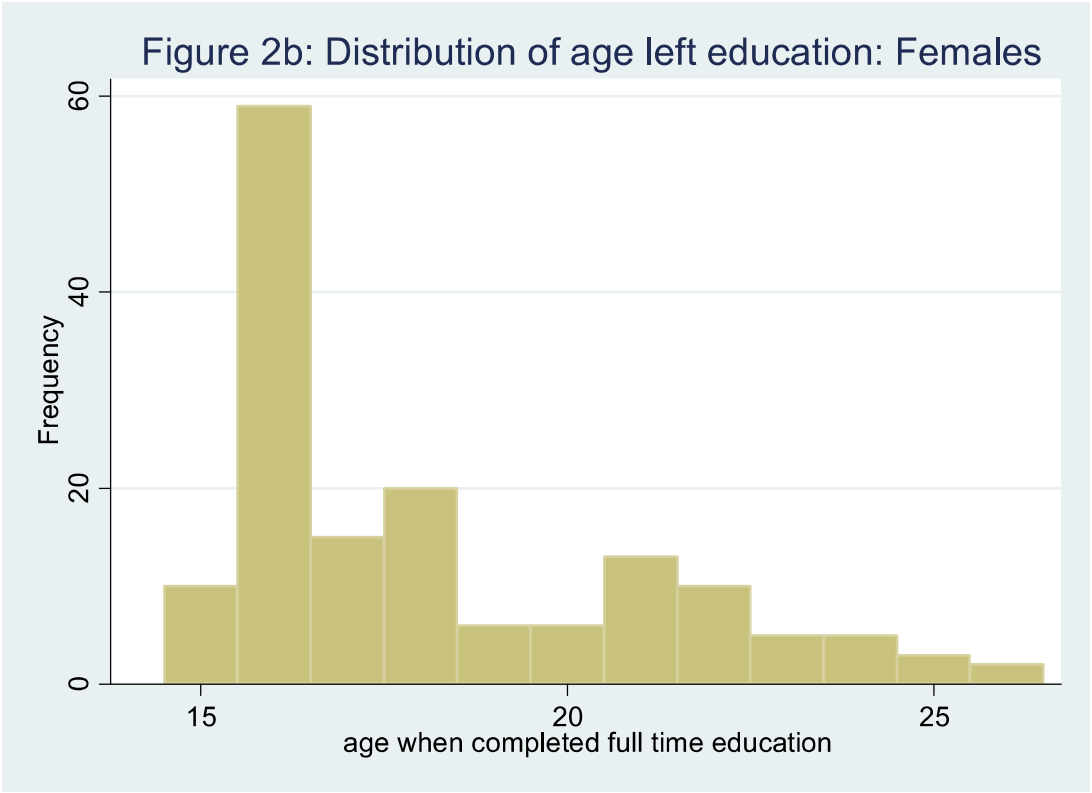


Figure 2b: Distribution of age left education: Females
Frequency
0
20
40
60
15
20
25
age when completed full time education

Figure 5: Symmetric distribution

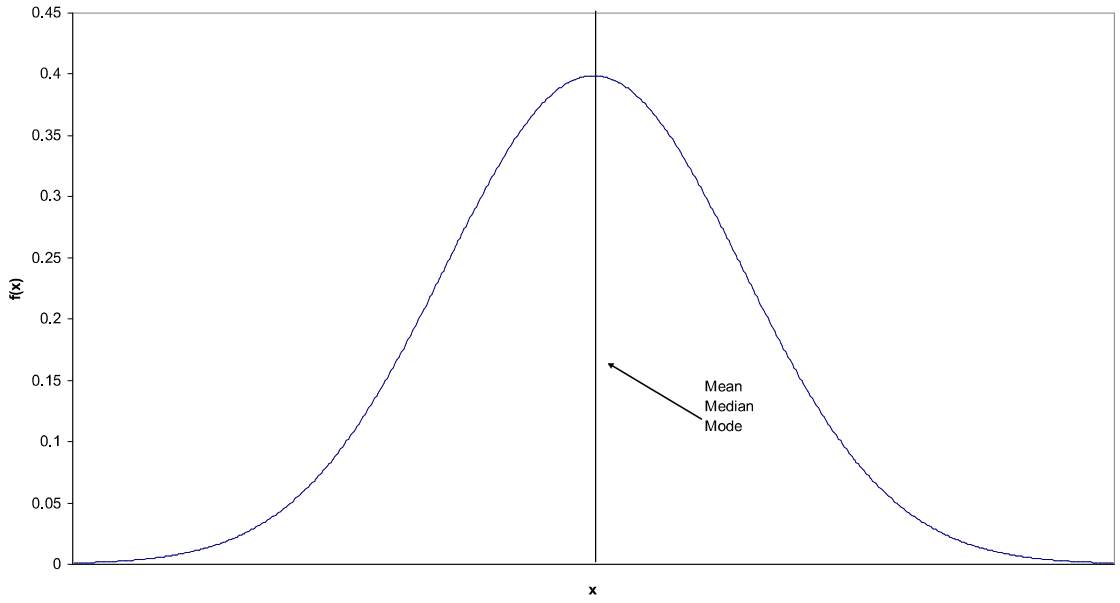


Figure 4: Cumulative distribution for age left education: Males

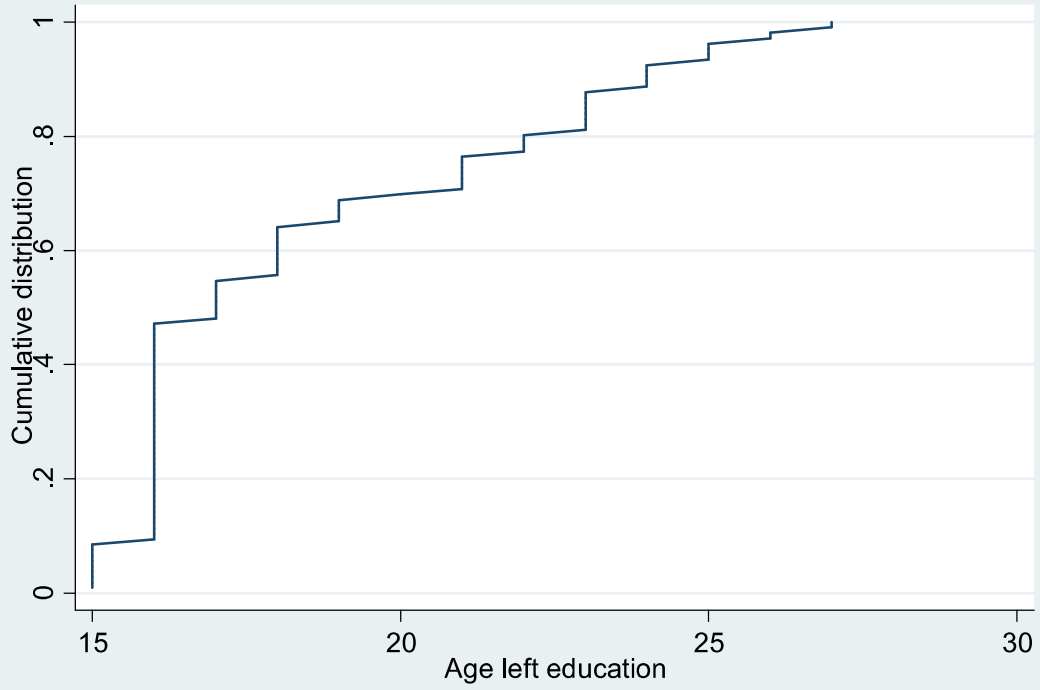


Figure 7: Negative skewed distribution

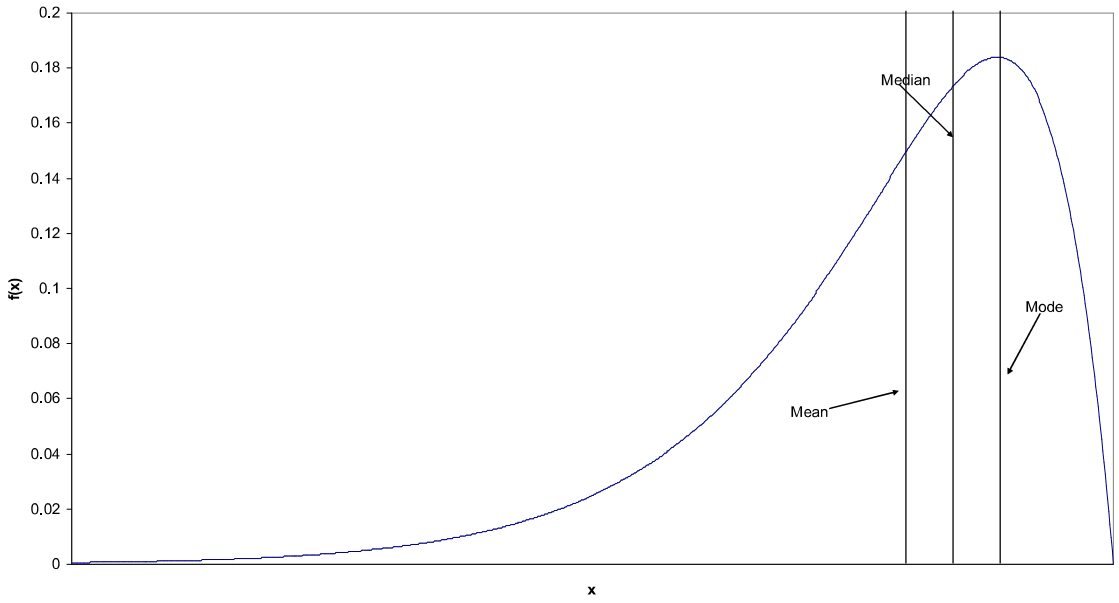


Figure 6: Positively skewed distribution

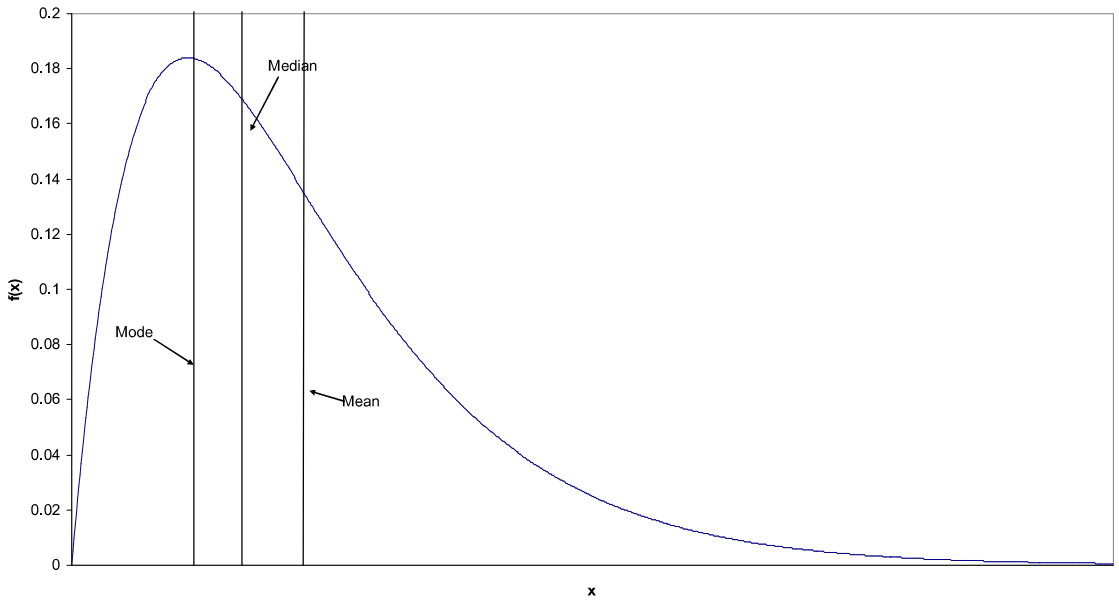


Figure 9: Positive correlation between y and x

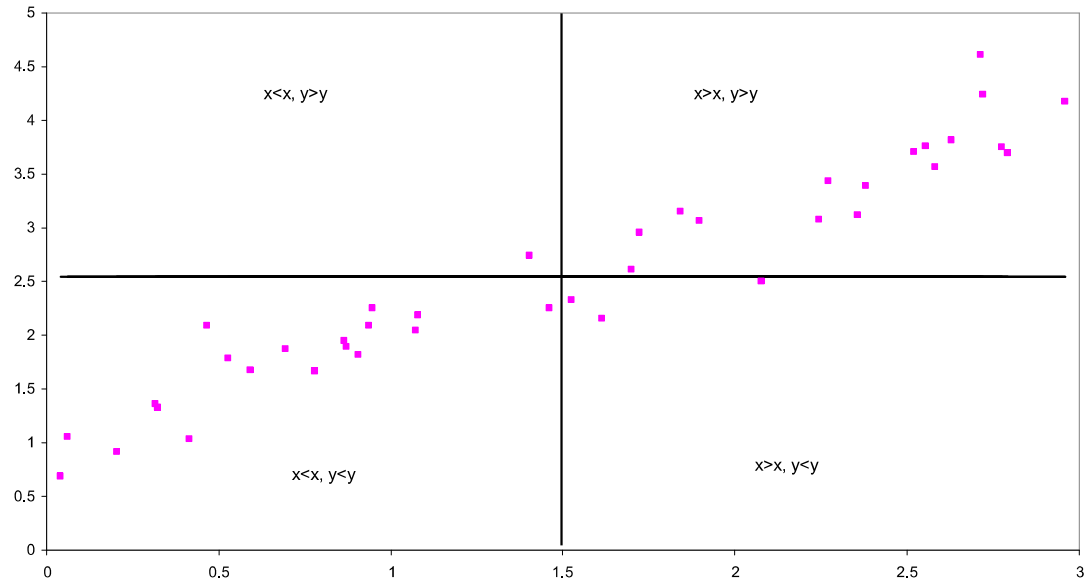


Figure 8: Distribution of birth weight

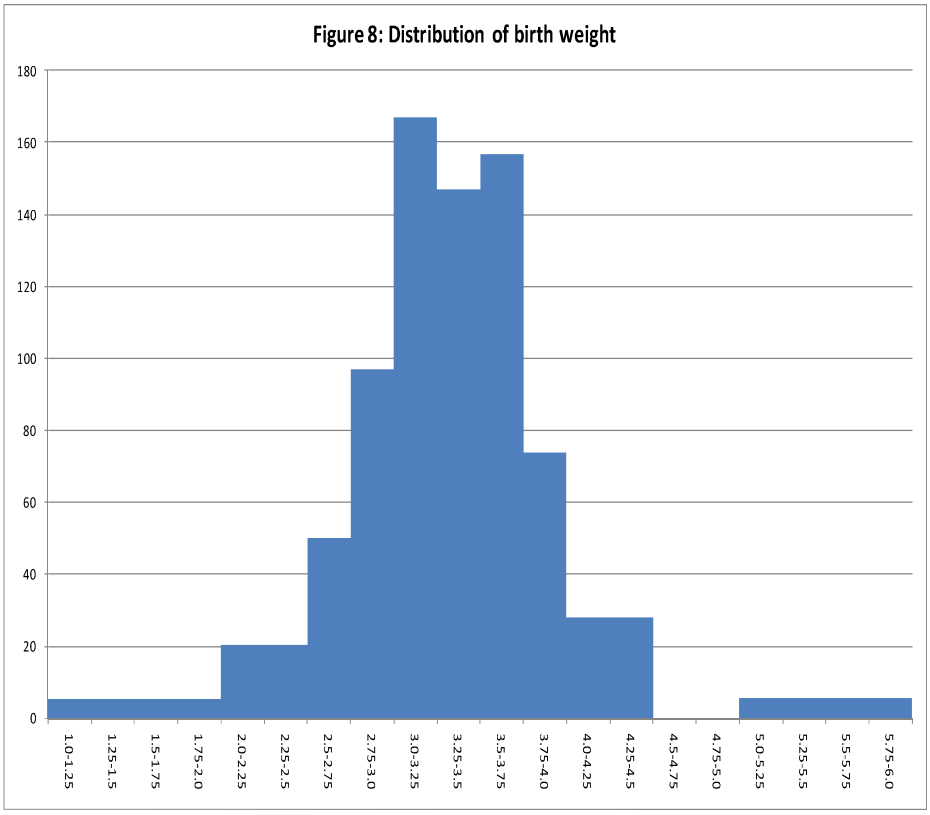


Figure 11: Zero correlation between y and x

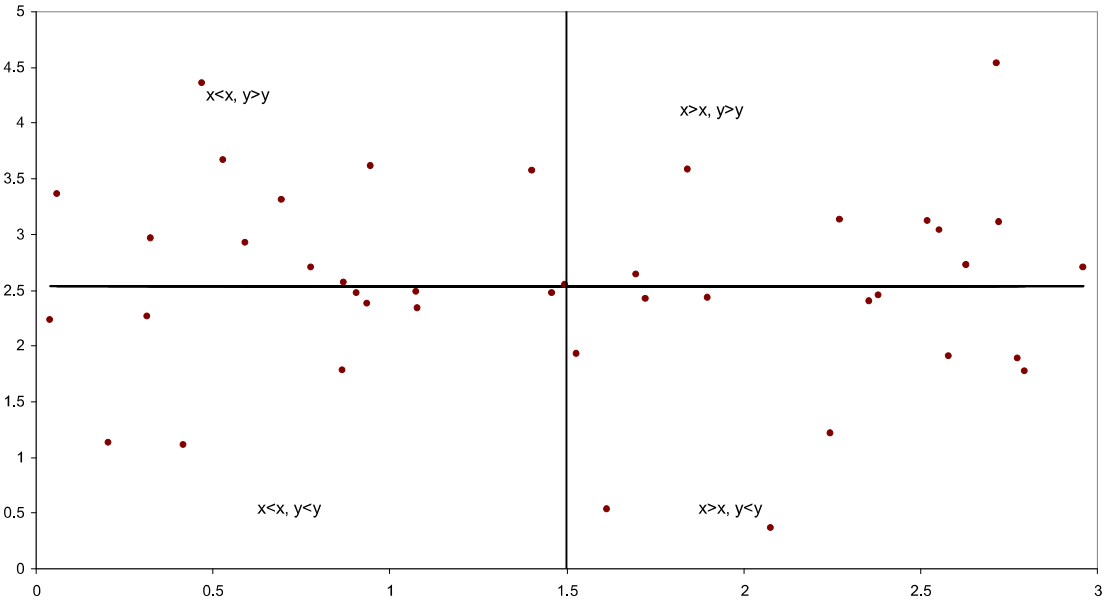
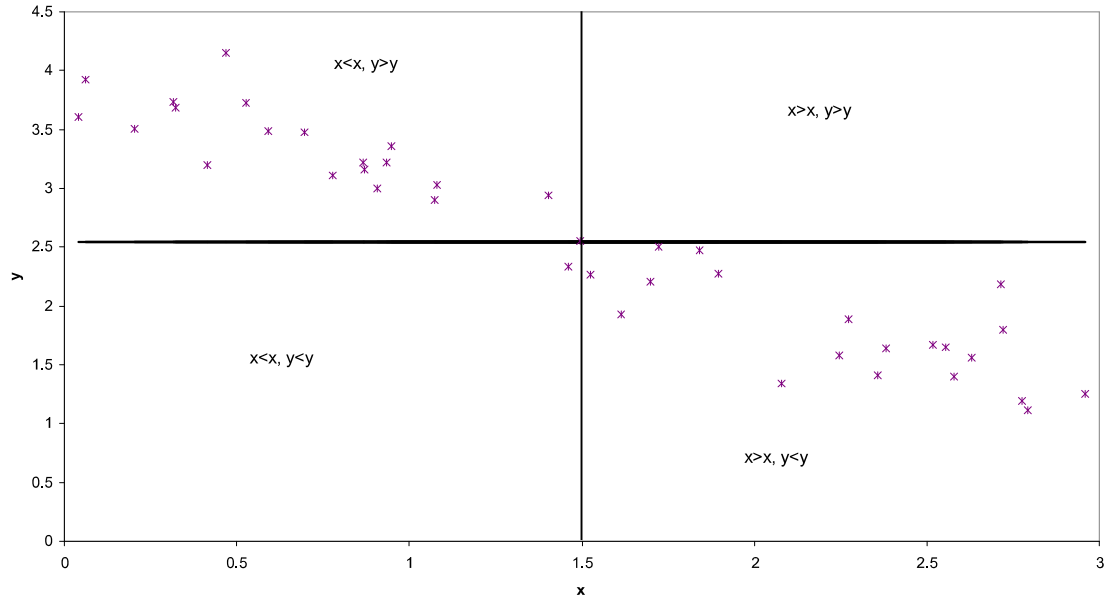


Figure 10: Negative correlation between y and x



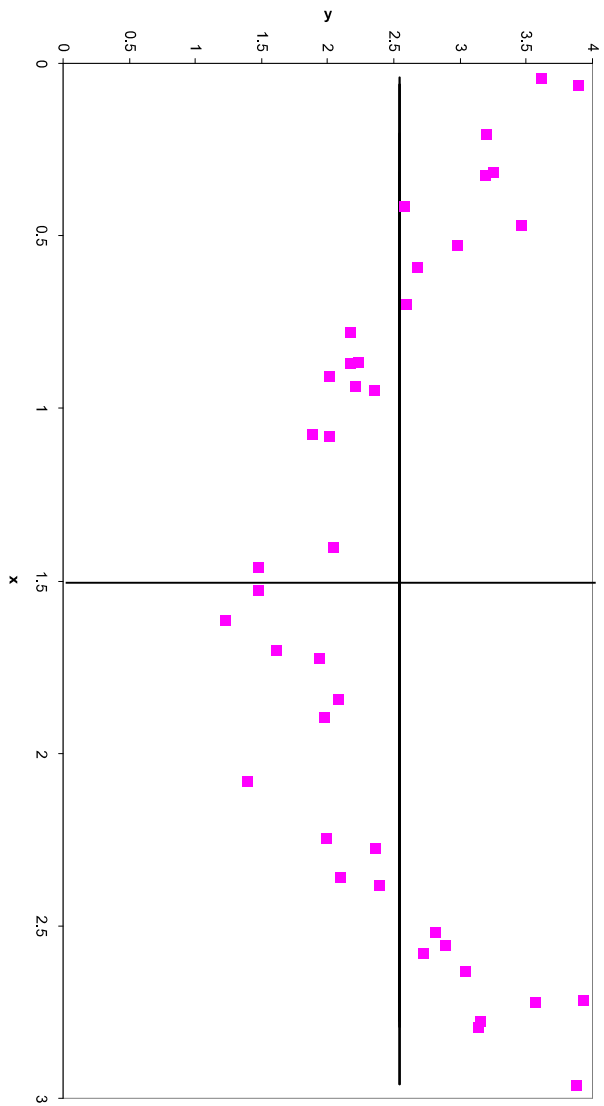


Figure 12: Zero correlation between y and x

Descriptive Statistics: Examples

1. The following list of numbers was the result of asking 30 people how many brothers and sisters they had:

2	1	1	2	0	1	0	3	2	0	1	1	5	1	2
2	2	1	1	3	0	2	4	0	0	2	1	0	1	2

(i) Find the mean, median and mode for this set of numbers.

(ii) Find the standard deviation for this set of numbers.

(iii) Assuming that each of the 30 people questioned were not related to each other, calculate the average number of children in the 30 families.

(iv) Calculate the standard deviation of the number of children in each family.

Answer

No. siblings	Freq	Cum. Freq.
0	7	7
1	10	17
2	9	26
3	2	28
4	1	29
5	1	30

(i) Mode=1, Median=1, $\sum_i x_i = 43$, Mean=1.43

(ii) $\sum_i x_i^2 = 105$, $s^2 = \frac{105 - 30 \times 1.43^2}{29} = 2.14$

(iii) No. of children=No. of siblings+1, therefore $Y = X + 1$, $E(Y) = E(X) + 1$, mean=2.43

(iv) $V(Y) = V(X)$, variance=2.14.

2. For the data given in Table 8 on the number of jobs for the sample of males and females, calculate (i) the sample mean, (ii) median, (iii) mode, (ii) the sample variance

Answer

(i) $\bar{x}^M = \frac{1 \times 13 + 2 \times 15 + 3 \times 21 + 4 \times 12 + 5 \times 9 + 6 \times 5 + 7 \times 4 + 8 \times 4 + 14 \times 2}{85} = \frac{317}{85} = 3.73$

or

$\bar{x}^M = 1\times0.153+2\times0.176+3\times0.247+4\times0.141+5\times0.106+6\times0.059+7\times0.047+$
$8\times0.047+14\times0.024=3.73$

$$\bar{x}^F = \frac{1\times13+2\times7+3\times17+4\times23+5\times21+6\times15}{96} = \frac{365}{96} = 3.80$$

or

$\bar{x}^F = 1\times0.135+2\times0.073+3\times0.177+4\times0.240+5\times0.219+6\times0.156=3.80$

(ii) median^M=3, median^F=4

(iii) mode^M=3, mode^F=4

(iv)

$$s_M^2 = \frac{1^2\times13+2^2\times15+3^2\times21+4^2\times12+5^2\times9+6^2\times5+7^2\times4+8^2\times4+14^2\times2-85\times3.73^2}{84}$$

$$s_M^2 = \frac{1703-85\times3.73^2}{84} = 6.20$$

or

$s_M^2 = 1^2\times0.153+2^2\times0.176+3^2\times0.247+4^2\times0.141+5^2\times0.106+6^2\times0.059+7^2\times0.047+$
$8^2\times0.047+14^2\times0.024-3.73^2=6.20$

$$\bar{x}^F = \frac{1^2\times13+2^2\times7+3^2\times17+4^2\times23+5^2\times21+6^2\times15-96\times3.80^2}{95}$$

$$s_F^2 = \frac{1627-96\times3.80^2}{95} = 2.52$$

or

$s_F^2 = 1^2\times0.135+2^2\times0.073+3^2\times0.177+4^2\times0.240+5^2\times0.219+6^2\times0.156-3.80^2=2.52$

3. Draw a histogram of the amount of time spent studying on a particular module outside the usual lecture and tutorial hours in an average (typical) week for the sample of 35 students, given in the Table below.

Amount of time spent studying on a particular module

Minutes	No. students
<20	2
20-<40	5
40-<60	4
60-<90	6
90-<120	5

120-<180	7
180-<240	3
240-<360	2
≥360	1
Total	35

Answer

Neither the bottom nor the top class is not closed. For the bottom class the obvious lowest bound is zero, the upper bound is less obvious and we determine a likely upper bound (8 hours: 540 minutes). (see sheet).

4. Using the data given in the Table above for the sample of 35 students, calculate the (i) sample mean, (ii) sample standard deviation, (iii) the median and (iv) the inter-quartile range.

Answer

Amount of time spent studying on a particular module

Mid point	No. students	Cum. Freq.	Rel Freq
10	2	2	0.057
30	5	7	0.143
50	4	11	0.114
75	6	17	0.171
105	5	22	0.143
150	7	29	0.200
210	3	32	0.086
300	2	24	0.057
420	1	35	0.029
Total	35		1.000

(i) $\bar{x}=115.6$, (ii) $s=91.1$, (iii) For the median sometimes people actual calculate the median point in the interval as:

$$F_{50} = 90+\left[\frac{18-17}{5}\right]30 = 95 \text{ minutes.}$$

(iv) $F_{25} = 40+\left[\frac{9-7}{4}\right]20 = 50$, $F_{75} = 120+\left[\frac{27-22}{7}\right]60 = 162.86$

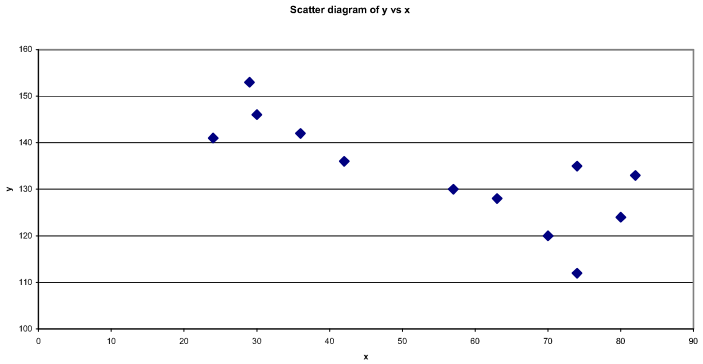
5. Consider the following sample of bivariate data

x	42	24	82	74	70	36	57	29	63	74	80	30
y	136	141	133	135	120	142	130	153	128	112	124	146

(a) plot a scatter graph of y against x.

(b) Calculate the covariance between y and x.
(c) Calculate the correlation between y and x.

Answer



(b) $\text{cov}(x,y) = \dfrac{\sum_{i=1}^{12} x_i y_i - n\overline{xy}}{n-1} = \dfrac{86003 - 12 \times 55.08 \times 133.33}{11} = -193.67$

(c) $corr(x,y) = \dfrac{\text{cov}(x,y)}{s_x s_y} = \dfrac{-193.67}{21.66 \times 11.48} = -0.779$.

Rules of summation

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

2. $\sum_{i=1}^{n} c = (c + c + c + \ldots + c) = nc$

3. $\sum_{i=1}^{n} cX_i = (cX_1 + cX_2 + cX_3 + \ldots + cX_n) = c(X_1 + X_2 + X_3 + \ldots + X_n) = c\sum_{i=1}^{n} X_i$

4. $\sum_{i=1}^{n} (X_i + Y_i) = (X_1 + Y_1) + (X_2 + Y_2) + (X_3 + Y_3) \ldots + (X_n + Y_n)$
$= (X_1 + X_2 + X_3 \ldots + X_n) + (Y_1 + Y_2 + Y_3 + \ldots + Y_n) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

5. $\sum_{i=1}^{n} (X_i + Y_i)^2 = (X_1 + Y_1)^2 + (X_2 + Y_2)^2 + (X_3 + Y_3)^2 \ldots + (X_n + Y_n)^2$
$= (x_1^2 + Y_1^2 + 2X_1Y_1) + (X_2^2 + Y_2^2 + 2X_2Y_2) + \ldots (X_n^2 + Y_n^2 + 2X_nY_n)$
$= \sum_{i=1}^{n} (X_i^2 + Y_i^2 + 2X_iY_i)$

6. $\sum_{i=1}^{n} (cX_i + cY_i)^2 = \sum_{i=1}^{n} (c^2X_i^2 + c^2Y_i^2 + 2c^2X_iY_i) = c^2 \sum_{i=1}^{n} (X_i^2 + Y_i^2 + 2X_iY_i)$
$= c^2 \sum_{i=1}^{n} (X_i + Y_i)^2$

Rules on expectations variances

Define $E(X)=\sum_{i=1}^{k} p_i x_i$ as the expected value of the random variable X and

$V(X)=E[X-E(X)]^2=E(X^2)-E(X)^2=\sum_{i=1}^{k} p_i(x_i-E(X))^2$.

1. $E(a+X)=\sum_{i=1}^{k} p_i(a+x_i)=a+\sum_{i=1}^{k} p_i x_i=a+E(X)$

2. $E(aX)=\sum_{i=1}^{k} p_i(ax_i)=a\sum_{i=1}^{k} p_i x_i=aE(X)$

3. $$V(a+X)=\sum_{i=1}^{k} p_i\left[(a+x_i)-E(a+X)\right]^2=\sum_{i=1}^{k} p_i\left[(a+x_i)-a-E(X)\right]^2$$
$$=\sum_{i=1}^{k} p_i\left[x_i-E(X)\right]^2=V(X)$$

4. $$V(aX)=\sum_{i=1}^{k} p_i\left[ax_i-E(aX)\right]^2=\sum_{i=1}^{k} p_i\left[ax_i-aE(X)\right]^2$$
$$=a^2\sum_{i=1}^{k} p_i\left[x_i-E(X)\right]^2=a^2V(X)$$

Suppose $E(X)=\mu$ and $V(X)=\sigma^2$ and define $Z=\frac{X-\mu}{\sigma}$, then

$$E(Z)=E\left[\frac{X-\mu}{\sigma}\right]=\frac{1}{\sigma}E\left[X-\mu\right]=\frac{1}{\sigma}\left[\underbrace{E(X)}_{\mu}-\mu\right]=0$$

(using rules (2) and (1))

$$V(Z)=V\left[\frac{X-\mu}{\sigma}\right]=\frac{1}{\sigma^2}V(X-\mu)=\frac{1}{\sigma^2}\underbrace{V(X)}_{\sigma^2}=1$$

(using rules (4) and (3)) therefore $E(Z)=0$ and $V(Z)=1$, this is a standardised variable.

While these rules have been derived for a discrete distribution on the random variable, X, similar arguments would hold if X was a continuous random variable. The exception being that the summation sign would be replaced by an integral

Log transformation as an approximation to a growth rate

Define g as the growth rate, that is: $g=\frac{(X_t-X_{t-4})}{X_{t-4}}\equiv\frac{X_t}{X_{t-4}}-1$. Now rearranging we have:

$1+g=\frac{X_t}{X_{t-4}}$. Consider taking natural logs of both sides in which case we have:

$$\ln(1+g)=\ln\left(\frac{X_t}{X_{t-4}}\right)\equiv\ln(X_t)-\ln(X_{t-4})$$

As the log of a ratio is the same as the difference in logs.

Consider the following table:

g	(1+g)	ln(1+g)	$\approx \ln(1+g)$
0.01	1.01	0.00995	0.01
0.02	1.02	0.01980	0.02
0.03	1.03	0.02956	0.03
0.04	1.04	0.03922	0.04
0.08	1.08	0.07696	0.08
0.12	1.12	0.11333	0.11
0.20	1.20	0.18232	0.18
0.50	1.50	0.33647	0.34
-0.01	0.99	-0.01005	-0.01
-0.02	0.98	-0.02020	-0.02
-0.03	0.97	-0.03046	-0.03
-0.04	0.96	-0.04082	-0.04
-0.08	0.92	-0.08338	-0.08
-0.12	0.88	-0.12783	-0.13
-0.20	0.80	-0.22314	-0.22
-0.50	0.50	-0.51083	-0.51

in which case: $\ln(1+g)\approx g=\ln(X_t)-\ln(X_{t-4})$.

Comparing alternate measures of UK GDP growth - 1960-2009 (quarterly data)

STATISTICAL TECHNIQUES B

Probability

1. Introduction

There is an experiment whose *outcome* is random and which can be repeated. Define Ω as the *sample space*, which is the set of all possible outcomes of the experiment and the basic (or elementary) outcomes are defined as C_i, i=1,…k, and these are the list of all possible outcomes and only one of the list can be the outcome in any particular experiment.

Example:

(i) If the experiment is rolling a dice then $\Omega = \{1,2,3,4,5,6\}$ and C_i =the number on the face of the die; $C_1 = \{1\}$, $C_2 = \{2\} \ldots C_6 = \{6\}$

(ii) If the experiment was rolling two die then $\Omega = \{(1,1),(1,2),\ldots(1,6)\ldots(6,6)\}$ and C_i is the number on the faces combined.

The *events*, A, are a collection of one or more outcomes of the basic outcomes (C_i) and are a subset of Ω.

Example:

(i) Let A = the event that an odd face is rolled = $\{1,3,5\}$ and $A \subset \Omega$.

(ii) Let A = Ordered pair of Ω which sum to 7 = $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$.

Generally the events A and B are said to be mutually exclusive events if they share no common element, that is, $A \cap B = \varnothing$. The events C_1, C_2, C_3 are said to be all the elementary events of Ω if:

(i) Each pair of the elementary events is mutually exclusive, that is,
$C_i \cap C_j = \varnothing$ for $i \neq j$

(ii) The events are exhaustive, that is $C_1 \cup C_2 \cup C_3 = \Omega$.

2. Set Theory and Venn Diagrams

A set is a selection of objects. Members of a set are referred to as elements and are written inside brackets, {}. Some notation:

(1) $\in$ ($\notin$) means belongs (does not belong) to

(a) The set $A = \{1,2,3,4\}$ and $1 \in A$

(b) The set $B = \{x \mid 0 \le x \le 1\}$ and $0 \in B$

(c) The set $C = \{\text{Heads,Tails}\}$ and $\text{Heads} \in C$

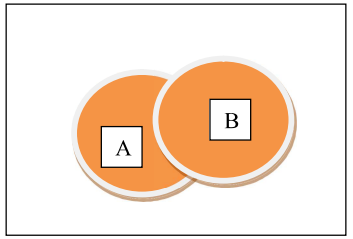
(d) The set $D = \{\text{real numbers} | x^2 = -1\}$ and $D = \varnothing$

(2) $E \subset A$, meaning the set E is a subset of the set A, implying that all elements that belong to E also belong to A.

(a) $A = \{1,2,3,4\}$ and $E = \{1,2\}$ then $E \subset A$

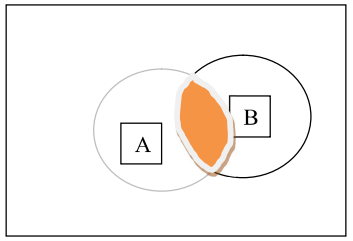
(3) $A \cup E$ is the set of all elements which are in A or E or both

(a) $A = \{0,0.5,1\}$, $E = \{1,2,3\} \Rightarrow A \cup E = \{0,0.5,1,2,3\}$



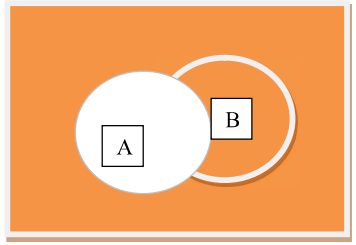
(4) $A \cap E$ is the set of all elements which belong to both A and E

(a) $A = \{0,0.5,1\}$, $E = \{1,2,3\} \Rightarrow A \cap E = \{1\}$



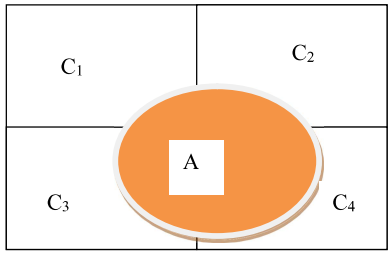
(5) $\overline{A}$ is the complementary event to A, and $\overline{A} = \{a \in \Omega; a \notin A\}$

(a) $\Omega = \{1,2,3,4,5,6,7,8,9,10\}$, $A = \{1,2,3,4\} \Rightarrow \overline{A} = \{5,6,7,8,9,10\}$



(6) If $C_1, C_2, \ldots C_k$ are mutually exclusive and exhaustive events, and if A is some other event then:

$A = (C_1 \cap A) \cup (C_2 \cap A) \cup \ldots (C_k \cap A)$



The is the "union-intersection" rule.

2.1 Rules

(1) $A \cup B = B \cup A$

(2) $A \cap B = B \cap A$

(3) $A \cup (B \cup C) = (A \cup B) \cup C$

(4) $A \cap (B \cap C) = (A \cap B) \cap C$

(5) $A \cap \varnothing = \varnothing$

(6) $A \cup \varnothing = A$

(7) $\overline{\overline{A}} = A$

(8) $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

(9) $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

(10) $\overline{(A \cap B)} = \overline{A} \cup \overline{B}$

(11) $\overline{(A \cup B)} = \overline{A} \cap \overline{B}$

3. Three types of probability

1. Classical:

When basic outcomes are equally likely (and there are N of these) and an experiment, A, has N_A of these, then:

$$P(A) = \frac{N_A}{N}$$

2. Frequentist

When the basic outcomes are not equally likely, an empirical approach to determine the probability of an event A occurring is to observe how often in n trials the event A actually occurred

$$P(A) \approx \frac{n_A}{n}$$

3. Prior posterior

A Bayesian approach to probability is to base the probability of an event A on your belief of the likelihood of the event occurring.

4. Probability

The probability of an event *A*, *P(A)*, is the probability that an outcome of the experiment is *A*.

Example:

(i) *A* is made up of 3 outcomes and Ω of 6, then *P(A)*=3/6=1/2.

(ii) *A* is made up of 6 outcomes, Ω of 36, then *P(A)*=6/36=1/6.

Let *P* be a function which assigns a real number, *P(A)*, to *A*, $\forall A \subset \Omega$. Then *P* is a probability measure if:

(i) *P(A)*≥0

(ii) If $C_i \cap C_j = \varnothing$, for *i*≠*j*, then $P(C_1 \cup C_2 \cup C_3 \ldots) = P(C_1) + P(C_2) + \ldots$

(iii) *P(Ω)*=1

4.1 Useful properties of probability

(i) Let $\overline{A}$ be the complementary event to *A*, then $\overline{A} = \{a \in \Omega; a \notin A\}$ (where $\in (\notin)$ belongs to (does not belong to)), $P(\overline{A}) = 1 - P(A)$

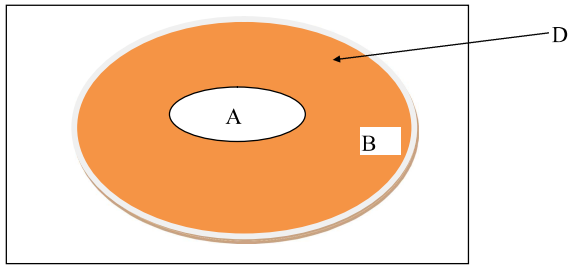
Proof

$A \cap \overline{A} = 0 \Rightarrow P(A \cup \overline{A}) = P(A) + P(\overline{A})$

but, $A \cup \overline{A} = \Omega \Rightarrow P(A \cup \overline{A}) = P(A) + P(\overline{A}) = P(\Omega) = 1 \Rightarrow P(\overline{A}) = 1 - P(A)$.

(ii) $P(0) = 0$

(iii) If $A \subset B$ then $P(A) \leq P(B)$



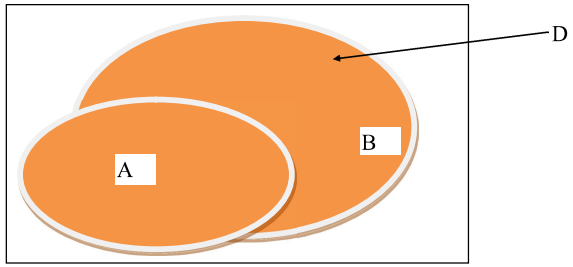
Proof

Let $D = B \cap \bar{A} \Rightarrow D \cap A = 0$ Therefore, $P(D \cup A) = P(B) = P(D) + P(A)$ as

$P(D) \geq 0$ this implies $P(B) \geq P(A)$

(iv) For each event $C_i \in \Omega$, $0 \leq P(C_i) \leq 1$

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



Proof

$P(A) = P(A \cap B) + P(A \cap \bar{B})$ and $P(B) = P(A \cap B) + P(\bar{A} \cap B)$

$$P(A \cup B) = P(A) + P(D) = P(A) + \underbrace{P(\bar{A} \cap B)}_{P(B) - P(A \cap B)} = P(A) + P(B) - P(A \cap B)$$

If A and B are mutually exclusive, such that $A \cap B = \varnothing$ then,

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

Example: Univariate distribution

	C_1	C_2	C_3	C_4	C_5	C_1
Highest Qual	No Qual	Other	GCSE	A-level	HE	Degree
Pr(.)	0.080	0.103	0.221	0.217	0.109	0.271

Then: $C_j \quad j = 1,2,3,4,5,6$ are elementary events, $C_i \cap C_j = 0 \quad i \neq j$ and

$C_1 \cup C_2 \cup C_3 \cup C_4 \cup C_5 \cup C_6 = \Omega$.

Define

$A = \text{Post compulsory qualifications}\{\text{A-level,HE, Degree}\}$

$\Pr(A) = \Pr(C_4) + \Pr(C_5) + \Pr(C_6) = 0.217 + 0.109 + 0.271 = 0.597$

$B = \text{HE qualifications}\{\text{HE, Degree}\}$

$\Pr(B) = \Pr(C_5) + \Pr(C_6) = 0.109 + 0.271 = 0.38$

$C = \{\text{No Quals}\}$, $\Pr(C) = \Pr(C_1) = 0.080$

(i) $\Pr(\bar{A}) = \Pr\{\text{No qual, Other, GCSE}\} = \Pr(C_1) + \Pr(C_2) + \Pr(C_3)$
$= 0.403 = 1 - \Pr(A)$

(ii) If $\Pr(D) = 0 \Rightarrow D$ is an empty set.

(iii) If $\Pr(E) = 1 \Rightarrow E = \Omega$

(iv) As $B \subset A \Rightarrow \Pr(B) \leq \Pr(A)$

(v) $\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B) = 0.597 + 0.38 - 0.38 = 0.597$ (Note: $\Pr(A \cap B) = \Pr(B)$

(vi) $\Pr(A \cup C) = \Pr(A) + \Pr(C) - \Pr(A \cap C) = 0.597 + 0.080 - 0 = 0.677$ (Note: $\Pr(A \cap C) = 0$

5. Bivariate probabilities

Consider the two events A and B, which have elementary events $A_1, A_2, \ldots A_h$ and $B_1, B_2, \ldots B_k$, such that:

(i) $A_i \cap A_j = \varnothing$, for $i \neq j$ and $B_i \cap B_j = \varnothing$, for $i \neq j$, and

(ii) $A_1 \cup A_2 \cup \ldots \cup A_h = \Omega$ and $B_1 \cup B_2 \cup \ldots \cup B_k = \Omega$

Each event A_i can occur jointly with any B_j and these joint outcomes can be thought of as the basic outcomes.

Under this scenario we can observe the following sets of possible outcomes as shown in Table 1.

Table 1: Outcomes table

	B_1	B_2	…	B_j	…	B_k
A_1	$A_1 \cap B_1$	$A_1 \cap B_2$		$A_1 \cap B_i$		$A_1 \cap B_k$
A_2	$A_2 \cap B_1$	$A_2 \cap B_2$	…	$A_2 \cap B_j$	…	$A_2 \cap B_k$
…	…					…
A_i	$A_i \cap B_1$	$A_i \cap B_2$		$A_i \cap B_j$		$A_i \cap B_k$
…	…					…
A_h	$A_h \cap B_1$	$A_h \cap B_2$	…	$A_h \cap B_j$	…	$A_h \cap B_k$

These have associated probabilities described in Table 2.

Table 2: Probabilities table

	B_1	B_2	…	B_j	…	B_k	Total
A_1	$P(A_1 \cap B_1)$	$P(A_1 \cap B_2)$		$P(A_1 \cap B_j)$		$P(A_1 \cap B_k)$	$P(A_1)$
A_2	$P(A_2 \cap B_1)$	$P(A_2 \cap B_2)$	…	$P(A_2 \cap B_j)$	…	$P(A_2 \cap B_k)$	$P(A_2)$
…	…						…
A_i	$P(A_i \cap B_1)$	$P(A_i \cap B_2)$		$P(A_i \cap B_j)$		$P(A_i \cap B_k)$	$P(A_i)$
…	…						…
A_h	$P(A_h \cap B_1)$	$P(A_h \cap B_2)$	…	$P(A_h \cap B_j)$	…	$P(A_h \cap B_k)$	$P(A_h)$
Total	$P(B_1)$	$P(B_2)$		$P(B_j)$		$P(B_k)$	1

where $P(A_i \cap B_j)$ are the joint probabilities of the event (A_i and B_j) and $P(A_i)$ is the probability of event A_i occurring irrespective of outcome of B. Similarly $P(B_j)$ is the probability of event B_j occurring irrespective of the outcome of A. $P(A_i)$ and $P(B_j)$ are the *marginal probabilities*. From the union-intersection rule we know that:

$$P(A_1) = P(A_1 \cap B_1) + P(A_1 \cap B_2) \cdots + P(A_1 \cap B_j) \ldots + P(A_1 \cap B_h)$$

6. Conditional Distributions

Consider now the probability that any one of the outcomes associated with experiment A occurs given that the outcome from experiment B was B_k, this is written as $P(A_i \mid B_k),\ i = 1, \ldots h$. We are saying that event B_k has occurred and so there is no longer any uncertainty associated with this, in which case the outcomes tables corresponding to Table 1 is:

Table 3: Outcomes table given that $B=B_k$

	B_k
A_1	$A_1 \cap B_k$
A_2	$A_2 \cap B_k$
…	…
A_i	$A_i \cap B_k$
…	…
A_h	$A_h \cap B_k$

We are interested in the probability associated with each outcome of experiment A, given that $B = B_k$, that is, we want to consider, $P(A_i \mid B = B_k),\ i = 1, \ldots h$.

Correspondingly we only need the penultimate column of Table 2. However, these probabilities (which define all of the possible outcomes of the experiment A, given that B_k occurred) sum to $P(B_k)$, rather than unity, scaling each probability in the table by $P(B_k)$, gives

$$P(A_i \mid B_k) = \frac{P(A_i \cap B_k)}{P(B_k)}.$$

Table 4: Probabilities table given that $B=B_k$

	B_k	$P(A_i \mid B=B_k)$
A_1	$P(A_1 \cap B_k)$	$P(A_1 \cap B_k)/P(B_k)$
A_2	$P(A_2 \cap B_k)$	$P(A_2 \cap B_k)/P(B_k)$
…		
A_i	$P(A_i \cap B_k)$	$P(A_i \cap B_k)/P(B_k)$
…		
A_h	$P(A_h \cap B_k)$	$P(A_h \cap B_k)/P(B_k)$
Total	$P(B_k)$	1

Similarly we can find

$$P(B_k \mid A_i) = \frac{P(A_i \cap B_k)}{P(A_i)}.$$

From the above we therefore have that:

$$P(A_i \cap B_k) = P(A_i \mid B_k).P(B_k)$$

$$P(A_i \cap B_k) = P(B_k \mid A_i).P(A_i)$$

and

$$P(A_i \mid B_k) = \frac{P(B_k \mid A_i).P(A_i)}{P(B_k)}$$

7. Statistical Independence

The events A_i and B_j are said to be statistically, if

$$P(A_i \mid B = B_j) = P(A_i)$$

that is, the events are independent if probability of A_i occurring, conditional on B_j having occurred is simply the marginal probability of A_i (the conditioning has no effect).

Independence implies:

$$P(A_i \cap B_j) = P(A_i).P(B_j)\,.$$

This idea can be extended to 3 or more events where,

$$P(A_i, B_j, D_l) = P(A_i).P(B_j).P(D_l)\,, \text{ and}$$

$$P(A_i \mid B_j, D_l) = \frac{P(A_i, B_j, D_l)}{P(B_j, D_l)}$$

Rearranging and using the rule that $P(B_j, D_l) = P(B_j \mid D_l).P(D_l)$, we have

$$P(A_i, B_j, D_l) = P(A_i \mid B_j, D_l).P(B_j, D_l) = P(A_i \mid B_j, D_l).P(B_j \mid D_l).P(D_l)$$

8. Bayes Theorem

Define an event A and some mutually exclusive and exhaustive basic events, $C_1,\ldots C_k$.

From earlier we had that

$$P(A_1) = P(A_1 \cap C_1) + P(A_1 \cap C_2) \cdots + P(A_1 \cap C_j) \ldots + P(A_1 \cap C_k)$$

$$P(A_1) = P(A_1 \mid C_1).P(C_1) + P(A_1 \mid C_2).P(C_2) + \ldots + P(A_1 \mid C_k).P(C_k)$$

$$P(A_1) = \sum_{i=1}^{k} P(A_1 \mid C_i).P(C_i)$$

Above we wrote that the joint probability of two events can be written as the product of the conditional probability and the marginal probability, that is,

$$P(A_1 \cap C_j) = P(A_1 \mid C_j).P(C_j) = P(C_j \mid A_1).P(A_1)$$

substituting in the formula for A_1 above we get Bayes Theorem:

$$P(C_j \mid A_1) = \frac{P(A_1 \mid C_j).P(C_j)}{P(A_1)} = \frac{P(A_1 \mid C_j).P(C_j)}{\sum_{i=1}^{k} P(A_1 \mid C_i).P(C_i)}$$

Example: Bivariate distribution

		Highest Qual (X_1)						
		NQ	Other	GCCE	A-lev	HE	Degree	
Gender	Male	0.039	0.055	0.088	0.121	0.045	0.130	0.479
(X_2)	Female	0.041	0.048	0.132	0.096	0.064	0.141	0.521
		0.080	0.103	0.221	0.217	0.109	0.271	1.000

Joint probabilities: $\Pr(NQ \cap Male) = 0.039$, $\Pr(Degree \cap Male) = 0.130$

Marginal probabilities:

$$\Pr(\text{M}) = \Pr(\text{M} \cap \text{NQ}) + \Pr(\text{M} \cap \text{Oth}) + \Pr(\text{M} \cap \text{GCSE}) + \Pr(\text{M} \cap \text{A-lev}) + \Pr(\text{M} \cap \text{HE}) + \Pr(\text{M} \cap \text{Deg})$$
$$\Pr(\text{M}) = 0.039 + 0.055 + 0.088 + 0.121 + 0.045 + 0.130 = 0.479$$

$$\Pr(\text{F}) = \Pr(\text{F} \cap \text{NQ}) + \Pr(\text{F} \cap \text{Oth}) + \Pr(\text{F} \cap \text{GCSE}) + \Pr(\text{F} \cap \text{A-lev}) + \Pr(\text{F} \cap \text{HE}) + \Pr(\text{F} \cap \text{Deg})$$
$$\Pr(\text{F}) = 0.041 + 0.048 + 0.132 + 0.096 + 0.064 + 0.141 = 0.521$$

Conditional probabilities:

$$\Pr(\text{Degree}|\text{M}) = \frac{\Pr(\text{M} \cap \text{Degree})}{\Pr(\text{M})} = \frac{0.130}{0.479} = 0.271;$$

$$\Pr(\text{NQ}|\text{M}) = \frac{\Pr(\text{M} \cap \text{NQ})}{\Pr(\text{M})} = \frac{0.039}{0.479} = 0.081;$$

$$\Pr(\text{Degree}|\text{F}) = \frac{\Pr(\text{F} \cap \text{Degree})}{\Pr(\text{F})} = \frac{0.141}{0.521} = 0.271;$$

$$\Pr(\text{NQ}|\text{F}) = \frac{\Pr(\text{F} \cap \text{NQ})}{\Pr(\text{F})} = \frac{0.041}{0.521} = 0.079.$$

As $\Pr(\text{NQ}) \equiv 0.080 \neq \Pr(\text{NQ}|\text{M}) \equiv 0.081$ - not independent

Alternatively consider:

		Highest Qual (X_1)						
		NQ	Other	GCCE	A-lev	HE	Degree	
Wages	<£10	0.064	0.072	0.143	0.115	0.035	0.051	0.480
(X_3)	£10-16	0.012	0.023	0.055	0.070	0.041	0.078	0.279
	≥£16	0.003	0.008	0.023	0.032	0.033	0.141	0.241
		0.080	0.103	0.221	0.217	0.109	0.271	1.000

$$\Pr(<£10|\text{NQ}) = \frac{\Pr(<£10 \cap \text{NQ})}{\Pr(\text{NQ})} = \frac{0.064}{0.080} = 0.800$$

$$\Pr(<£10|\text{A-lev}) = \frac{\Pr(<£10 \cap \text{A-lev})}{\Pr(\text{A-lev})} = \frac{0.115}{0.217} = 0.530$$

$$\Pr(<£10|\text{Deg}) = \frac{\Pr(<£10 \cap \text{Deg})}{\Pr(\text{Deg})} = \frac{0.051}{0.271} = 0.188$$

Example: Trivariate distribution

Males		Highest Qual (X_1)						
		NQ	Other	GCCE	A-lev	HE	Degree	
Wages	<£10	0.028	0.034	0.048	0.052	0.011	0.021	0.194
(X_3)	£10-16	0.008	0.015	0.026	0.045	0.016	0.031	0.142
	≥£16	0.003	0.006	0.014	0.024	0.019	0.077	0.143
		0.039	0.055	0.088	0.121	0.045	0.130	0.479

Females		Highest Qual (X_1)						
		NQ	Other	GCCE	A-lev	HE	Degree	
Wages	<£10	0.036	0.038	0.095	0.063	0.024	0.030	0.286
(X_3)	£10-16	0.004	0.008	0.028	0.024	0.025	0.047	0.137
	≥£16	0.001	0.003	0.009	0.008	0.015	0.064	0.098
		0.041	0.048	0.132	0.096	0.064	0.141	0.521

Joint probabilities: $\Pr(<£10 \cap \text{NQ} \cap \text{M}) = 0.028$ and $\Pr(<£10 \cap \text{Deg} \cap \text{M}) = 0.021$

$$\Pr(<£10 \cap \text{M}) = \Pr(<£10 \cap \text{M} \cap \text{NQ}) + \Pr(<£10 \cap \text{M} \cap \text{Oth}) + \Pr(<£10 \cap \text{M} \cap \text{GCSE}) + \Pr(<£10 \cap \text{M} \cap \text{A-lev}) + \Pr(<£10 \cap \text{M} \cap \text{HE}) + \Pr(<£10 \cap \text{M} \cap \text{Deg}) = 0.194$$

$$\Pr(\text{NQ} \cap \text{M}) = \Pr(<£10 \cap \text{M} \cap \text{NQ}) + \Pr(£10-16 \cap \text{M} \cap \text{NQ}) + \Pr(\geq £16 \cap \text{M} \cap \text{NQ}) = 0.039$$

Marginal probabilities:

$$\Pr(M) = \Pr(M\cap<£10\cap NQ)+\Pr(M\cap<£10\cap Oth)\ldots+\Pr(M\cap<£10\cap Deg)+$$
$$\Pr(M\cap £10\text{-}16\cap NQ)+\Pr(M\cap £10\text{-}16\cap Oth)\ldots+\Pr(M\cap £10\text{-}16\cap Deg)+$$
$$\Pr(M\cap \geq £16\cap NQ)+\Pr(M\cap \geq £16\cap Oth)\ldots+\Pr(M\cap \geq £16\cap Deg)=0.479$$

$$\Pr(NQ) = \Pr(NQ\cap<£10\cap M)+\Pr(NQ\cap £10\text{-}16\cap M)+\ldots+\Pr(NQ\cap \geq £16\cap M)+$$
$$\Pr(NQ\cap<£10\cap F)+\Pr(NQ\cap £10\text{-}16\cap F)+\ldots+\Pr(NQ\cap \geq £16\cap F)=0.080$$

$$\Pr(<£10) = \Pr(<£10\cap NQ\cap M)+\Pr(<£10\cap Oth\cap M)+\ldots+\Pr(<£10\cap Deg\cap M)$$
$$\Pr(<£10\cap NQ\cap F)+\Pr(<£10\cap Oth\cap F)+\ldots+\Pr(<£10\cap Deg\cap F)$$
$$=0.480$$

Conditional probabilities:

$$\Pr(<£10\cap NQ\,|\,M)=\frac{\Pr(<£10\cap NQ\cap M)}{\Pr(M)}=\frac{0.028}{0.479}=0.058$$

$$\Pr(<£10\,|\,NQ\cap M)=\frac{\Pr(<£10\cap NQ\cap M)}{\Pr(M\cap NQ)}=\frac{0.028}{0.039}=0.718$$

$$\Pr(£10-16\,|\,NQ\cap M)=\frac{\Pr(£10-16\cap NQ\cap M)}{\Pr(M\cap NQ)}=\frac{0.008}{0.039}=0.205$$

$$\Pr(\geq £16\,|\,NQ\cap M)=\frac{\Pr(\geq £16\cap NQ\cap M)}{\Pr(M\cap NQ)}=\frac{0.003}{0.039}=0.077$$

$$\Pr(<£10\,|\,Deg\cap M)=\frac{\Pr(<£10\cap Deg\cap M)}{\Pr(M\cap Deg)}=\frac{0.021}{0.130}=0.164$$

$$\Pr(£10-16\,|\,Deg\cap M)=\frac{\Pr(£10-16\cap Deg\cap M)}{\Pr(M\cap Deg)}=\frac{0.031}{0.130}=0.239$$

$$\Pr(\geq £16\,|\,Deg\cap M)=\frac{\Pr(\geq £16\cap Deg\cap M)}{\Pr(M\cap Deg)}=\frac{0.077}{0.130}=0.594$$

9. Combinations and permutations

There are *n* objects to be arranged in order: how many different ways are there of doing this?

${}_nP_n = n(n-1)(n-2)\ldots 1 = n!$ - permutations

There are *n* different objects and you choose *r* of them, how many ways can you order these r objects?

$${}_nP_r = n! = n(n-1)(n-2)\ldots(n-r+1) = \frac{n!}{(n-r)!}$$

There are *n* different objects and you choose *r* of them, how many ways can you choose *r* (without ordering) ${}_nC_r = \frac{n!}{r!(n-r)!} = \frac{{}_nP_r}{r!}$

Probability - Examples

1. Let $A=(2,4,6,8)$, $B=(1,3,5,9)$ and $S=(1,2,3,4,5,6,7,8,9)$

   Evaluate the sets:

   (a) $\overline{A}$, (b) $\overline{B}$, (c) $A\cup B$, (d) $\overline{A\cup B}$, (e) $A\cap B$, (f) $\overline{A}\cup\overline{B}$, (g) $\overline{A}\cap\overline{B}$

Answer

(a) $\overline{A}=(1,3,5,7,9)$

(b) $\overline{B}=(2,4,6,7,8)$

(c) $A\cup B=(1,2,3,4,5,6,8,9)$

(d) $\overline{A\cup B}=(7)$

(e) $A\cap B=(\varnothing)$

(f) $\overline{A}\cup\overline{B}=(1,2,3,4,5,6,7,8)=\overline{A\cap B}$

(g) $\overline{A}\cap\overline{B}=(7)=\overline{A\cup B}$

2. If $P(A)=1/3$, $P(B)=1/2$ and $P(A\cup B)=3/4$.

   Find (a) $P(A\cap B)$, (b) $P(\overline{A\cap B})$, (c) $P(\overline{A\cup B})$, (d) $P(A\cap\overline{B})$ (e) $P(\overline{A}\cap B)$,

   (f) $P(\overline{A}\cap\overline{B})$, (g) $P(\overline{A}\cup\overline{B})$

Answer

(a) $P(A\cup B)=3/4=1/3+1/2-P(A\cap B)\Rightarrow P(A\cap B)=1/12$

(b) $P(\overline{A\cap B})=1-1/12=11/12$

(c) $P(\overline{A\cup B})=1-3/4=1/4$

(d) $P(A)=P(A\cap B)+P(A\cap\overline{B})\Rightarrow P(A\cap\overline{B})=1/3-1/12=1/4$

(e) $P(B)=P(A\cap B)+P(\overline{A}\cap B)\Rightarrow P(\overline{A}\cap B)=1/2-1/12=5/12$

(f) $P(\overline{A})=P(\overline{A}\cap B)+P(\overline{A}\cap\overline{B})\Rightarrow P(\overline{A}\cap\overline{B})=2/3-5/12=1/4=P(\overline{A\cup B})$

(g) $P(\overline{A}\cup\overline{B})=P(\overline{A})+P(\overline{B})-P(\overline{A}\cap\overline{B})=2/3+1/2-1/4=11/12=P(\overline{A\cap B})$

3. A fair octagonal (eight sided) die, with faces marked 1 to 8, is thrown as an experiment, the result being the number on the face of the die. Define the following events: $E_1=(1,2,3,4,5)$, $E_2=(2,4,6,8)$, $E_3=(1,3,5,7)$. Find the following:

   (a) $\Pr(E_1\mid E_2)$, (b) $\Pr(E_1\mid E_3)$, (c) $\Pr(\overline{E}_1\mid E_2)$, (d) $\Pr(E_2\mid E_1)$,

   (e) $\Pr(E_3\mid E_1\cup E_2)$.

Answer

(a) $\Pr(E_1\mid E_2)=\dfrac{\Pr(E_1\cap E_2)}{\Pr(E_2)}=\dfrac{\Pr(2,4)}{\Pr(2,4,6,8)}=\dfrac{0.25}{0.5}=0.5$

(b) $\Pr(E_1\mid E_3)=\dfrac{\Pr(E_1\cap E_3)}{\Pr(E_3)}=\dfrac{\Pr(1,3,5)}{\Pr(1,3,5,7)}=\dfrac{0.375}{0.5}=0.75$

(c) $\Pr(\overline{E}_1\mid E_2)=\dfrac{\Pr(\overline{E}_1\cap E_2)}{\Pr(E_2)}=\dfrac{\Pr(6,8)}{\Pr(2,4,6,8)}=\dfrac{0.25}{0.5}=0.5$

(d) $\Pr(E_2\mid E_1)=\dfrac{\Pr(E_1\cap E_2)}{\Pr(E_1)}=\dfrac{\Pr(2,4)}{\Pr(1,2,3,4,5)}=\dfrac{0.25}{0.625}=0.4$

(e) $\Pr(E_3\mid E_1\cup E_2)=\dfrac{\Pr(E_3\cap(E_1\cup E_2))}{\Pr(E_1\cup E_2)}=\dfrac{\Pr(1,3,5)}{\Pr(1,2,3,4,5,6,8)}=\dfrac{0.375}{0.875}=\dfrac{3}{7}$

4. A town has three buses A, B and C. In the "rush hour", A has twice as many buses on its route as both B and C. Over a period of time it has been found that, along a certain stretch of road, where the three buses converge, the buses on these routes run more than five minutes late 0.5, 0.2 and 0.1 of the time, respectively. If an inspector (standing near this stretch of road) finds that the first bus is more than five minutes late, find the probability that it is route B bus.

Answer

$\Pr(A)=0.5$, $\Pr(B)=\Pr(C)=0.25$

In addition, we know

$\Pr(L\mid A)=0.5$, $\Pr(L\mid B)=0.2$, $\Pr(L\mid C)=0.1$

we want to know:

$$\Pr(B \mid L) = \frac{\Pr(L \mid B).\Pr(B)}{\Pr(L)} = \frac{0.2(0.25)}{0.325} = \frac{0.05}{0.325} = 0.154$$

and from the union-intersection rule:

$$\Pr(L) = \Pr(L \mid A).\Pr(A) + \Pr(L \mid B).\Pr(B) + \Pr(L \mid C).\Pr(C)$$

$$\Pr(L) = 0.5(0.5) + 0.2(0.25) + 0.1(0.25) = 0.325$$

5. A child uses a home-made metal detector to look for valuable metallic objects on a beach. There is fault in the machine which causes it to signal the presence of only 95% of metallic objects over which it passes and to signal the presence of 6% of non-metallic objects. Of the objects over which the machine passes, 20% are metallic.

(a) Find the probability that a given object is metallic and the machine gives a signal.

(b) Find the probability of a signal being received by the child for any given object.

(c) Find the probability that the child has found a metallic object when they receive a signal.

(d) Given that 10% of metallic objects found on the beach are valuable, find the proportion of objects, discovered by a signal from the detector, that are valuable.

Answer

$\Pr(S \mid M) = 0.95$, $\Pr(S \mid NM) = 0.06$, $\Pr(M) = 0.2$

(a) $\Pr(M \cap S) = \Pr(S \mid M).\Pr(M) = 0.95(0.2) = 0.19$.

(b) $\Pr(S) = \Pr(S \cap M) + \Pr(S \cap NM) = 0.19 + 0.06(0.8) = 0.238$

(c) $\Pr(M \mid S) = \frac{\Pr(M \cap S)}{\Pr(S)} = \frac{0.19}{0.238} = 0.798$

(d) $P(V \mid M \cap S) = 0.1 \Rightarrow P(V \cap M \cap S) = 0.1(0.19) = P(V \cap S)$

as $P(V \cap S \cap \bar{M}) = 0$, therefore $\Pr(V \mid S) = \frac{\Pr(V \cap S)}{P(S)} = \frac{0.1(0.19)}{0.238} = 0.08$

6. In "Pop Band: The Rivals" a group has to consist of 5 singers. If there a total of 10 suitable singers, but due to telephone rigging the band must include a particular pair of singers, in many ways can the "band" be made up?

Answer

As two band members are already selected, we only have to choose 3 from 8 and as the order does not matter, we have ${}^{8}C^{3} = 56$

7. A passenger compartment on a train has six seats, three facing forwards and three facing backwards. Three men and two women enter the compartment and seat themselves randomly.

(a) In how many ways can they be seated?

(b) In how many ways will the women be seated opposite each other?

(c) In how many ways can two men be seated opposite each other?

Answer

(a) In how many ways can we arrange 5 people in 6 seats - ${}_{6}P_{6} = 720$.

(b) If the two women sit opposite one another (next to the window) then how many ways can the three men occupy the remaining 4 seats - ${}_{4}P_{4} = 24$. In total the women can sit opposite each other in 3 ways (window, aisle or middle seats and can swap places) – therefore we have 6*24=144.

(c) Clearly if there we only 2 men then the answer would be the same as (b), but there are three men and these can sit opposite each other as man 1 and man 2, man1 and man 3, or man 2 and man 3 (and can swap around). Therefore we have 432.

STATISTICAL TECHNIQUES B

Univariate and Bivariate Distributions

1. Introduction to Univariate Distributions

For a random experiment, with a sample space, Ω, a function *X*, which assigns to each element of *C*, a real number *X*(*C*)=*x*, is called a random variable. We must distinguish between the random variable, *X*, and the possible outcomes, $x \in \Omega$.

Example 1:

Consider rolling a dice then $\Omega = \{1,2,3,4,5,6\}$, define a random variable, which considers only odd or even numbers and $X(C) = \begin{cases} 1 & \text{if even number} \\ 0 & \text{if odd number} \end{cases}$, then

x	0	1
$P(X=x)$	½	½

Example 2:

Toss two coins $\Omega = \{HH, HT, TH, TT\}$. Let X(C)=Number of tails

$X(HH) = 0, X(HT) = 1, X(TH) = 1, X(TT) = 2$ then $\begin{cases} \Omega = \{HH, HT, TH, TT\} \\ A = \{0,1,2\} \end{cases}$ and

P(*X*=1)=P(HT,TH)=1/2

x	0	1	2
$P(X=x)$	¼	½	¼

Example 3:

Probability of having an MMR vaccine is 0.8. In a random sample of 4 people, how many would have had the MMR vaccine:

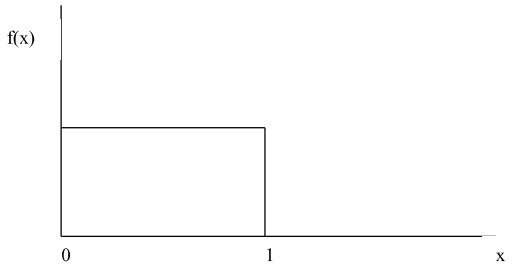
x	NNNN	YNNN	NYNN	NNYN	NNNY	YYNN	YNYN	YNNY
$P(X=x)$	0.2^4	$0.8(0.2)^3$	$0.8(0.2)^3$	$0.8(0.2)^3$	$0.8(0.2)^3$	$0.8^2 0.2^2$	$0.8^2 0.2^2$	$0.8^2 0.2^2$

x	NYYN	NYNY	NNYY	NYYY	YNYY	YYNY	YYYN	YYYY
$P(X=x)$	$0.8^2 0.2^2$	$0.8^2 0.2^2$	$0.8^2 0.2^2$	$0.8^3 0.2$	$0.8^3 0.2$	$0.8^3 0.2$	$0.8^3 0.2$	0.8^4

x	0	1	2	3	4
$P(X=x)$	0.0016	0.0256	0.1536	0.4096	0.4096

Example 4:

$f(x) = 1 \quad 0 \le x \le 1 \quad \Omega = \{x; 0 \le x \le 1\}$



2. Discrete Univariate Distributions

Suppose X is a scalar random variable with a finite number of values - this is a discrete set of points. Let $p_X(x)$ be a function such that, (i) $p_X(x) \geq 0$, and (ii) $\sum_x p_X(x) = 1$, then X is a discrete random variable with probability density (mass) function, $p_X(x)$ and $P(a \leq X \leq b) = \sum_{x=a}^{b} f(x)$, so using example 3

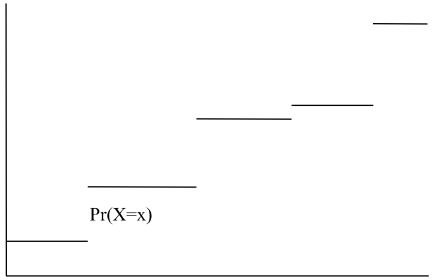
$$P(2 \leq X \leq 3) = \sum_{x=2}^{3} f(x) = 0.5632$$

The cumulative distribution (mass) function of X is such that for each

$F_X(x_0) = P(X \leq x_0) = \sum_{x \leq x_0} p_X(x)$, using example 3,

$$F_X(2) = P(X \leq 2) = \sum_{x \leq 2} p_X(x) = 0.1808.$$

(a) cdf of a discrete random variable



(iii) $\lim_{x \to -\infty} F(x) = 0$

(iv) $\lim_{x \to \infty} F(x) = 1$

(v) $\Pr(a < X \leq b) = F(b) - F(a)$

2.1 Measures of central tendency and dispersion

2.1.1 Median

For a discrete random variable X, with cdf F, there are 2 possibilities:

(i) There is NO x, such that $F(x)$=1/2

(ii) There is a set of x, such that $F(x)$=1/2, $\{x, a \leq x < b\}$. Consider a $y \in [a,b)$, where $P(X < y) \leq 1/2$, $P(X \leq y) = 1/2$. Hence all $y \in [a,b)$ are median values, usually call $\frac{(a+b)}{2}$ the median of the random variable.

2.1.2 Mode

That value of x such that $p(x)$ is maximised.

2.1.3 Expectations and variances

X takes on a finite number of outcomes $x = x_1, x_2, \ldots, x_n$ and each has an associated probability:

X	x_1	x_2	…	x_n
$P(X=x)$	p_1	p_2	…	p_n

$$E(X) = \sum_x p_X(x)x = \mu_X$$

$$V(X) = E\left[(X - \mu_X)^2\right] = \sum_x p_X(x)(x - \mu_X)^2$$

$$= \sum_x p_X(x)x^2 - 2\mu_X \underbrace{\sum_x p_X(x)x}_{\mu_X} + \mu_X^2 \underbrace{\sum_x p_X(x)}_{1} = \sum_x p_X(x)x^2 - \mu_X^2 = E(X^2) - \mu_X^2$$

$$V(X) = E\left[(X - \mu_X)^2\right] = E(X^2) - E(X)^2$$

in general, $E\left[g(X)\right] = \sum_x p_X(x)g(x)$

for example,

$$E\left[X^2\right] = \sum_x p_X(x)x^2$$

Example

	1	2	3	4	5	6
Highest Qual	No Qual	Other	GCSE	A-level	HE	Degree
Pr(.)	0.080	0.103	0.221	0.217	0.109	0.271

$$E(X) = 0.080(1) + 0.103(2) + 0.221(3) + 0.217(4) + 0.109(5) + 0.271(6) = 3.988$$

$$E(X^2) = 0.080(1^2) + 0.103(2^2) + 0.221(3^2) + 0.217(4^2) + 0.109(5^2) + 0.271(6^2) = 18.434$$

$$V(X) = E(X^2) - \left[E(X)\right]^2 = 18.434 - 3.988^2 = 2.530$$

3. Introduction to Discrete Bivariate Distributions

Let $p_{X_1,X_2}(x_1,x_2) = P(X_1 = x_1, X_2 = x_2)$ be the joint probability density (mass) function for the discrete random variables X_1 and X_2. The joint probability density function defines a probability for all values of (x_1,x_2). Suppose that the sample space of X_1 is $\Omega_1 = \{A_1, A_2 \ldots A_h\}$ and the sample space of X_2 is $\Omega_2 = \{B_1, B_2 \ldots B_k\}$. Then we can define all possible outcomes and the joint probability density function using the Table 1 below (reproduced from Handout 2 - Probability):

Table 1: Joint Probability Table

	B_1	B_2	…	B_j	…	B_k	Total
A_1	P($A_1 \cap B_1$)	P($A_1 \cap B_2$)		P($A_1 \cap B_j$)		P($A_1 \cap B_k$)	P(A_1)
A_2	P($A_2 \cap B_1$)	P($A_2 \cap B_2$)	…	P($A_2 \cap B_j$)	…	P($A_2 \cap B_k$)	P(A_2)
…	…						…
A_i	P($A_i \cap B_1$)	P($A_i \cap B_2$)		P($A_i \cap B_j$)		P($A_i \cap B_k$)	P(A_i)
…	…						…
A_h	P($A_h \cap B_1$)	P($A_h \cap B_2$)	…	P($A_h \cap B_j$)	…	P($A_h \cap B_k$)	P(A_h)
Total	P(B_1)	P(B_2)		P(B_j)		P(B_k)	1

The table defining a probability for each pair of events, A_i, B_j for i=1,…h and j=1, …k, such that $P(A_i, B_j) \geq 0$ and $\sum_{x_1}\sum_{x_2} p_{X_1,X_2}(x_1,x_2) = 1$ then it is a valid probability density function.

In general,

$$p_{X_1,X_2\ldots X_k}(x_1,x_2,\ldots,x_k) = P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$$

and to be a valid pdf we require:

(i) $p_{X_1,X_2\ldots X_k}(x_1,x_2,\ldots,x_k) \geq 0$

(ii) $\sum_{x_1}\sum_{x_2}\cdots\sum_{x_k} p_{X_1,X_2\ldots X_k}(x_1,x_2,\ldots,x_k) = 1$

3.1 Marginal Distributions

For the bivariate case above, consider the event, $X_1 = A_i$, where $A_1 \leq A_i$. This event occurs when $X_1 = A_i$ and X_2 takes any possible value. This probability is

$$P(X_1 = A_i, B_1 \leq X_2 \leq B_k) = \sum_{x_2} p_{X_1,X_2}(A_i, x_2)$$

$$P(X_1 = A_i, B_1 \leq X_2 \leq B_k) = P(A_i \cap B_1) + P(A_i \cap B_2) + \ldots + P(A_i \cap B_k)$$

Now $\sum_{x_2} p(x_1,x_2)$ removes the variable X_2 out of the probability formula, leaving the marginal probability, denoted as $p_1(x_1)$, and this is a probability density function of X_1, x_2 having been summed out, and is known as the MARGINAL PROBABILITY DENSITY (MASS) FUNCTION of X_1.

Similarly we have the marginal probability of X_2 calculated as:

$$P(X_2 = B_j, A_1 \leq X_1 \leq A_h) = \sum_{x_1} p_{X_1,X_2}(x_1, B_j)$$

$$P(X_2 = B_j, A_1 \leq X_1 \leq A_h) = P(A_1 \cap B_j) + P(A_2 \cap B_j) + \ldots + P(A_h \cap B_j).$$

From these marginal distributions we can calculate the moments of the random variables X_1 and X_2, that is, $E(X_1)$, $E(X_2)$, $V(X_1)$, $V(X_2)$ and $\text{cov}(X_1,X_2)$ as we did before. For the discrete random variable X_1, with marginal probability density function $p_1(x_1)$:

$$E(X_1) = \sum_{x_1} x_1 p_1(x_1)$$

$$V(X_1) = \sum_{x_1} x_1^2 p_1(x_1) - E(X_1)^2$$

similarly for X_2,

$$E(X_2) = \sum_{x_2} x_2 p_2(x_2)$$

$$V(X_2) = \sum_{x_2} x_2^2 p_2(x_2) - E(X_2)^2$$

and

$$\text{cov}(X_1,X_2) = E(X_1 - E(X_1))(X_2 - E(X_2)) = E(X_1X_2) - E(X_1)E(X_2)$$

$$\text{cov}(X_1,X_2) = \sum_{x_2}\sum_{x_1} x_1x_2 p(x_1,x_2) - E(X_1)E(X_2)$$

Rules on expectations and variance for bivariate and higher order distributions are in Appendix 2.

3.2 Conditional Distributions

The conditional probability density function for random variables, X_1 and X_2 with a joint probability density (mass) function $p(x_1,x_2)$ and marginals $p_1(x_1)$ and $p_2(x_2)$, is written as:

$$p(x_1 \mid x_2) = \frac{p(x_1,x_2)}{p_2(x_2)}.$$

This is a valid p.d.f. as

$$\sum_{x_1} p(x_1 \mid x_2) = \sum_{x_1} \frac{p(x_1, x_2)}{p_2(x_2)} = \frac{1}{p_2(x_2)} \sum_{x_1} p(x_1, x_2) = \frac{1}{p_2(x_2)} \sum_{x_1} p(x_1, x_2) = \frac{p_2(x_2)}{p_2(x_2)} = 1$$

and $p(x_1 \mid x_2) = \dfrac{p(x_1, x_2)}{p_2(x_2)} > 0$

Rearranging the above expression we also have that

$p(x_1, x_2) = p(x_1 \mid x_2).p_2(x_2)$.

This idea can be extended to more than two events, in which case we have

$$p(x_1 \mid x_2, x_3) = \frac{p(x_1, x_2, x_3)}{p_{2,3}(x_2, x_3)}$$

Rearranging and using the rule that $p(x_2, x_3) = p(x_2 \mid x_3).p_3(x_3)$, we have

$$p(x_1, x_2, x_3) = p(x_1 \mid x_2, x_3).p_{2,3}(x_2, x_3) = p(x_1 \mid x_2, x_3).p(x_2 \mid x_3).p_3(x_3)$$

and so in general

$$p(x_1, x_2, x_3, \ldots x_n) = p(x_1 \mid x_2, x_3, \ldots x_n).p(x_2 \mid x_3, \ldots x_n).p_3(x_3 \mid \ldots x_n) \ldots p_n(x_n)$$

Example

			Highest Qual (X_1)						
			1	2	3	4	5	6	
			NQ	Other	GCCE	A-lev	HE	Deg	
Wages	1	<£10	0.064	0.072	0.143	0.115	0.035	0.051	0.480
(X_3)	2	£10-16	0.012	0.023	0.055	0.070	0.041	0.078	0.279
	3	≥£16	0.003	0.008	0.023	0.032	0.033	0.141	0.241
			0.080	0.103	0.221	0.217	0.109	0.271	1.000

$E(X_1) = 3.988; V(X_1) = 2.530$

$E(X_3) = 0.480(1) + 0.279(2) + 0.241(3) = 1.761;$

$E(X_3^2) = 0.480(1^2) + 0.279(2^2) + 0.241(3^3) = 3.765 \Rightarrow V(X_3) = 0.664$

$E(X_1.X_3) = 0.064(1)(1) + 0.072(1)(2) + 0.141(1)(3) + \ldots 0.051(1)(6) +$
$\ldots 0.141(3)(6) = 7.611$

$\text{cov}(X_1, X_3) = E(X_1X_3) = E(X_1).E(X_3) = 7.611 - 3.988(1.761) = 0.588$

Below we report the probability of being in each of the alternative wage groups conditional on a given level of education. So, for example, $\Pr(< £10 \mid NQ) = 0.810$ and $\Pr(<£10 \mid \text{Deg}) = 0.189$ (see Probability handout)

			Highest Qual (X_1)					
			1	2	3	4	5	6
			NQ	Other	GCCE	A-lev	HE	Deg
Wages	1	<£10	0.810	0.699	0.647	0.530	0.321	0.189
(X_3)	2	£10-16	0.152	0.223	0.249	0.323	0.376	0.289
	3	≥£16	0.038	0.078	0.104	0.147	0.303	0.522
			1.000	1.000	1.000	1.000	1.000	1.000

$E(X_3 \mid X_1 = 1) = 0.810(1) + 0.152(2) + 0.038(3) = 1.228;$

$E(X_3 \mid X_1 = 2) = 0.699(1) + 0.223(2) + 0.078(3) = 1.379;$

$E(X_3 \mid X_1 = 3) = 0.647(1) + 0.249(2) + 0.104(3) = 1.457;$

$E(X_3 \mid X_1 = 4) = 0.530(1) + 0.323(2) + 0.147(3) = 1.618$

$E(X_3 \mid X_1 = 5) = 0.321(1) + 0.376(2) + 0.303(3) = 1.982;$

$E(X_3 \mid X_1 = 6) = 0.189(1) + 0.289(2) + 0.522(3) = 2.333;$

4. Continuous Univariate Distributions

Suppose X is a scalar random variable along the real line and (i) $f_X(x) \geq 0$ and (ii) $\int_x f_X(x)dx = 1$, then X is a continuous random variable with probability density function, $f_X(x)$ and $P(a \leq X \leq b) = \int_a^b f_X(x)dx$. Moreover $P(X = a) = \int_a^a f_X(x)dx = 0$, therefore $P(a < X < b) = P(a \leq X \leq b)$.

The cumulative distribution function of X is such that for each

$$F_X(x_0) = P(X \leq x_0) = \int_{x \leq x_0} f_X(x)dx,$$

The cdf of a random variable X is:

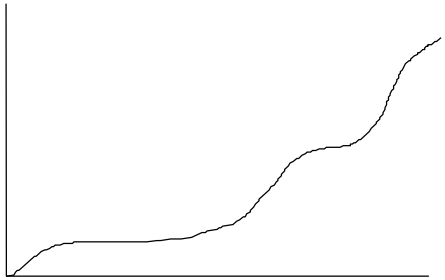
$$F(x) = P(X \leq x_0) = \int_{-\infty}^{x_0} f(y)dy \quad \text{or} \quad \sum_{-\infty}^{x_0} f(y)$$

From cdf can determine all the relevant probability statements.

(i) F is non-decreasing, that is, if $y \leq x$ then $F(y) \leq F(x)$

(ii) cdfs are everywhere continuous from the right, that is, $\lim_{h \to 0} F(x+h) = F(x)$

(b) cdf of a continuous random variable



4.1 Measures of central tendency and dispersion

4.1.1 Median

For a continuous random variable X, with cdf F, the median is the point x, such that $F(x)=1/2$.

4.1.2 Mode

That value of x such that $f(x)$ is maximised.

4.1.3 Expectations and variances

For continuous random variables

$$E(X) = \int_x x f_X(x)dx = \mu_X$$

$$V(X) = E(X^2) - E(X)^2 = \int_x x^2 f_X(x)dx - \mu_x^2$$

5. Introduction to Continuous Bivariate Distributions

Let $f_{X_1,X_2}(x_1,x_2)$ be the joint probability density function for the continuous random variables X_1 and X_2. This function is a valid probability density function, if,

(i) $f_{X_1,X_2}(x_1,x_2) \geq 0$

(ii) $\int_{x_1}\int_{x_2} f_{X_1,X_2}(x_1,x_2)\partial x_2 \partial x_1 = 1$

5.1 Marginal Distributions

For the continuous random variables X_1 and X_2, with $f_{X_1,X_2}(x_1,x_2)$ as the joint probability density function, the marginal probability of the marginal probability of X_1 is:

$$P(X_1, -\infty < X_2 < \infty) = \int_{-\infty}^{\infty} f_{X_1X_2}(x_1,x_2)\,dx_2 = f_1(x_1)$$

and this removes the variable X_2 out of the formula, leaving the marginal probability density function of X_1, $f_1(x_1)$, a function of x_1 alone. Similarly to above the marginal probability of X_2 is:

$$P(X_2, -\infty < X_2 < \infty) = \int_{-\infty}^{\infty} f_{X_1X_2}(x_1,x_2)\,dx_1 = f_2(x_2).$$

$$E(X_1) = \int_{x_1} x_1 f_{X_1}(x_1)dx_1 = \mu_{X_1}$$

$$V(X_1) = E(X_1^2) - E(X_1)^2 = \int_{x_1} x_1^2 f_{X_1}(x_1)dx_1 - \mu_{x_1}^2$$

similarly for X_2,

and

$$\text{cov}(X_1,X_2) = E(X_1 - E(X_1))(X_2 - E(X_2)) = E(X_1X_2) - E(X_1)E(X_2)$$

$$\text{cov}(X_1,X_2) = \int_{x_1}\int_{x_2} x_1x_x f(x_1,x_2)\partial x_2\partial x_1 - E(X_1)E(X_2)$$

Univariate and Bivariate Distributions – Examples

1. For the following discrete distribution:

x	1	3	5	8	9
Pr(X=x)	0.1	0.4	0.3	0.15	0.05

Find (a) E(X), (b) E(X-1), (c) E[3(X-1)], (d) V(X), (e) V(X-1), (f) V[3(X-1)]

Answer

(a) $E(X) = 1(0.1) + 3(0.4) + 5(0.3) + 8(0.15) + 9(0.05) = 4.45$

(b)

x-1	0	2	4	7	8
Pr[(X-1)=x-1]	0.1	0.4	0.3	0.15	0.05

$$E(X-1) = 0(0.1) + 2(0.4) + 4(0.3) + 7(0.15) + 8(0.05) = 3.45 = E(X) - 1$$

(c)

3(x-1)	0	6	12	21	24
Pr[3(X-1)=3(x-1)]	0.1	0.4	0.3	0.15	0.05

$$E[3(X-1)] = 0(0.1) + 6(0.4) + 12(0.3) + 21(0.15) + 24(0.05) = 10.35 = 3E(X) - 3$$

(d) $V(X) = E(X^2) - E(X)^2$

$$= 1^2(0.1) + 3^2(0.4) + 5^2(0.3) + 8^2(0.15) + 9^2(0.05) - 4.45^2 = 5.05$$

(e) $V(X-1) = 0^2(0.1) + 2^2(0.4) + 4^2(0.3) + 7^2(0.15) + 8^2(0.05) - 3.45^2 = 5.05$

$$V(X-1) = E[(X-1)^2] - E[(X-1)]^2 = E[X^2 - 2X + 1] - [E(X) - 1]^2$$

$$V(X-1) = E(X^2) - 2E(X) + 1 - E(X)^2 + 2E(X) - 1 = E(X^2) - E(X)^2 = V(X)$$

(f) $V[3(X-1)] = 0^2(0.1) + 6^2(0.4) + 12^2(0.3) + 21^2(0.15) + 24^2(0.05) - 10.35^2 = 45.45$

$$V[3(X-1)] = E[\{3(X-1)\}^2] - E[3(X-1)]^2 = E[9X^2 - 18X + 9] - [3E(X) - 3]^2$$

$$V[3(X-1)] = 9E(X^2) - 18E(X) + 9 - 9E(X)^2 + 18E(X) - 9 = 9[E(X^2) - E(X)^2] = 9V(X)$$

2. For the discrete random variable defined by the p.d.f.

x	-3	2	4
Pr(X=x)	0.4	0.3	0.3

Find $E(Y)$ if $Y = 2(X-1)^2 + 3(X-1) - 5$

Answer

$$E(X) = -3(0.4) + 2(0.3) + 4(0.3) = 0.6$$

y	15	0	22
Pr(Y=y)	0.4	0.3	0.3

$E(Y) = 15(0.4) + 0(0.3) + 22(0.3) = 12.6$

$E(Y) = 2E(X^2) - 4E(X) + 2 + 3E(X) - 3 - 5 = 2E(X^2) - E(X) - 6$

$E(Y) = 2E(X^2) - E(X) - 6 = 19.2 - 0.6 - 6 = 12.6$

3. Consider the following bivariate distribution for X_1 and X_2.

		x_2			
		1	2	3	4
x_1	1	0.10	0.05	0.00	0.10
	2	0.10	0.00	0.20	0.00
	3	0.05	0.10	0.25	0.05

(a) Write out the marginal distributions for X_1 and X_2.

(b) Calculate E(X_1) and E(X_2), V(X_1) and V(X_2)

(c) Calculate cov(X_1,X_2).

(d) Write out the distribution of (X_1|X_2=2) and calculate E(X_1|X_2=2).

Answer

(a)

x_1	1	2	3
$P(X_1 = x_1)$	0.25	0.30	0.45

x_2	1	2	3	4
$P(X_2 = x_2)$	0.25	0.15	0.45	0.15

(b) $E(X_1) = 1(0.25) + 2(0.3) + 3(0.45) = 2.2$

$E(X_2) = 1(0.25) + 2(0.15) + 3(0.45) + 4(0.15) = 2.5$

$V(X_1) = E(X_1^2) - E(X_1)^2 = 1^2(0.25) + 2^2(0.3) + 3^2(0.45) - (2.2)^2 = 0.66$

$\mathrm{V(X_2)} = \mathrm{E(X_2^2)} - \mathrm{E(X_2)}^2 = 1^2(0.25) + 2^2(0.15) + 3^2(0.45) + 4^2(0.15) - (2.5)^2 = 0.66$

(c) $\mathrm{cov}(X_1, X_2) = E(X_1X_2) - E(X_2)E(X_2)$

$\mathrm{cov}(x_1, x_2) = 1(1)(0.1) + 1(2)(0.05) + 1(3)(0.0) + 1(4)(0.1)$
$+2(1)(0.1) + 2(2)(0.0) + 2(3)(0.2) + 2(4)(0.0) + 3(1)(0.05) + 3(2)(0.1)$
$+3(3)(0.25) + 3(4)(0.05) - (2.2)(2.5) = 5.6 - 5.5 = 0.1$

(d)

$x_1 \mid X_2 = 2$	1	2	3
$P(X_1 = x_1 \mid X_2 = 2)$	0.05/0.15	0.00	0.1/0.15

$x_1 \mid X_2 = 2$	1	2	3
$P(X_1 = x_1 \mid X_2 = 2)$	0.3333	0.00	0.6666

(e) $E(X_1 \mid X_2 = 2) = 1(0.3333) + 2(0.0) + 3(0.66666) = 2.3333$

4. The random variables X_1 is distributed with a mean of 50 and variance of 10, while is independently X_2 is distributed with mean of 50 and variance 5. Find the mean and variance of (a) X_1+X_2, (b) X_1-2X_2, (c) X_2-0.4X_1.

Answer

(a) X_1 and X_2 are assumed to be independent.

$E(X_1 + X_2) = E(X_1) + E(X_2) = 50 + 50 = 100$

$V(X_1 + X_2) = V(X_1) + V(X_2) + 2\,\mathrm{cov}(X_1, X_2) = 10 + 5 = 15$

(b) $E(X_1 - 2X_2) = E(X_1) - 2E(X_2) = 50 - 2(50) = -50$

$V(X_1 - 2X_2) = V(X_1) + V(-2X_2) + 2\,\mathrm{cov}(X_1, -2X_2) = V(X_1) + 4V(X_2) - 4\,\mathrm{cov}(X_1, X_2)$

$V(X_1 - 2X_2) = 10 + 4(5) = 30$

(c) $E(X_2 - 0.4X_1) = E(X_2) - 0.4E(X_2) = 50 - 0.4(50) = 30$

$V(X_2 - 0.4X_1) = V(X_2) + V(-0.4X_1) + 2\,\mathrm{cov}(X_2, -0.4X_1) = V(X_2) + 0.16V(X_1) - 0.8\,\mathrm{cov}(X_1, X_2)$
$V(X_2 - 0.4X_1) = 5 + 0.16(10) = 6.6$

5. The random variables X_1, X_2 and X_3 are a random sample from a population with a mean of μ and variance σ^2. Find the mean and variance of (a) X_1+X_2+X_3, (b) X_1-X_2+X_3, (c) (X_1+X_2+X_3)/3

Answer

(a) Assuming independence

$E(X_1 + X_2 + X_3) = E(X_1) + E(X_2) + E(X_3) = \mu + \mu + \mu = 3\mu$

$V(X_1 + X_2 + X_3) = V(X_1) + V(X_2) + V(X_3) + 2\,\mathrm{cov}(X_1, X_2) + 2\,\mathrm{cov}(X_1, X_3) + 2\,\mathrm{cov}(X_2, X_3)$
$V(X_1 + X_2 + X_3) = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2$

(b) $E(X_1 - X_2 + X_3) = E(X_1) - E(X_2) + E(X_3) = \mu - \mu + \mu = \mu$

$$V(X_1 - X_2 + X_3) = V(X_1) + V(X_2) + V(X_3) - 2\operatorname{cov}(X_1, X_2) + 2\operatorname{cov}(X_1, X_3) - 2\operatorname{cov}(X_2, X_3)$$
$$V(X_1 + X_2 + X_3) = \sigma^2 + \sigma^2 + \sigma^2 = 3\sigma^2$$

(c) $E[(X_1 + X_2 + X_3)/3] = 1/3[E(X_1) + E(X_2) + E(X_3)] = 1/3[\mu + \mu + \mu] = \mu$

$$V[(X_1 + X_2 + X_3)/3] = 1/9[V(X_1) + V(X_2) + V(X_3)]$$

$$V[(X_1 + X_2 + X_3)/3] = 1/9[\sigma^2 + \sigma^2 + \sigma^2] = \sigma^2/3$$

6. A random variable X has the p.d.f, f(x), where:

$$f(x) = \begin{cases} 4(x-1)(2-x)(3-x) & \text{if } 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

Find the mean of X. What is the prob that X taken at random exceeds 1.8?

Answer

$$E(X) = \int_1^2 xf(x)\,dx = \int_1^2 x.4(x^3 - 6x^2 + 11x - 6)\,dx$$

$$E(X) = 4\int_1^2 x^4 - 6x^3 + 11x^2 - 6x\,dx == 4\left[\frac{x^5}{5} - \frac{6x^4}{4} + \frac{11x^3}{3} - \frac{6x^2}{2}\right]_1^2$$

$$E(X) = 4\left\{\left[\frac{32}{5} - \frac{96}{4} + \frac{88}{3} - \frac{24}{2}\right] - \left[\frac{1}{5} - \frac{6}{4} + \frac{11}{3} - \frac{6}{2}\right]\right\} = 1.47$$

$$\Pr(X > 1.8) = \int_{1.8}^{2} f(x)dx = 4\int_{1.8}^{2} (x^3 - 6x^2 + 11x - 6)dx$$

$$\Pr(X > 1.8) = 4\left[\frac{x^4}{4} - \frac{6x^3}{3} + \frac{11x^2}{2} - 6x\right]_{1.8}^{2} = -8 - (-8.0784) = 0.0784$$

7. The continuous random variable X has p.d.f., $f(x)$, where

$$f(x) = \begin{cases} 0 & x < 2 \\ k(3-x) & 2 \le x \le 3 \\ 0 & x > 3 \end{cases}$$

Calculate (a) the constant, k, (b) the median of X.

Answer

(a) $\int_2^3 k(3-x)\,dx = 1 \Rightarrow k[3x - x^2/2]_2^3 = 1 \Rightarrow k\{4.5 - 4\} = 1 \Rightarrow k = 2$

(b) $\int_2^d 2(3-x)\,dx = 0.5 \Rightarrow 2[3x - x^2/2]_2^d = 0.5 \Rightarrow 2\{3d - d^2/2 - 4\} = 0.5$

$$d^2 - 6d + 8.5 = 0 \Rightarrow d = 2.29$$

More rules on expectations variances

Define the probability density function for the random variables X, Y as $p(X,Y)$, such that, $\sum_x \sum_y p(x,y) = 1$. Define the marginal density of X as $p(x)$ and the marginal probability density for Y as $p(y)$, such that $p(X) = \sum_y p(x,y)$ and $p(Y) = \sum_x p(x,y)$.

Then define $E(X) = \sum_y xp(y)$, $E(Y) = \sum_y yp(y)$, $V(X) = \sum_x (x - E(X))^2 p(x)$, $V(Y) = \sum_y (y - E(Y))^2 p(y)$ and $\operatorname{cov}(X,Y) = \sum_x \sum_y (x - E(X))(y - E(Y))p(x,y)$.

Then we can show that

1. $E(X+Y) = \sum_x \sum_y (x+y)p(x,y) = \sum_x \sum_y xp(x,y) + \sum_x \sum_y yp(x,y)$

$$\sum_x x \underbrace{\sum_y p(x,y)}_{p(x)} + \sum_y y \underbrace{\sum_x p(x,y)}_{p(y)} = \sum_x xp(x) + \sum_y yp(y) = E(X) + E(Y)$$

2. $\operatorname{cov}(a+X,Y) = \sum_x \sum_y ((a+x) - E(a+X))(y - E(Y))p(x,y)$

$$= \sum_x \sum_y (a + x - a - E(X))(y - E(Y))p(x,y) = \operatorname{cov}(X,Y).$$

3. $\operatorname{cov}(aX,Y) = \sum_x \sum_y (ax - E(aX))(y - E(Y))p(x,y)$

$$\operatorname{cov}(aX,Y) = \sum_x \sum_y (ax - aE(X))(y - E(Y))p(x,y)$$

$$\operatorname{cov}(aX,Y) = a\sum_x \sum_y (x - E(X))(y - E(Y))p(x,y) = a\operatorname{cov}(X,Y)$$

4. $V(X+Y) = \sum_x \sum_y \left[(x+y) - E(X+Y)\right]^2 p(x,y) = \sum_x \sum_y \left[(x - E(X)) + (y - E(Y))\right]^2 p(x,y)$

$$\sum_x \sum_y \left[(x - E(X)) + (y - E(Y))\right]^2 p(x,y)$$

$$\sum_x \sum_y (x - E(X))^2 p(x,y) + \sum_x \sum_y (y - E(Y))^2 p(x,y) + 2\sum_x \sum_y (x - E(X)(y - E(Y))p(x,y)$$

$$= \sum_x (x - E(X))^2 \underbrace{\sum_y p(x,y)}_{p(x)} + \sum_x (y - E(Y))^2 \underbrace{\sum_y p(x,y)}_{p(y)} + 2\operatorname{cov}(X,Y)$$

$$= V(X) + V(Y) + 2\operatorname{cov}(X,Y)$$

5. $E(X - Y) = E(X) - E(Y)$
6. $V(X - Y) = V(X) + V(Y) - 2\operatorname{cov}(X,Y)$

7. $E(aX-bY)=aE(X)-bE(Y)$
8. $V(aX-bY)=a^2V(X)+b^2V(Y)-2ab\,\mathrm{cov}(X,Y)$
9. $E(X+Y+Z)=E(X)+E(Y)+E(Z)$
10. $V(X+Y+Z)=V(X)+V(Y)+V(Z)+2\,\mathrm{cov}(X,Y)+2\,\mathrm{cov}(X,Z)+2\,\mathrm{cov}(Y,Z)$
11. $E(X-Y-Z)=E(X)-E(Y)-E(Z)$
12. $V(X-Y-Z)=V(X)+V(Y)+V(Z)-2\,\mathrm{cov}(X,Y)-2\,\mathrm{cov}(X,Z)+2\,\mathrm{cov}(Y,Z)$

Suppose $E(X_i)=\mu$, $V(X_i)=\sigma^2$ for all i and $\mathrm{cov}(X_i,X_j)=\gamma_{|i-j|}$ for all i and j, $i\neq j$.

Now define $\bar{X}=\dfrac{\sum_{i=1}^{n}X_i}{n}=\dfrac{(X_1+X_2+\ldots+X_n)}{n}$

$$E(\bar{X})=E\left[\frac{(X_1+X_2+\ldots+X_n}{n}\right]=\frac{1}{n}E(X_1+X_2+\ldots+X_n)$$

$$=\frac{1}{n}\left[E(X_1)+\ldots+E(X_n)\right]=\frac{1}{n}\left[\mu+\mu+\ldots+\mu\right]=\mu$$

$$V(\bar{X})=V\left[\frac{(X_1+X_2+\ldots+X_n)}{n}\right]=\frac{1}{n^2}[V(X_1)+V(X_2)+\ldots+V(X_n)$$
$$+2\,\mathrm{cov}(X_1,X_2)+2\,\mathrm{cov}(X_1,X_3)+\ldots+2\,\mathrm{cov}(X_1,X_n)$$
$$+2\,\mathrm{cov}(X_2,X_3)+\ldots+2\,\mathrm{cov}(X_2,X_n)$$
$$+\ldots+$$
$$\ldots+2\,\mathrm{cov}(X_{n-1},X_n)]$$

assuming $X_1,\ldots,X_n$ are a random sample (with or without replacement, providing n is sufficiently large) then $\mathrm{cov}(X_i,X_j)=0$ for $i\neq j$ and

$V(\bar{X})=\dfrac{1}{n^2}[\sigma^2+\sigma^2+\ldots+\sigma^2]=\dfrac{\sigma^2}{n}$ (therefore the variance of the sample mean falls as the number of points in the sample increases – this is the effect of smoothing see Figure 1a-1d).

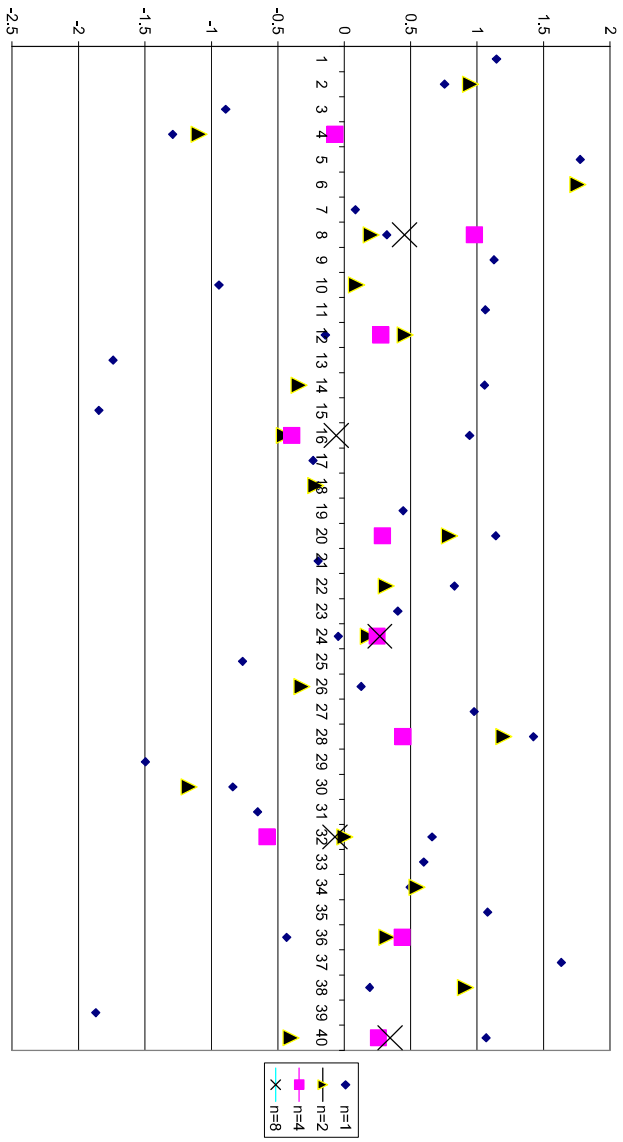


Figure 1: Effects of averaging of the standard deviation

Figure 1b: Dispersion with an average of n=2

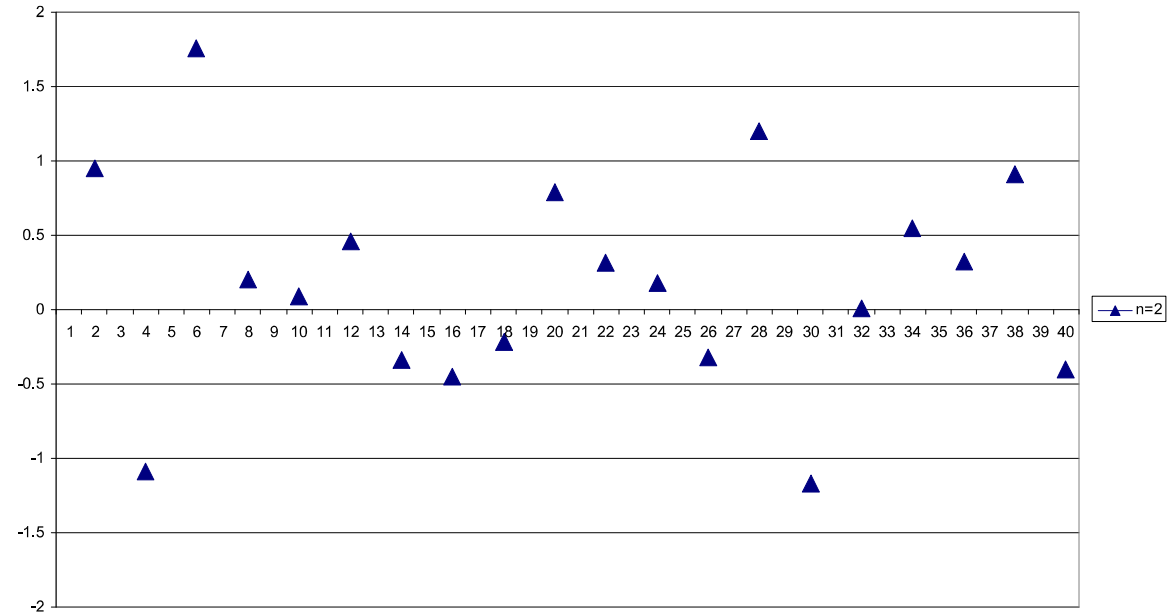


Figure 1a: Dispersion with an average of n=1

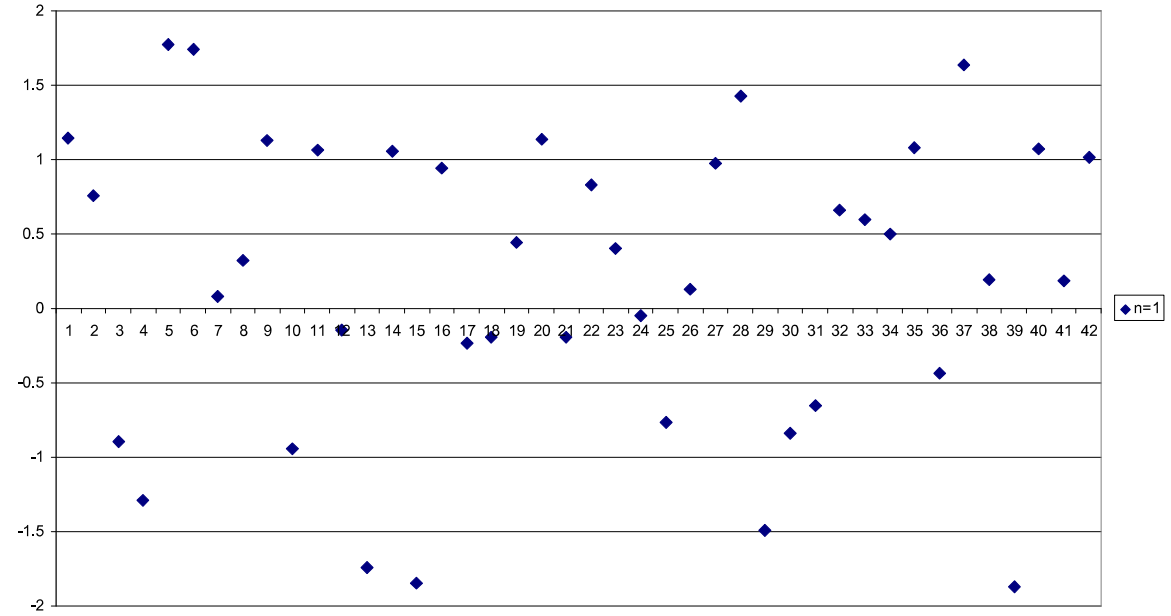


Figure 1d: Dispersion with an average of n=8

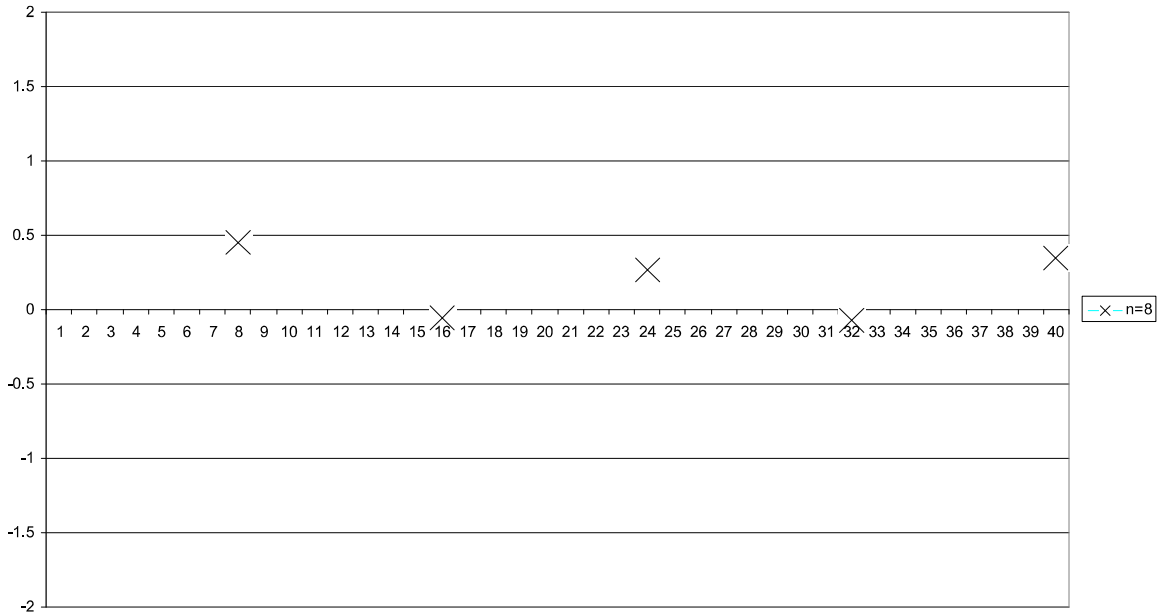
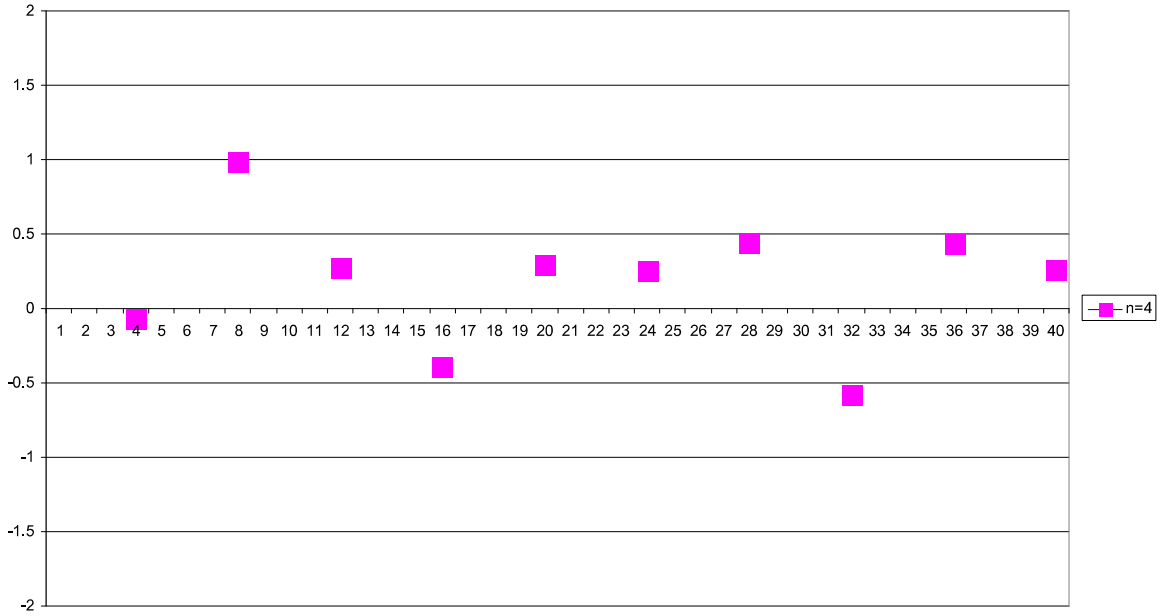


Figure 1c: Dispersion with an average of n=4



Basic Integration

Integration can be approximated as a summation of the function over small changes in x, that is,

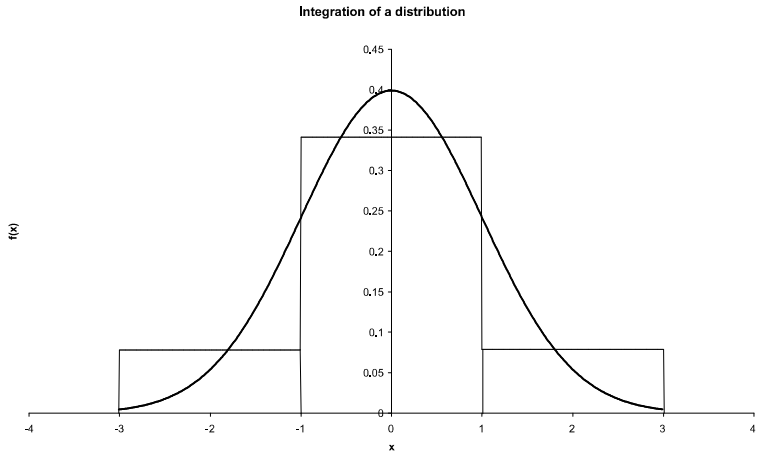
$\int_x f(x)\partial x \approx \sum_{i=1} f(x_i)(x_i - x_{i-1})$, the approximation is best as the changes in x become infinitesimal (very small) – see figures overleaf – where we can see that as the changes in x, ∂x, become smaller so the approximation of the areas of the rectangles become a better approximation to the area under the continuous line.

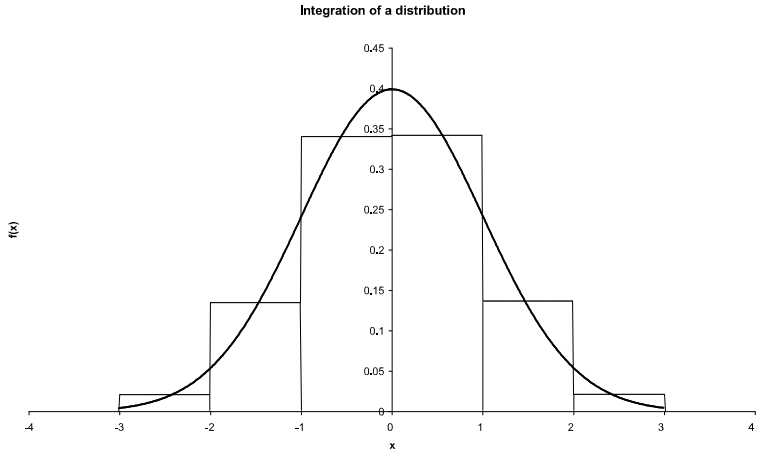
Some basic rules of integration:

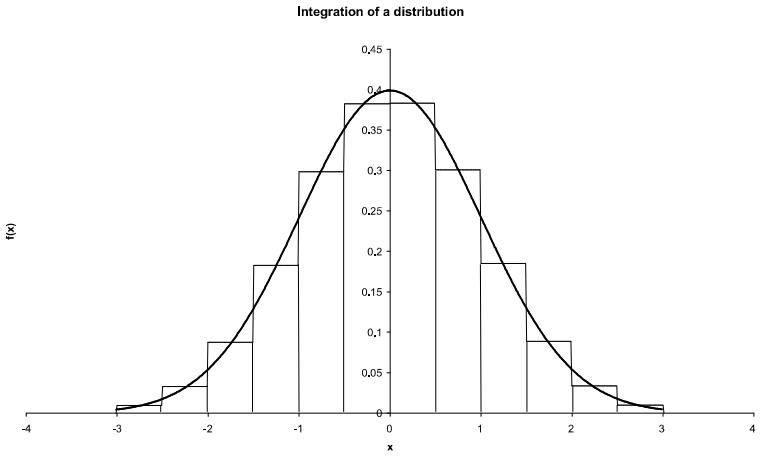
1. $\int_c^d x^n \partial x = \left.\frac{x^{n+1}}{n+1}\right|_c^d = \frac{d^{n+1}}{n+1} - \frac{c^{n+1}}{n+1}$ for all $n \neq -1$
2. $\int_c^d x^{-1} \partial x = \ln(x)\Big|_c^d = \ln(d) - \ln(c)$
3. $\int_c^d e^x \partial x = e^x\Big|_c^d = e^d - e^c$

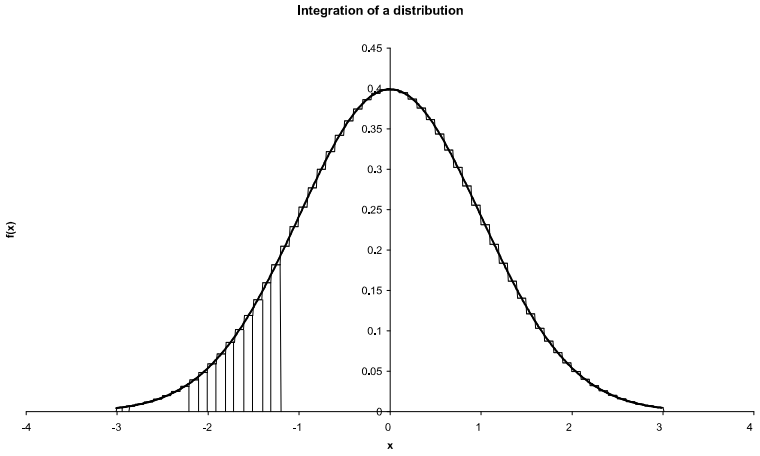
Remember integration is simply the inverse function of differentiation, so

$$\frac{\partial x^n}{\partial x} = nx^{n-1} \Rightarrow \int nx^{n-1}\partial x = \frac{nx^n}{n} = x^n$$






Integration of a distribution
f(x)
x
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4


Integration of a distribution
f(x)
x
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-4
-3
-2
-1
0
1
2
3
4

STATISTICAL TECHNIQUES B

Special Distributions

1. Bernoulli distribution

An experiment leading to only two outcomes – a 'success' and a 'failure' – called a Bernoulli trial x=0 (=failure) with probability, 1-p, and 1 (=success) with probability, p.

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

This is a valid pdf as:

$$\sum_{x=0} p_X(x) = (1-p) + p = 1$$

1.1 Mean

$$E(X) = 0(1-p) + 1p = p$$

$$E(X^2) = 0^2(1-p) + 1^2 p = p$$

1.2 Variance

$$V(X) = E(X^2) - E(X)^2 = p - p^2 = p(1-p)$$

2. Binomial distribution

This consists of having n independent Bernoulli trials, where we define X=number of 'successes' in n trials and X=0,1,2,…n.

We can then calculate the probability of a specific outcome such as:

$$P(S_1 \cap S_2 \ldots \cap S_x \cap F_{x+1} \cap F_{x+2} \ldots \cap F_n) = p^x(1-p)^{n-x}$$

However, as there are ${}_nC_x = \frac{n!}{(n-x)!x!}$ in which we can get x successes in n trials, given that the order is unimportant, we have:

$$p_X(x) = {}_nC_x p^x (1-p)^{n-x}$$

This is a valid probability density function as:

$$\sum_{x=0}^{n} p_X(x) = \sum_{x=0}^{n} {}_nC_x p^x (1-p)^{n-x} = (p + (1-p))^n = 1$$

For, n=3

$$p_X(0) + p_X(1) + p_X(2) + p_X(3) = p^0(1-p)^3 + 3p(1-p)^2 + 3p^2(1-p) + p^3(1-p)^0$$

2.1 Mean

$$E(X) = \sum_{x=0}^{n} x p_X(x) = \sum_{x=0}^{n} x \;\; {}_nC_x p^x (1-p)^{n-x} = np \sum_{x=1}^{n} {}_{n-1}C_{x-1} p^{x-1} (1-p)^{n-x}$$

defining m=n-1 and y=x-1 then,

$$E(X) = np \underbrace{\sum_{y=0}^{m} {}_mC_y p^y (1-p)^{m-y}}_{1} = np$$

2.2 Variance

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E[X(X-1)] =$$

$$\sum_{x=0}^{n} x(x-1) p_X(x) = \sum_{x=0}^{n} x(x-1) \;\; {}_nC_x p^x (1-p)^{n-x} = n(n-1)p^2 \sum_{x=2}^{n} {}_{n-2}C_{x-2} p^{x-2} (1-p)^{n-x}$$

$$= n(n-1)p^2 \sum_{y=0}^{m} {}_mC_y p^y (1-p)^{m-y} = n(n-1)p^2$$

$$E(X^2) = n(n-1)p^2 + np$$

$$V(X) = n(n-1)p^2 + np - n^2p^2 = np\left[np - p + 1 - np\right] = np(1-p)$$

2.3 Probability values (see Table 7)

$$P(a \le X \le b) = P(a) + P(a+1) + \ldots + P(b)$$

$$P(a \le X \le b) = {}_nC_a p^a (1-p)^{n-a} + {}_nC_{a+1} p^{a+1} (1-p)^{n-a-1} + \ldots + {}_nC_b p^b (1-p)^{n-b}$$

3. Poisson

We are interested in the number of occurrences of an event during a period of time, where the period of time, *T*, is divided into *n* unit time intervals (and n is very large). The probability of an event in an interval of time is *p* (which is small) and we assume that the events are independent occurrences.

λ = mean rate of occurrence

$$p_X(x) = \frac{e^{-\lambda}\lambda^x}{x!} \quad x = 0,1,2\ldots$$

This is a valid probability density function as:

$$\sum_{x=0} p_X(x) = \sum_{x=0} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda} \underbrace{\sum_{x=0} \frac{\lambda^x}{x!}}_{e^{\lambda}} = 1$$

as,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots$$

3.1 Mean

$$E(X) = \sum_{x=0} x p_X(x) = \sum_{x=0} x \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left[\frac{1\lambda}{1!} + \frac{2\lambda^2}{2!} + \frac{3\lambda^3}{3!} + \frac{4\lambda^4}{4!} + \ldots\right]$$

$$= e^{-\lambda}\lambda\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] = \lambda e^{-\lambda} \underbrace{\sum_{y=0} \frac{\lambda^y}{y!}}_{e^{\lambda}} = \lambda$$

3.2 Variance

$$V(X) = E(X^2) - [E(X)]^2$$

$$E(X^2) = E[X(X-1)] + E(X)$$

$$E(X(X-1)) = \sum_{x=0} x(x-1) p_X(x) = \sum_{x=0} x(x-1) \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left[\frac{2\lambda^2}{2!} + \frac{3(2)\lambda^3}{3!} + \frac{4(3)\lambda^4}{4!} + \ldots\right]$$

$$= e^{-\lambda}\lambda^2\left[1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \ldots\right] = \lambda^2$$

$$E(X^2) = \lambda^2 + \lambda$$

$$V(X) = E(X^2) - [E(X)]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

3.3 Probability values (see Table 8)

$$P(a \le X \le b) = P(a) + P(a+1) + \ldots + P(b)$$

$$P(a \le X \le b) = e^{-\lambda}\left[\frac{\lambda^a}{a!} + \frac{\lambda^{a+1}}{(a+1)!} + \ldots \frac{\lambda^b}{b!}\right]$$

NOTE:

The Poisson distribution can be used as an approximation to the Binomial distribution having the same mean. If a binomial distribution has a large, n (n>50) and a small p (p<0.1), then the probabilities of 0, 1, 2, … successes given by a Poison distribution with parameter $\lambda = np$ approximates well to the true probabilities given by the defined binomial distribution.

4. Uniform

Define the probability density function, such that all points in the interval (a,b) have an equal likelihood of occurring,

$$f(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{elsewhere} \end{cases}$$

This is a valid probability density function as:

$$\int_a^b f(x)\,dx = \int_a^b \frac{1}{b-a}\,dx = \frac{1}{b-a}\int_a^b 1\,dx = \frac{1}{b-a}\left[x\right|_a^b = 1$$

4.1 Mean

$$E(X) = \int_a^b xf(x)\,dx = \frac{1}{b-a}\int_a^b x\,dx = \frac{1}{b-a}\left[x^2/2\right|_a^b = \frac{1}{b-a}\left[b^2/2 - a^2/2\right]$$

$$= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2}$$

4.2 Variance

$$V(X) = E(X^2) - E(X)^2$$

$$E(X^2) = \int_a^b x^2 f(x)\,dx = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{b-a}\left[x^3/3\right|_a^b = \frac{1}{b-a}\left[b^3/3 - a^3/3\right]$$

$$= \frac{(b^3-a^3)}{3(b-a)} = \frac{(b^2+a^2)(b-a)+ab(b-a)}{3(b-a)} = \frac{(b^2+a^2)+ab}{3}$$

$$V(X) = E(X^2) - [E(X)]^2 = \frac{(b^2+a^2)+ab}{3} - \frac{(b+a)^2}{4}$$

$$V(X) = \frac{4(b^2+a^2)+4ab-3(b^2+a^2+2ab)}{12} = \frac{b^2+a^2-2ab}{12} = \frac{(b-a)^2}{12}$$

4.3 Probability values

$$P(c \le X \le d) = \int_c^d \frac{1}{b-a}\,dx = \frac{x}{b-a}\Big|_c^d = \frac{d-c}{b-a}$$

5. Normal Distribution

Define the probability density function

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right), \quad -\infty < x < \infty$$

This is a valid probability density function as:

$$\int_{-\infty}^{\infty} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) dx = 1$$

although, showing this is non-trivial.

5.1 Mean

$$E(X) = \mu$$

although, showing this is non-trivial.

5.2 Variance

$$V(X) = \sigma^2$$

although, showing this is non-trivial.

5.3 Probability values

$$P(a \le X \le b) = \int_{a}^{b} (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(x-\mu)^2}{2\sigma^2} \right) dx$$

and this is non-trivial. However, statistical tables are available for the standard normal distribution, Z, where $E(Z) = 0$ and $V(Z) = 1$, such that:

$$P(Z \le c) = (2\pi)^{-1/2} \int_{-\infty}^{c} \exp\left(\frac{-z^2}{2} \right) dz$$

As we know that

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

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

and this can be calculated from the standard normal statistical tables as:

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

(see Figure 1 and Table 1).

NOTE: As the area under the pdf is unity, $P(Z > c) = 1 - P(Z < c)$. In addition due to symmetry we have that $P(Z < -c) = P(Z > c) = 1 - P(Z < c)$ where $c > 0$.

NOTE: If X_1 and X_2 are both normally distributed then any linear combination of them is also normally distributed.

6. Negative exponential

Define the pdf as:

$$f(x) = \frac{1}{\theta} e^{-x/\theta}, \; x \geq 0$$

This is a valid probability density function as:

$$\int_0^\infty \frac{1}{\theta} e^{-x/\theta} dx = -e^{-x}\Big|_0^\infty = -0 - (-1) = 1$$

6.1 Mean

$$E(X) = \int_0^\infty x \frac{1}{\theta} e^{-x/\theta} dx = \left[-xe^{-x/\theta} - \int -e^{-x/\theta} dx \right] = -xe^{-x/\theta} - \theta e^{-x/\theta}\Big|_0^\infty = 0 - (-\theta) = \theta$$

6.2 Variance

$$V(X) = E(X^2) - E(X)^2$$

$$E(X^2) = \int_0^\infty x^2 \frac{1}{\theta} e^{-x/\theta} dx = -x^2 e^{-x/\theta} - 2x\theta e^{-x/\theta} - 2\theta^2 e^{-x/\theta}\Big|_0^\infty = 0 - (-2\theta^2) = 2\theta^2$$

$$V(X) = 2\theta^2 - \theta^2 = \theta^2$$

6.3 Probability values

$$P(a \leq X \leq b) = \int_a^b \frac{1}{\theta} e^{-x/\theta} dx = -\theta e^{-x/\theta}\Big|_a^b = -\theta \left[e^{-b/\theta} - e^{-a/\theta} \right]$$

7. Chi-squared distribution

Define the pdf as:

$$f(x) = \frac{1}{\Gamma(\upsilon/2)} \left(\frac{1}{2} \right)^{\frac{\upsilon}{2}} x^{\upsilon/2-1} e^{-x/2} \quad x > 0$$

This is a valid probability density function and is denoted as χ^2_υ, where υ are the degrees of freedom.

NOTE: $N(0,1)^2 = \chi^2_1$ and if $W_i \sim \chi^2_1$ and these are independent, then $\sum_{i=1}^{n} W_i = \chi^2_n$

7.1 Mean

$$E(X) = \upsilon$$

7.2 Variance

$$V(X) = E(X^2) - E(X)^2 = 2\upsilon$$

7.3 Probability values (see Table 3)

These are tabulated in the statistical tables.

8. F-distribution

Define the pdf as:

$$f(x)=\frac{\Gamma[(m+n)/2]}{\Gamma(m/2)\Gamma(n/2)}\left(\frac{m}{n}\right)^{m/2}\frac{x^{(m-2)/2}}{[1+(m/n)x]^{(m+n)/2}} \quad x>0$$

This is a valid probability density function and is denoted as an F-distribution with degrees of freedom m and n.

NOTE: An F distribution is formed as the ratio of 2 independent chi-squared distributions, $\frac{\chi_m^2/m}{\chi_n^2/n}\sim F_{m,n}$. As $n\to\infty$ so $F_{m,n}=\frac{\chi_m^2/m}{\chi_\infty^2/\infty}\sim\chi_m^2/m$.

8.1 Mean (for a $F_{m,n}$)

$$E(X)=\frac{n}{n-2} \qquad \text{for } n>2$$

8.2 Variance (for a $F_{m,n}$)

$$V(X)=\frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad \text{for } n>4$$

8.3 Probability values (see Table 5)

These are tabulated in the statistical tables.

9. Student t-distribution

Define the pdf as:

$$f(x)=\frac{\Gamma[(n+1)/2]}{\Gamma(n/2)}\frac{1}{\sqrt{n\pi}}\frac{1}{(1+x^2/n)^{(n+1)/2}}$$

This is a valid probability density function and is denoted as a t-distribution with degrees of freedom n.

NOTE: A t-distribution is formed as the ratio of a N(0,1) to a chi-square distribution, $\frac{N(0,1)}{\sqrt{\chi_n^2/n}}\sim t_n$. As $n\to\infty$ so $t_n\sim N(0,1)$.

9.1 Mean (for a t_n)

$$E(X)=0$$

9.2 Variance (for a t_n)

$$V(X)=\frac{n}{(n-2)} \qquad \text{for } n>2$$

9.3 Probability values (see Table 2)

These are tabulated in the statistical tables.

Table 7 (from Statistical tables): Binomial Distribution (cont'd)

$\Pr(X \le k)$

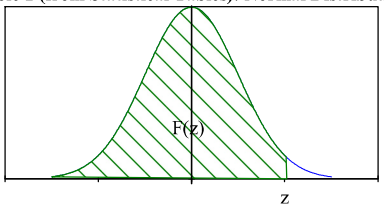
	n=8								n=9					
p	k=0	1	2	3	4	5	6	7	k=0	1	2	3	4	5
0.01	0.923	0.997	1.000						0.914	0.997	1.000			
0.02	0.851	0.990	1.000						0.834	0.987	0.999	1.000		
0.03	0.784	0.978	0.999	1.000					0.760	0.972	0.998	1.000		
0.04	0.721	0.962	0.997	1.000					0.693	0.952	0.996	1.000		
0.05	0.663	0.943	0.994	1.000					0.630	0.929	0.992	0.999	1.000	
0.06	0.610	0.921	0.990	0.999	1.000				0.573	0.902	0.986	0.999	1.000	
0.07	0.560	0.897	0.985	0.999	1.000				0.520	0.873	0.979	0.998	1.000	
0.08	0.513	0.870	0.979	0.998	1.000				0.472	0.842	0.970	0.996	1.000	
0.09	0.470	0.842	0.971	0.997	1.000				0.428	0.809	0.960	0.994	0.999	1.000
0.10	0.430	0.813	0.962	0.995	1.000				0.387	0.775	0.947	0.992	0.999	1.000
0.11	0.394	0.783	0.951	0.993	0.999	1.000			0.350	0.740	0.933	0.988	0.999	1.000
0.12	0.360	0.752	0.939	0.990	0.999	1.000			0.316	0.705	0.917	0.984	0.998	1.000
0.13	0.328	0.721	0.926	0.987	0.999	1.000			0.286	0.670	0.899	0.979	0.997	1.000
0.14	0.299	0.689	0.911	0.983	0.998	1.000			0.257	0.634	0.880	0.973	0.996	1.000
0.15	0.272	0.657	0.895	0.979	0.997	1.000			0.232	0.599	0.859	0.966	0.994	0.999
0.16	0.248	0.626	0.877	0.973	0.996	1.000			0.208	0.565	0.837	0.958	0.993	0.999
0.17	0.225	0.594	0.859	0.967	0.995	1.000			0.187	0.532	0.814	0.949	0.990	0.999
0.18	0.204	0.563	0.839	0.960	0.993	0.999	1.000		0.168	0.499	0.790	0.938	0.988	0.998
0.19	0.185	0.533	0.819	0.952	0.992	0.999	1.000		0.150	0.467	0.764	0.927	0.984	0.998
0.20	0.168	0.503	0.797	0.944	0.990	0.999	1.000		0.134	0.436	0.738	0.914	0.980	0.997
0.21	0.152	0.474	0.775	0.934	0.987	0.998	1.000		0.120	0.407	0.711	0.901	0.976	0.996
0.22	0.137	0.446	0.751	0.924	0.984	0.998	1.000		0.107	0.378	0.684	0.886	0.971	0.995
0.23	0.124	0.419	0.728	0.912	0.981	0.997	1.000		0.095	0.351	0.657	0.870	0.965	0.994
0.24	0.111	0.392	0.703	0.900	0.977	0.997	1.000		0.085	0.325	0.629	0.852	0.958	0.992
0.25	0.100	0.367	0.679	0.886	0.973	0.996	1.000		0.075	0.300	0.601	0.834	0.951	0.990
0.26	0.090	0.343	0.653	0.872	0.968	0.995	1.000		0.067	0.277	0.573	0.815	0.943	0.988
0.27	0.081	0.319	0.628	0.857	0.962	0.994	0.999	1.000	0.059	0.255	0.545	0.795	0.934	0.985
0.28	0.072	0.297	0.603	0.841	0.956	0.992	0.999	1.000	0.052	0.234	0.517	0.774	0.924	0.982
0.29	0.065	0.276	0.577	0.824	0.949	0.991	0.999	1.000	0.046	0.214	0.490	0.752	0.913	0.979
0.30	0.058	0.255	0.552	0.806	0.942	0.989	0.999	1.000	0.040	0.196	0.463	0.730	0.901	0.975
0.31	0.051	0.236	0.526	0.787	0.934	0.987	0.998	1.000	0.035	0.179	0.436	0.706	0.888	0.970
0.32	0.046	0.218	0.501	0.768	0.925	0.984	0.998	1.000	0.031	0.163	0.411	0.683	0.875	0.965
0.33	0.041	0.201	0.476	0.748	0.915	0.981	0.998	1.000	0.027	0.148	0.385	0.658	0.860	0.960
0.34	0.036	0.184	0.452	0.728	0.905	0.978	0.997	1.000	0.024	0.134	0.361	0.634	0.845	0.953
0.35	0.032	0.169	0.428	0.706	0.894	0.975	0.996	1.000	0.021	0.121	0.337	0.609	0.828	0.946
0.36	0.028	0.155	0.404	0.685	0.882	0.971	0.996	1.000	0.018	0.109	0.314	0.584	0.811	0.939
0.37	0.025	0.141	0.381	0.663	0.869	0.966	0.995	1.000	0.016	0.098	0.292	0.558	0.793	0.930
0.38	0.022	0.129	0.359	0.640	0.856	0.961	0.994	1.000	0.014	0.088	0.271	0.533	0.774	0.921
0.39	0.019	0.117	0.337	0.617	0.841	0.956	0.993	0.999	0.012	0.079	0.251	0.508	0.754	0.911
0.40	0.017	0.106	0.315	0.594	0.826	0.950	0.991	0.999	0.010	0.071	0.232	0.483	0.733	0.901
0.41	0.015	0.096	0.295	0.571	0.810	0.944	0.990	0.999	0.009	0.063	0.213	0.458	0.712	0.889
0.42	0.013	0.087	0.275	0.547	0.794	0.937	0.988	0.999	0.007	0.056	0.196	0.433	0.690	0.877
0.43	0.011	0.078	0.256	0.524	0.776	0.929	0.986	0.999	0.006	0.049	0.180	0.409	0.668	0.863
0.44	0.010	0.070	0.238	0.500	0.758	0.921	0.984	0.999	0.005	0.044	0.164	0.385	0.645	0.849
0.45	0.008	0.063	0.220	0.477	0.740	0.912	0.982	0.998	0.005	0.039	0.150	0.361	0.621	0.834
0.46	0.007	0.057	0.203	0.454	0.720	0.902	0.979	0.998	0.004	0.034	0.136	0.339	0.598	0.818
0.47	0.006	0.050	0.187	0.431	0.700	0.891	0.976	0.998	0.003	0.030	0.123	0.316	0.573	0.801
0.48	0.005	0.045	0.172	0.408	0.680	0.880	0.973	0.997	0.003	0.026	0.111	0.295	0.549	0.784
0.49	0.005	0.040	0.158	0.385	0.658	0.868	0.969	0.997	0.002	0.023	0.100	0.274	0.525	0.765
0.50	0.004	0.035	0.145	0.363	0.637	0.855	0.965	0.996	0.002	0.020	0.090	0.254	0.500	0.746

Table 8 (from Statistical Tables): Poisson Distribution

$\Pr(X \le k)$

λ	k=0	1	2	3	4	5	6	7	8	9	10
0.1	0.905	0.995	1.000	1.000							
0.2	0.819	0.982	0.999	1.000							
0.3	0.741	0.963	0.996	1.000							
0.4	0.670	0.938	0.992	0.999	1.000						
0.5	0.607	0.910	0.986	0.998	1.000						
0.6	0.549	0.878	0.977	0.997	1.000						
0.7	0.497	0.844	0.966	0.994	0.999	1.000					
0.8	0.449	0.809	0.953	0.991	0.999	1.000					
0.9	0.407	0.772	0.937	0.987	0.998	1.000					
1.0	0.368	0.736	0.920	0.981	0.996	0.999	1.000				
1.1	0.333	0.699	0.900	0.974	0.995	0.999	1.000				
1.2	0.301	0.663	0.879	0.966	0.992	0.998	1.000				
1.3	0.273	0.627	0.857	0.957	0.989	0.998	1.000				
1.4	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000			
1.5	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000			
1.6	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000			
1.7	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000			
1.8	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000		
1.9	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000		
2.0	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000		
2.1	0.122	0.380	0.650	0.839	0.938	0.980	0.994	0.999	1.000		
2.2	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000		
2.3	0.100	0.331	0.596	0.799	0.916	0.970	0.991	0.997	0.999	1.000	
2.4	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	
2.5	0.082	0.287	0.544	0.758	0.891	0.958	0.986	0.996	0.999	1.000	
2.6	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	
2.7	0.067	0.249	0.494	0.714	0.863	0.943	0.979	0.993	0.998	0.999	1.000
2.8	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000
2.9	0.055	0.215	0.446	0.670	0.832	0.926	0.971	0.990	0.997	0.999	1.000
3.0	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000
3.1	0.045	0.185	0.401	0.625	0.798	0.906	0.961	0.986	0.995	0.999	1.000
3.2	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000
3.3	0.037	0.159	0.359	0.580	0.763	0.883	0.949	0.980	0.993	0.998	0.999
3.4	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999
3.5	0.030	0.136	0.321	0.537	0.725	0.858	0.935	0.973	0.990	0.997	0.999
3.6	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999
3.7	0.025	0.116	0.285	0.494	0.687	0.830	0.918	0.965	0.986	0.995	0.998
3.8	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998
3.9	0.020	0.099	0.253	0.453	0.648	0.801	0.899	0.955	0.981	0.993	0.998
4.0	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997
4.1	0.017	0.085	0.224	0.414	0.609	0.769	0.879	0.943	0.976	0.990	0.997
4.2	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996
4.3	0.014	0.072	0.197	0.377	0.570	0.737	0.856	0.929	0.968	0.987	0.995
4.4	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994
4.5	0.011	0.061	0.174	0.342	0.532	0.703	0.831	0.913	0.960	0.983	0.993
4.6	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992
4.7	0.009	0.052	0.152	0.310	0.495	0.668	0.805	0.896	0.950	0.978	0.991
4.8	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990
4.9	0.007	0.044	0.133	0.279	0.458	0.634	0.777	0.877	0.938	0.972	0.988
5.0	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986

Table 1 (from Statistical Tables): Normal Distribution



$\Pr(Z \le z) = F(z)$

z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)	z	F(z)
0.00	0.500	0.50	0.691	1.00	0.841	1.50	0.933	2.00	0.977	2.50	0.994	3.00	0.999
0.01	0.504	0.51	0.695	1.01	0.844	1.51	0.934	2.01	0.978	2.51	0.994	3.01	0.999
0.02	0.508	0.52	0.698	1.02	0.846	1.52	0.936	2.02	0.978	2.52	0.994	3.02	0.999
0.03	0.512	0.53	0.702	1.03	0.848	1.53	0.937	2.03	0.979	2.53	0.994	3.03	0.999
0.04	0.516	0.54	0.705	1.04	0.851	1.54	0.938	2.04	0.979	2.54	0.994	3.04	0.999
0.05	0.520	0.55	0.709	1.05	0.853	1.55	0.939	2.05	0.980	2.55	0.995	3.05	0.999
0.06	0.524	0.56	0.712	1.06	0.855	1.56	0.941	2.06	0.980	2.56	0.995	3.06	0.999
0.07	0.528	0.57	0.716	1.07	0.858	1.57	0.942	2.07	0.981	2.57	0.995	3.07	0.999
0.08	0.532	0.58	0.719	1.08	0.860	1.58	0.943	2.08	0.981	2.58	0.995	3.08	0.999
0.09	0.536	0.59	0.722	1.09	0.862	1.59	0.944	2.09	0.982	2.59	0.995	3.09	0.999
0.10	0.540	0.60	0.726	1.10	0.864	1.60	0.945	2.10	0.982	2.60	0.995	3.10	0.999
0.11	0.544	0.61	0.729	1.11	0.867	1.61	0.946	2.11	0.983	2.61	0.995	3.11	0.999
0.12	0.548	0.62	0.732	1.12	0.869	1.62	0.947	2.12	0.983	2.62	0.996	3.12	0.999
0.13	0.552	0.63	0.736	1.13	0.871	1.63	0.948	2.13	0.983	2.63	0.996	3.13	0.999
0.14	0.556	0.64	0.739	1.14	0.873	1.64	0.949	2.14	0.984	2.64	0.996	3.14	0.999
0.15	0.560	0.65	0.742	1.15	0.875	1.65	0.951	2.15	0.984	2.65	0.996	3.15	0.999
0.16	0.564	0.66	0.745	1.16	0.877	1.66	0.952	2.16	0.985	2.66	0.996	3.16	0.999
0.17	0.567	0.67	0.749	1.17	0.879	1.67	0.953	2.17	0.985	2.67	0.996	3.17	0.999
0.18	0.571	0.68	0.752	1.18	0.881	1.68	0.954	2.18	0.985	2.68	0.996	3.18	0.999
0.19	0.575	0.69	0.755	1.19	0.883	1.69	0.954	2.19	0.986	2.69	0.996	3.19	0.999
0.20	0.579	0.70	0.758	1.20	0.885	1.70	0.955	2.20	0.986	2.70	0.997	3.20	0.999
0.21	0.583	0.71	0.761	1.21	0.887	1.71	0.956	2.21	0.986	2.71	0.997	3.21	0.999
0.22	0.587	0.72	0.764	1.22	0.889	1.72	0.957	2.22	0.987	2.72	0.997	3.22	0.999
0.23	0.591	0.73	0.767	1.23	0.891	1.73	0.958	2.23	0.987	2.73	0.997	3.23	0.999
0.24	0.595	0.74	0.770	1.24	0.893	1.74	0.959	2.24	0.987	2.74	0.997	3.24	0.999
0.25	0.599	0.75	0.773	1.25	0.894	1.75	0.960	2.25	0.988	2.75	0.997	3.25	0.999
0.26	0.603	0.76	0.776	1.26	0.896	1.76	0.961	2.26	0.988	2.76	0.997	3.26	0.999
0.27	0.606	0.77	0.779	1.27	0.898	1.77	0.962	2.27	0.988	2.77	0.997	3.27	0.999
0.28	0.610	0.78	0.782	1.28	0.900	1.78	0.962	2.28	0.989	2.78	0.997	3.28	0.999
0.29	0.614	0.79	0.785	1.29	0.901	1.79	0.963	2.29	0.989	2.79	0.997	3.29	0.999
0.30	0.618	0.80	0.788	1.30	0.903	1.80	0.964	2.30	0.989	2.80	0.997	3.30	1.000
0.31	0.622	0.81	0.791	1.31	0.905	1.81	0.965	2.31	0.990	2.81	0.998	3.31	1.000
0.32	0.626	0.82	0.794	1.32	0.907	1.82	0.966	2.32	0.990	2.82	0.998	3.32	1.000
0.33	0.629	0.83	0.797	1.33	0.908	1.83	0.966	2.33	0.990	2.83	0.998	3.33	1.000
0.34	0.633	0.84	0.800	1.34	0.910	1.84	0.967	2.34	0.990	2.84	0.998	3.34	1.000
0.35	0.637	0.85	0.802	1.35	0.911	1.85	0.968	2.35	0.991	2.85	0.998	3.35	1.000
0.36	0.641	0.86	0.805	1.36	0.913	1.86	0.969	2.36	0.991	2.86	0.998	3.36	1.000
0.37	0.644	0.87	0.808	1.37	0.915	1.87	0.969	2.37	0.991	2.87	0.998	3.37	1.000
0.38	0.648	0.88	0.811	1.38	0.916	1.88	0.970	2.38	0.991	2.88	0.998	3.38	1.000
0.39	0.652	0.89	0.813	1.39	0.918	1.89	0.971	2.39	0.992	2.89	0.998	3.39	1.000
0.40	0.655	0.90	0.816	1.40	0.919	1.90	0.971	2.40	0.992	2.90	0.998	3.40	1.000
0.41	0.659	0.91	0.819	1.41	0.921	1.91	0.972	2.41	0.992	2.91	0.998	3.41	1.000
0.42	0.663	0.92	0.821	1.42	0.922	1.92	0.973	2.42	0.992	2.92	0.998	3.42	1.000
0.43	0.666	0.93	0.824	1.43	0.924	1.93	0.973	2.43	0.992	2.93	0.998	3.43	1.000
0.44	0.670	0.94	0.826	1.44	0.925	1.94	0.974	2.44	0.993	2.94	0.998	3.44	1.000
0.45	0.674	0.95	0.829	1.45	0.926	1.95	0.974	2.45	0.993	2.95	0.998	3.45	1.000
0.46	0.677	0.96	0.831	1.46	0.928	1.96	0.975	2.46	0.993	2.96	0.998	3.46	1.000
0.47	0.681	0.97	0.834	1.47	0.929	1.97	0.976	2.47	0.993	2.97	0.999	3.47	1.000
0.48	0.684	0.98	0.836	1.48	0.931	1.98	0.976	2.48	0.993	2.98	0.999	3.48	1.000
0.49	0.688	0.99	0.839	1.49	0.932	1.99	0.977	2.49	0.994	2.99	0.999	3.49	1.000

Calculating the probability Pr[(a-μ)/σ>z>(b-μ)/σ]

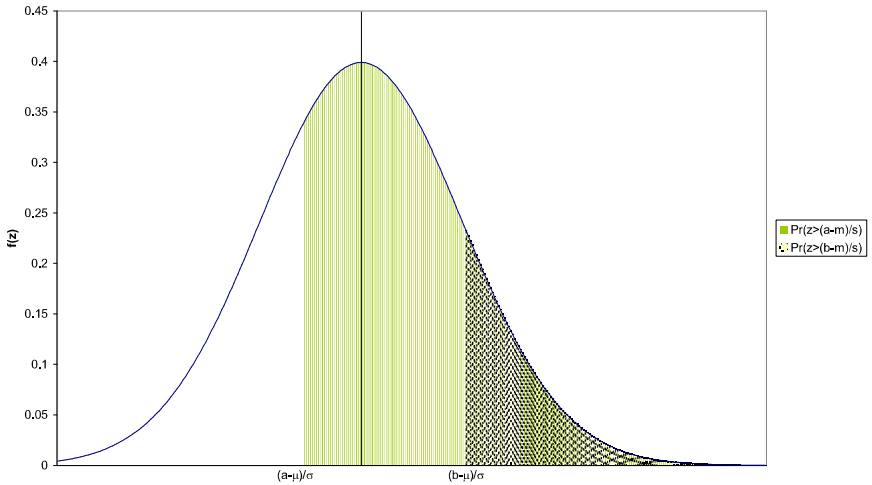
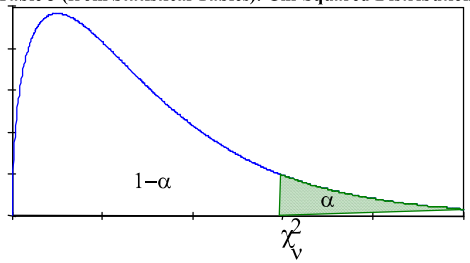


Table 3 (from Statistical Tables): Chi-Squared Distribution

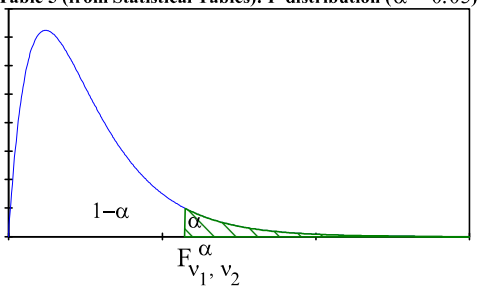


$\Pr(X \ge \chi^2_{\nu,\alpha})$

	α									
ν	0.995	0.99	0.975	0.95	0.9	0.1	0.05	0.025	0.01	0.005
1	0.00	0.00	0.00	0.00	0.02	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	6.25	7.81	9.35	11.34	12.84
4	0.21	0.30	0.48	0.71	1.06	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16

19	6.84	7.63	8.91	10.12	11.65	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	21.43	41.42	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	22.27	42.58	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	23.11	43.75	47.40	50.73	54.78	57.65
34	16.50	17.79	19.81	21.66	23.95	44.90	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	24.80	46.06	49.80	53.20	57.34	60.27
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	63.17	67.50	71.42	76.15	79.49
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
70	43.28	45.44	48.76	51.74	55.33	85.53	90.53	95.02	100.43	104.21
80	51.17	53.54	57.15	60.39	64.28	96.58	101.88	106.63	112.33	116.32
90	59.20	61.75	65.65	69.13	73.29	107.57	113.15	118.14	124.12	128.30
100	67.33	70.06	74.22	77.93	82.36	118.50	124.34	129.56	135.81	140.17

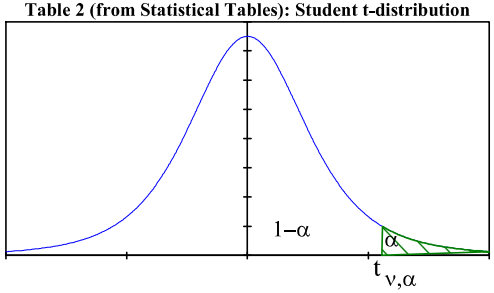
Table 5 (from Statistical Tables): F-distribution ($\alpha = 0.05$)



$$\Pr(X > F^{\alpha}_{\nu_1,\nu_2}) = \alpha$$

	ν_1									
ν_2	1	2	3	4	5	6	7	8	9	10
1	161.55	199.71	215.95	224.84	230.42	234.25	237.04	239.16	240.82	242.16
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.73
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
31	4.16	3.30	2.91	2.68	2.52	2.41	2.32	2.25	2.20	2.15
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19	2.14
33	4.14	3.28	2.89	2.66	2.50	2.39	2.30	2.23	2.18	2.13
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17	2.12
35	4.12	3.27	2.87	2.64	2.49	2.37	2.29	2.22	2.16	2.11
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08
45	4.06	3.20	2.81	2.58	2.42	2.31	2.22	2.15	2.10	2.05
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83

Table 2 (from Statistical Tables): Student t-distribution



$\Pr(X > t_{\nu,\alpha}) = \alpha$

	α				
ν	0.10	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
31	1.309	1.696	2.040	2.453	2.744
32	1.309	1.694	2.037	2.449	2.738
33	1.308	1.692	2.035	2.445	2.733
34	1.307	1.691	2.032	2.441	2.728
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
100	1.290	1.660	1.984	2.364	2.626
∞	1.282	1.645	1.960	2.327	2.576

Special Distributions - Examples

1. A large batch of clay pots is moulded and fired. After firing, a random sample of 10 pots is inspected for flaws before glazing, decoration and final firing. If 20% of pots in the batch have flaws, calculate the probability that the random sample contains: (a) no pots with flaws, (b) exactly one pot with a flaw, (c) exactly two pots with flaws, (d) less than three pots with flaws.

Answer

$X \sim B(10, 0.2)$

(a) $\Pr(X = 0) = 0.2^0 0.8^{10} = 0.107$, (b) $\Pr(X = 1) = 10(0.2^1)0.8^9 = 0.268$

(c) $\Pr(X = 2) = 45(0.2^2)0.8^8 = 0.302$

(d) $\Pr(X \le 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) = 0.678$ (from Statistical Tables)

2 A manufacturer sets up a 'double sampling' scheme as follows. A sample of 8 items is taken from a large lot ready for dispatch to customers. If there are no defectives, the lot is accepted and if there are 3 or more defectives, the lot is rejected. If there is either 1 or 2 defectives in the sample, a second sample is taken from the same lot, and the lot is rejected only if there are 3 or more defectives in the 2 samples combined. 12% of items are defective. (a) What proportion of lots will be accepted using only a single sampling scheme (with 3 or more defectives per sample causing a lot rejection, (b) What proportion of lots will be accepted using the 'double sampling' scheme.

Answer

(a) $\Pr(X < 2) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2)$

As $X \sim B(8, 0.12)$, then $\Pr(X \le 2) = 0.939$.

(b) Proportion of rejections is

$\Pr(X_1 \ge 3) + \Pr(X_1 = 1).\Pr(X_2 \ge 2) + \Pr(X_1 = 2).\Pr(X_2 \ge 1)$

$\Pr(X_1 \ge 3) + \Pr(X_1 = 1).[1 - \Pr(X_2 \le 1)] + \Pr(X_1 = 2).[1 - \Pr(X_2 = 0)]$

$0.061 + 0.392(0.248) + 0.187(0.64) = 0.278$.

Therefore proportion of acceptances=0.722.

3. A large batch of items is known to have a proportion 0.03 defective. If a sample of 200 is taken, what is the probability that the sample will contain (a) no defectives, (b) 4 defectives or less, (c) more than 5 defectives.

Answer

Poisson approximation to binomial distribution.

$X \sim B(200, 0.03) \approx X \sim P(200(0.03))$

(a) $\Pr(X = 0) = \frac{6^0}{0!} e^6 = 0.0025$

(b) $\Pr(X \le 4) = \Pr(X = 0) + \Pr(X = 1) + \Pr(X = 2) + \Pr(X = 3) + \Pr(X = 4)$

$e^{-6} + \frac{6^1}{1!} e^{-6} + \frac{6^2}{2!} e^{-6} + \frac{6^3}{3!} e^{-6} + \frac{6^4}{4!} e^{-6} = 0.0025 + 0.0149 + 0.0446 + 0.0892 + 0.1339 = 0.285$

Alternatively, from Statistical Tables $\lambda = 6$ and k=4.

(c)

$\Pr(X > 5) = 1 - \Pr(X \le 5) = 1 - [\Pr(X \le 4) + \Pr(X = 5)] = 1 - [0.285 + 0.161] = 1 - 0.446 = 0.554$

(also from Statistical Tables).

4. A random variable, Y is N(3,16). Find the probability that a value of Y taken at random will be negative. If 20 values are taken randomly, what is the probability that at least 3 have negative values?

Answer

(a) $\Pr(Y < 0) = \Pr\left(\frac{Y-3}{4} < \frac{0-3}{4}\right) = \Pr(z < -0.75) = 0.227$

(b) $X \sim B(20, 0.227) \Rightarrow \Pr(X \ge 3) = 1 - \Pr(X \le 2)$

$1 - (0.0058 + 0.0341 + 0.0951) = 0.865$

5. As a result of tests on electric light bulbs, it was found that the lifetime of a particular make of bulb was distributed normally, with a mean of 2040 hours and standard deviation of 60 hours. What proportion of bulbs can be expected to burn (a) For more than 2150 hours, (b) for more than 1960 hours?

Answer

(a) $\Pr(X > 2150) = \Pr\left(\frac{X-2040}{60} > \frac{2150-2040}{60}\right) = \Pr(z > 1.83) = 0.034$

(b) $\Pr(X > 1960) = \Pr\left(\frac{X-2040}{60} > \frac{1960-2040}{60}\right) = \Pr(z > -1.33) = 0.908$

6. If the random variables X_1, X_2, and X_3 are distributed as χ_1^2, χ_5^2 and χ_{10}^2, respectively, find the distribution of (a) X_1+X_2, (b) X_1+X_3.

Answer

(a) χ_6^2, (b) χ_{11}^2

7. Use the chi-squared tables such that (a) $\Pr(\chi_9^2 > 19.02) = p$, (b) $\Pr(\chi_{40}^2 > 24.43) = p$, (c) $\Pr(\chi_{29}^2 > x) = 0.005$, (d) $\Pr(\chi_4^2 > x) = 0.99$

Answers

(a) 0.025, (b) 0.975, (c) 52.34, (d) 0.30.

8. Use the F tables such that (a) $\Pr(\chi_{5,7}^2 > 7.46) = p$, (b) $\Pr(F_{1,60} > 2.79) = p$, (c) $\Pr(F_{10,1} > x) = 0.10$, (d) $\Pr(F_{15,20} > x) = 0.05$

Answer

(a) 0.01, (b) 0.10, (c) 60.24, (d) 2.20

8. If $X \sim B(3, 0.667)$ and $Y \sim P(1)$, find, (a) $\Pr(X + Y = 4)$, (b) $\Pr(X + Y \le 2)$

Answer

X and Y are independent

(a) $\Pr(X+Y=4) = \Pr(X=0).\Pr(Y=4) + \Pr(X=1).\Pr(Y=3) + \Pr(X=2).\Pr(Y=4) + \Pr(X=3).\Pr(Y=1)$

(as Pr(X=4)=0).

$\Pr(X + Y = 4) = 0.0006 + 0.0136 + 0.0818 + 0.1090 = 0.2049$

(b) $\Pr(X+Y\le 2) = \Pr(X=0).\Pr(Y=0) + \Pr(X=0).\Pr(Y=1) + \Pr(X=0).\Pr(Y=2) + \Pr(X=1).\Pr(Y=0) + \Pr(X=1).\Pr(Y=1) + \Pr(X=2).\Pr(Y=0)$

$\Pr(X + Y \le 2) = 0.0136 + 0.0136 + 0.0068 + 0.0817 + 0.0817 + 0.1635 = 0.3611$

Binomial Distribution

Binomial expansion of $(x+y)^d$

d											
1										x	y
2									x^2	$2xy$	y^2
3								x^3	$3x^2y$	$3xy^2$	y^3
4							x^4	$4x^3y$	$6x^2y^2$	$4xy^3$	y^4
5						x^5	$5x^4y$	$10x^3y^2$	$10x^2y^3$	$5xy^4$	y^5
6					x^6	$6x^5y$	$15x^4y^2$	$20x^3y^3$	$15x^2y^4$	$6xy^5$	y^6
7				x^7	$7x^6y$	$21x^5y^2$	$35x^4y^3$	$35x^3y^4$	$21x^2y^5$	$7xy^6$	y^7
8			x^8	$8x^7y$	$28x^6y^2$	$56x^5y^3$	$70x^4y^4$	$56x^3y^5$	$28x^2y^6$	$8xy^7$	y^8
9		x^9	$9x^8y$	$36x^7y^2$	$84x^6y^3$	$126x^5y^4$	$126x^4y^5$	$84x^3y^6$	$36x^2y^7$	$9xy^8$	y^9
10	x^{10}	$10x^9y$	$45x^8y^2$	$120x^7y^3$	$210x^6y^4$	$252x^5y^5$	$210x^4y^6$	$120x^3y^7$	$45x^2y^8$	$10xy^9$	y^{10}

Coefficients on the binomial expansion of $(x+y)^d$

d											
1										1	1
2									1	2	1
3								1	3	3	1
4							1	4	6	4	1
5						1	5	10	10	5	1
6					1	6	15	20	15	6	1
7				1	7	21	35	35	21	7	1
8			1	8	28	56	70	56	28	8	1
9		1	9	36	84	126	126	84	36	9	1
10	1	10	45	120	210	252	210	120	45	10	1

Coefficients on the binomial expansion of $(x+y)^d$

d											
1										${}_1C_1$	${}_1C_0$
2									${}_2C_2$	${}_2C_1$	${}_2C_0$
3								${}_3C_3$	${}_3C_2$	${}_3C_1$	${}_3C_0$
4							${}_4C_4$	${}_4C_3$	${}_4C_2$	${}_4C_1$	${}_4C_0$
5						${}_5C_5$	${}_5C_4$	${}_5C_3$	${}_5C_2$	${}_5C_1$	${}_5C_0$
6					${}_6C_6$	${}_6C_5$	${}_6C_4$	${}_6C_3$	${}_6C_2$	${}_6C_1$	${}_6C_0$
7				${}_7C_7$	${}_7C_6$	${}_7C_5$	${}_7C_4$	${}_7C_3$	${}_7C_2$	${}_7C_1$	${}_7C_0$
8			${}_8C_8$	${}_8C_7$	${}_8C_6$	${}_8C_5$	${}_8C_4$	${}_8C_3$	${}_8C_2$	${}_8C_1$	${}_8C_0$
9		${}_9C_9$	${}_9C_8$	${}_9C_7$	${}_9C_6$	${}_9C_5$	${}_9C_4$	${}_9C_3$	${}_9C_2$	${}_9C_1$	${}_9C_0$
10	${}_{10}C_{10}$	${}_{10}C_9$	${}_{10}C_8$	${}_{10}C_7$	${}_{10}C_6$	${}_{10}C_5$	${}_{10}C_4$	${}_{10}C_3$	${}_{10}C_2$	${}_{10}C_1$	${}_{10}C_0$

STATISTICAL TECHNIQUES B

Central Limit Theorem

1. Introduction

Many random variables can be characterised as either the sum or the average of a fairly large number of independent random variables. Let, $X_1, X_2, \ldots X_n$, be n independent random variables having identical distributions with mean, μ, and variance, σ^2. Denote their sum by:

$$X = X_1 + X_2 + \ldots + X_n$$

Now we know that from Appendix 1 in Handout 5,

$$E(X) = E(X_1 + X_2 + \ldots + X_n) = \underbrace{E(X_1)}_{\mu} + \underbrace{E(X_2)}_{\mu} + \ldots + \underbrace{E(X_n)}_{\mu} = n\mu$$

$$V(X) = V(X_1 + X_2 + \ldots + X_n) = V(X_1) + V(X_2) + \ldots + V(X_n)$$
$$+2\underbrace{\operatorname{cov}(X_1, X_2)}_{0} + 2\underbrace{\operatorname{cov}(X_1, X_3)}_{0} + \ldots + 2\underbrace{\operatorname{cov}(X_{n-1} X_n)}_{0} = n\sigma^2$$

The central limit theorem states, that whatever the distribution of X_i (provided that σ^2 is finite) as the number of terms in the sum become large, the distribution of X tends to a normal distribution, that is,

$$X \xrightarrow{a} N(n\mu, n\sigma^2).$$

Therefore the normal distribution will provide a satisfactory approximation to the true distribution for many statistical problems as these involve either sums or averages. This results applies regardless of whether the underlying distribution is continuous and symmetric like the uniform distribution, continuous and asymmetric like the chi-squared distribution, or even discrete such as the Binomial distribution. In fact, figures 1-5 show the shape of Binomial distribution when n=1, n=3, n=10, n=30 and n=100 – it is clear by the last graph (n=100) the distribution of the trial is normal.

2. Normal approximation to the Binomial distribution

If n independent trial, each with probability of success, p, are carried out, then the number X of successes resulting has a Binomial distribution, with mean, $E(x) = np$ and variance, $V(x) = np(1-p)$.

Clearly the random variable X can be written as the sum of n independent Bernoulli random variables, that is,

$$X = X_1 + X_2 + \ldots + X_n$$

where X_i takes the value 1 with probability, p, if the ith trial is a success and zero otherwise. It therefore follows that

$$X \xrightarrow{a} N(np, np(1-p))$$

This enables us to calculate with reasonable ease the probability that the number of successes lies in some given range as

$$P(a \le X \le b) = P\left(\frac{a-np}{\sqrt{np(1-p)}} \le \underbrace{\frac{X-np}{\sqrt{np(1-p)}}}_{Z} \le \frac{b-np}{\sqrt{np(1-p)}} \right)$$

for large n. For reasonably small n a finite correction may be needed as the binomial distribution is discrete and the normal distribution is continuous we may want to do a continuity correction:

$$P(a \le X \le b) = P\left(\frac{a-0.5-np}{\sqrt{np(1-p)}} \le \underbrace{\frac{X-np}{\sqrt{np(1-p)}}}_{Z} \le \frac{b+0.5-np}{\sqrt{np(1-p)}} \right)$$

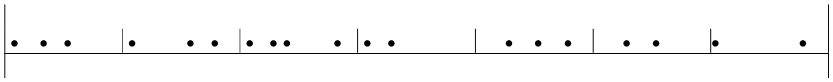
3. Normal approximation to the Poisson distribution

Let the random variable X denote the number of occurrences of an event in a particular interval of time and denote by λ the expected number of occurrences in that interval of time. Then X will follow a Poisson distribution with $E(X)=\lambda$ and $V(X)=\lambda$. Assume now that the mean number of occurrences is large[1] and the interval of time is broken down into sub-intervals of equal width as in Figure 6 below. Then the total number of occurrences (X) is the sum of the number of occurrences in each sub-interval, that is,

$$X = X_1 + X_2 + \ldots + X_n$$

where X_1 denotes the number of occurrences in the 1st sub-interval and X_n the number of occurrences in the n^{th} sub-interval.

Figure 6: Occurrences (•) in the interval broken down into equal sub-intervals



It therefore follows that

$$X \overset{a}{\rightarrow} N(\lambda,\lambda)$$

This enables us to calculate with reasonable ease the probability that the number of successes lies in some given range as

$$P(a \le X \le b) = P\left(\frac{a-\lambda}{\sqrt{\lambda}} \le \underbrace{\frac{X-\lambda}{\sqrt{\lambda}}}_{Z} \le \frac{b-\lambda}{\sqrt{\lambda}} \right)$$

As the Poisson distribution is a discrete random distribution and the normal distribution is continuous we may want to do a continuity correction

$$P(a \le X \le b) = P\left(\frac{a-0.5-\lambda}{\sqrt{\lambda}} \le \underbrace{\frac{X-\lambda}{\sqrt{\lambda}}}_{Z} \le \frac{b+0.5-\lambda}{\sqrt{\lambda}} \right).$$

[1] If λ is small we cannot divide the time interval into sufficient sub-intervals with some occurrences in each.

Binomial Distribution

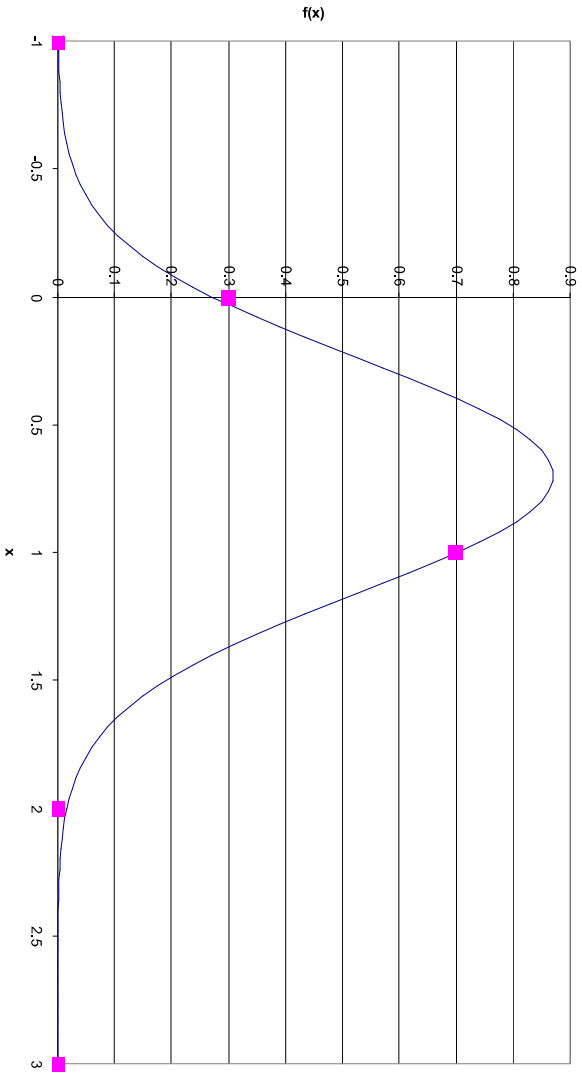


Figure 1: Distribution for a binomial with p=0.7 and n=1

Figure 3: Distribution for a binomial with p=0.7 and n=10

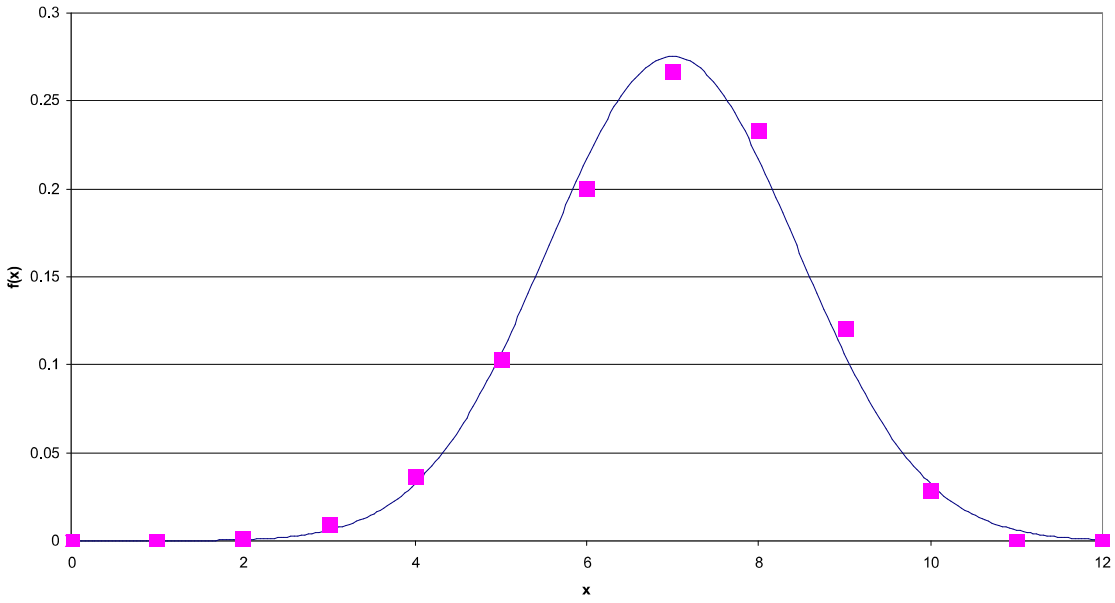


Figure 2: Distribution for a binomial with p=0.7 and n=3

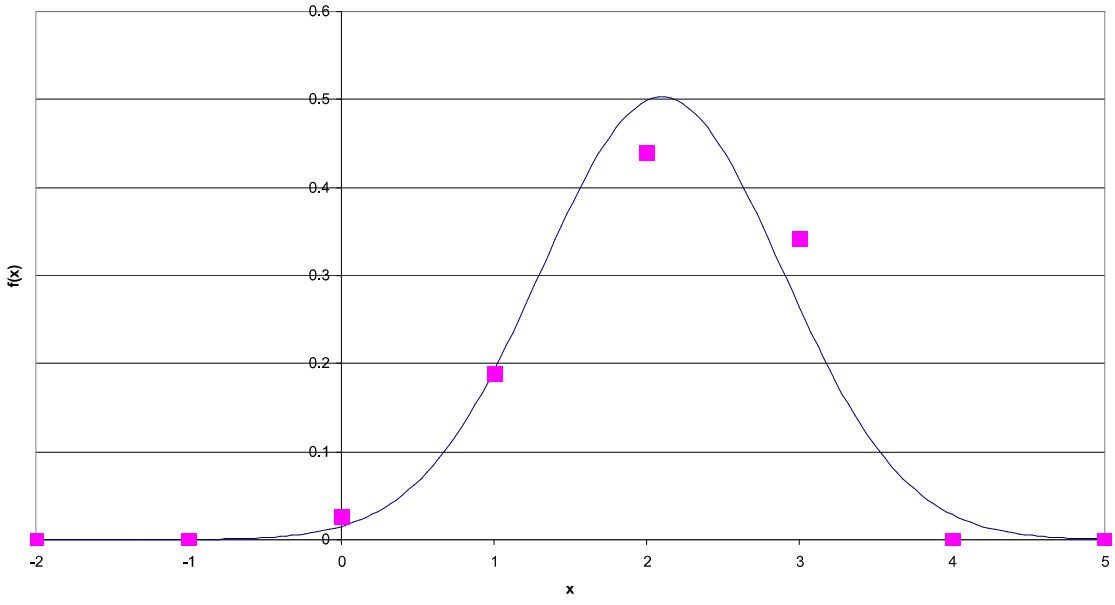


Figure 5: Distribution for a binomial with p=0.7 and n=30

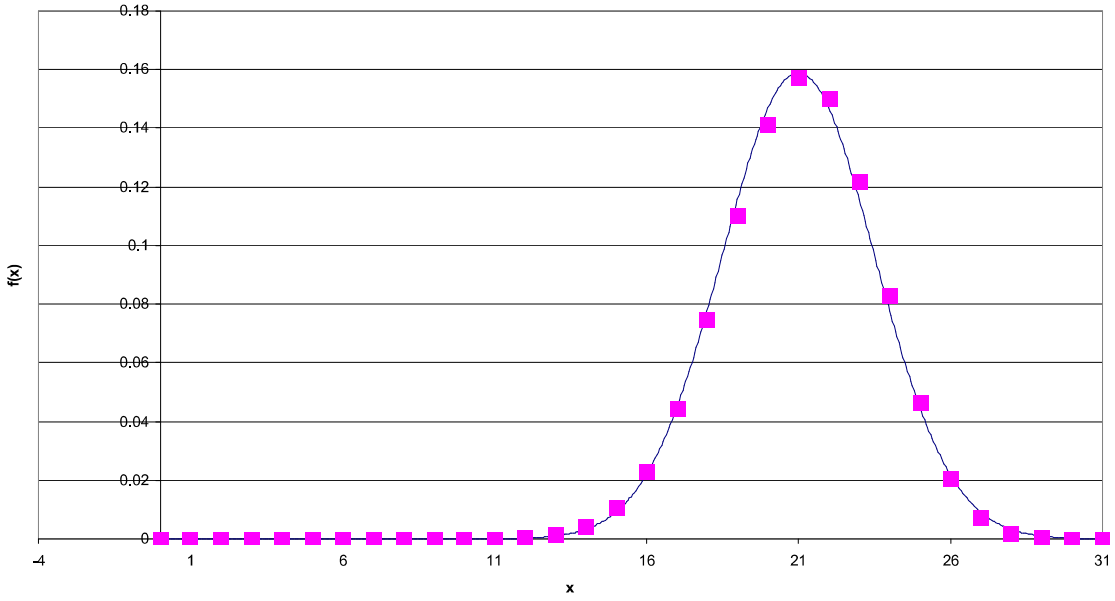


Figure 4: Distribution for a binomial with p=0.7 and n=20

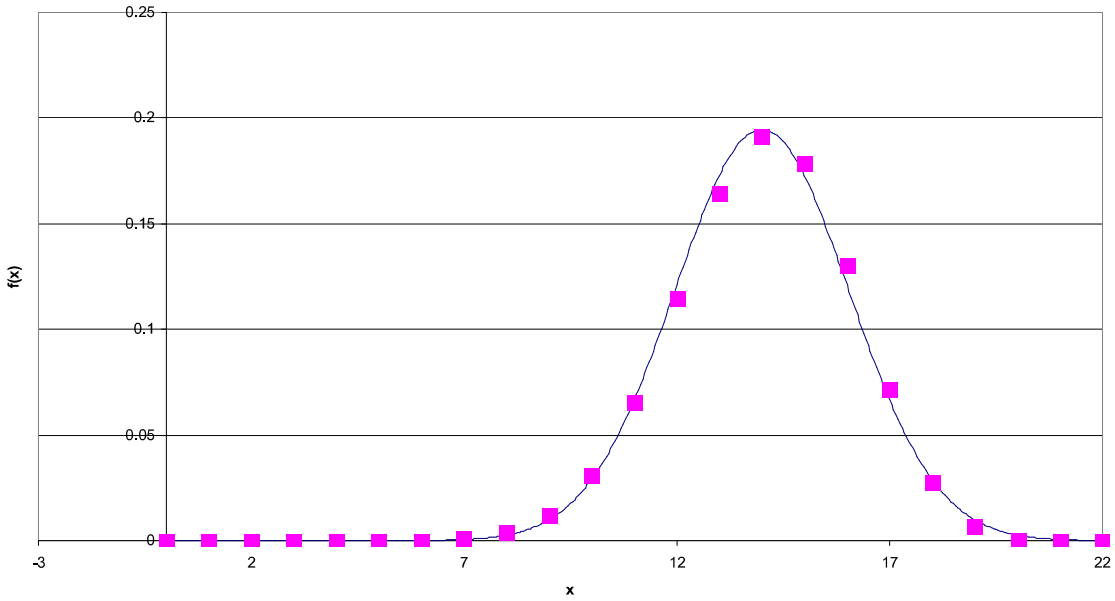


Figure 7: Poisson Distribution (lambda=1)

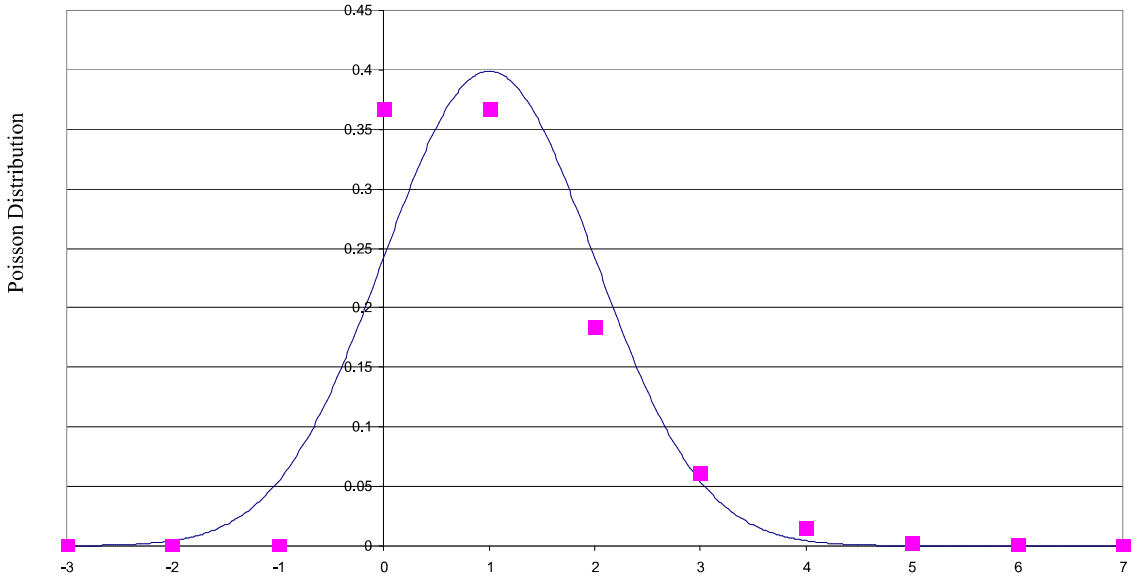


Figure 6: Distribution for a binomial with p=0.7 and n=100

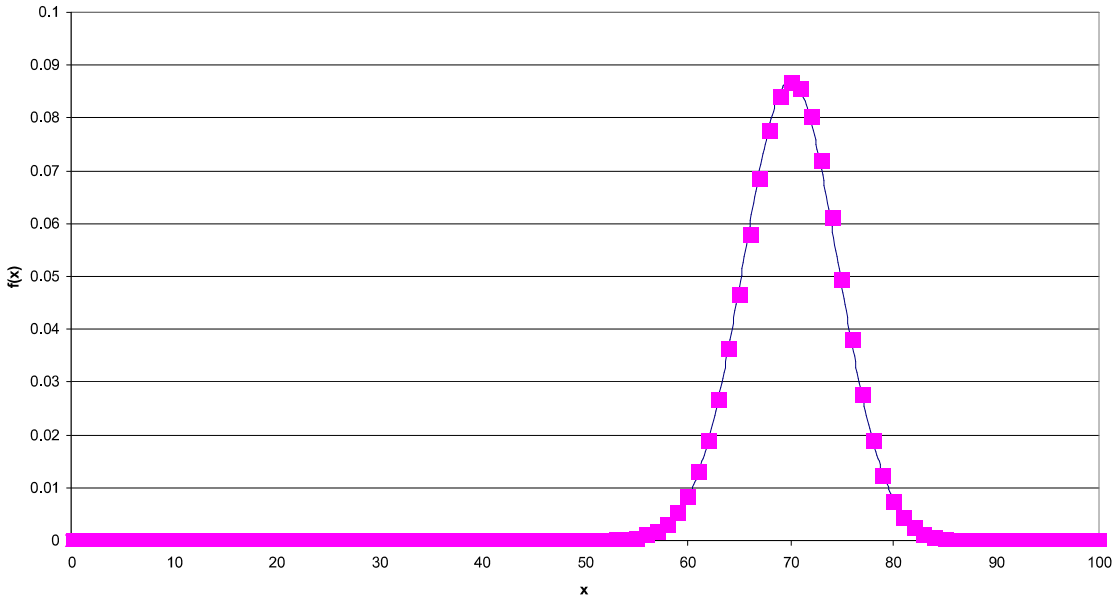


Figure 9: Poisson Distribution (lambda=4)

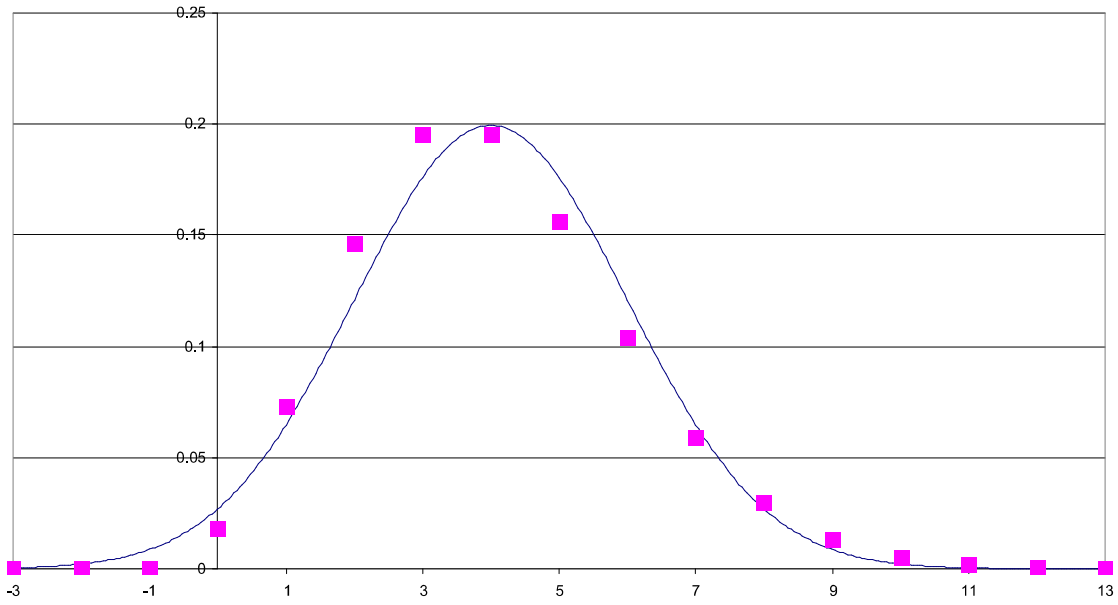


Figure 8: Poisson Distribution (lambda=2)

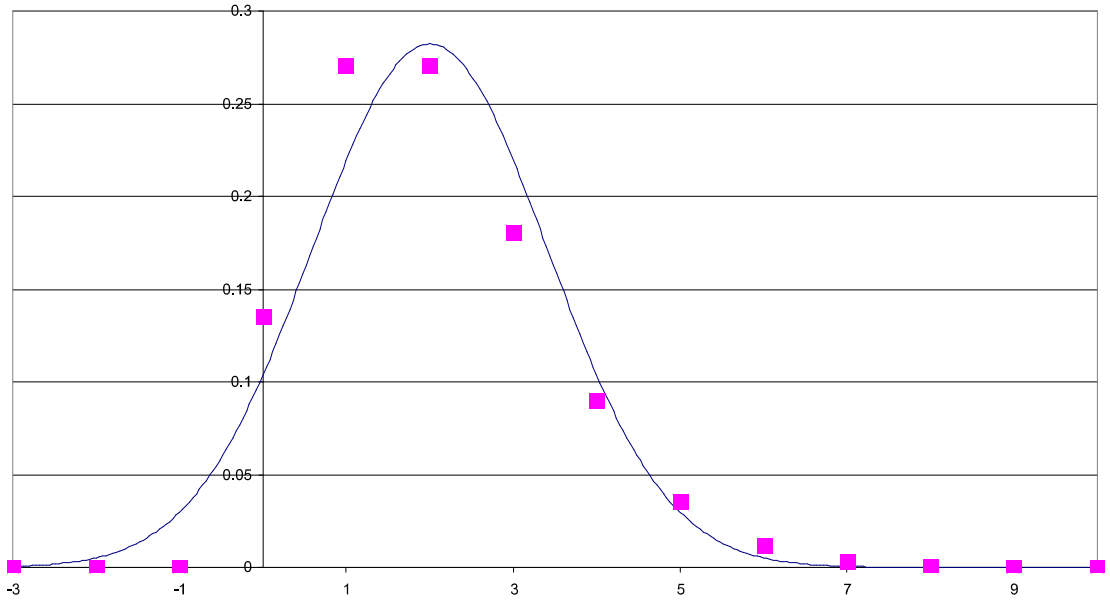


Figure 11: Poisson Distribution (lambda=20)

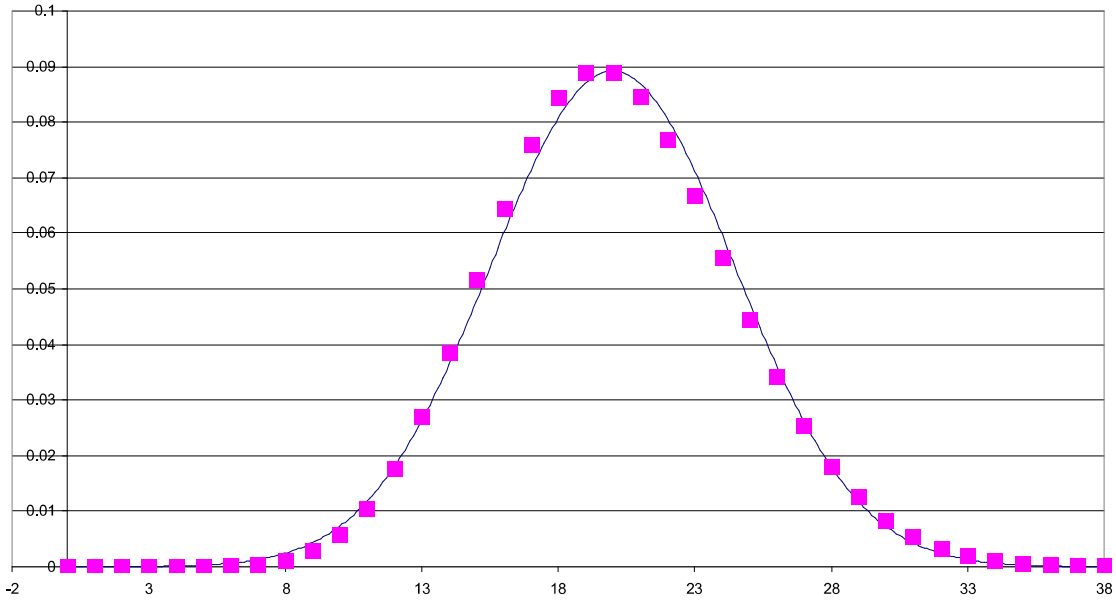
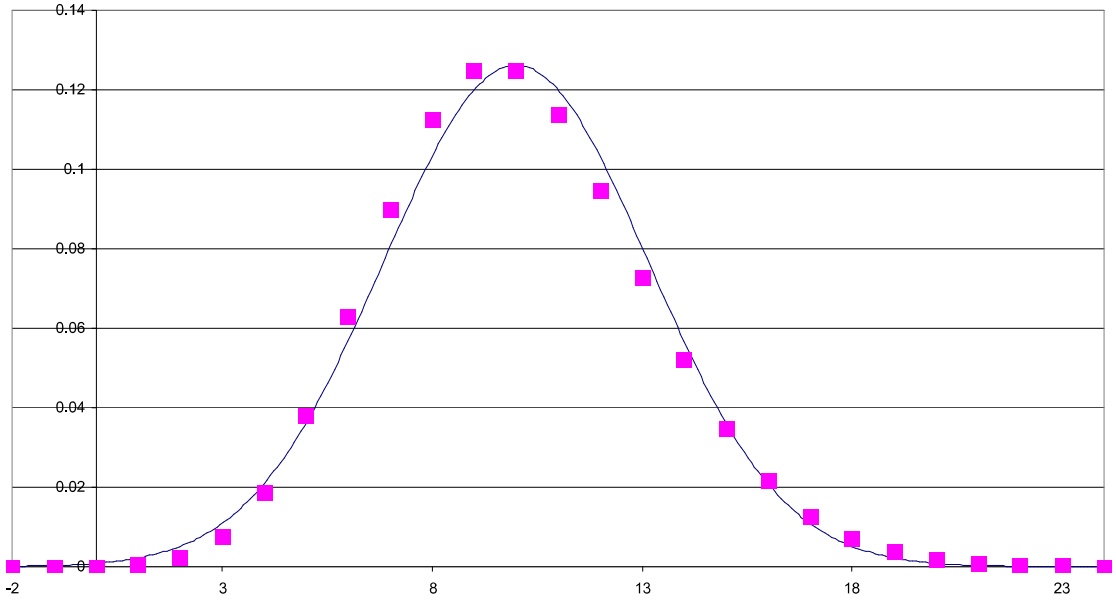


Figure 10: Poisson Distribution (lambda=10)



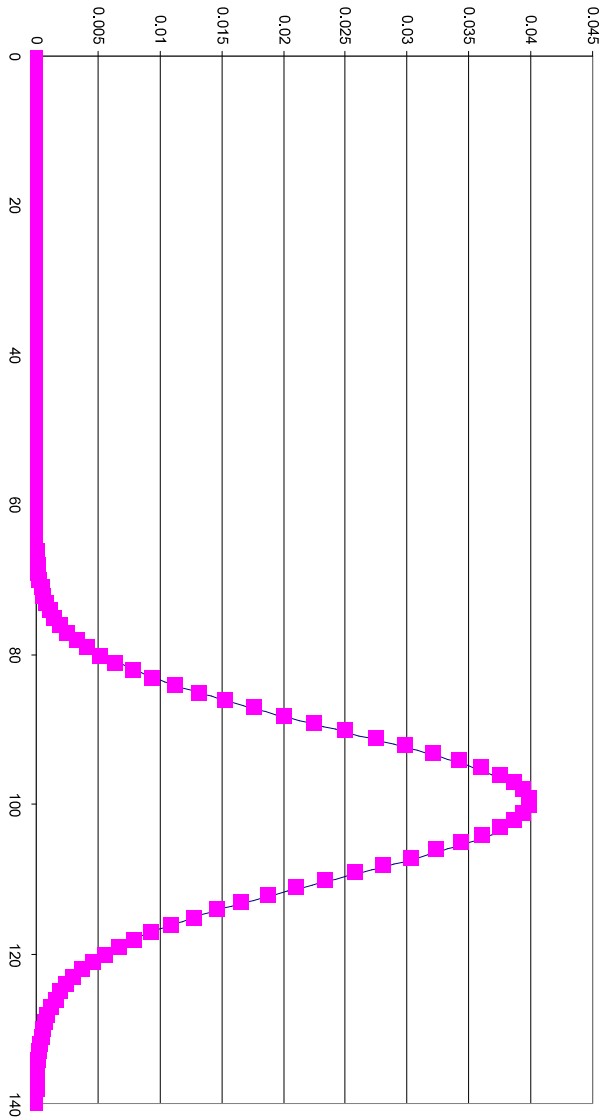


Figure 12: Poisson Distribution (lambda=100)

Central Limit Theorem: Examples

Normal Approximation to Binomial Distribution

$X \sim B(3,0.7)$

$\Pr(X \geq 3) = 0.343$

$X \overset{a}{\sim} N(2.1, 0.63)$

$$\Pr(X \geq 3) = \Pr(X \geq 2.5) = \Pr(Z \geq \frac{2.5-2.1}{0.794}) = \Pr(Z \geq 0.504) = 0.309$$

$X \sim B(10,0.7)$

$\Pr(X \geq 7) = 0.65$

$X \overset{a}{\sim} N(7, 2.1)$

$$\Pr(X \geq 7) = \Pr(X \geq 6.5) = \Pr(Z \geq \frac{6.5-7}{1.449}) = \Pr(Z \geq -0.345) = 0.633$$

$X \sim B(20,0.7)$

$\Pr(X \geq 14) = 0.608$

$X \overset{a}{\sim} N(14, 4.2)$

$$\Pr(X \geq 14) = \Pr(X \geq 13.5) = \Pr(Z \geq \frac{13.5-14}{2.049}) = \Pr(Z \geq -0.244) = 0.595$$

$X \sim B(30,0.7)$

$\Pr(X \geq 21) = 0.589$

$X \overset{a}{\sim} N(21, 6.3)$

$$\Pr(X \geq 21) = \Pr(X \geq 20.5) = \Pr(Z \geq \frac{20.5-21}{2.510}) = \Pr(Z \geq -0.199) = 0.579$$

$X \sim B(100,0.7)$

$\Pr(X \geq 70) = 0.549$

$X \overset{a}{\sim} N(70, 21)$

$$\Pr(X \geq 70) = \Pr(X \geq 69.5) = \Pr(Z \geq \frac{69.5-70}{4.5826}) = \Pr(Z \geq -0.109) = 0.544$$

$X \sim B(200,0.7)$

$\Pr(X \geq 140) = 0.534$

$X \overset{a}{\sim} N(140, 42)$

$$\Pr(X \geq 140) = \Pr(X \geq 139.5) = \Pr(Z \geq \frac{139.5-140}{6.4807}) = \Pr(Z \geq -0.077) = 0.532$$

Normal Approximation to Poisson Distribution

$X \sim P(1)$

$\Pr(X \geq 2) = 0.264$

$X \overset{a}{\sim} N(1,1)$

$$\Pr(X \geq 2) = \Pr(X \geq 1.5) = \Pr(Z \geq \frac{1.5-1}{1}) = \Pr(Z \geq 0.50) = 0.309$$

$X \sim P(4)$

$\Pr(X \geq 5) = 0.371$

$X \overset{a}{\sim} N(4,4)$

$$\Pr(X \geq 5) = \Pr(X \geq 4.5) = \Pr(Z \geq \frac{4.5-4}{2}) = \Pr(Z \geq 0.25) = 0.401$$

$X \sim P(10)$

$\Pr(X \geq 11) = 0.417$

$X \overset{a}{\sim} N(10,10)$

$$\Pr(X \geq 11) = \Pr(X \geq 10.5) = \Pr(Z \geq \frac{10.5-10}{3.1623}) = \Pr(Z \geq 0.158) = 0.436$$

$X \sim P(20)$

$\Pr(X \geq 21) = 0.441$

$X \overset{a}{\sim} N(20,20)$

$$\Pr(X \geq 21) = \Pr(X \geq 20.5) = \Pr(Z \geq \frac{20.5-20}{4.472}) = \Pr(Z \geq 0.112) = 0.456$$

$X \sim P(30)$

$\Pr(X \geq 31) = 0.452$

$X \overset{a}{\sim} N(30,30)$

$$\Pr(X \geq 31) = \Pr(X \geq 30.5) = \Pr(Z \geq \frac{30.5-30}{5.477}) = \Pr(Z \geq 0.091) = 0.464$$

$X \sim P(100)$

$\Pr(X \geq 101) = 0.473$

$X \overset{a}{\sim} N(100,100)$

$$\Pr(X \geq 101) = \Pr(X \geq 100.5) = \Pr(Z \geq \frac{100.5-100}{10}) = \Pr(Z \geq 0.050) = 0.480$$

STATISTICAL TECHNIQUES B

Point Estimation

1. Estimators and estimates

An *estimator* of a population parameter is a random variable based on random observations. Whereas an *estimate* is a particular realisation, or an actual value, based on a specific sample of data points.

2. Point estimators

A POINT ESTIMATOR is a function of the sample information and yields a single number. A realisation is called a POINT ESTIMATE,

For example,

$\bar{X}$ is a point estimator of μ and $\bar{x} = 4$ is the point estimate

s_X^2 is a point estimator of σ_X^2 and $s^2 = 0.05$ is the point estimate

$\hat{p}_X$ is a point estimator of p_X and $\hat{p} = 0.54$ is the point estimate

3. Unbiasedness

An estimator, θ, is said to be UNBIASED if:

$E(\hat{\theta}) = \theta$, that is, if the mean of the sampling distribution of $\hat{\theta}$ is centred on θ. The ESTIMATORS, $\bar{X}$, s_X^2 and $\hat{p}_X$ are all unbiased:

3.1 Unbiasedness of $\bar{X}$

$$E(\bar{X}) = E(X_1 + X_2 + \ldots + X_n)/n = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}\left[\mu + \mu + \ldots + \mu\right] = \mu$$

3.2 Unbiasedness of s_X^2

$$s_X^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} = \frac{\sum_{i=1}^{n}\left[(X_i - \mu) - (\mu - \bar{X})\right]^2}{n-1} = \frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2}{n-1}$$

$$E(s_X^2) = E\left[\frac{\sum_{i=1}^{n}(X_i - \mu)^2 - n(\bar{X} - \mu)^2}{n-1}\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2\right]$$

$$E(s_X^2) = \frac{1}{n-1}\left[n\sigma_X^2 - n\left(\frac{\sigma_X^2}{n}\right)\right] = \sigma_X^2$$

3.3 Unbiasedness of $\hat{p}_X$

$$E(\hat{p}_X) = E(X_1 + X_2 + \ldots + X_n)/n = \frac{1}{n}\left[E(X_1) + E(X_2) + \ldots + E(X_n)\right]$$

$$= \frac{1}{n}\left[p_X + p_X + \ldots + p_X\right] = p_X$$

4. Bias

Bias is defined as:

$$Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

5. Efficiency

Consider two alternative estimators of θ, $\hat{\theta}_1$ and $\hat{\theta}_2$, based on the same information, we say that $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$, if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$. One possible measure of this is relative efficiency constructed as: $\frac{V(\hat{\theta}_1)}{V(\hat{\theta}_2)}$.

In general if we are choosing between two unbiased estimators then we choose the estimator with the smaller variance.

For example, consider 3 alternative estimators for the population parameter, μ,

Estimator	E(.)	V(.)
$\bar{X}_1 = X_1$	μ	σ^2
$\bar{X}_2 = \frac{(X_1 + X_2)}{2}$	μ	$\sigma^2/2$
$\bar{X}_3 = \frac{(X_1 + X_2 + X_3 + \ldots X_n)}{n}$	μ	σ^2/n
$\bar{X}_4 = 3$	3	0

Last estimator is biased, but with very small variance.

6. Mean square error

How might one select between an estimator which is unbiased with a large variance and a biased estimator which is biased with a small variance? On solution is to use the Mean Square Error (MSE), which is calculated as:

$$MSE(\hat{\theta}) = E\left[\left(\hat{\theta} - \theta\right)^2\right] = V(\hat{\theta}) + Bias(\hat{\theta})^2$$

This offers a trade-off between the measure of bias and the variance.

STATISTICAL TECHNIQUES B

Hypothesis Testing

1. Introduction

Hypothesis testing involves assessing the validity of some conjecture or hypothesis. Within statistics some hypothesis is made about some unknown population parameter, θ, This is referred to as the maintained or NULL HYPOTHESIS and is denoted as H_0. If the null hypothesis is NOT true, then some alternative is TRUE. The investigator then formulates as ALTERNATIVE HYPOTHESIS (H_1) against which to test the null hypothesis. This alternative hypothesis is invariably a composite hypothesis (encompassing many values of θ). The null hypothesis is always assumed to be true until counter evidence forces us to reject this working hypothesis.

For example,

$H_0 : \theta = \theta_0$ simple null

$H_1 : \theta \neq \theta_0$ composite 2-sided alternative

$H_0 : \theta \leq \theta_0$ composite null $\qquad$ $H_0 : \theta \geq \theta_0$ composite null

$H_1 : \theta > \theta_0$ composite 1-sided alternative $\qquad$ $H_1 : \theta < \theta_0$ composite 1-sided alternative

Now any null hypothesis can be TRUE or FALSE (as the population parameter, θ, is unknown). Based on the sample evidence we are going to draw conclusions about the population parameters.

2. Types of errors

Needless to say one can clearly make errors when testing a particular hypothesis. In particular, there are two types of errors one can make:

Type I error – Rejecting a TRUE H_0

$\Pr(\text{Type I error}) = \alpha = \text{significance level}$

Type II error – Accept a FALSE H_0

$\Pr(\text{Type II error}) = \beta$

$Power = 1 - \beta = \Pr(\text{Correctly rejecting a FALSE } H_0)$ (see Figure 1).

Suppose, we believe that that a random variable X is normally distributed with a mean of zero and a variance of unity, that is, $X \sim N(0,1)$ and we wish to test the hypothesis

$H_0 : \mu = 0$

$H_1 : \mu > 0$

Now if a randomly selected individual had a value of x=1.2. Test the hypothesis that this came from a distribution with a mean of zero.

You proceed by asking the question: What is the probability of observing a number as big as (as small as, for a negative number) the one observed, given $\mu = 0$?

$$\Pr(X \geq 1.2) = \Pr\left(Z \geq \frac{(1.2-0)}{1}\right) = \Pr(Z \geq 1.2) = 0.115$$

so there is an 11.51% chance of observing x>1.2. This probability is known as the p-value, the probability of observing a sample mean as big (or as small) as the one actually observed). However, the question remains:

At what point would you start to question H_0?

The answer depends on the significance level. If you are prepared to only reject H_0 for a p-value of say 0.001 (0.1%), then you really have a low Pr(Type I error) – you must strongly believe in H_0 (naturally conservative). If you are prepared to reject H_0 at say 0.20, then you are prepared to have a high Pr(Type I error) – naturally prepared to overthrow prior beliefs. In statistics the significance level (the probability at which you are prepared to reject H_0) are generally set at $\alpha = 0.01, 0.05, 0.10$, that is, 1%, 5% or 10% and this should be determined before undertaking the test. If we choose to use a significance level of 5%, this implies that you are accepting that 1 time in 20 will incorrectly reject H_0:

Why do we not make Pr(Type I error)≈0.000? Because there is a trade-off between type I and type II errors. So that by choosing a very low type I error probability – that

is, minimising the probability of rejecting a true null, you increase the probability of accepting a false null, compare Figures 1 and 2.Do we regard this as sufficiently rare to reject H_0?

In hypothesis testing, for a given significance level, as we are only interested in the dichotomous decision of either rejecting, or not rejecting, H_0 we do not need to calculate the p-value, but simply calculate test statistic (z=1.2, in the example above) and compare this to a critical value – where the critical value is that value associated with a probability of α (significance level), under the hypothesised distribution.

Table 1: Critical values from a standard normal distribution

a	Pr(X>a)
1.280	0.100
1.645	0.050
1.960	0.025
2.320	0.010
2.575	0.005

In our example as the test statistics of1.2 is less than the critical value of 1.645 (at the 5% significance level) we would not reject H_0.

3. Testing the mean of a normal distribution

Procedure

1. Specify a null hypothesis.
2. Specify an alternative hypothesis.
3. Choose a significance level and corresponding critical region.
4. Calculate the test under the null hypothesis, by calculating how far the sample statistic is from the hypothesised value.
5. Compare the test statistic with the critical value and formulate a decision.

Suppose $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a normal distribution with unknown mean, μ, and known variance, σ^2. Then, we know that $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and by standardising, $Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$. We are interested in testing the hypothesis that the population mean equals μ_0 against an alternative, e.g. $\mu > \mu_0$, the 5-step procedure is:

1. $H_0 : \mu = \mu_0$
2. $H_1 : \mu > \mu_0$. This alternative hypothesis is a 1-sided alternative, implying we reject H_0 only when we observe a sample mean a long way above the hypothesised value, μ_0.
3. We choose some appropriate significance level of α, and find the corresponding critical value from a NORMAL distribution (as distribution of sample mean is normal), denoted z_α - this is the value which occurs with exactly $100\alpha\%$ probability.
4. Under the null hypothesis: $\bar{X} \sim N(\mu_0, \sigma^2 / n)$, hence $\Pr(\bar{X} > \bar{x}) = \Pr\left(Z > \frac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}\right)$. In which case we calculate test statistic as $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$, this essentially measures how far the sample mean is from the hypothesised population mean, μ_0 and scales this distance by the standard error of the sample mean.

5. Then if z is greater than z_α then we have observed an event which occurs with a probability of less than α and should therefore reject H_0. The decision rule is Reject H_0 if $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > z_\alpha$ Do not reject H_0 if $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < z_\alpha$. (Figure 3 shows the appropriate acceptance and rejection regions)

If the alternative hypothesis had been $H_1 : \mu < \mu_0$ ($H_1 : \mu \neq \mu_0$), then the corresponding critical-value would have been $-z_\alpha$ ($-z_{\alpha/2}$ and $z_{\alpha/2}$) and the decision rule is: Reject H_0 if $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} < -z_\alpha$ $\left(z = \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| > z_{\alpha/2} \right)$; Do not reject H_0 if $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} > -z_\alpha$ $\left(z = \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right| < z_{\alpha/2} \right)$ (Figures 4 and 5 shows the appropriate rejection regions).

3.1 Variants of this basic hypothesis test case:

(1) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a distribution which is NOT NORMAL, with an unknown mean, μ, but σ^2 known. If *n*>30 then by a central limit theorem (CLT) we can say that

$\bar{X} \overset{a}{\sim} N(\mu, \sigma^2 / n)$

in which case, the 5 step procedure is as above.

(2) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a distribution which is NOT NORMAL, with σ^2 unknown but with a sample variance of s^2. If *n*>30 then by a CLT the 5-step procedure is as above, except in step 4, we have $z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$. (Question (1) in Examples)

(3) As a specific example of the case above suppose now that $X_1, X_2, \ldots, X_n$ comes from a Bernoulli distribution, that is,

x	0	1
Pr(X=x)	1-π	π

$E(X) = \mu = \pi$ and $V(X) = \sigma^2 = \pi(1-\pi)$, with a unknown mean, $\mu (= \pi)$, and an unknown population variance, $\sigma^2 (= \pi(1-\pi))$. If *n*>30 then by a CLT and under the null hypothesis $H_0 : \mu = \pi_0$, $\bar{X} \overset{a}{\sim} N(\pi_0, \pi_0(1-\pi_0)/n)$ then the test statistic in step 4 is $z = \frac{\bar{x} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$. (Question (2) in Examples).

(4) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of *n* observations from a distribution which is NORMAL, with unknown σ^2 and sample variance equal to s_X^2. Then $\bar{X} \sim N(\mu, \sigma^2/n)$, implying $\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$

and $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$. In which case

$$\frac{\frac{\bar{X} - \mu}{\sigma / \sqrt{n}}}{\sqrt{\frac{(n-1)s_X^2}{\sigma^2 (n-1)}}} \equiv \frac{\bar{X} - \mu}{s_X^2 / \sqrt{n}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}} \sim t_{n-1}$$

where $t_{n-1}^{\alpha/2}$ is the critical value from a t-distribution with n-1 degrees of freedom. A t-distribution looks similar to a normal distribution (symmetric and bell-shaped); however, this distribution has a higher proportion of points in its tails, see Figure 6, which compares the distributions of a t_8 with a N(0,1).

Table 3: Normal compared with t-distributions

	a	Pr(*X*>*a*)	*B*	Pr(*X*>*b*)	*c*	Pr(*X*>*c*)
Normal	2.32	0.010	1.96	0.025	1.645	0.050
t-dist(5)	2.32	0.034	1.96	0.054	1.645	0.080
t-dist(10)	2.32	0.021	1.96	0.039	1.645	0.065
t-dist(15)	2.32	0.017	1.96	0.034	1.645	0.060
t-dist(20)	2.32	0.016	1.96	0.032	1.645	0.058
t-dist(30)	2.32	0.014	1.96	0.030	1.645	0.055
t-dist(50)	2.32	0.012	1.96	0.028	1.645	0.053
t-dist(100)	2.32	0.011	1.96	0.026	1.645	0.052

and so $t_\infty \to N(0,1)$. Many people argue for n>30 the t-distribution can be reasonably well approximated by a standard normal distribution, but this is only an approximation.

In which case the 5 step procedure for testing the null and alternative below is:

1. $H_0 : \mu = \mu_0$.
2. $H_1 : \mu \neq \mu_0$.
3. The critical values come from a t-distribution, denoted $-\mathrm{t}_{\alpha/2,n-1}$ and $\mathrm{t}_{\alpha/2,n-1}$.
4. The test statistic is $t = \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}}$.
5. The decision rule is: Reject H_0 if $t = \left| \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} \right| > t_{\alpha/2,n-1}$; Do not reject H_0 if $t = \left| \dfrac{\bar{x} - \mu_0}{s_x / \sqrt{n}} \right| < t_{\alpha/2,n-1}$. (Question (3) in Examples).

4. Test for the difference in means

4.1 Independent samples:

Assume we have two samples of size, n_1 and n_2, on the random variables X_1 and X_2, respectively. The sample means are $\bar{X}_1$ and $\bar{X}_2$ and we know that: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ and $V(\bar{X}_1 - \bar{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$.

If the underlying distributions are NORMAL and the population variances are KNOWN then:

$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} \right)$. In which case the 5-step procedure is:

1. $H_0 : \mu_1 - \mu_2 = D_0$ and so under H_0 $\bar{X}_1 - \bar{X}_2 \sim N\left(D_0, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2} \right)$
2. $H_1 : \mu_1 - \mu_2 \neq D_0$
3. The critical values are $-\mathrm{z}_{\alpha/2}$ and $\mathrm{z}_{\alpha/2}$.
4. The test statistic is $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$.
5. The decision rule is: Reject H_0 if $z = \left| \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right| > z_{\alpha/2}$; Do not reject H_0 if $z = \left| \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \right| < z_{\alpha/2}$.

4.1.2 Variants of the difference in means hypothesis test

(1) If the underlying distribution of X_1 and X_2 is NOT NORMAL, and the populations variances are KNOWN, providing n_1 and n_2>30, we can apply a CLT and follow the 5-steps above.

(2) If the underlying distribution of X_1 and X_2 is NOT NORMAL and σ_1^2 and σ_2^2 are UNKNOWN, then we follow the 5-step procedure above except that in step 4 we have $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$ (Question (6) in Examples).

(3) As an example of the above if X_1 and X_2 are both Bernoulli distributions then

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

1. $H_0 : \pi_1 - \pi_2 = 0$ and so under H_0 $\bar{X}_1 - \bar{X}_2 \overset{a}{\sim} N\left(0, \dfrac{\pi_0(1-\pi_0)}{n_1} + \dfrac{\pi_0(1-\pi_0)}{n_2}\right)$, where π_0 is the true overall proportion.
2. $H_1 : \pi_1 - \pi_2 \neq 0$
3. The critical values are $-z_{\alpha/2}$ and $z_{\alpha/2}$.
4. The test statistic is $z = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{(p_0(1-p_0)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$. As under H_0 the population proportions are equal the standard error is based on $p_0 = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$.
5. The decision rule is: Reject H_0 if $z = \left|\dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(p_0(1-p_0)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\right| > z_{\alpha/2}$; Do not reject H_0 if $z = \left|\dfrac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{(p_0(1-p_0)\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}\right| < z_{\alpha/2}$ (Question (7) in Examples).

(4) If the underlying distribution of X_1 and X_2 is NORMAL and the population variances are UNKNOWN, but EQUAL, i.e. $\sigma_1^2 = \sigma_2^2$, then

1. $H_0 : \mu_1 - \mu_2 = D_0$
2. $H_1 : \mu_1 - \mu_2 \neq D_0$

3. The critical values are from a t-distribution, denoted $-t_{\alpha/2,(n_1+n_2-2)}$ and $t_{\alpha/2,(n_1+n_2-2)}$.
4. The test statistic is $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$, where $s_x^2 = \dfrac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{(n_1 + n_2 - 2)}$. Note: To use this test it MUST be the case that there is no evidence that the population variances are different.
5. The decision rule is: Reject H_0 if $t = \left|\dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right| > t_{\alpha/2,(n_1+n_2-2)}$; Do not reject H_0 if $t = \left|\dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{s_x\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}\right| < t_{\alpha/2,(n_1+n_2-2)}$. (Question (8) in Examples).

(5) If the underlying distribution of X_1 and X_2 is NORMAL and the population variances are UNKNOWN and UNEQUAL, then

1. $H_0 : \mu_1 - \mu_2 = D_0$
2. $H_1 : \mu_1 - \mu_2 \neq D_0$
3. The critical values are from a t-distribution, denoted $-t_{\alpha/2,DoF}$ and $t_{\alpha/2,DoF}$.
4. The test statistic is $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}}}$, where $DoF = \dfrac{\left[s_{x_1}^2/n_1 + s_{x_2}^2/n_2\right]^2}{\dfrac{\left(s_{x_1}^2/n_1\right)^2}{(n_1-1)} + \dfrac{\left(s_{x_2}^2/n_2\right)^2}{(n_2-1)}}$.
5. The decision rule is: Reject H_0 if $t = \left|t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}}}\right| > t_{\alpha/2,DoF}$; Do not reject H_0 if $t = \left|t = \dfrac{(\bar{x}_1 - \bar{x}_2) - D_0}{\sqrt{\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}}}\right| < t_{\alpha/2,DoF}$. (Question (9) in Examples).

5. Test of the variance of a distribution

To formulate a hypothesis testing on a sample variance, *X*, MUST be normally distributed, $X_i \sim N(\mu,\sigma^2)$. In which case, $w=\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$. In which case the 5-step procedure is:

1. $H_0 : \sigma^2 = \sigma_0^2$
2. $H_1 : \sigma^2 > \sigma_0^2$
3. The critical value is from a χ^2-distribution, denoted $\chi^2_{\alpha,n-1}$
4. $\chi^2 = \frac{(n-1)s_X^2}{\sigma_0^2}$
5. The decision rule is: Reject *H0* if $\chi^2 = \frac{(n-1)s_x^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$; Do not reject *H0* if $\chi^2 = \frac{(n-1)s_x^2}{\sigma_0^2} < \chi^2_{\alpha,n-1}$ (Question (5) in Examples).

6. Testing equality of variances

This must be done before you can use the 4th option from section 4.1.2 (variants of the difference in means hypothesis test)

1. $H_0 : \sigma_1^2 = \sigma_2^2$
2. $H_1 : \sigma_1^2 \neq \sigma_2^2$
3. The critical value is from the F-distribution, denoted $F^{\alpha}_{n_1-1,n_2-1}$.
4. The test statistic as $F=\frac{s_{x_1}^2}{s_{x_2}^2}$, when $s_{x_1}^2 > s_{x_2}^2$ (or $F=\frac{s_{x_2}^2}{s_{x_1}^2}$ when $s_{x_2}^2 > s_{x_1}^2$).
5. The decision rule is: Reject *H0* if $F=\frac{s_{x_1}^2}{s_{x_2}^2} > F^{\alpha}_{n_1-1,n_2-1}$; Do not reject *H0* if $F=\frac{s_{x_1}^2}{s_{x_2}^2} < F^{\alpha}_{n_1-1,n_2-1}$ (Question (8) in Examples).

In Appendix 2, we include a reference table for the different hypothesis test formulas for alternative distributions.

7. Matched pairs

W are interested in formulating tests about $\mu_1 - \mu_2$, when the two experiments (X_1 and X_2) are undertaken with the same sample and are not therefore independent. Given the outcomes of the two trials for the same population we form the difference in the outcomes of the two random variables, that is, $D = X_1 - X_2$. For a given sample,

$d_1 = x_1^1 - x_{21}^1$, $d_2 = x_1^2 - x_2^2$, $d_3 = x_1^3 - x_2^3$, … $d_n = x_1^n - x_2^n$

we then calculate $\bar{d} = \dfrac{\sum_{i=1}^{n} d_i}{n}$ and $s_d^2 = \dfrac{\sum_{i=1}^{n} (d_i - \bar{d})^2}{n-1}$. If the underlying distribution of X_1 and X_2 is normal and the population variances are unknown then:

1. $H_0 : \mu_d = D_0$
2. $H_1 : \mu_d \neq D_0$
3. The critical values from the t-distribution are $-t_{\alpha/2,n-1}$ and $t_{\alpha/2,n-1}$.
4. The test statistic as $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}}$.
5. The decision rule is : Reject H_0 if $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}} > t_{\alpha/2,n-1}$; Do not reject H_0 if $t = \dfrac{\bar{d} - D_0}{s_d / \sqrt{n}} < t_{\alpha/2,n-1}$

8. Calculating the power of a test

Power = Pr(Rejecting $H_0 \mid H_0$ false) (see Figure 1). This is calculated as three steps, for

$H_0 : \mu = \mu_0$

$H_1 : \mu \neq \mu_0$

given $\mu = \mu_1$.

(1) Define the critical value as the point at which you just reject H_0, for example, $\pm z_{\alpha/2}$?

(2) Find the sample mean, $\bar{x}^c$, corresponding to the critical value, that is,

$$\frac{\bar{x}^c - \mu_0}{s/\sqrt{n}} = \pm z_{\alpha/2} \Rightarrow \bar{x}_1^c = z_{\alpha/2}(s/\sqrt{n}) + \mu_0,\ \bar{x}_1^c = z_{\alpha/2}(s/\sqrt{n}) + \mu_0$$

$$\Rightarrow \bar{x}_2^c = -z_{\alpha/2}(s/\sqrt{n}) + \mu_0$$

(3) Calculate $\Pr(\bar{X} > \bar{x}_1^c \mid \mu = \mu_1) + \Pr(\bar{X} < \bar{x}_2^c \mid \mu = \mu_1)$

$$\Rightarrow \Pr(Z > \frac{\bar{x}_1^c - \mu_1}{s/\sqrt{n}}) + \Pr(Z < \frac{\bar{x}_2^c - \mu_1}{s/\sqrt{n}})$$

Figure 7-9 show the effect on power as the true mean, μ_1, moves increasingly further away from the null hypothesis, μ_0. Figure 10 shows that as $\mu_1 \to \mu_0$, then power approaches the significance level, α. (Question (4) in Examples).

9. Testing correlation

1. $H_0 : \rho(X,Y) = 0$
2. $H_1 : \rho(X,Y) \neq 0$
3. The critical values are from the t-distribution, denoted $-t_{\alpha/2,n-2}$ and $t_{\alpha/2,n-2}$.
4. $t = \dfrac{r_{XY}}{\sqrt{(1-r_{XY}^2)/(n-2)}}$
5. The decision rule is :Reject H_0 if $\left|\dfrac{r_{XY}}{\sqrt{(1-r_{XY}^2)/(n-2)}}\right| > t_{\alpha/2,n-2}$; Do not reject H_0 if

   $\left|\dfrac{r_{XY}}{\sqrt{(1-r_{XY}^2)/(n-2)}}\right| < t_{\alpha/2,n-2}$ (Question (1) in Regression Example).

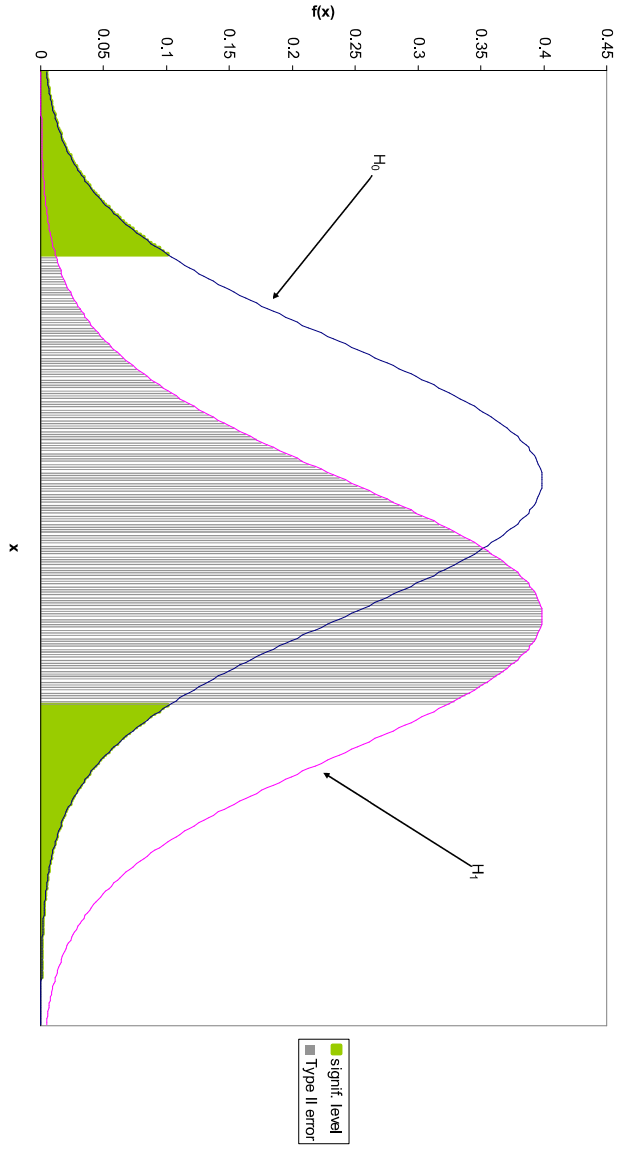


Figure 1: Pr(Type I error), Pr(Type II error) and Power (α=5%)

Figure 3: Significance level and critical region for a one-sided alternative $H_1: \mu > \mu_0$

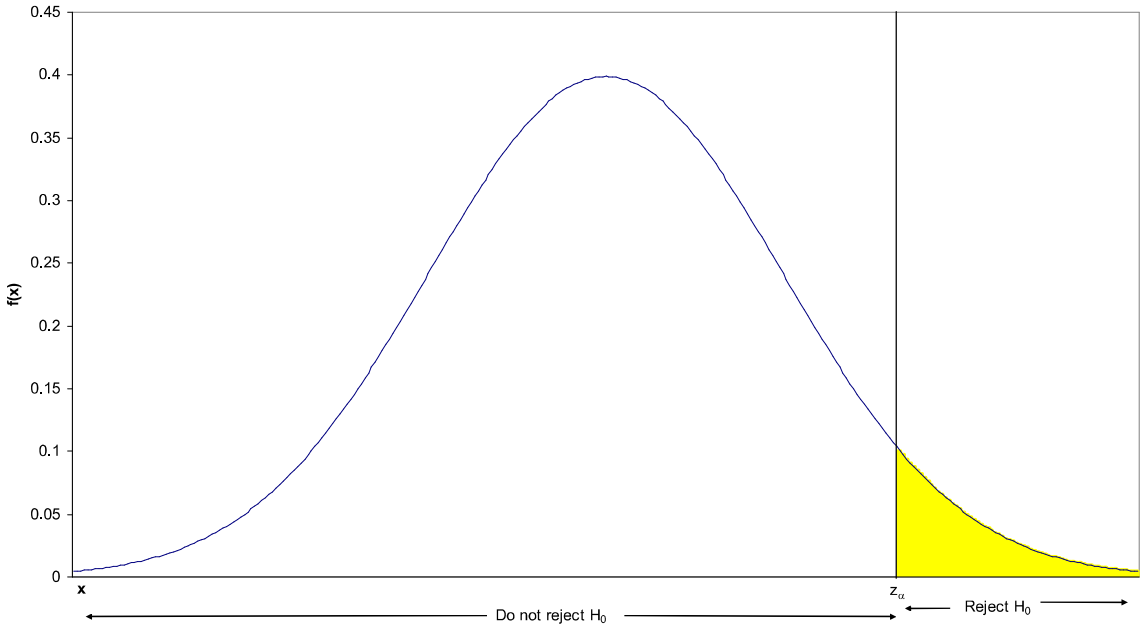


Figure 2: Pr(Type I error), Pr(Type II error) and Power (α=1%)

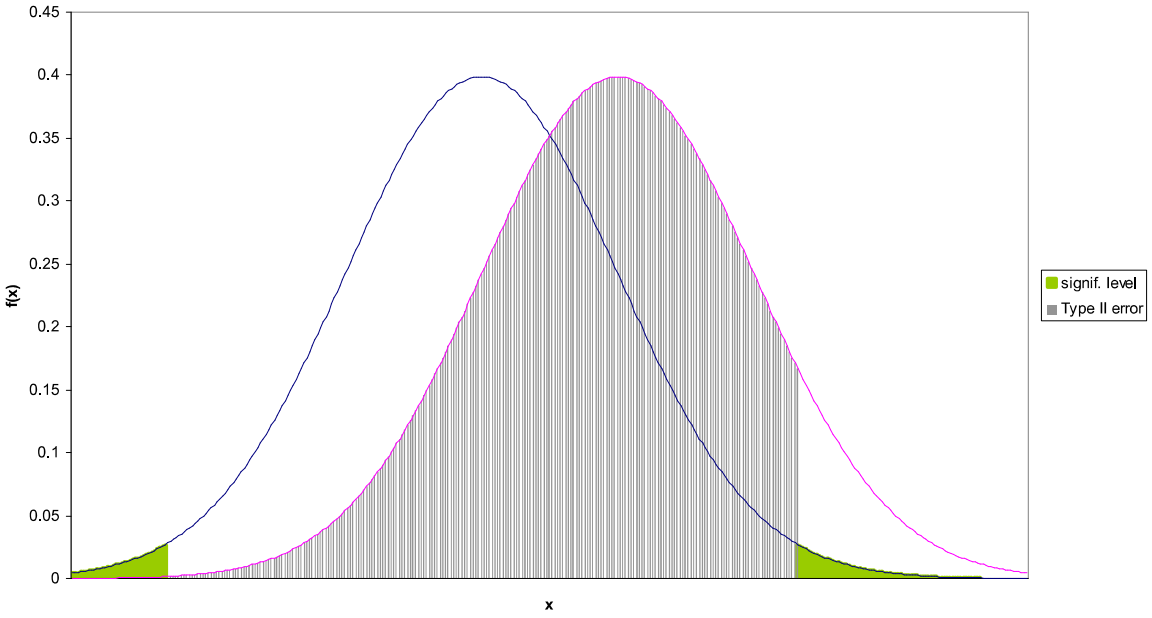


Figure 5: Significance level and critical regions for a two-sided alternative H_1: $\mu=\mu_0$

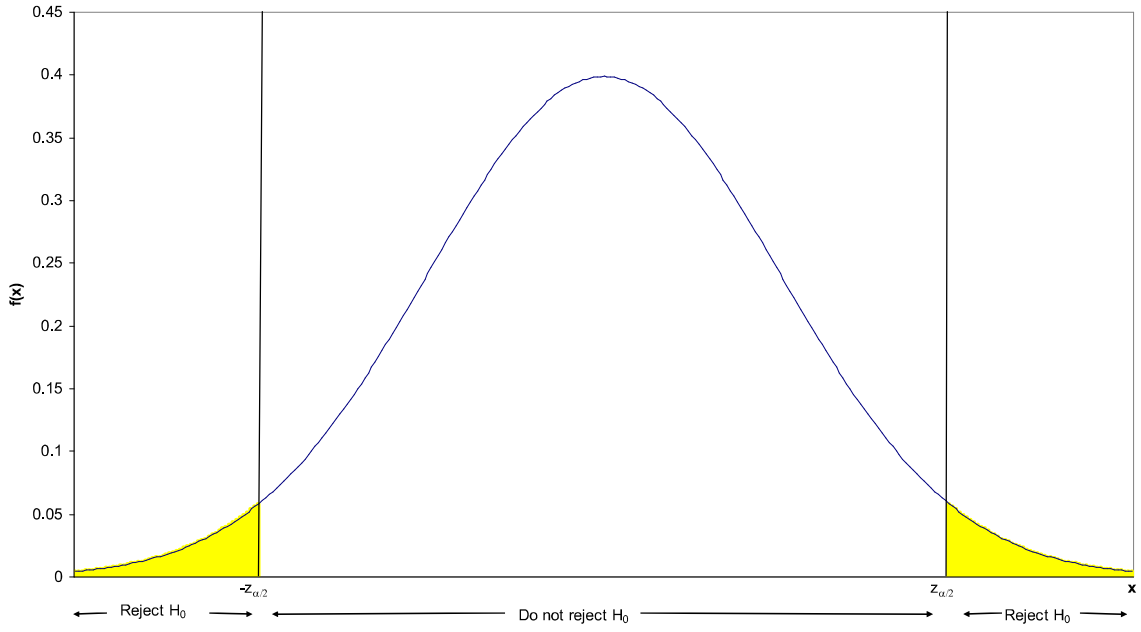


Figure 4: Significance level and critical region for a one-sided alternative H_1 $\mu<\mu_0$

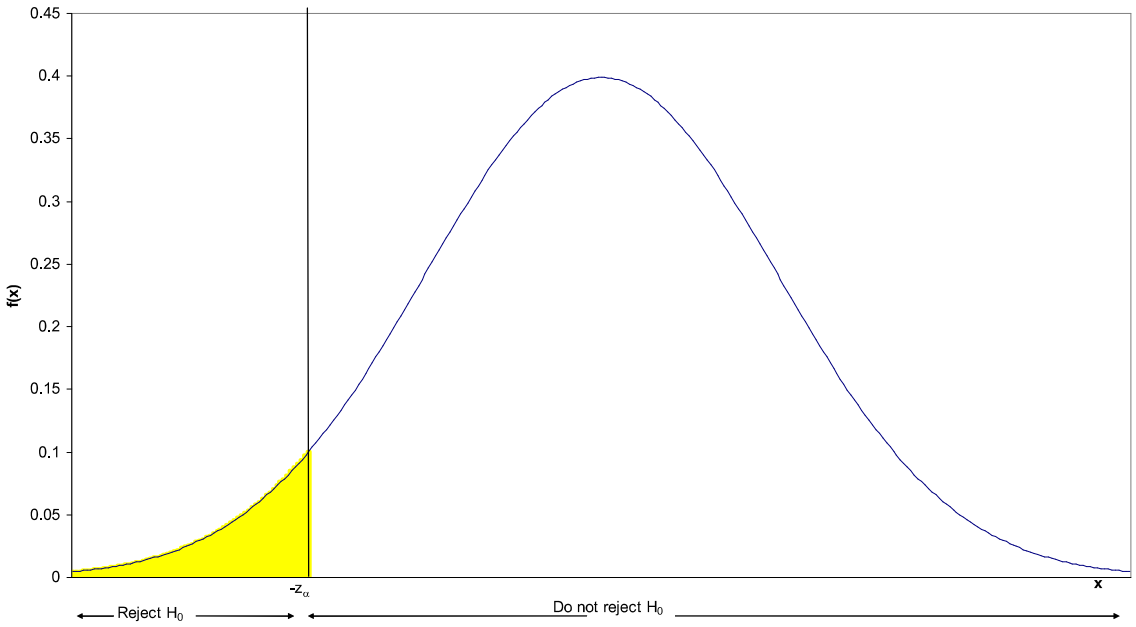


Figure 7: Size and Power: H_0:μ=0 when μ=1.0

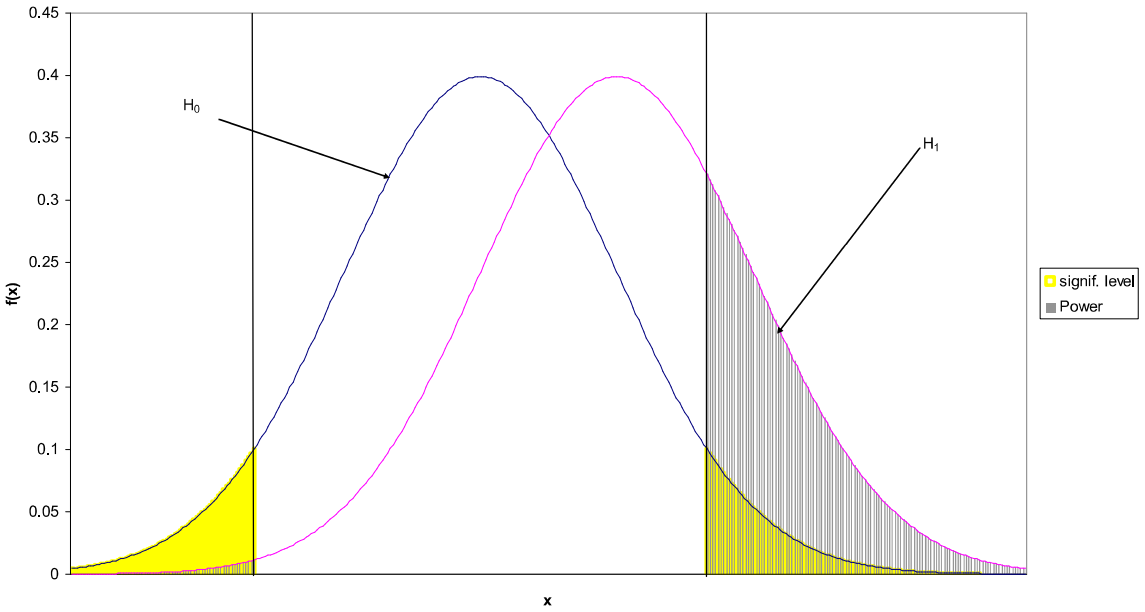


Figure 6: t-distribution vs standard normal distribution

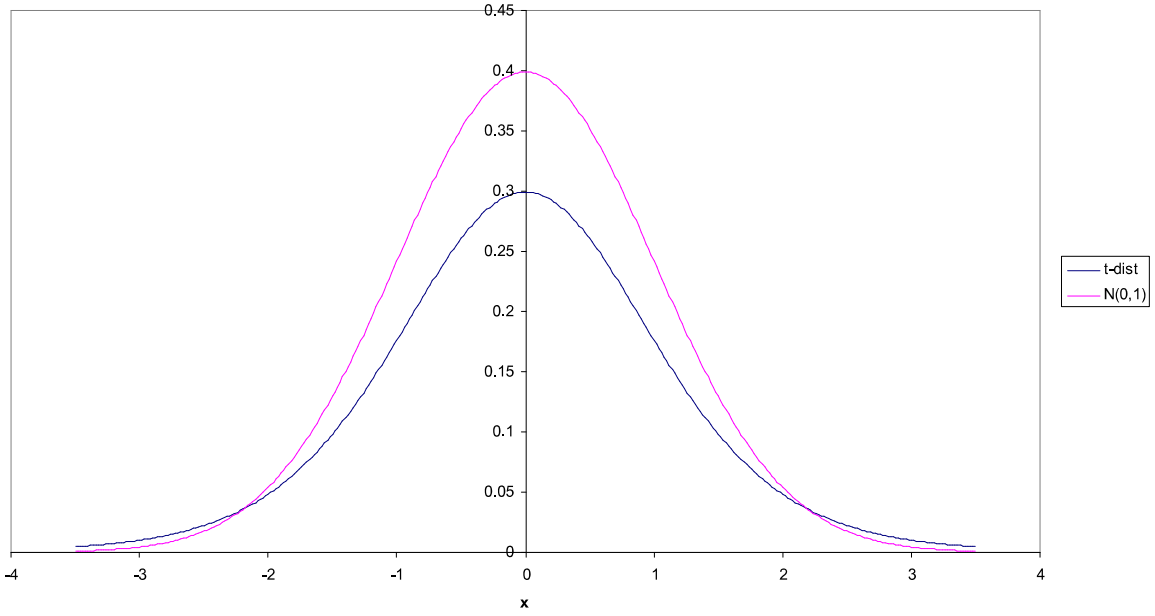


Figure 9: Size and Power: H_0:μ=0 when μ=2.5

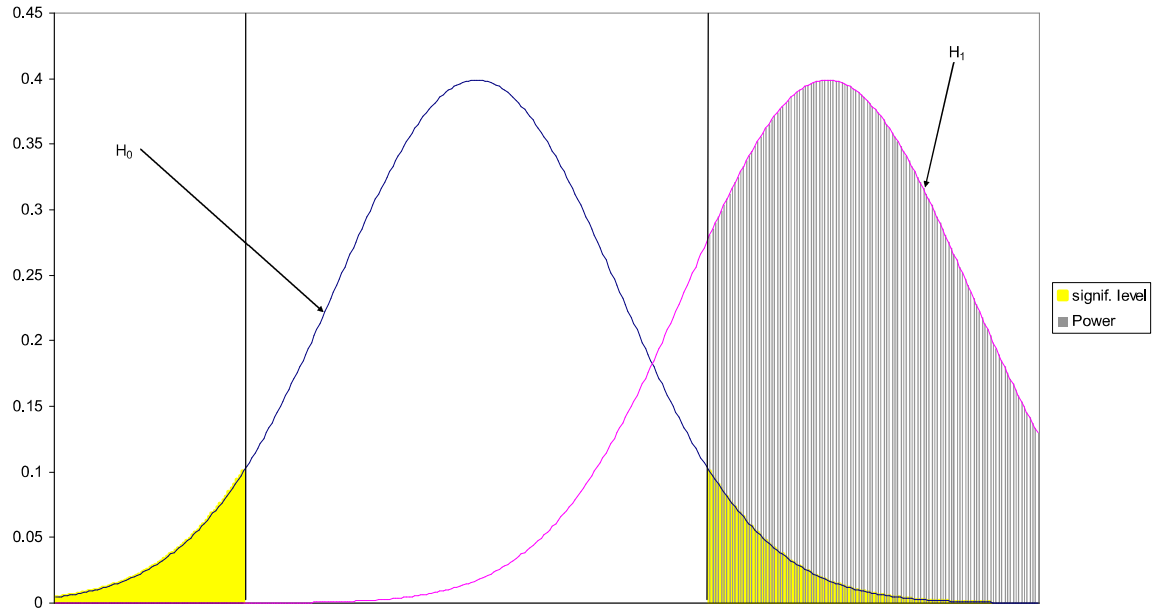
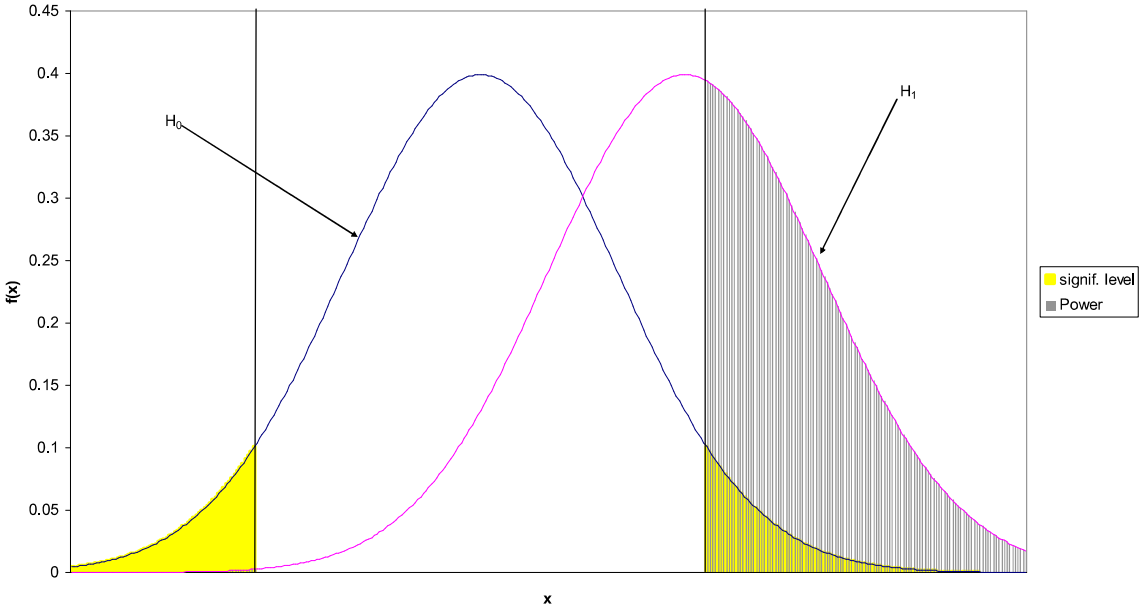


Figure 8: Size and Power: H_0:μ=0 when μ=1.5



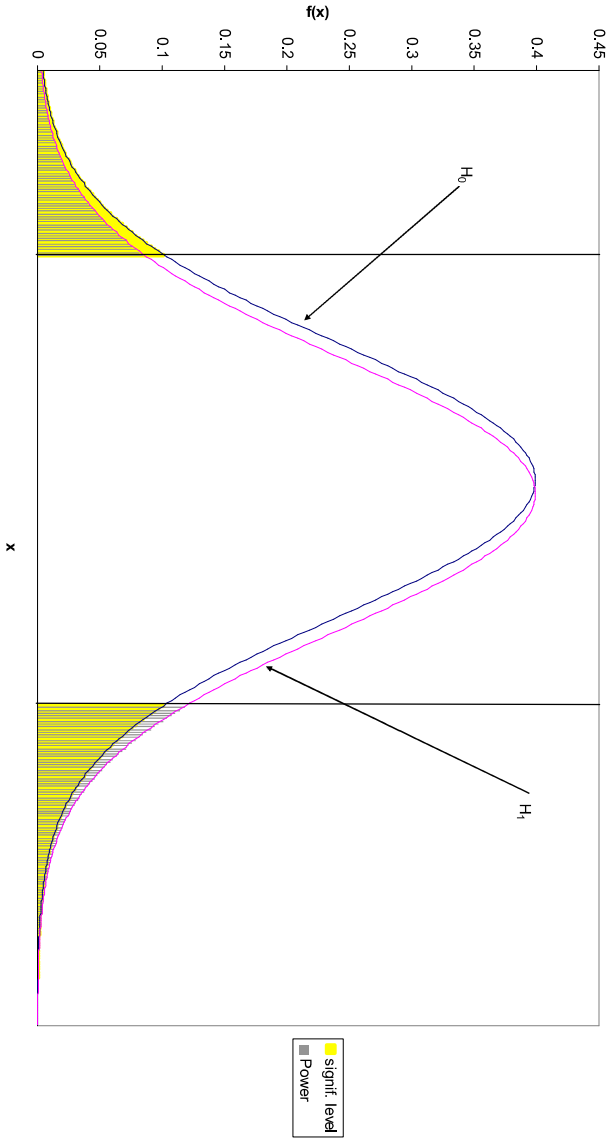


Figure 10: Size and Power: H_0:μ=0 when μ=0.2

Hypothesis Testing – Examples

1. A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was

used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. Test the null hypothesis that the population mean is at least 3 miles per gallon. Find the p-value of this test.

Answer

$$\bar{X} \xrightarrow{a} N(\mu, s_X^2/n) \Rightarrow \frac{(\bar{X}-\mu)}{s_X/\sqrt{n}} \sim N(0,1)$$

$$H_0 : \mu \geq 3$$

$$H_1 : \mu < 3$$

$$z = \frac{(2.6-3.0)}{1.8/\sqrt{100}} = -2.22 \Rightarrow P(Z < -2.22) = 0.013$$

and so we reject the null hypothesis at the 1.3% significance level. The probability of observing a value as low as 2.6 miles (assuming the null hypothesis is true, that is, $\mu \geq 3$) is 1.3%.

2. A mayor in a major city claims that in one particularly depressed neighbourhood, at least 20% of all males between the ages of 18 and 65 are unemployed. A random sample of 120 people from this population contained 20 unemployed people. Test the mayor's claim.

Answer

The underlying series is a Bernoulli trial. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as n→∞.

In particular:

$$\bar{X} \xrightarrow{a} N(\pi, \pi(1-\pi)/n)$$

$$H_0 : \pi \geq 0.2 \qquad z_{0.05} = -1.645$$

$$H_1 : \pi < 0.2 \qquad z_{0.01} = -2.323$$

Now under H_0 we have: $\Rightarrow \bar{X}_1 \xrightarrow{a} N(0.2, 0.2(0.8)/120)$ $z = \dfrac{(0.1666-0.2)}{\sqrt{\frac{(0.2)(0.8)}{120}}} = -0.913$

and so we are unable to reject the null hypothesis at the 5% significance level.

3. A beer manufacturer claims that a new display featuring a life-size picture of a well-known footballer will increase product sales in supermarkets by an average of 50 cases. For a random sample of 20 supermarkets, the average sales increase was 44.3 cases with a sample standard deviation of 12.2 cases. Test at the 5% significance level the null hypothesis that the population mean sales increase is at least 50 cases, stating any assumptions you make.

Answer

This question can only be answered if we are prepared to assume that the increase in product sales will be normally distributed.

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\bar{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

we also know that

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X}-\mu)}{s_X/\sqrt{n}} \sim t_{n-1}$$

$H_0 : \mu \geq 50$ $\qquad$ $t_{19,0.05} = -1.729$

$H_1 : \mu < 50$

$$t = \frac{(44.3-50)}{12.2/\sqrt{20}} = -2.089$$

A t-value of –1.729 occurs with probability of 5%. By using a significance level of 5%, we are saying that an event which occurs with a probability of 5%, or less, is sufficiently rare that we should question the assumption under which the test was undertaken. As we obtained a test statistic of –2.089 this occurs with a probability of less than 5% and we therefore reject H_0.

4. A manufacturer claims that through the use of a fuel additive, automobiles should achieve on average an additional 3 miles per gallon of gas. A random sample of 100 automobiles was used to evaluate this product. The sample mean increase in miles per gallon achieved was 2.6 miles and a sample standard deviation was 1.8 miles per gallon. At the 5% significance level, calculate the power of the test that the population mean is at least 3 miles per gallon.

a. Given that the true mean increase in miles per gallon is 2.0
b. Given that the true mean increase in miles per gallon is 2.2
c. Given that the true mean increase in miles per gallon is 2.4
d. Given that the true mean increase in miles per gallon is 2.6
e. Given that the true mean increase in miles per gallon is 2.8
f. Given that the true mean increase in miles per gallon is 2.9
g. Given that the true mean increase in miles per gallon is 3.0

Answer

$H_0 : \mu \geq 3$ $\qquad$ $H_1 : \mu < 3$ $\qquad$ $z_{0.05} = -1.645$

$$z = \frac{(\bar{x}^c - 3.0)}{1.8/\sqrt{100}} = -1.645 \Rightarrow \bar{x}^c = -1.645(0.18) + 3 \Rightarrow \bar{x}^c = 2.7039$$

Power=Pr(Reject H_0|H_0 false)=Pr(Reject $\mu \geq 3 | \mu = 2.0$)

(a) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.0) = \Pr(Z < \frac{2.7039-2.0}{0.18}) = \Pr(Z < 3.91) = 1.00$

(b) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.2) = \Pr(Z < \frac{2.7039-2.2}{0.18}) = \Pr(Z < 2.80) = 0.997$

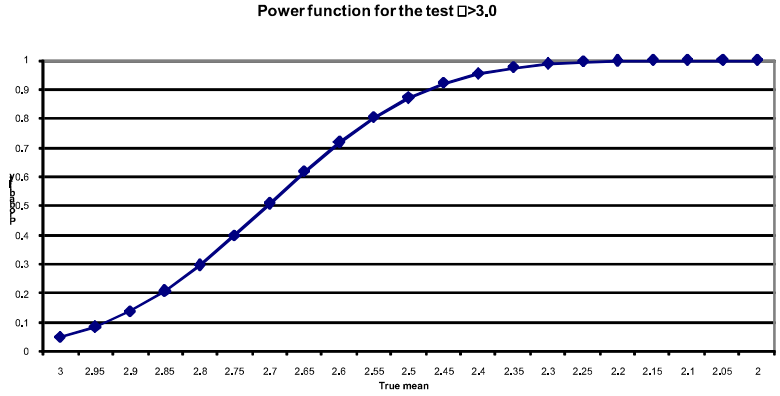
(c) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.4) = \Pr(Z < \frac{2.7039-2.4}{0.18}) = \Pr(Z < 1.69) = 0.954$

(d) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.6) = \Pr(Z < \frac{2.7039-2.6}{0.18}) = \Pr(Z < 0.58) = 0.718$

(e) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.8) = \Pr(Z < \frac{2.7039-2.8}{0.18}) = \Pr(Z < -0.53) = 0.297$

(f) $\Pr(\bar{X} < 2.7039 \mid \mu = 2.9) = \Pr(Z < \frac{2.7039-2.9}{0.18}) = \Pr(Z < -1.09) = 0.138$

(g) $\Pr(\bar{X} < 2.7039 \mid \mu = 3.0) = \Pr(Z < \frac{2.7039-3.0}{0.18}) = \Pr(Z < -1.645) = 0.050$



5. A company produces electric devices operated by a thermostat control. The standard deviation of the temperature at which these controls actually operate should not exceed 2°F. For a random sample of 20 of these controls, the sample standard deviation of operating temperatures was 2.36°F. Stating any assumptions you need to make, test at the 5% significance level the null hypothesis that the population standard deviation is 2°F against the alternative that it is bigger.

Answer

Assuming the underlying distribution of the temperature at which the controls operate is normal, then we have:

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$$

$H_0 : \sigma^2 = 4$ $\qquad\qquad$ $\chi^2_{19,0.05} = 30.14$

$H_1 : \sigma^2 > 4$

$$\chi^2 = \frac{19(2.36)^2}{4} = 26.46$$

A chi-squared value of 30.14 occurs with probability of 5%. As we obtained a test statistic of 26.46 this occurs with a probability of more than 5% and we therefore we are unable to reject H_0.

6. The MATWES procedure was designed to measure attitudes toward women as managers. High scores indicate negative attitudes and low scores indicate positive attitudes. Independent random samples were taken of 151 male MBA students and 108 female MBA students. For the former group, the sample mean and standard deviation MATWES scores were 75.8 and 19.3, while the corresponding figures for the latter group were 71.5 and 12.2. Test the hypothesis that the two population means are equal against the alternative that the true mean MATWES score is higher for male than for female MBA students.

Answer

The underlying series has an unknown distribution, but the sample means will both be normally distributed

$\bar{X}_1 \sim N(\mu_1, s^2_{X_1}/n_1)$ and $\bar{X}_2 \sim N(\mu_2, s^2_{X_2}/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \left[\frac{s^2_{X_1}}{n_1} + \frac{s^2_{X_2}}{n_2}\right]\right)$$

$H_0 : \mu_1 - \mu_2 = 0$ $\qquad\qquad$ $z_{0.05} = 1.645$

$H_1 : \mu_1 - \mu_2 > 0$ $\qquad\qquad$ $z_{0.01} = 2.323$

$$z = \frac{(75.8 - 71.5) - 0.0}{\sqrt{\frac{19.3^2}{151} + \frac{12.2^2}{108}}} = 2.193$$

and so we are able to reject the null hypothesis at the 5% significance level, but not at the 1% significance level..

7. Of a random sample of 381 investment grade corporate bonds, 191 had sinking funds. Of an independent random sample of 166 speculative-grade corporate bonds, 98 had sinking funds. Test a 2-sided alternative against the null hypothesis that the two population proportions are equal.

Answer

The underlying series follows a Bernoulli trial and hence the distribution cannot be normal, however, the distribution of the sample mean (sample proportion) will be approximately normally distributed as n→∞. In particular:

$\bar{X}_1 \overset{a}{\to} N(\pi_1, \pi_1(1-\pi_1)/n_1)$ and $\bar{X}_2 \overset{a}{\to} N(\pi_2, \pi_2(1-\pi_2)/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\pi_1 - \pi_2, \left[\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]\right)$$

$H_0 : \pi_1 - \pi_2 = 0$ $\qquad\qquad$ $z_{0.025} = \pm 1.96$

$H_1 : \pi_1 - \pi_2 \neq 0$

$$\text{Under } H_0 \Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(0, \left[\frac{p_0(1-p_0)}{n_1} + \frac{p_0(1-p_0)}{n_2}\right]\right),$$

where $p_0 = \dfrac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$

$$z = \frac{(0.5013 - 0.5904) - 0.0}{\sqrt{\frac{0.5283(0.4717)}{381} + \frac{0.5283(0.4717)}{166}}} = -1.919$$

and so we are just unable to reject the null hypothesis at the 5% significance level.

8. A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2107. For books with a plain cover average first year sales were 8167, with a standard deviation of 1681. Assuming the population distributions are normal, test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Answer

The underlying distribution is normal, i.e.

$\bar{X}_1 \sim N(\mu_1, \sigma^2/n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2/n_2)$

if in addition we assume the variances are equal, we have:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

we also know that

$$\frac{(n_1+n_2-2)s_X^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{s_X\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $s_x^2 = \frac{(n_1-1)s_{x_1}^2+(n_2-1)s_{x_2}^2}{n_1+n_2-2}$.

We must first test the assumption the variances are equal:

$H_0 : \sigma_1^2 = \sigma_2^2$ $\qquad F_{9,9}^{0.05} = 3.18$

$H_1 : \sigma_1^2 > \sigma_2^2$

$$F = \frac{2107^2}{1681^2} = 1.571$$

An F value of 3.18 occurs with a probability of 5%. As we obtained a test statistic of 1.571 this occurs with a probability of more than 5% and we therefore we are unable to reject H_0. In which case our assumption is reasonable:

$$s^2 = \frac{(9)2107^2+(9)1681^2}{18} \Rightarrow s = 1905.94$$

$H_0 : \mu_1 - \mu_2 = 0$ $\qquad t_{18}^{0.05} = 1.734$

$H_1 : \mu_1 - \mu_2 > 0$

$$t = \frac{(9254-8167)-0}{1905.94\sqrt{\frac{1}{10}+\frac{1}{10}}} = 1.275$$

and so we are unable to reject the null hypothesis at the 5% significance level.

9. A publisher is interested in the effect on sales of university textbooks of cover design. The publisher is planning to bring out 20 texts in the area of business and randomly chooses 10 text to have expensive designs, with the remaining texts having a plain cover. For those with expensive cover designs average sales in the first year were 9254 with a sample standard deviation of 2107. For books with a plain cover average first year sales were 8167, with a standard deviation of 1081. Assuming the population distributions are normal, test the hypothesis that the population means are equal against the alternative that the true mean is higher for books with an expensive cover.

Answer

The underlying distribution is normal and we have:

$\bar{X}_1 \sim N(\mu_1, \sigma^2/n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma^2/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1-\mu_2, \left[\frac{\sigma_{x_1}^2}{n_1}+\frac{\sigma_{x_2}^2}{n_2}\right]\right).$$

The greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that the distribution of the standardized statistic will be a t-distribution , but the DoF of this distribution depends on whether it is reasonable to assume the sample variance are equal. We must first test the assumption the variances are equal:

$H_0 : \sigma_1^2 = \sigma_2^2$ $\qquad F_{9,9}^{0.05} = 3.18$

$H_1 : \sigma_1^2 > \sigma_2^2$

$$F = \frac{2107^2}{1081^2} = 3.799$$

An F value of 3.18 occurs with probability of 5%. As we obtained a test statistic of 3.799 this occurs with a probability of less than 5% and we therefore reject H_0. Consequently,

$$\frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\left(\frac{s_{x_1}^2}{n_1}+\frac{s_{x_2}^2}{n_2}\right)}} \sim t_{DoF} \quad \text{where} \quad DoF = \frac{\left[\frac{2107^2}{10}+\frac{1081^2}{10}\right]^2}{\left[\left(\frac{2107^2}{10}\right)^2/9+\left(\frac{1081^2}{10}\right)^2/9\right]} = 13.43.$$

$H_0 : \mu_1 - \mu_2 = 0$ $\qquad t_{13}^{0.05} = 1.771$

$H_1 : \mu_1 - \mu_2 > 0$

$$t = \frac{(9254-8167)-0}{\sqrt{\frac{2107^2}{10}+\frac{1081^2}{10}}} = 1.452$$

and so we are unable to reject the null hypothesis at the 5% significance level.

Hypothesis Testing Sheet

Hypothesis Testing: Test Statistics for Tests of Means

One Population

Sample Hypothesis Distrib of X_i σ^2 Known

σ^2 Not Known

Large/Small $H_0 : \mu = \mu_0$ Normal $z = \dfrac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}$

$$t = \frac{(\bar{x} - \mu_0)}{s_x / \sqrt{n}}$$

Large $H_0 : \mu = \mu_0$ Non-Normal $z = \dfrac{(\bar{x} - \mu_0)}{\sigma / \sqrt{n}}$

$$z = \frac{(\bar{x} - \mu_0)}{s_x / \sqrt{n}}$$

Small $H_0 : \mu = \mu_0$ Non-Normal ?

?

Two Populations

Large/Small $H_0 : \mu_1 - \mu_2 = \delta$ Normal $z = \dfrac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2\}}}$

Large $H_0 : \mu_1 - \mu_2 = \delta$ Non-Normal $z = \dfrac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(\sigma_1^2 / n_1 + \sigma_2^2 / n_2)}}$

$$z = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2)}}$$

Large/Small $H_0 : \mu_1 - \mu_2 = \delta$ Normal ---

$$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{s_0 \sqrt{(1/n_1 + 1/n_2)}}$$

where $s_0^2 = \{(n_1 - 1)s_{x_1}^2 + (n_2 - 1)s_{x_2}^2\} / (n_1 + n_2 - 2)$

Large/Small $H_0 : \mu_1 - \mu_2 = \delta$ Normal ---

$$t = \frac{(\bar{x}_1 - \bar{x}_2 - \delta)}{\sqrt{(s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2)}}$$

where $DoF = \dfrac{\left[s_{x_1}^2 / n_1 + s_{x_2}^2 / n_2\right]^2}{\left(s_{x_1}^2 / n_1\right)^2 / (n_1 - 1) + \left(s_{x_2}^2 / n_2\right)^2 / (n_2 - 1)}$

Small Sample $H_0 : \mu_1 - \mu_2 = \delta$ Non-Normal ?

?

Tests on Proportions

One Population

	Hypothesis	Distrib of X_i	Test
Large sample	$H_0 : \pi = \pi_0$	Non-normal	

$$z = \frac{\bar{x} - \pi_0}{\sqrt{\pi_0(1 - \pi_0) / n}}$$

Small sample $H_0 : \pi = \pi_0$ Non-normal ?

Two Populations

Large sample $H_0 : \pi_1 - \pi_2 = 0$ Non-normal

$$z = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{p_0(1 - p_0)(1/n_1 + 1/n_2)}} \text{ where } p_0 = \frac{\bar{x}_1 n_1 + \bar{x}_2 n_2}{n_1 + n_2}$$

Small sample $H_0 : \pi_1 - \pi_2 = 0$ Non-normal ?

Tests on variances

One Population

Large/Small $H_0 : \sigma^2 = \sigma_o^2$ Normal

$$u = \frac{(n-1)s_x^2}{\sigma_0^2} \sim \chi_{n-1}^2$$

Two Populations

Large/Small $H_0 : \sigma_1^2 = \sigma_2^2$ Normal

$$F = s_{x_1}^2 / s_{x_2}^2 \sim F_{(n_1 - 1, n_2 - 1)}$$

STATISTICAL TECHNIQUES B

Confidence Intervals

1. Introduction

Confidence intervals give a range of likely values for the TRUE (but unknown) population parameter, together with a measure of the confidence (or likelihood) that the range contains the true value. For some unknown population parameter, θ, based on sample data we find two values a and b, such that,

$\Pr\{a < \theta < b\} = 1-\alpha \quad for \quad 0 < \alpha < 1$

then we can say with $100(1-\alpha)\%$ confidence that θ lies in the range a to b. That means in repeated samples, $100(1-\alpha)\%$ of the time, θ would lie within intervals calculated this way.

Consider a N(0,1) distribution we know that

$\Pr\{-1.645 < Z < 1.645\} = 0.90$.

We take symmetric points around zero as this minimises the range for the interval (compare Figures 1 and 2).

Table 1: Range for a N(0,1) for a 90% interval

p_a	A	p_b	b	range
0.05	-1.645	0.05	1.645	3.29
0.04	-1.74	0.06	1.56	3.30
0.01	-2.32	0.09	1.34	3.66

Table 2: Critical values for a N(0,1)

CI(%)	Lower limit	Upper limit
90	-1.645	1.645
95	-1.96	1.96
99	-2.575	2.575

A diagrammatic illustration of this is provided by figure 3. To be more confident of a statement or value our degree of uncertainty or range of possible values has to increase.

2. Confidence Interval for mean of a distribution

Let $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a normal distribution with unknown mean, μ, and known variance, σ^2. Then, we know that

$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and this implies (by standardising) that, $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.

As,

$\Pr(-1.645 < Z < 1.645) = 0.9$.

Therefore,

$$\Pr(-1.645 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.645) = 0.9$$

$$\Pr(-1.645(\sigma/\sqrt{n}) < \bar{X} - \mu < 1.645(\sigma/\sqrt{n})) = 0.9$$

$$\Pr(-\bar{X} - 1.645(\sigma/\sqrt{n}) < -\mu < -\bar{X} + 1.645(\sigma/\sqrt{n})) = 0.9$$

$$\Pr(\bar{X} - \underbrace{1.645(\sigma/\sqrt{n})}_{D} < \mu < \bar{X} + \underbrace{1.645(\sigma/\sqrt{n})}_{D}) = 0.9.$$

such that we expect the interval $\{\bar{X}-D, \bar{X}+D\}$ to contain μ on 90% of occasions.

However, after taking a sample and calculating the actual sample mean, $\bar{x}$, we can say that we are 90% confident that the interval $\{\bar{x}-D, \bar{x}+D\}$ contain the (unknown) population mean, μ.

The width of the confidence interval 2D depends three factors:

(i) The level of confidence, that is, 90%, 95% or 99%. The interval for 99% being far wider than that for 90%.

(ii) The variability in the underlying distribution, σ, the greater the variability the wider the interval.

(iii) The number of observation in the sample, *n*, as this effects the standard error of the sample mean, $\sigma/\sqrt{n}$, as n increases the standard error of the sample mean falls and hence the interval narrows. (Question (1) in Examples)

2.1 Variants of this basic confidence interval case:

(1) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a distribution which is NOT NORMAL, with an unknown mean, μ, but the population variance is KNOWN, σ^2. If n is large enough (n>30) then by appealing to the central limit theorem (CLT) we can say that

$\bar{X} \overset{a}{\sim} N(\mu, \sigma^2/n)$

in which case, the $100(1-\alpha)\%$ confidence interval is still written as:

$\bar{x} \pm z_{\alpha/2}(\sigma/\sqrt{n})$, where $z_{\alpha/2}$ is the critical value from a N(0,1).

(2) In the previous case, if the σ^2 is UNKNOWN, then if n>30 $\bar{X} \overset{a}{\sim} N(\mu, s^2/n)$ the confidence interval is written as: $\bar{x} \pm z_{\alpha/2}(s_x/\sqrt{n})$, where $z_{\alpha/2}$ is the critical value from a N(0,1) (Question (2) in examples).

(3) A specific example of the previous case is when $X_1, X_2, \ldots, X_n$ denotes a random sample from a Bernoulli (NOT NORMAL) distribution, that is,

X	0	1
Pr(X)	1-p	p

E(X)= p and V(X)= p(1-p)

with a unknown mean, μ(=p), and an unknown population variance, σ^2(=p(1-p)). Then if n>30 by a CLT $\bar{X} \overset{a}{\sim} N(p, p(1-p)/n)$ in which case, the $100(1-\alpha)\%$ confidence interval is written as: $\hat{p} \pm z_{\alpha/2}(\sqrt{\hat{p}(1-\hat{p})/n})$ (Question (3) in Examples).

(4) Suppose now that $X_1, X_2, \ldots, X_n$ denote a random sample of n observations from a distribution which is NORMAL, with a unknown mean, μ, and an UNKNOWN σ^2. Then: $\dfrac{\bar{X}-\mu}{s_X/\sqrt{n}} \sim t_{n-1}$ in which case, the $100(1-\alpha)\%$ confidence interval is written as:

$\bar{x} \pm t_{n-1}^{\alpha/2}(s_x/\sqrt{n})$, where $t_{n-1}^{\alpha/2}$ is the critical value from a t-distribution with n-1 degrees of freedom. (Question (4) in Examples).

3. Confidence Interval for the difference in means

3.1 Independent samples:

Assume we have two ***independent*** samples of size, n_1 and n_2, on X_1 and X_2, respectively. The sample means are $\bar{X}_1$ and $\bar{X}_2$: $E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2$ and $V(\bar{X}_1 - \bar{X}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$. If the underlying distribution of X_1 and X_2 are normal and the population variances are KNOWN, then: $\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$ and the confidence interval for $\mu_1 - \mu_2$ is: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right]}$

3.1.2 Variants on the CI for the difference in means (independent samples)

(1) If the underlying distributions are NOT NORMAL and the population variances are KNOWN, providing $n_1 > 30$ and $n_2 > 30$, then from a CLT the CI is: $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right]}$.

(2) If the underlying distributions is NOT NORMAL and the population variances are UNKNOWN, providing $n_1 > 30$ and $n_2 > 30$, then from a CLT the CI is $(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\left[\dfrac{s_{x_1}^2}{n_1} + \dfrac{s_{x_2}^2}{n_2}\right]}$ (Question (6) in Examples)

(3) Difference in sample proportions, applying CLT we get CI:

$(\hat{p}_1 - \hat{p}_2) \pm z_{\alpha/2}\sqrt{\left[\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]}$ (Question (7) in Examples)

(4) If the underlying distribution are NORMAL, and the population variances are UNKNOWN (and EQUAL), the CI is: $(\bar{x}_1 - \bar{x}_2) \pm t_{n_1+n_2-2,\alpha/2}\sqrt{\left[\dfrac{s_x^2}{n_1} + \dfrac{s_x^2}{n_2}\right]}$, where,

$s_x^2 = \dfrac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n_1+n_2-2}$ (Question (8) in Examples).

(5) If the underlying distribution are NORMAL, and the population variances are UNKNOWN (and UNEQUAL), the CI is: $(\bar{x}_1 - \bar{x}_2) \pm t_{DoF,\alpha/2}\sqrt{\left[\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right]}$, where,

$$DoF = \frac{\left[s_{x_1}^2/n_1 + s_{x_2}^2/n_2\right]^2}{\left(s_{x_1}^2/n_1\right)^2/(n_1-1) + \left(s_{x_2}^2/n_2\right)^2/(n_2-1)}$$ (Question (9) in Examples).

Appendix 2 is a reference table for the different confidence interval formulas.

3.2 Matched pairs

Again we are interested in formulating the confidence interval for $\mu_1 - \mu_2$, however, in this case, the two experiments are with the same sample of individuals and cannot therefore be independent. Given the outcomes of the two trials for the same population we form the difference in the outcomes of the two random variables, that is, $D = X_1 - X_2$. For a given sample,

$d_1 = x_1^1 - x_{21}^1$, $d_2 = x_1^2 - x_2^2$, $d_3 = x_1^3 - x_2^3$, … $d_n = x_1^n - x_2^n$

we calculate $\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$ and $s_d^2 = \frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}$. If the underlying distributions of X_1 and X_2 are normal, the $100(1-\alpha)\%$ confidence interval is written as

$\bar{d} \pm t_{n-1}^{\alpha/2}(s_d/\sqrt{n})$.

4. Confidence Interval for the variance of a distribution

Similarly to producing confidence intervals for the population mean, we wish to produce an interval estimate of the population variance on the basis of a sample of data. To formulate a confidence interval the random variable, *X*, MUST be normally distributed. In which case, $w = \frac{(n-1)s_X^2}{\sigma^2} \sim \chi_{n-1}^2$.

Now, $\Pr\left\{\chi_{n-1,0.95}^2 < w < \chi_{n-1,0.05}^2\right\} = 0.90$. Note that for a χ^2 distribution, a symmetric confidence interval does not necessarily minimise the range – in particular, a smaller range will be obtained by having only 1% in the left tail and 9% in the right tail for a 90% confidence interval (see Figure 4). Substituting for *w* and rearranging gives:

$\Pr\left\{\chi_{n-1,0.95}^2 < \frac{(n-1)s_X^2}{\sigma^2} < \chi_{n-1,0.05}^2\right\} = 0.90$. Rearranging, we get:

$$\Pr\left\{\frac{\chi_{n-1,0.95}^2}{(n-1)s_X^2} < \frac{1}{\sigma^2} < \frac{\chi_{n-1,0.05}^2}{(n-1)s_X^2}\right\} \Rightarrow \Pr\left\{\frac{(n-1)s_X^2}{\chi_{n-1,0.05}^2} < \sigma^2 < \frac{(n-1)s_X^2}{\chi_{n-1,0.95}^2}\right\} = 0.90$$

Figure 5 compares the distribution of a χ_4^2 to that of a χ_8^2 (Question (5) in Examples).

Figure 2: Alternate 90% confidence interval for N(0,1)

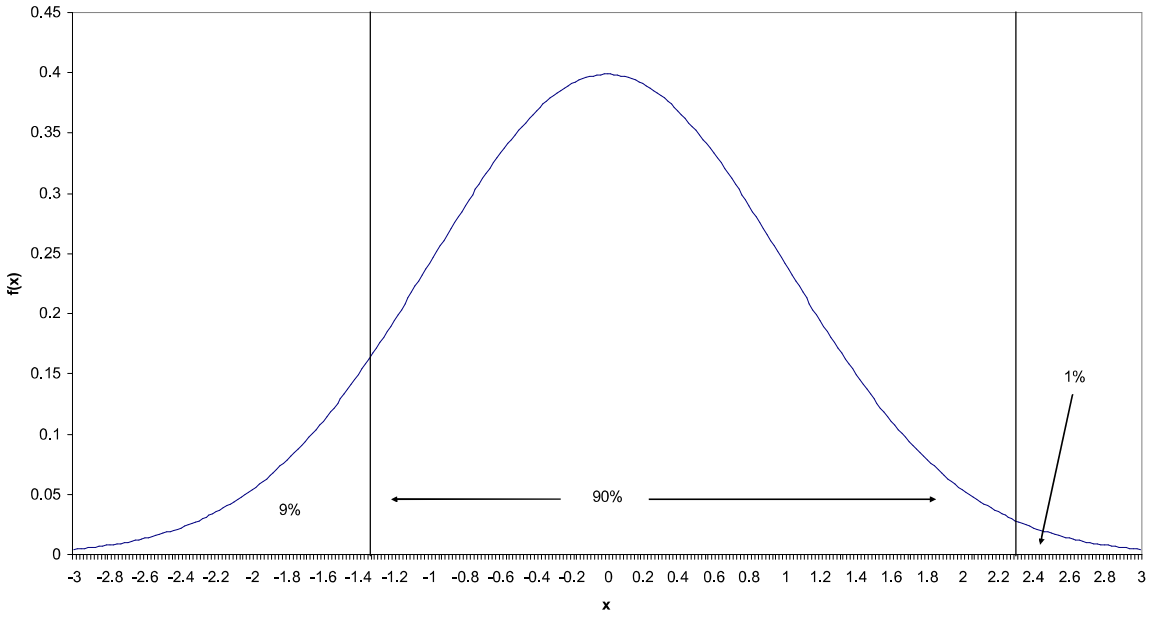


Figure 1: 90% Confidence interval for N(0,1) (symmetric)

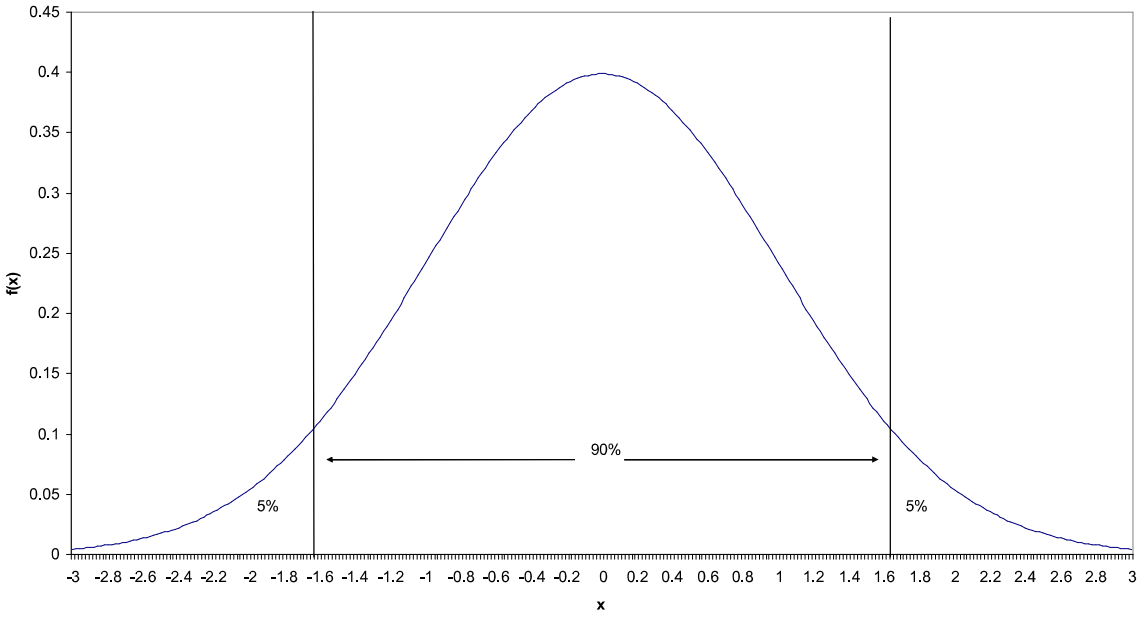


Figure 4: Confidence intervals for Chi-squared

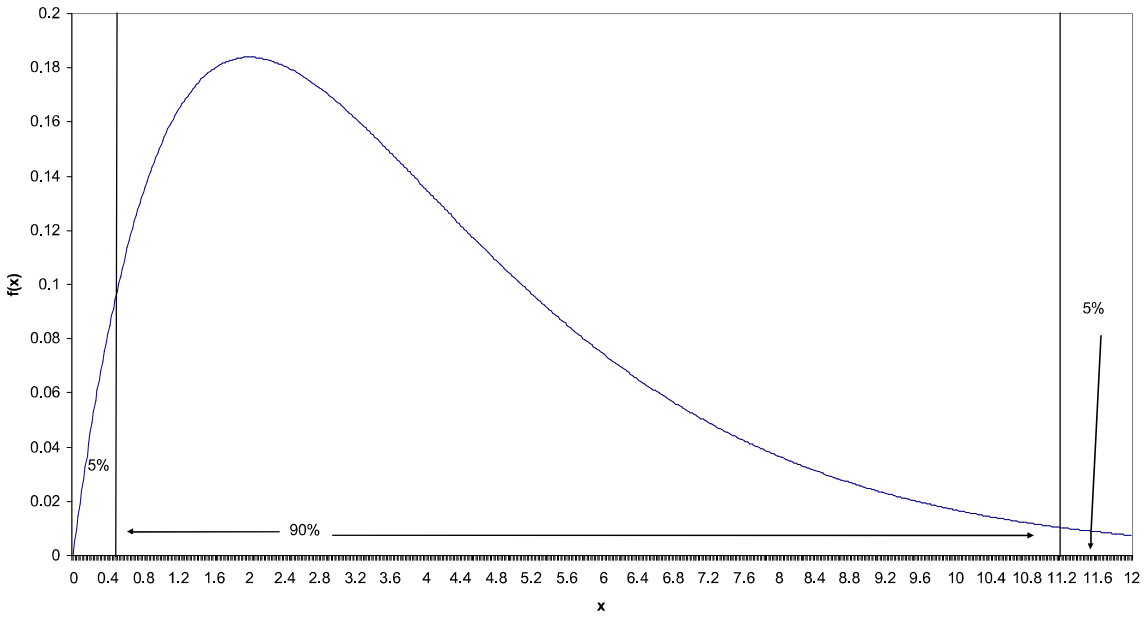
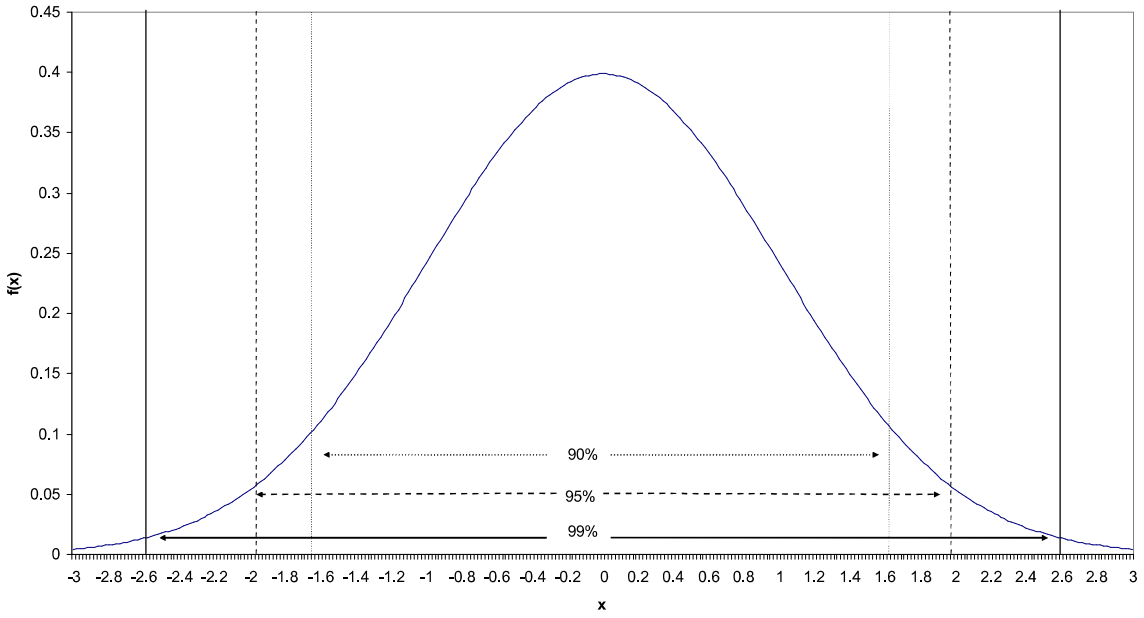


Figure 3: Confidence limits for a N(0,1)



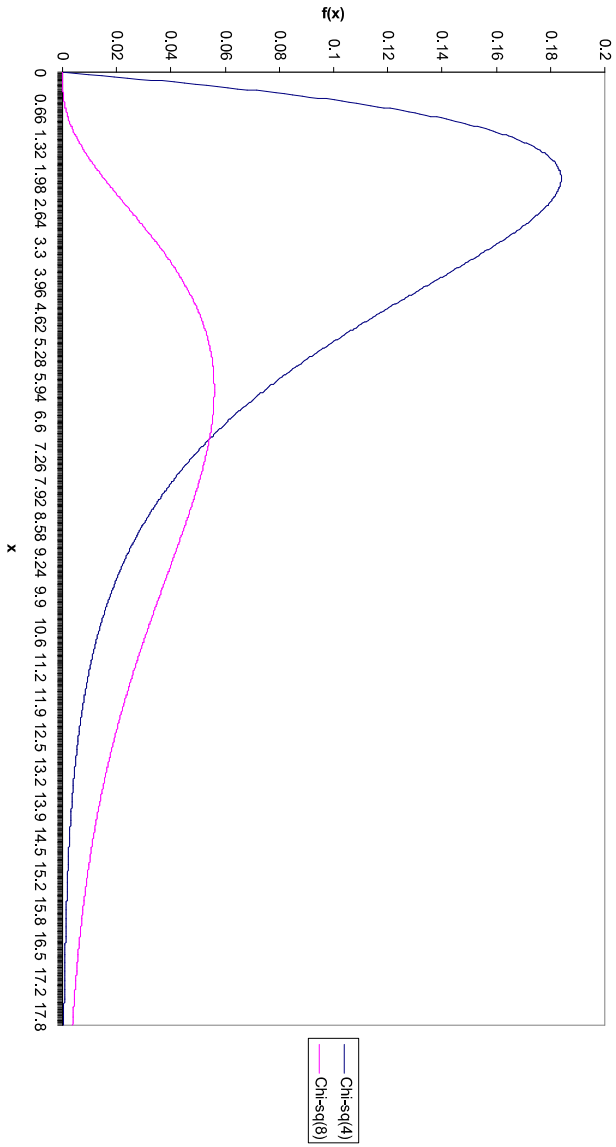


Figure 5: Chi-squared(4) and Chi-squared(8)

Confidence Intervals – Examples

1. A personnel manager found that historically, the scores on aptitude tests given to applicants are normal with a standard deviation of 32.4 points. A random sample of 9 scores produced a mean of 187.9 points. Based on these results a statistician found a population mean confidence interval of 165.8-210.0 points.
   (a) Find the probability of this interval.
   (b) Find an 80% confidence interval for the population mean score.

Answer

The underlying distribution is normal with a known population variance, therefore the distribution of the sample mean will be normal:

$$\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \bar{X} \sim N(\mu, 32.4^2/9)$$

(a) Now, $187.9 - z_{\alpha/2}\left(\frac{32.4}{3}\right) \le \mu \le 187.9 + z_{\alpha/2}\left(\frac{32.4}{3}\right) \Rightarrow 165.8 \le \mu \le 210.0$

$$D = 22.1 \Rightarrow z_{\alpha/2}\left(\frac{32.4}{3}\right) = \pm 22.1 \Rightarrow z_{\alpha/2} = \pm 2.046 \Rightarrow \alpha/2 = 0.0204$$

$\alpha = 0.0408 \Rightarrow$CI is 95.92%

(b) To construct an 80% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 = 0.10 \Rightarrow z_{0.10} = 1.28$$

Therefore the 80% confidence interval is:

$$\bar{x} - z_{0.10}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{0.10}\frac{\sigma}{\sqrt{n}}$$

where $\bar{x} = 187.9$.

The required 80% confidence interval is therefore:

$$187.9 - 1.28\left(\frac{32.4}{3}\right) \le \mu \le 187.9 + 1.28\left(\frac{32.4}{3}\right) \Rightarrow 174.05 \le \mu \le 201.75$$

and this is our 80% confidence interval for the population mean rating.

2. A random sample of 125 economics students were asked to rate the importance of particular job characteristics on a scale from 1 (not important) to 5 (extremely important). For the question on job security the sample mean rating was 4.18 and the sample standard deviation 0.80. Find a 99% confidence interval for the population mean.

Answer

The underlying series has outcomes taking one of five integer values between 1 and 5. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as $n \to \infty$. In particular:

$$\bar{X} \xrightarrow{a} N(\mu, s_X^2/n) \Rightarrow \bar{X} \xrightarrow{a} N(\mu, 0.64/125)$$

To construct a 99% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 = 0.005 \Rightarrow z_{0.005} = 2.575$$

Therefore the 99% confidence interval is:

$$\overline{x} - z_{0.005}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + z_{0.005}\frac{s}{\sqrt{n}}$$

where $\overline{x} = 4.18$.

The required 99% confidence interval is therefore:

$$4.18 - 2.575\left(\frac{0.80}{\sqrt{125}}\right) \le \mu \le 4.18 + 2.575\left(\frac{0.80}{\sqrt{125}}\right) \Rightarrow 3.996 \le \mu \le 4.364$$

and this is our 99% confidence interval for the population mean rating.

3. A random sample of 850 voters were asked, "If there was a referendum tomorrow on Britain joining the ERM, how would you vote?" 391 voters reported support for Britain joining the ERM. Find a 95% confidence interval for the population proportion of all voters supporting Britain joining the ERM.

Answer

While the underlying series follows a Bernoulli trial and hence the distribution cannot be normal, the distribution of the sample mean (sample proportion) will be approximately normally distributed as $n \to \infty$. In particular:

$$\overline{X} \xrightarrow{a} N(p, \hat{p}(1-\hat{p})/n) \Rightarrow \overline{X} \xrightarrow{a} N(p, (0.46)(0.54)/850)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 = 0.025 \Rightarrow z_{0.025} = 1.96$$

Therefore the 95% confidence interval is:

$$\hat{p} - z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{0.025}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

where $\hat{p} = 0.46$.

The required 95% confidence interval is therefore:

$$0.46 - 1.96\sqrt{\left(\frac{0.46(0.54)}{850}\right)} \le p \le 0.46 + 1.96\sqrt{\left(\frac{0.46(0.54)}{850}\right)} \Rightarrow 0.426 \le p \le 0.493$$

4. GAP is interested in the expenditure on clothes of university students in the first month of the academic year. For a random sample of 15 students, the mean expenditure was £89.56 and the sample standard deviation was £20.13. Assuming that the population distribution is normal, find a 95% confidence interval for population mean expenditure.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown.

$$\overline{X} \sim N(\mu, \sigma^2/n) \Rightarrow \frac{(\overline{X}-\mu)}{\sigma/\sqrt{n}} \sim N(0,1)$$

we also know that

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\overline{X}-\mu)}{s_X/\sqrt{n}} \sim t_{n-1}$$

To construct a 95% confidence interval we need to find a point $t_{14,\alpha/2}$, such that:

$$P(t_{14} > t_{14,\alpha/2}) = \alpha/2 = 0.025 \Rightarrow t_{14,0.025} = 2.145$$

Therefore the 95% confidence interval is:

$$\overline{x} - t_{14,0.025}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + t_{14,0.025}\frac{s}{\sqrt{n}}$$

where $\overline{x} = 89.56$ and $s = 20.13$.

The required 95% confidence interval is therefore:

$$89.56 - 2.145\left(\frac{20.13}{\sqrt{15}}\right) \le \mu \le 89.56 + 2.145\left(\frac{20.13}{\sqrt{15}}\right) \Rightarrow 78.41 \le \mu \le 100.71.$$

5. A random sample of 15 financial analysts' forecasts of next years' earnings per share for a large corporation was taken. The sample standard deviation was £0.88. Find a 95% confidence interval for the variance of predicted earnings per share for all analysts.

Answer

Assuming the underlying distribution of predicted earnings per share is normal. Then the sample variance of the predicted earnings will follow a chi-squared distribution.

$$\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$$

To construct a 95% confidence interval we need to find a the points $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$, such that:

$$P(\chi^2_{1-\alpha/2} > \chi^2_{n-1} > \chi^2_{\alpha/2}) = \alpha/2 = 0.025 \Rightarrow \chi^2_{0.025} = 26.12 \Rightarrow \chi^2_{0.975} = 5.63$$

Therefore the 95% confidence interval is:

$$\frac{(n-1)s_x^2}{\chi^2_{0.025}} \le \sigma^2 \le \frac{(n-1)s_x^2}{\chi^2_{0.975}}$$

where $s^2 = 0.88$.

The required 95% confidence interval is therefore:

$$\frac{14(0.88)^2}{26.12} \le \sigma^2 \le \frac{14(0.88)^2}{5.63} \Rightarrow 0.415 \le \sigma^2 \le 1.925$$

6. Independent samples of Vice-Chancellors (VCs) and Chief Executive Officers (CEOs) in large private companies were asked the importance of salary to their job satisfaction on a scale of 1 (not important at all) to 10 (the most important aspect). A random sample of 42 VCs had a mean rating of 4.01 and sample standard deviation of 1.2. For an independent random sample of 68 CEOs the mean rating was 5.43 and a sample standard deviation of 1.7. Find a 95% confidence interval for the difference in the population mean responses.

Answer

The underlying series has outcomes taking values between 1 and 10. The distribution cannot therefore be normal. Nevertheless the distribution of the sample mean will be approximately normal as $n \to \infty$. In particular:

$$\bar{X}_1 \overset{a}{\to} N(\mu_1, s_{x_1}^2 / n_1) \Rightarrow \bar{X}_1 \overset{a}{\to} N(\mu_1, 1.44/42)$$

$$\bar{X}_2 \overset{a}{\to} N(\mu_2, s_{x_2}^2 / n_2) \Rightarrow \bar{X}_2 \overset{a}{\to} N(\mu_2, 2.89/68)$$

and

$$\bar{X}_1 - \bar{X}_2 \overset{a}{\to} N\left(\mu_1 - \mu_2, \frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right) \Rightarrow \bar{X}_1 - \bar{X}_2 \overset{a}{\to} N(\mu_1 - \mu_2, 0.0768)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 = 0.025 \Rightarrow z_{0.025} = 1.96$$

Therefore the 95% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - z_{0.025}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + z_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where $\bar{x}_1 - \bar{x}_2 = -1.42$.

The required 95% confidence interval is therefore:

$$-1.42 - 1.96(0.277) \le \mu_1 - \mu_2 \le -1.42 + 1.96(0.277) \Rightarrow -1.963 \le \mu_1 - \mu_2 \le -0.877.$$

7. Of a random sample of 150 Economics students 105 said that teaching as a career was very unappealing. For an independent sample of 120 English Literature students 72 had the same reaction to teaching as a career. Find a 95% confidence interval for the difference between the population proportions regarding teaching as an unappealing career.

Answer

The underlying series follows a Bernoulli trial and hence the distribution cannot be normal, however, the distribution of the sample mean (sample proportion) will be approximately normally distributed as $n \to \infty$. In particular:

$$\bar{X}_1 \overset{a}{\to} N(p_1, \hat{p}_1(1-\hat{p}_1)/n_1) \text{ and } \bar{X}_2 \overset{a}{\to} N(p_2, \hat{p}_2(1-\hat{p}_2)/n_2)$$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(p_1 - p_2, \left[\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right]\right)$$

To construct a 95% confidence interval we need to find a point $z_{\alpha/2}$, such that:

$$P(Z > z_{\alpha/2}) = \alpha/2 = 0.025 \Rightarrow z_{0.025} = 1.96$$

Therefore the 95% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - z_{0.025}\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)} \le p_1 - p_2 \le (\bar{x}_1 - \bar{x}_2) + z_{0.025}\sqrt{\left(\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}\right)}$$

The required 95% confidence interval is therefore:

$$0.1 - 1.96(0.0583) \le \mu_1 - \mu_2 \le 0.1 + 1.96(0.0583) \Rightarrow -0.014 \le p_1 - p_2 \le 0.214.$$

8. A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 4.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 5.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown, assuming they are equal:

$$\bar{X}_1 \sim N(\mu_1, \sigma^2/n_1) \text{ and } \bar{X}_2 \sim N(\mu_2, \sigma^2/n_2)$$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left[\frac{1}{n_1} + \frac{1}{n_2}\right]\right)$$

we also know that

$$\frac{(n_1 + n_2 - 2)s_X^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}$$

the greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_X\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2}$$

where $s_x^2 = \dfrac{(n_1-1)s_{x_1}^2 + (n_2-1)s_{x_2}^2}{n_1+n_2-2}$, $s_1^2 = 4.2^2$ and $s_{21}^2 = 5.2^2$.

To construct a 90% confidence interval we need to find a point $t_{17,\alpha/2}$, such that:

$$P(t_{17} > t_{17,\alpha/2}) = \alpha/2 = 0.05 \Rightarrow t_{17,0.05} = 1.74$$

Therefore the 90% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - t_{17,0.05}s_x\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{17,0.05}s_x\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

where $\bar{x}_1 - \bar{x}_2 = -5.32$ and $s = 4.7557$.

The required 95% confidence interval is therefore:

$$-5.32 - 1.74(4.7557)\sqrt{\frac{1}{9}+\frac{1}{10}} \le \mu_1 - \mu_2 \le -5.32 + 1.74(4.7557)\sqrt{\frac{1}{9}+\frac{1}{10}}$$
$$\Rightarrow -9.121 \le \mu_1 - \mu_2 \le -1.519.$$

9. A researcher intends to estimate the effect of a drug on scores of human subjects performing a task of psychomotor coordination. The members of a random sample of 9 subjects were given the drug prior to testing, their mean score was 9.78 and the sample standard deviation was 3.2. An independent sample of 10 subjects were given a placebo prior to testing, the mean score for this group was 15.10 and the sample standard deviation was 7.2. Assuming the population distributions are normal, find a 90% confidence interval for the difference in the population means.

Answer

The underlying series has a normal distribution, but the population standard deviation is unknown:

$\bar{X}_1 \sim N(\mu_1, \sigma_{x_1}^2/n_1)$ and $\bar{X}_2 \sim N(\mu_2, \sigma_{x_2}^2/n_2)$

implying:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \left[\frac{\sigma_{x_1}^2}{n_1} + \frac{\sigma_{x_2}^2}{n_2}\right]\right)$$

The greater uncertainty associated having to use the sample variance as opposed to the unknown population variance, means that

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}\right)}} \sim t_{DoF}, \text{ where } DoF = \frac{\left[3.2^2/9 + 7.2^2/10\right]^2}{\left(3.2^2/9\right)^2/8 + \left(7.2^2/10\right)^2/9} = 12.70.$$

To construct a 90% confidence interval we need to find a point $t_{12,\alpha/2}$, such that:

$$P(t_{12} > t_{12,\alpha/2}) = \alpha/2 = 0.05 \Rightarrow t_{12,0.05} = 1.782$$

Therefore the 90% confidence interval is:

$$(\bar{x}_1 - \bar{x}_2) - t_{12,0.05}\sqrt{\frac{s_{x_1}^2}{n_1}+\frac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{12,0.05}\sqrt{\frac{s_{x_1}^2}{n_1}+\frac{s_{x_2}^2}{n_2}},$$ where

$\bar{x}_1 - \bar{x}_2 = -5.32$.

The required 90% confidence interval is therefore:

$$-5.32 - 1.782\sqrt{\frac{3.2^2}{9}+\frac{7.2^2}{10}} \le \mu_1 - \mu_2 \le -5.32 + 1.782\sqrt{\frac{3.2^2}{9}+\frac{7.2^2}{10}}$$
$$\Rightarrow -9.800 \le \mu_1 - \mu_2 \le -0.839$$

Confidence Intervals Sheet

Confidence Intervals for Means

Parameter	Distrib. of X_i	Sample	Variance	Confidence Interval
One Population				
μ	Normal	Large/Small	Known	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
μ	Normal	Large/Small	Not Known	$\bar{x} - t_{n-1}^{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{n-1}^{\alpha/2}\frac{s}{\sqrt{n}}$
μ	Non-Normal	Large	Known	$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
μ	Non-Normal	Large	Not Known	$\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}$
μ	Non-Normal	Small	Known/Not Known	$? \le \mu \le ?$

Two Populations

$\mu_1 - \mu_2$ Normal Large/Small Known

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

$\mu_1 - \mu_2$ Normal Large/Small Not Known (Equal)

$$(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2}^{\alpha/2}\sqrt{\frac{s_0^2}{n_1} + \frac{s_0^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2}^{\alpha/2}\sqrt{\frac{s_0^2}{n_1} + \frac{s_0^2}{n_2}}$$

where, $s_0^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

$\mu_1 - \mu_2$ Normal Large/Small Not Known (Unequal)

$$(\bar{x}_1 - \bar{x}_2) - t_{DoF}^{\alpha/2}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + t_{DoF}^{\alpha/2}\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}$$

where, $DoF = \dfrac{\left[s_{x_1}^2/n_1 + s_{x_2}^2/n_2\right]^2}{\left(s_{x_1}^2/n_1\right)^2/(n_1-1) + \left(s_{x_2}^2/n_2\right)^2/(n_2-1)}$

$\mu_1 - \mu_2$ Non-Normal Large Known

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

$\mu_1 - \mu_2$ Non-Normal Large Not Known

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \le \mu_1 - \mu_2 \le (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$\mu_1 - \mu_2$ Non-Normal Small Known/Not Known

$$? \le \mu_1 - \mu_2 \le ?$$

Confidence Intervals for Proportions

p Non-Normal Large Not Known

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

p Non-Normal Small Not Known $? \le p \le ?$

$p_1 - p_2$ Non-normal Large Not Known

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \le p_1 - p_2 \le (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$p_1 - p_2$ Non-normal Small Not Known $? \le p_1 - p_2 \le ?$

Confidence Intervals on variances

σ^2 Normal Large/Small Not Known $\dfrac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \le \sigma^2 \le \dfrac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}$

STATISTICAL TECHNIQUES B

Nonparametric Tests

1. Introduction

In Handout 8 (Hypothesis Testing) a number of our tests were reliant on the assumption of normality of the underlying distribution of X_i. For example, we learnt that if the underlying distribution is normal: (i) and the population variance is known the distribution of the resultant test statistic of the sample mean would be normal; (ii) and the population variance is unknown the distribution of the resultant test statistic of the sample mean would be a t-distribution. By virtue of a Central Limit Theorem, the distribution of the test statistic of the sample mean will be approximately normal, in large samples, even if the population distribution is not normal. So for example, we might have: The Thomas Pink Gold Cup (held in November) and the Cheltenham Gold Cup (in March) are 2-mile National Hunt jumps races for horses at Cheltenham Racecourse. The same nine two-year old horses were timed in each race of these races and we are interested in testing the hypothesis $H_0 : \mu_d = 0$ (mean time difference is zero) against a 2-sided alternative at the 5% significance level. The times taken were as follows (in minutes):

PARAMETRIC TEST

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
Difference	-0.2	-0.2	-0.3	-0.5	-0.1	+0.4	-0.4	-0.1	-0.1

Matched pairs: $\bar{x}_d = -0.167, s_d = 0.255$ if underlying distribution is normal then $\bar{X}_d \sim N(\mu_d, \sigma_d^2/n)$ and with σ_d^2 unknown we have: $\dfrac{\bar{X}_d - \mu_d}{\sqrt{s_d^2/n}} \sim t_{n-1}$. In which case: $\Pr\left(t_{n-1} < \dfrac{(-0.167-0)}{\sqrt{0.255^2/9}}\right) = \Pr\left(t_{n-1} < -1.960\right) = 0.043$ and for a 2-sided test this probability is 0.086 and we do NOT reject H_0.

However, it often is the case that the normality assumption is not reasonable and/or the sample size is not large. In these cases it is desirable to base inference on tests which are valid over a wide range of distributions of X_i (although they do require certain assumptions to be valid, e.g. independent random samples). These tests are often referred to as nonparametric tests.

2. Sign Test (Wilcoxon)

This is the simplest test to undertake and is used for testing hypotheses about the central location of a population distribution. This is most frequently used in analysing matched pairs data and is based on assigning a sign (plus(+) if the original sample value is greater than the subsequent sample value and minus(-) if the original sample value is less than the subsequent sample value and 0 if the two values are equal). The null hypothesis is that in the population the two values have the same mean. In undertaking this test we discard the 0 values, i.e. those individuals who did not have a strictly higher value in one sample compared to the other.

Based on the reduced sample of n observations, for which there was a preference stated, under the null hypothesis the number of + and - values should be a random sample from a population in which the Pr(+)=0.5 and the Pr(-)=0.5. Consider only + values and denote p as the true proportion of +'s in the population, then $H_0 : p = 0.5$ and the distribution for the number of say W (the number + values) follows a Binomial distribution, $W \sim B(n, 0.5)$. For an alternative hypothesis $H_1 : p < 0.5 \left(p > 0.5\right)$, we want $\Pr(W \le w)$ $\left(\Pr(W \ge w)\right)$, for a 1-sided test and for an alternative hypothesis $H_1 : p \neq 0.5$, we want $2 \times \Pr(W \le w)$ (see example 1)

Note that for large n (>25) $W/n \overset{a}{\sim} N(0.5, 0.25/n) \Rightarrow \dfrac{W/n - 0.5}{\sqrt{0.25/n}} \sim N(0,1)$ (see example 2).

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
	-	-	-	-	-	+	-	-	-

$\Pr(W \le 1) = {}_9C_0 (0.5)^9 + {}_9C_1 (0.5)^9 = 0.020$ and for a 2-sided test this probability is 0.040 and we reject H_0.

3. Wilcoxon Signed Rank Test

The problem with the Sign Test is that it only uses a very limited amount of information (namely the sign of the difference) and therefore ignores the strength of preference of one value over the other. As a result the test can lack power in small samples. The signed rank test, uses not only the sign, but also the magnitudes of the differences. This test is also applied to matched pairs and is testing the null hypothesis $H_0 : \mu_d = 0$, where μ_d is the population mean difference in scores across the matched pairs. As with the Sign Test differences of 0 are ignored. The nonzero absolute differences are then ranked in ascending order of magnitude (where equal values are assigned the average rank). The ranks of positive and negative differences are then summed separately as $W_+ = \sum_{i=1}^{n} \phi_i^+ R_i$ and $W_- = \sum_{i=1}^{n} \phi_i^- R_i$, where

$\phi_i^+ = \begin{cases} 1 & \text{if difference positive} \\ 0 & \text{otherwise} \end{cases}$, $\phi_i^- = \begin{cases} 1 & \text{if difference negative} \\ 0 & \text{otherwise} \end{cases}$ and R_i is the rank of the absolute value of the difference in the scores for the i^{th} value.

We denote the Wilcoxon signed rank test as $T = \min(W_+, W_-)$. For a 1-sided alternative hypotheses $H_1 : \mu_d < 0 \; (H_1 : \mu_d > 0)$, you want $\Pr(T \le t)$, but for a 2-sided alternative hypothesis $H_1 : \mu_d \neq 0$, you want $2 \times \Pr(T \le t)$ and for small samples these probabilities are based on critical values reported in the Table below.

Under the null hypothesis that the true population difference in scores across the matched pairs is zero, it can be shown that:

$E(T) = \frac{n(n+1)}{4}$ and $V(T) = \frac{n(n+1)(2n+1)}{24}$ and for n>25 then

$T \overset{a}{\sim} N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$ and therefore $\frac{T - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \overset{a}{\sim} N(0,1)$ (see example 4).

Table 1: Critical values of the Wilcoxon Signed Rank Test (n<30)

1-tailed	α=0.05	α=0.025	α=0.01	α=0.005
2-tailed	α=0.10	α=0.05	α=0.02	α=0.01
n				
6	2	0	-	-
7	3	2	0	-
8	5	3	1	0
9	8	5	3	1
10	10	8	5	3
11	13	10	7	5
12	17	13	9	7
13	21	17	12	9
14	25	21	15	12
15	30	25	19	15
16	35	29	23	19
17	41	34	27	23
18	47	40	32	27
19	53	46	37	32
20	60	52	43	37
21	67	58	49	42
22	75	65	55	48
23	83	73	62	54
24	91	81	69	61
25	100	89	76	68

Thomas Pink	8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9
Cheltenham	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
Difference	-0.2	-0.2	-0.3	-0.5	-0.1	+0.4	-0.4	-0.1	-0.1
Rank	(-)4.5	(-)4.5	(-)6	(-)9	(-)2	(+)7.5	(-)7.5	(-)2	(-)2

$W_- = 37.5$, $W_+ = 7.5$, so $T = 7.5$ and this is greater than the critical value 5 and so we do not reject H_0.

4. Mann-Whitney Test

This test compares the central location of two populations, but in this case the samples come from independent random samples. Suppose that n_1 observations are available from the first population and n_2 from the second. All observations (n_1+ n_2) are then ranked in ascending order of magnitude (where equal values are assigned the average rank) and we denote R_1 as the sum of ranks of observations from the first population. The Mann-Whitney test is then defined as:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

In which case:

$$E(U) = \frac{n_1 n_2}{2} \text{ and } V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

and for relatively large samples (n>25) $U \overset{a}{\sim} N\left(\frac{n_1 n_2}{2}, \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}\right)$ implying

$$\frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} \overset{a}{\sim} N(0,1).$$

8.1	8.2	8.0	8.0	8.4	8.6	8.5	8.4	8.9	8.3	8.4	8.3	8.5	8.5	8.2	8.9	8.5	9.0
8.0	8.0	8.1	8.2	8.2	8.3	8.3	8.4	8.4	8.4	8.5	8.5	8.5	8.5	8.6	8.9	8.9	9.0
1.5	1.5	3	4.5				9	9		12.5				15	16.5		

where red is Thomas Pink and Green is Cheltenham. R_1=72.5, in which case:

U=81+45-72.5=53.5. We know, $E(U)$ = 40.5 and $V(U)$ = 128.5. In which case

$\Pr\left(z > \frac{(53.5-40.5)}{\sqrt{128.5}}\right) = \Pr(z > 1.148) = 0.125$ and for a 2-sided test this probability is 0.250, in which case we do NOT reject H_0.

5. Goodness-of-fit test

Suppose that we are given a random sample of n observations, each of which can be classified into exactly one of K categories. Denote the observed number of cases in each category as $O_1, O_2, O_3, \ldots, O_K$. If a null hypothesis ($H_0$) specifies probabilities $p_1, p_2, p_3, \ldots, p_K$ for an observation falling into each of these categories, the expected numbers in each category, under H_0, would be $E_i = np_i \quad (i = 1, 2, \ldots, K)$

We then test whether the actual data is a close fit to the expect data (based on some assumed population distribution for probabilities) – and this is done by looking at the magnitude of the discrepancy between the observed and expected values. Where large (absolute) values ought to make one increasingly suspicious of the null hypothesis. The test is constructed as (see Appendix 1 for a proof of this equivalence):

$$\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{K} \frac{O_i^2}{E_i} - n \sim \chi^2_{(K-1)}$$

And H_0 is rejected at significance level α, if $\sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{(K-1),\alpha}$. (see example 6)

In November 2009, leading Economists were asked about inflation expectations for December 2012. The results showed that 10% thought inflation would be <1%, 40% thought inflation would be 1-2%, 40% thought inflation would be 2-3% and 10% thought inflation would be >3%. In November 2011, 40 Economists reported their inflation expectations for December 2012 as follows:

Range	<1%	1-2%	2-3%	>3%
Frequency	2	10	18	10

Test the hypothesis the distribution has not changed.

Range	<1%	1-2%	2-3%	>3%
Frequency	2	10	18	10
Expected	0.1*40=4	0.4*40=16	0.4*40=16	0.1*40=4

$\frac{2^2}{4} + \frac{10^2}{16} + \frac{18^2}{16} + \frac{10^2}{4} - 40 = 12.5$, $\chi^2_{3,0.05} = 7.81$ and so we reject H_0.

6. Contingency Tables

Suppose we have two attributes A and B. There are K categories in A and H in B so that there are KH cross-classifications in total. The number of sample observations belonging to the i^{th} category of A and the j^{th} category of B is denoted as O_{ij} and there are n observations in total. To test the null hypothesis of no association (independence) between the two attributes, we want to know how many observations were would expect to find in each cross-classification. Under the null hypothesis of independence between the two attributes A and B we know that the joint probability (p_{ij}) is equal to the product of the marginal probabilities ($p_{i.}p_{j}$), in other words:

$$p_{ij} \equiv \Pr(A=i, B=j) = \Pr(A=i).\Pr(B=j) \equiv p_{i.}p_{j}.$$

Now $p_{i} \equiv \Pr(A=i) = \sum_{j=1}^{H} O_{ij}/n$ and $p_{j} \equiv \Pr(B=j) = \sum_{i=1}^{K} O_{ij}/n$. In which case under the null hypothesis of independence the expected number of observations is

$$E_{ij} = np_{ij} = \sum_{j=1}^{H} O_{ij} \sum_{i=1}^{K} O_{ij}/n.$$

We then test whether the actual data is a close fit to the expect data and this is done by looking at the magnitude of the discrepancy between the observed and expected values. Where large (absolute) values ought to make one increasingly suspicious of the null hypothesis. The test is constructed as:

$$\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = \sum_{i=1}^{K}\sum_{j=1}^{H}\frac{O_{ij}^{\,2}}{E_{ij}} - n \sim \chi^2_{(K-1)(H-1)}$$

And H_0 is rejected at some significance level α, if $\sum_{i=1}^{K}\sum_{j=1}^{H}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} > \chi^2_{(K-1)(H-1),\alpha}$.

(see example 7)

50 leading Economists were asked about inflation expectations for December 2012 in both November 2009 and in November 2011 and the results were reported as follows:

		Nov 2009			
		<1%	1-2%	2-3%	>3%
Nov 2011	<1%	2	0	1	0
	1-2%	4	8	1	0
	2-3%	2	8	12	0
	>3%	2	2	2	6

Test the hypothesis that there is no association between these two series.

		Nov 2009				
		<1%	1-2%	2-3%	>3%	
Nov 2011	<1%	2	0	1	0	3
	1-2%	4	8	1	0	13
	2-3%	2	8	12	0	22
	>3%	2	2	2	6	12
		10	18	16	6	50

Therefore under independence

$$P(<1\%,<1\%) = \frac{10}{50}\times\frac{3}{50} = \frac{30}{2500} = 0.012,\quad P(<1\%,1-2\%) = \frac{10}{50}\times\frac{13}{50} = \frac{130}{2500} = 0.052,$$

$$P(<1\%,2-3\%) = \frac{10}{50}\times\frac{22}{50} = \frac{220}{2500} = 0.088,\quad P(<1\%,>3\%) = \frac{10}{50}\times\frac{12}{50} = \frac{120}{2500} = 0.048$$

$$P(1-2\%,<1\%) = \frac{18}{50}\times\frac{3}{50} = \frac{54}{2500} = 0.0216,\quad P(1-2\%,1-2\%) = \frac{18}{50}\times\frac{13}{50} = \frac{234}{2500} = 0.0936,$$

$$P(1-2\%,2-3\%) = \frac{18}{50}\times\frac{22}{50} = \frac{396}{2500} = 0.1584,\quad P(1-2\%,>3\%) = \frac{18}{50}\times\frac{12}{50} = \frac{216}{2500} = 0.0864$$

$P(2-3\%,<1\%)=\frac{16}{50}\times\frac{3}{50}=\frac{48}{2500}=0.0192$, $P(2-3\%,1-2\%)=\frac{16}{50}\times\frac{13}{50}=\frac{208}{2500}=0.0832$,

$P(2-3\%,2-3\%)=\frac{16}{50}\times\frac{22}{50}=\frac{352}{2500}=0.1408$, $P(2-3\%,>3\%)=\frac{16}{50}\times\frac{12}{50}=\frac{192}{2500}=0.0768$

$P(>3\%,<1\%)=\frac{6}{50}\times\frac{3}{50}=\frac{18}{2500}=0.0072$, $P(>3\%,1-2\%)=\frac{6}{50}\times\frac{13}{50}=\frac{78}{2500}=0.0312$,

$P(>3\%,2-3\%)=\frac{6}{50}\times\frac{22}{50}=\frac{132}{2500}=0.0528$, $P(>3\%,>3\%)=\frac{6}{50}\times\frac{12}{50}=\frac{72}{2500}=0.0288$

Expected numbers

		Nov 2009				
		<1%	1-2%	2-3%	>3%	
Nov 2011	<1%	0.012×50=0.6	0.0216×50=1.08	0.0192×50=0.96	0.0072×50=0.36	3
	1-2%	0.052×50=2.6	0.0936×50=4.68	0.0832×50=4.16	0.0312×50=1.56	13
	2-3%	0.088×50=4.4	0.1584×50=7.92	0.1408×50=7.04	0.0528×50=2.64	22
	>3%	0.048×50=2.4	0.0864×50=4.32	0.0768×50=3.84	0.0288×50=1.44	12
		10	18	16	6	50

$$\frac{2^2}{0.6}+\frac{4^2}{2.6}+\frac{2^2}{4.4}+\frac{2^2}{2.4}+\frac{0^2}{1.08}+\frac{8^2}{4.68}+\frac{8^2}{7.92}+\frac{2^2}{4.32}+\frac{1^2}{0.96}+\frac{1^2}{4.16}+\frac{12^2}{7.04}+\frac{2^2}{3.84}+\frac{0^2}{0.36}+$$
$$\frac{0^2}{1.56}+\frac{0^2}{2.64}+\frac{6^2}{1.44}-50=35.86$$

$\chi^2_{9,0.05}=16.92$, therefore we reject H_0.

Equivalence of Contingency Tests

$$\sum_{i=1}^{K}\frac{(O_i-E_i)^2}{E_i}=\sum_{i=1}^{K}\frac{O_i^2+E_i^2-2O_iE_i}{E_i}=\sum_{i=1}^{K}\left(\frac{O_i^2}{E_i}+E_i-2O_i\right)$$
$$=\sum_{i=1}^{K}\frac{O_i^2}{E_i}+\sum_{i=1}^{K}E_i-2\sum_{i=1}^{K}O_i$$

As $\sum_{i=1}^{K}E_i=\sum_{i=1}^{K}O_i=n$

$$=\sum_{i=1}^{K}\frac{O_i^2}{E_i}+n-2n=\sum_{i=1}^{K}\frac{O_i^2}{E_i}-n$$

Non-Parametric Test Examples

1. A random sample of twelve financial analysts was asked to predict the percentage increases in the prices of two common stocks over the next year. The results obtained are shown in the table below. Use the sign test to test the null hypothesis that for the population of analysts, there is no overall preference for one stock over the other:

Analyst	Stock 1	Stock 2	Analyst	Stock 1	Stock 2
A	6.8	7.1	G	9.3	10.1
B	9.8	12.3	H	1.0	2.7
C	2.1	5.3	I	-0.2	1.3
D	6.2	6.8	J	9.6	9.8
E	7.1	7.2	K	12.0	12.0
F	6.5	6.2	L	6.3	8.9

Answer

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

n=11 with 1 + value and 10 – values, we want

$2 \times \Pr(W \leq 1) = 2 \times [\Pr(W = 0) + \Pr(W = 1)] = 2 \times [0.0005 + 0.0054] = 0.0118$

and so we reject H0 at significance levels in excess of 1.18%.

2. In a random sample of 130 voters, 44 favoured tax increases to raise funding for education, 68 opposed the tax increase, and 18 expressed no opinion. Test against a 2-sided alternative the null hypothesis that voters in the state are evenly divided on the issue of a tax increase.

Answer

$H_0 : p = 0.5$

$H_1 : p \neq 0.5$

n=130-18=112, $W / n = \hat{p} = 44/112 = 0.3929$

$$z = \frac{0.3929 - 0.5}{\sqrt{0.25/112}} = -2.27$$

p-value=2[1-Φ(2.27)]=0.0232 and so we reject H_0 at significance levels in excess of 2.32%

3. Using the data in (1), test the null hypothesis that for the population of analysts, there is no difference in the mean performance of one stock over the other, using the Wilcoxon Signed Rank Test.

Answer

$H_0 : \mu_d = 0$

$H_1 : \mu_d \neq 0$

n=11 with 1 + value and 10 – values, we want:

Analyst	Stock 1	Stock 2	Stk1-Stk2	$\phi_i^+ R_i$	$\phi_i^+ R_i$
A	6.8	7.1	-0.3		3.5
B	9.8	12.3	-2.5		9
C	2.1	5.3	-3.2		11
D	6.2	6.8	-0.6		5
E	7.1	7.2	-0.1		1
F	6.5	6.2	+0.3	3.5	
G	9.3	10.1	-0.8		6
H	1.0	2.7	-1.7		8
I	-0.2	1.3	-1.5		7
J	9.6	9.8	-0.2		2
K	12.0	12.0	0.0		
L	6.3	8.9	-2.6		10
				3.5	125.5

From this we have T=3.5, with critical value of 10 (at the 5% significance level for a 2-sided test), we reject H_0.

4. A consultant id interested in the impact of the introduction of a quality management program on job satisfaction of employees. A random sample of 30 employees was asked to assess level of satisfaction on a scale of 1 (very dissatisfied) to 10 (very satisfied) 3 months before the introduction of the program. These same individuals were then asked to make this assessment again 3 months after the introduction of the program. The 30 differences in the pairs of rating were calculated and the absolute differences ranked. The smaller of the

rank sums, which was for those more satisfied before the introduction of the program was 169. What can be concluded from these findings?

Answer

$H_0 : \mu_d = 0$

$H_1 : \mu_d < 0$

T=169, $E(T) = \frac{30(31)}{4} = 232.5, V(T) = \frac{30(31)(61)}{24} = 2363.75$

$z = \frac{169 - 232.5}{\sqrt{2363.5}} = -1.31 \Rightarrow$ p-value=1-Φ(1.31)=0.0951 and so we reject H_0 at significance levels in excess of 9.51%.

5. A random sample of 15 male and an independent random sample of 15 female students were asked to write essays at the conclusion of their writing module. Essays were then ranked from 1 (best) to 30 (worst) by the module leader as:

Males	26	24	15	16	8	29	12	6	18	11	13	19	10	28	7
Females	22	2	17	25	14	21	5	30	3	9	4	1	27	23	20

Answer

n_m=15, R_m=242, n_f=15 R_f=223.

$H_0 : \mu_d = 0$

$H_1 : \mu_d \neq 0$

$U = 15(15) + \frac{15(16)}{2} - 242 = 103$

$E(U) = \frac{15(15)}{2} = 112.5, V(U) = \frac{15(15)(15+15+1)}{12} = 581.25$

$z = \frac{103 - 112.5}{\sqrt{581.25}} = -0.39 \Rightarrow$ p-value=2[1-Φ(0.39)]=0.6966 and so we reject H_0 at significance levels in excess of 69.66%.

6. A random sample of 520 customers were asked about the importance of quality of food as a factor in choosing a hospital. Sample members were asked to respond as "not important", "important", or "very important". Respective numbers selecting these answers were: 199, 136 and 167. Test the null hypothesis that a randomly chosen consumer is equally likely to select each of these answers.

Answer

H_0: All outcomes equally likely

H_1: otherwise

	Not imp	Imp	Very imp	Total
Observed	199	136	167	502
Prob (under H_0)	0.333	0.333	0.333	1
Expected number	167.33	167.33	167.33	502

$\sum_{i=1}^{K} \frac{O_i^2}{E_i} - 502 = \frac{199^2}{167.33} + \frac{136^2}{167.33} + \frac{167^2}{167.33} - 502 = 11.86$

$\chi^2_{2,0.01} = 9.21 \Rightarrow$ Reject H_0 at 1% signficance level.

7. In a series of surveys, 55 forecasters were asked whether they thought inflation would increase over the next 12 months from its current level. It was also noted whether or not actual inflation increased. The results are reported in the table below:

Outcome	Forecast	
	Increase	No increase
Increase	18	11
No increase	6	20

Test the null hypothesis of no association between forecast and outcome.

Answer

H_0: No association between forecast and outcome

H_1: otherwise

Under H_0 (independence) the probability of being in each category is:

$$P(Increase, Increase) = \frac{24}{55} \times \frac{29}{55} = 0.230,$$

$$P(Increase, No\ Increase) = \frac{24}{55} \times \frac{26}{55} = 0.206,$$

$$P(No\ Increase, Increase) = \frac{31}{55} \times \frac{29}{55} = 0.297,$$

$$P(Increase, No\ Increase) = \frac{31}{55} \times \frac{26}{55} = 0.266,$$

and expected number of observations in each category is:

Outcome	Forecast		
	Increase	No increase	
Increase	0.230×55	0.206×55	29
No increase	0.297×55	0.266×55	26
	24	31	55

Outcome	Forecast		
	Increase	No increase	
Increase	12.65	16.35	29
No increase	11.35	14.65	26
	24	31	55

$$\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{O_{ij}^{\ 2}}{E_{ij}} - 55 = \frac{18^2}{12.65} + \frac{11^2}{16.35} + \frac{6^2}{11.35} + \frac{20^2}{14.65} - 55 = 8.48$$

$\chi^2_{1,0.01} = 6.63 \Rightarrow$ Reject H_0 at 1% signficance level.