Handbook to accompany the Blended Learning course “World of Development and the Role of Statistic”

Contents

Foreword

This Handbook was compiled to assist the users and participants of the course “World of Development and the Role of Statistics”. The course is Web based and has the goal to make participants interested in the history and selected essentials of development policy as well as to make professional statisticians be aware of the different actors, scenarios and milestones in this area. No expertise of statistics is expected of the participants and the exercises are prepared that basic knowledge of EXCEL and reading and understanding scientific textbooks should suffice for a successful participation.

It may be that a participant might want to know more of statistics and misses a more systematic approach to the subject, because introduction to probability and sampling, to name just two subjects have been limited to the very basic. Here comes a more systematic approach and deals more with statistics theory and less with “World of Development”. This part of the Handbook concentrates on Regression Theory because most what has been said in the WBT about regression has close links to the other subjects. The Regression theory will be discussed using a very simple example of yield influenced by fertilizer input and rainfall, something very similar to what we use in the WBT. The Factor and Principal Component Analysis are rather specific cases of the general Regression Model. Again a summary of the different techniques and tools of methods will be exposed and another example is exposed rather verbally than technically. Factor analysis is by no means the “best” method for analysis complex and abundant data material: Linear discriminate analysis, ANOVA and cluster analysis are others to name a few and would merit to be included into this course. However the outcome and the relevant example determined the choice for Factor analysis. So principal emphasis is laid on explaining the theory and the reason why to use it without getting too theoretical, I hope.

Still this handbook would not substitute a statistics course. Many of the contents have been inspired by two great textbooks [6] and [37] and which are standards of teaching statistics to students and academics. The references you will find in the Bibliography. Please study these text books, especially the very practice related exercises which may help to understand many details and subject of statistics better. As a final word from the author of the Web based training let me summarize: Statistics should be regarded for a vast majority as a supporting science, except for those who deal with it for scientific purposes. The great use and benefit is that reality can be simplified and answers been given to the questions like: Which are the outcomes of my policy, how do I measure well-being in my region, which measures have proven to be successful to reduce poverty, which groups of people experience the highest resilience against economic changes with more authority than a rule of thumb. Never forget that there is not only one response to these questions and not all responses given are equally valid and scientifically based. The approach of the WBT should help to pick the better choice.

Introduction

The word "statistics" originally meant the collection of population and economic information vital to the state. From that modest beginning, statistics has grown into a scientific method of analysis that now is applied to all the social and natural sciences. The present aims and methods of statistics are best illustrated by a familiar example.

Before every presidential election, the pollsters try to pick the winner; specifically, they try to guess the proportion of the population that will vote for each candidate. Clearly, canvassing all voters would be an impossible task. As the only alternative, pollsters survey a sample of a few thousand in the hope that the sample proportion will constitute a good estimate of the total population proportion. This is a typical example of *statistical inference* or *statistical induction:* the (voting) characteristics of an unknown population are inferred from the (voting) characteristics of an observed sample.

As any pollster will admit, it is an uncertain business. To be sure of the population, we have to wait until Election Day, when all votes are counted. Yet if the sampling is done fairly and adequately, we can have high hopes that the sample proportion will be close to the population proportion. This will allow us to estimate the unknown population T proportion from the observed sample proportion P, as follows:
T = P ± a small error
With the crucial questions being, "How small is this error?" and "How sure are we that we are right?" Since this typifies the very core of this handbook, we state the precise formula which will be explained later.

If the sampling is random, we can state with 95% confidence that:

(Formula 1.1) $T = P \pm 1.96\sqrt{\frac{P(1-P)}{n}}$, with T being the population, P the sample proportion and n the sample size

Example: Just before an presidential election, a poll of 2,000 voters shows 760 for Candidate A and 1,240 for Candidate B. Calculate the 95% confidence interval for the population proportion T that will vote for Candidate Solution: $T \cong 0.38 \pm 0.2$ for candidate A, that is with 95% confidence, the proportion for candidate A among the whole population of voters will be between 36% and 40%.
Remark:. In the actual election, this proportion of the voting population has to be confirmed and there have been spectacular miscalculations and predictions. However, people like these types of forecasts and statisticians have to provide the tools of calculations to make the errors less likely and in general forecasts are better than their reputation at least if they are calculated

with the necessary professional diligence.

Before we illustrate this formula with an example, we repeat the warning that we gave in the preface: Every numbered example in this text is an exercise that you should actively work out yourself, rather than passively read. We therefore put each example in the form of a question for you to answer; if you get stuck, then you may read the solution. But in all cases remember that statistics is not a spectator sport. You cannot learn it by watching, any more than you can learn to ride a bike by watching. You have to jump on and take a few spills.

Constructing confidence intervals will be one of our major objectives. Another related objective is to test hypotheses. To use the same example, suppose that an ardent supporter of candidate A claimed that he/she would win the election. In mathematical terms, this hypothesis may be written: π > .50. On the basis of the information in above mentioned equation we would reject this hypothesis, of course. In general, there is a very close association of this kind between confidence intervals and hypothesis tests.

We can make several other crucial observations:.
1. The estimate is not made with certainty; we are only 95% confident. We must concede the possibility that we are wrong—simply because we were unlucky enough to draw a misleading sample. For example, if less than half the population is in fact supports candidate A it is still possible, ,though unlikely, for us to run into a string of supporters of candidate A in our sample. In such circumstances, our confidence interval would be wrong. Since this sort of bad luck is possible but not likely, we are just 95% confident.

2. As sample size n increases, we note that the error allowance in decreases. In Example 1, if the poll increased the sample to 10,000 voters and continued to observe a proportion of .38, the 95'% confidence interval would become more precise: $T = 0.38 \pm 0.1$. This also is intuitively correct: a larger sample contains more information, and hence allows a more precise conclusion.

3. Suppose that we feel that 95% confidence is not good enough, and that instead we want to be 99% sure of our conclusion. If the additional resources for further sampling are not available, then we can increase our confidence only by making a less precise statement. As we will be able to show later, for 99% confidence the formula must have the coefficient 1.96 enlarged to 2.58; this yields the 99% confidence interval: $T \cong 0.38 \pm 0.3$

This is broader and less precise than the 95% confidence interval; we must be less precise because we wish to be more certain of being right. In any case, we note that any statistical statement must be prefaced by some degree of uncertainty.

Deduction and Induction

Deduction in panel involves arguing from the general to the specific—i.e., from the population to the sample. Induction is the reverse—arguing from the specific to the general, i.e., from the sample to the population. The above Equation (1.1) represents inductive reasoning; we are arguing from a sample proportion to a population proportion. This is possible only if we study the simpler problem of deduction first. Specifically, in Equation (1.1) the inductive statement (that the population proportion can be inferred from the sample proportion) is based on a prior deduction (that the sample proportion is likely to be close to the population proportion).

Subsequent chapters are devoted to deduction. This involves probability theory, leading up to such questions as, "With a given population, how will a sample behave? Will the sample be on target?" Only when this deductive issue is resolved can we move to questions of statistical inference in later chapters. To keep these terms straight, remember that the population is the point of reference. The prefix “de” means "away from." Thus deduction is arguing away from the population. The prefix “in” means "into" or "towards." Thus induction is arguing towards the population. Finally, statistical inference is based on induction

Later we turn the argument around and ask, "From a given observed sample, what can we conclude about the unknown population?"

Sampling—why and how?

We draw a sample, rather than examine the whole population, for several reasons:

1. Limited resources. For example, in pre-election polls, neither funds nor time are available to observe the whole population.

2. Scarcity. Sometimes only a small sample is available. For example, in heredity versus environment controversies, identical twins provide ideal data because they have identical heredity. Yet very few such twins are available.

3. Destructive testing. For example, suppose that we wish to know the average life of all the light bulbs produced by a certain factory. It would be absurd to insist on observing the whole population of bulbs until they burn out.

If sampling is required, how should it be done? In statistics, as in business or any other profession, it is essential to distinguish

between bad luck and bad management.
If we now return to our original example of pre-election polls, we note that the sample proportion of candidate A may misrepresent the population proportion for either of these reasons. No matter how well-managed and designed our sampling procedure may be, we may be unlucky enough to turn up a sample favouring Candidate A from a population favouring Candidate B. The Equation (1.1) relates to this case; it is assumed that the only complication is the luck of the draw, and not mismanagement. From that equation we confirm that the best defence against bad luck is to "keep playing"; by increasing our sample size, we improve the reliability of our estimate.
The other problem is that sampling can be badly mismanaged or biased. For example, in sampling a population of voters, it is a mistake to take their names from a phone book, since poor voters who often cannot afford telephones are badly underrepresented.
Other examples of biased samples are easy to find. Informal polls of people on the street often are biased because the interviewer tends to select people who seem civil and well-dressed; a surly worker or harassed mother is overlooked.

The simplest way to ensure an unbiased sample is to give each member of the population an equal chance of being included in the sample. This, in fact, is essentially the definition of a "random" sample.' For a sample to be random it cannot be chosen in a sloppy or haphazard way; it must be designed carefully. One possibility is to number all the individuals in the population, and draw the sample by using a chance device such as a bowlful of numbered chips, a roulette wheel, or the random digits given by a computer. If a sample is random, not only will it be free of bias, but it also will satisfy the assumptions of probability theory, and allow us to make scientific inferences of induction.

In some circumstances, the only available sample will be a non-random one. While probability theory often cannot be applied strictly to such a sample, it still may provide the basis for a good educated guess—or what we might term the art of inference. Although this art is very important, it cannot be discussed here although we apply it in the WBT; therefore only scientific inference is considered based on the assumption that samples are random. The techniques for ensuring this are discussed later.

Descriptive Statistics for Samples

Discrete Example

In a sample of 50 families, let us record the number of children, X, which takes on the values 0, 1, 2, 3, ... We call X a "discrete" random variable because it can take on only a finite number of values.1 Suppose that the 50 values of X turn out to be:
0,2,2,3,5,1,2,0, 4,2.
To simplify, we keep a running tally of each of the possible outcomes. In column (3) we record, for example, that 13 is the frequency (*f*) that we observed for a two-child family. That is, we obtained this outcome on 13/50 of our sample observations; this proportion (.26 or 26%) is called relative frequency (f/n), and is recorded in the last column.

Calculation of the Frequency and Relative Frequency of the Number of Children in a Sample of 50 families.

$$\sum f = 50 \qquad \sum \frac{f}{n} = 1$$

(1) Number of Children	(2) Tally	(3) Frequency (*f*)	(4) Relative Frequency $\left(\frac{f}{n}\right)$
0	𝍸 𝍸 𝍸	15	.30
1	𝍸 𝍸	10	.20
2	𝍸 𝍸 |||	13	.26
3	𝍸 |	6	.12
4	|||	3	.06
5	|||	3	.06

Table 2-1

where $\sum$ means "the sum of." Thus, for example, $\sum f$ means "the sum of the frequencies."
Usually, for relative frequencies (and probabilities) mathematicians prefer decimals, while applied statisticians prefer percentages. Therefore, we usually do our calculations in decimal form and usually give the verbal interpretations in percentage form.

The information in column (3) is called a "frequency distribution," which is graphed in Figure 2-1. The "relative frequency distribution" in the last column could be graphed similarly; note that the two graphs are identical except for the vertical scale. Hence, a simple change of vertical scale transforms Figure 2-1 left side into a relative frequency distribution (right side).

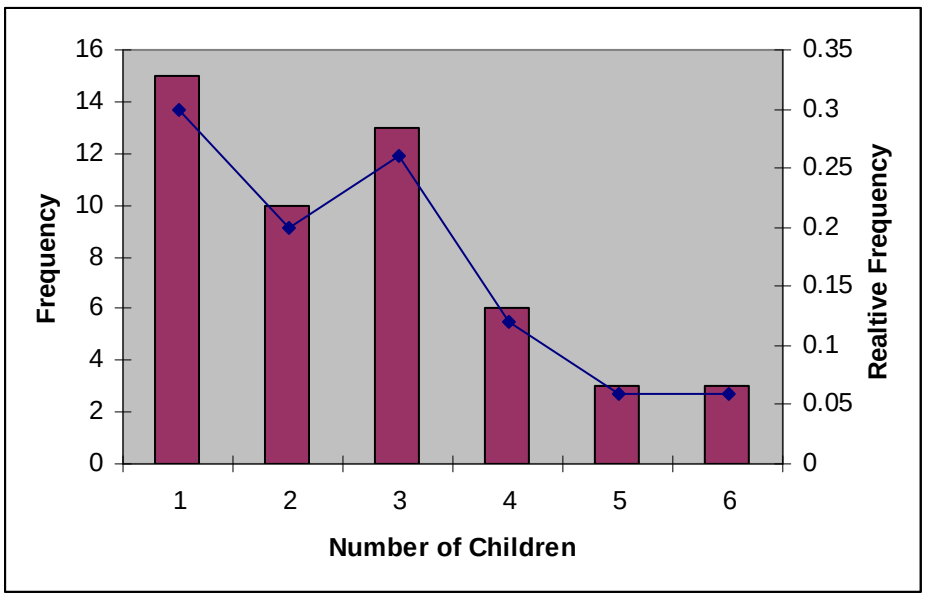


Figure 2-1

Continuous Example

Suppose that we take a sample of 200 men, each of whose height is recorded in inches. We call height X a "continuous" random variable, since an individual's height might be any value, such as 64.328 inches. It no longer makes sense to talk about the frequency of this specific value of X, since never again will we observe anyone who is exactly 64.328 inches tall. Instead we can tally the frequency of heights within a class or cell (e.g., 58.5" to 61.5"), as in Table 2-2. Then the frequency and relative frequency are tabulated, as before.
We have chosen the cells somewhat arbitrarily, but with the following conveniences in mind:

(1)	*(2)*	*(3)*	*(4)*	*(5)*
Cell No.	*Cel Boundaries*	*Cel Midpoints*	*Frequency*	*Relative Frequency*
1	58.5-61.5	60	2	0.01
2	61.5-64.5	63	10	0.05
		66	48	0.24
		69	64	0.32
		72	56	0.28
		75	16	0.08
7	76.5-79.5	78	4	0.02

1. The number of cells is a reasonable compromise between too much detail and too little. Usually, 5 to 15 cells is appropriate.
2. Each cell midpoint, which hereafter will represent all observations in the cell, is a convenient whole number. The grouping of the 200 observations into cells is illustrated.

Table 2-2

The grouped data are graphed in Figure 2-2. We use bars to represent frequencies as a reminder that the observations occurred throughout the cell, and not just at the midpoint. Such a graph is called a bar diagram or histogram.

We next turn to the question of how we may characterize a sample frequency distribution with a single descriptive number. There are two very useful concepts: the first is the centre of the distribution, and the second is the spread. These concepts will be illustrated with the continuous distribution of men's heights; but their application to discrete distributions (such as family size) is even more straightforward.

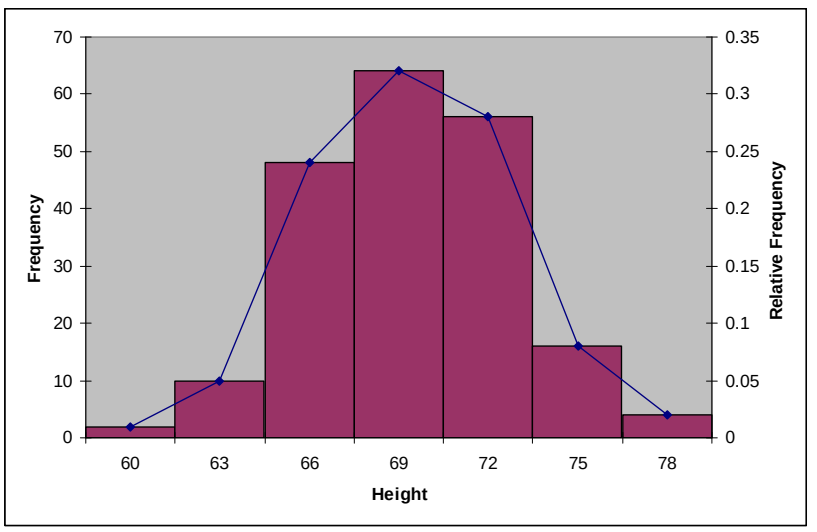


Figure 2-2

Centre of a Distribution

There are many different ways to measure the centre of distribution. Three of these—the mode, the median, and the mean—are discussed below, starting with the simplest.

- The Mode

Since mode is the French word for fashion, the mode of a distribution is defined as the most frequent (fashionable) value. In the example of men's heights, the mode is 69 inches, since this cell has the greatest frequency or highest bar in Figure 2-2. Generally, the mode is not a good measure of central tendency, since often it depends on the arbitrary grouping of the data. We also can draw a sample in which the largest frequency (highest bar in the group) occurs twice; this ambiguity is left unresolved, and the distribution is called "bimodal."

- The Median

The median is just the 50th percentile, i.e., the value below which 50% of the values in the sample fall. Since it splits the observations into two halves, it sometimes is called the middle value. In the sample of 200 detailed heights shown in Figure 2-2, the median (say, 69.3) easily is found by reading the 100th value from the left. Bui if the only information available is the grouped frequency distribution in Figure 2-2, the median can only be approximated, by choosing an appropriate value within the median cell.

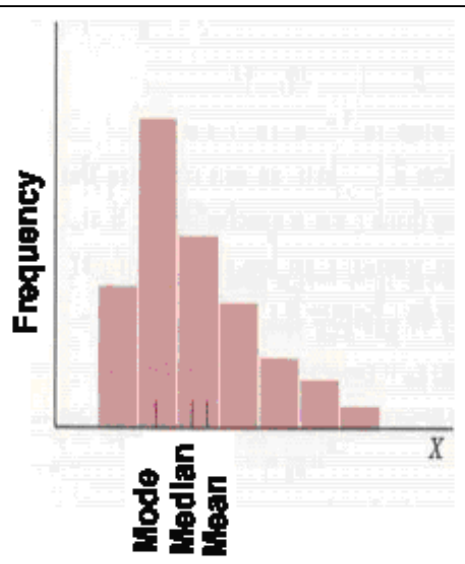


Figure 2-3

- The Mean

This sometimes is called the arithmetic mean, or simply the average, and is the most common central measure. The original observations (X_1, X_2, . . . , X_n,,,) simply are summed, then divided by n.

Comparison of Mean, Median, and Mode

These three measures of centre are compared in Figure 2-3. In Figure 2-2 we showed a distribution that has a single peak and is symmetric (i.e., one hall is the mirror image of the other); in this case, all three central measures coincide. But when the distribution is skewed to the right, as in Figure 2-3 the median falls to the right of the mode; with the long scatter of observations strung out in the right-hand tail, we have to move from the mode to the right to pick up half the observations.

Spread of a Distribution

Although average height may be the most important single statistic, it also is important to know how spread out or varied the observations are. As with measures of centre, we find that there are several measures of spread. We will start with the simplest.

- The Range

The range is simply the distance between the largest and smallest value: Range = largest — smallest observation

For men's heights, the range is 21 (i.e., 79.5-58.5). It may be criticized fairly on the grounds that it tells us nothing about the distribution except where it ends. And using only these two observations may be very unreliable. We therefore turn to measures of spread that take account of all observations.

- Mean Absolute Deviation (MAD)

The average deviation, as its name implies, is found by calculating the deviation of each observation from the mean; these deviations {Xi — Mean) then are averaged by summing and dividing by n. Although this sounds like a promising measure, in fact it is worthless; positive deviations always cancel negative deviations, leaving an average of zero. This sign problem can be avoided by ignoring all negative signs and taking the average of absolute values of the deviations:

$$MAD \equiv \frac{1}{n}\sum_{i=1}^{n}\left|X_i - \overline{X}\right| \qquad \text{where} \qquad \overline{X} \qquad \text{is the mean}$$

- Mean Squared Deviation (MSD)

Although MAD intuitively is a good measure of spread, it is mathematically intractable. We therefore turn to an alternative means of avoiding the sign problem, squaring each deviation:

$$MSD \equiv \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

- Variance and Standard Deviation

MSD is a good measure, provided that we only wish to describe the sample. But typically we shall want to go one step further and use this to make a statistical inference about the population. For this purpose it is better to use the divisor (n-1) rather than n

$$Variance, s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 \qquad \text{and the Standard Deviation} \qquad s \equiv \sqrt{Variance}$$

(Formula 2-1)

- Kurtosis

Kurtosis is based on the size of a distribution's tails. Distributions with relatively large tails are called "leptokurtic"; those with small tails are called "platykurtic." A distribution with the same kurtosis as the normal distribution is called "mesokurtic."

The following formula can be used to calculate the Kurtosis of a sample:

$$Kurtosis \equiv \frac{1}{(n-1)s^4}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4 - 3$$

The kurtosis for a standard normal distribution is 3. For this reason, most sources use the above definition of kurtosis, sometimes referred to as "excess kurtosis". This definition is used so that the standard normal distribution has a kurtosis of zero. In addition, with this definition positive kurtosis indicates a "peaked" distribution and negative kurtosis indicates a "flat" distribution.

- Skewness

Skewness is a measure of symmetry, or more precisely, the lack of symmetry. A distribution, or data set, is symmetric if it looks the same to the left and right of the centre point. Negative values for the Skewness indicate data that are skewed left and positive values for the Skewness indicate data that are skewed right. By skewed left, we mean that the left tail is long relative to the right tail, skewed right means that the right tail is long relative to the left tail.

The following formula can be used to calculate The Skewness of a sample:

$$Skewness \equiv \frac{1}{(n-1)s^3}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3$$

Probability

Introduction to Probability

In the next chapters, we make deductions about a sample from a known population. For example, if the population of American voters is 55% in favour of one candidate, we can hardly hope to draw exactly that same percentage in a random sample. Nevertheless, it is "likely" that "close to this percentage will turn up in our sample. Our objective is to define likely" and "close to" more precisely; in this way we shall be able to make useful predictions. First, however, we must lay a good deal of groundwork. Predicting in the face of uncertainty requires a knowledge of the laws of probability, and this chapter is devoted exclusively to their development. We shall begin with the simplest example - rolling dice - which was also the historical beginning of probability theory, several hundred years ago.

Concept of Probability

Suppose that a gambler has a die he suspects is loaded, and asks us the probability that it will come up an ace (one dot). One solution would be lo roll it over and over again, observing the relative frequency of aces is 1/6. Of course, rolling if five or ten tines would not be enough lo average out chance fluctuations. But over the long run, the relative frequency of aces would settle down to a limiting value, which is probability. That is:

Probability = proportion, in the long run.

or, more formally: $\Pr(e_1) = \lim \frac{n1}{n}$

where e_1 is the outcome ("ace")
n is the total number of times that the trial is repeated (die is thrown)
n_1 is the number of times that the outcome el occurs (also called the frequency, n_1 / n is therefore the relative frequency of e_1
lim is "the limit of . . . , as n approaches infinity."

Throughout this handbook, we shall continue to think of probabilities as proportions, because this is such a clear and intuitive concept. Strictly speaking, however the formula should be taken as a way to empirically determine or interpret probability.

Elementary Properties of Probability

We generalize by considering an experiment with N outcomes (ex, e_2, ..,e_i,…,e_N). The relative frequency n_i/n of any outcome e_i must be positive, since both the numerator and denominator are positive; moreover, since the numerator cannot exceed the denominator, relative frequency cannot exceed 1. So:

- $0 <= n_i/n <= 1$ and
- Any probability lies between 0 and 1
- The sum of all relative frequencies adds up to 1

Probability Distributions

Discrete Random Variables

Suppose that a couple is planning three children and is primarily interested in the number of boys. This is an example of a random variable and is usually denoted by a capital letter:
X = the number of boys
The possible values of X are 0, 1, 2, 3; however, they are not equally likely. To find the probabilities are, we must examine the original sample space. Thus, for example, the event "one boy" (X=1) consists of three of the equally probable outcomes and its probability is 3/8. Similarly the probability of each of the other events is computed

Original Sample Space

Pr (e)	e		
´1/8	G	G	G
´1/8	G	G	B
´1/8	G	B	G
´1/8	G	B	B
´1/8	B	G	G
´1/8	B	G	B
´1/8	B	B	G
´1/8	B	B	B

One Boy (B)

Smaller derived sample space

X	p (x)
0	´1/8
1	´3/8
2	´3/8
3	´1/8

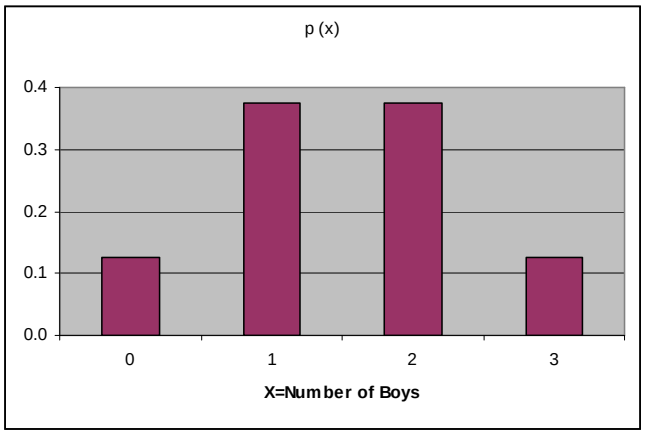


Figure 4-1 The random variable X = “number of boys”

So in Figure 4-1, we obtain the probability distribution of X. and one can state:
A discrete random variable takes on various values with probabilities specified by its probability distribution.

As shown in Figure 4-1, we begin in the original sample space by considering events such as (X — 0), (X = 1), . . , in general (X = x); note that capital X represents the random variable while small x represents a specific value that it may take. For these events we calculate the probabilities and denote them p(0), p(1), …, p(x), This probability distribution p(x) may be presented equally well in any of the customary forms for a function:

- Table form, as in the right-hand side of Figure 4-1.
- A Graph as in Figure 4-1
- A general formula, which we skip her for simplicity

In Figure 4-1, the original sample space (outcome set) is reduced to a much smaller and more convenient numerical sample space. The original sample space was introduced to enable us to calculate the probability distribution p(x) for the new space; having served its purpose, the old unwieldy space is then forgotten. The interesting questions can be answered very easily in the new space. For example from Figure 4-1, what is the probability of one boy or fewer? We simply add up the relevant probabilities in the new sample space:

$$\Pr(X \le 1) = p(0) + p(1) = 1/8 + 3/8 = 1/2$$

Mean and Variance

Notice the close relation between the relative frequency distribution observed and the probability distribution calculated in Figure 4-1 for planning 3 children: if the sample size were increased without limit, the relative frequency distribution would settle down to the probability distribution. This is an old story: relative frequency becomes probability in the limit.
From the relative frequency distribution, we calculated the mean $\bar{x}$ and the variance s^2 of the sample. It is natural to calculate analogous values from the probability distribution and call them the mean μ and variance σ^2 of the probability distribution p(x), or of the random variable X itself
So the population mean is $\mu \equiv \sum_x xp(x)$ (Formula 4-1)

and the population variance (Formula 4-2)

$$\sigma^2 \equiv \sum_x (x-\mu)^2 p(x)$$

We are following the usual custom of reserving Greek letters for ***population*** values. In Greek μ is the equivalent of m for mean, and σ is the Greek equivalent of s for standard deviation.
A clear distinction must be made between sample and population values: μ is called the population mean since it is based on the population of all possible repetitions of the experiment; on the other hand, we call $\bar{x}$ the sample mean since it is based

on a mere sample drawn from the parent population. Similarly, σ^2 and $\mathbf{s^2}$ represent population and sample variance, respectively. Since the definitions of μ and σ^2 are similar to those of mean $\mathbf{x,}^{-}$ and $\mathbf{s^{2,}}$ we find similar interpretations. We come to think of the mean μ as a weighted average using probability weights rather than relative frequency weights. The standard deviation σ is a measure spread, in a sense, a typical deviation.

Continuous Distributions

In Figure 2-2, we saw how a continuous quantity such as height could be nicely represented by a bar graph showing relative frequencies. This graph is reproduced in Figure 4-2 (a), below (with height now measured in feet, rather than inches; furthermore, the y-axis has been shrunk to the same scale as the x-axis.) The sum of all the relative frequencies (i.e., the sum of all the heights of the bars) in Figure 4-2 is of course 1, as we first noted in Table 2-2. We find it convenient to change the vertical scale to relative frequency density as in Figure 4-2 (b). This rescaling is designed specifically to make the total area equal to 1. We accomplish this by defining:

$$\text{relative frequency density} = \frac{\text{relative frequency}}{\text{cell width}}$$

$$= \frac{\text{relative frequency}}{1/4}$$

$$= 4(\text{relative frequency})$$

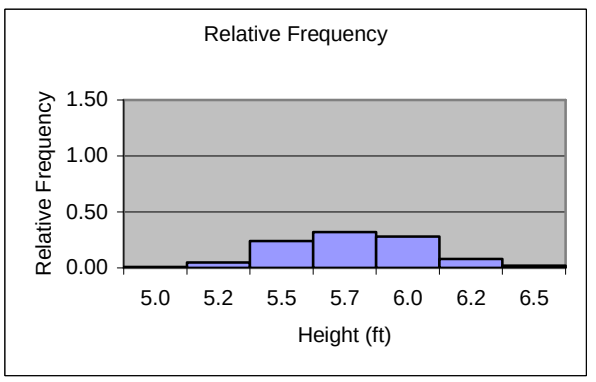


Figure 4-2 (a)

Total area of relative frequency density =1,
because 4* 0.25 (cell width) *1 =1

With a small sample, chance fluctuations influence the picture.
But as sample size increases, chance is averaged out.
And a relative frequency settles down to probabilities. At the same time, the increase in sample size allows a finer definition of cells. While the area remains fixed at 1 the relative frequency density becomes approximately a (red) curve, the so-called probability density function, which we shall refer to simply as the probability distribution, denoted by p(x).
II we wish to compute the mean and variance from Figure 4-2 (b) the discrete formulas (4-1) and (4-2) can be applied. But if we are working with the probability density function (the red line) then integration (which, as calculus students will recognize, is the limiting case of summation) must be used; if a and b are the limits of X, then the formulas will become

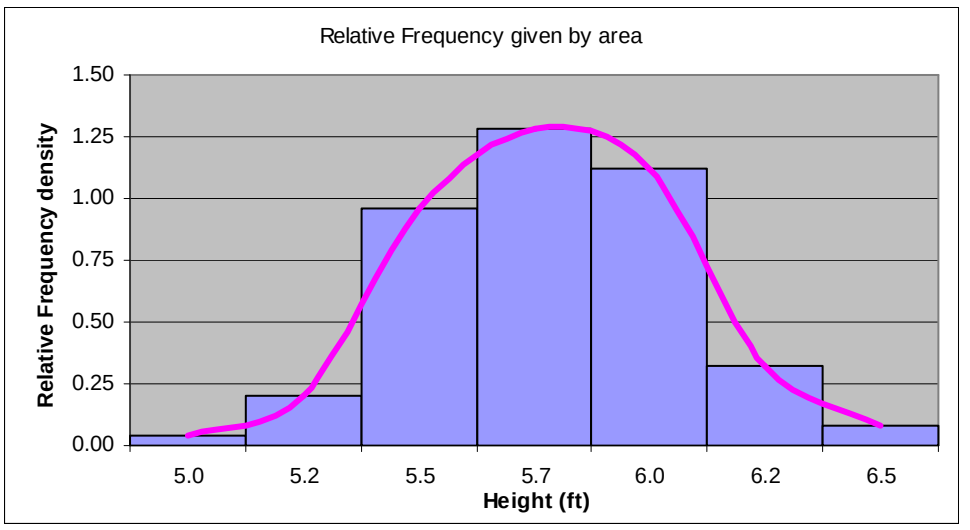


Figure 4-2 (b)

So the population mean $\mu = \int_a^b xp(x)d(x)$ becomes (Formula 4-3)

and the population variance $\sigma^2 = \int_a^b (x-\mu)^2 p(x)d(x$ (Formula 4-4)

All the theorems that we state about discrete random variables are equally valid for continuous random variables, with summations replaced by integrals. Therefore theorems are giving for discrete random variables only.

The Normal Distribution

For many random variables, the probability distribution is a specific bell-shaped curve, called the normal curve, or Gaussian curve. It is the most useful probability distribution in statistics. For example, errors made in measuring physical and economic phenomena often are distributed normally. In addition, many other probability distributions often can be approximated by the normal curve.

Standard Normal Distribution

A random variable Z is called standard normal if its probability distribution is:

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-(1/2)z^2}$$

The constant $1/\sqrt{2\pi}$ is a scale factor required to make the total area 1. The symbols "π" and "e" denote important mathematical constants, approximately 3.14 and 2.72 respectively. (Formula 4-5)

We draw the normal curve in Figure 4-3 to reach a maximum at z = 0; we confirm in (4-5) that this is so. As we move to the left or right of 0, z increases in the negative exponent; therefore p(z) decreases, approaching zero in both tails. This curve also is symmetric: since z appears only in squared form, -z generates the same probability in (4-5) as +z.
The mean and variance of Z can be calculated by integration using (4-3) and (4-4); since this requires calculus, we quote the results without proof:

$$\mu_z = 0$$
$$\sigma_z = 1$$

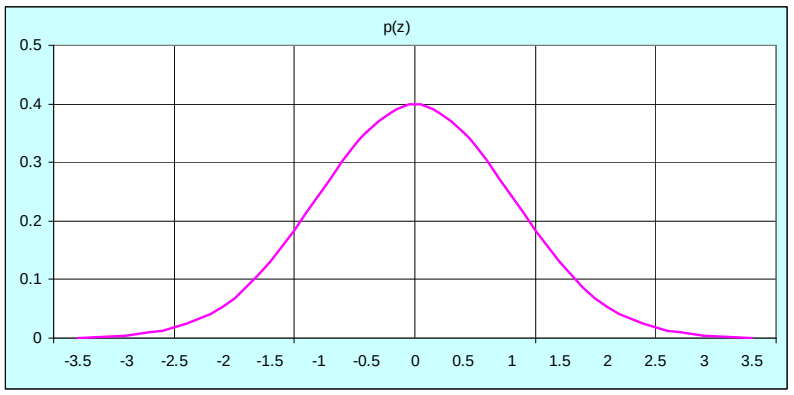


Figure 4-3

It is for this very reason, in fact, that Z is called a standard normal variable. Later when we speak of "standardizing" any variable, this is precisely what we mean: shifting it so that its mean is zero and rescaling it so that its standard deviation (or variance) is one.
The probability (area) enclosed by the normal curve above any specified value z_0 also requires calculus to evaluate precisely, but may be easily pictured, as in Figure 4-4. Without resorting to calculus you can think of this as accumulating the area of the approximating rectangles, as in Figure 4-2.

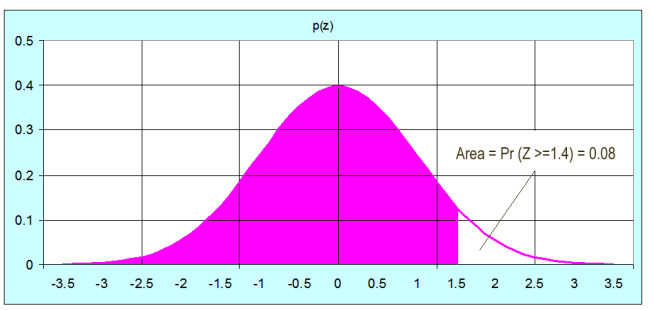


Figure 4-4

General Normal Distribution

In the previous section, we considered only a very special normal distribution the standard normal Z with mean 0 and standard deviation 1. Now consider the general form of the normal distribution, centred on any mean μ and with any standard deviation σ. It's probability distribution has the formula,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-(1/2)\left(\frac{x-\mu}{\sigma}\right)^2}$$

but for this introduction we will leave further explanations to individual studies and the mentioned textbooks.

Covariance and Correlation

Covariance

In this section, we shall develop a measure of how well two variables are linearly related. As an example, consider the joint distribution tabled in Figure 4-5. We notice some tendency for these two variables to move together: a large x tends to be associated with a large y, and a small x with a small y.
Our measure of how the variables move together should be independent of where the variables happen to be centred. We therefore consider the deviations from the mean:

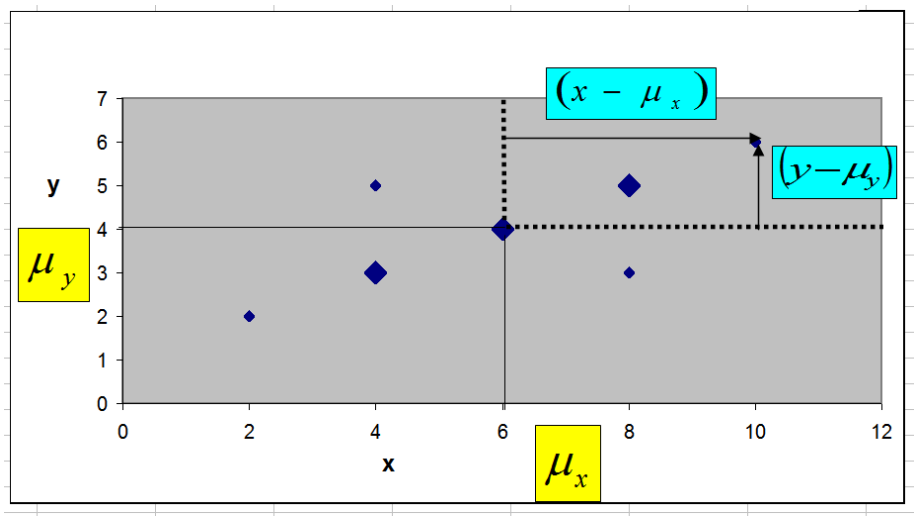


Figure 4-5

The Figure 4-5
Shows data points with two coordinates. Big diamonds mean 2 data points are located at this position or translation into relative frequency, their value is 0.2 whereas the expected frequency of the small diamonds is 0.1 (like always all p(x,y) adding up to 1) .
Graph of joint distribution p(x, y), showing new variables (X — μ_x) and (Y — μ_y) that translate axes into new dotted position (x - μ_x) and (y - μ_y)

Now let us multiply die deviations, obtaining the product:
(x - μ_x) (y - μ_y) (Formula 4-5)

For any point in the NE (North-East) quadrant of Figure 4-5, both deviations are positive, so their product is positive. This also holds for any point in the SW quadrant, since both deviations are negative. For points in the other two quadrants, the product is negative. We can obtain a good measure of how X and Y vary together if we sum all these products, attaching the appropriate probability weights to each. This is called the covariance:

$$\sigma_{xy} = \sum_x \sum_y (x-\mu_x)(y-\mu_y)p(x,y) \qquad \text{(Formula 4-6)}$$

For the distribution in Figure 4-5 (see data table on the right) the covariance is = +2

x	y
2	2
4	3
4	3
4	5
6	4
6	4
8	3
8	5
8	5
10	6

The covariance was positive in this case because the variables moved together; that is, the heavier probabilities occurred in the NE and SW quadrants. I f the heavy probabilities had occurred in the other two quadrants, the covariance would have been negative, indicating the tendency for X and Y to move in opposite directions. Finally, if the probabilities had been evenly distributed in all four quadrants, the covariance would have been zero, indicating no tendency for X and Y to move together.

Correlation

The covariance still can be improved. As it now stands, it depends upon the units in which X and Y are measured. If X were measured in feet instead of inches, each x-deviation in (4-5) and hence the covariance itself, would unfortunately change by a factor of 12. To eliminate this difficulty, consider a modified concept called the correlation ρ

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \qquad \text{(Formula 4-6)}$$

This does, in fact work, measuring X in terms of feet rather than inches still changes the numerator by a factor of 12, but this is exactly cancelled by a change in the denominator by the same factor 12. Since (4-6) similarly neutralizes any change in the scale of Y, the correlation coefficient ρ is, as desired, completely independent of the units of measurement of either variable. Another reason that ρ is a very useful measure of the relation between X and Y is that it is always bounded by:

$-1 < \rho < +1$

Whenever X and Y have a perfect positive linear relation (as would occur if the entire distribution in Figure 4-5 were located on a straight line with positive slope), then ρ takes on the limiting value of +1. If there is a perfect negative linear relation, then ρ would be -1. To illustrate these bounds, we calculate ρ for the data in Figure 4-5.

$$\rho_{xy} = \frac{2}{\sqrt{5.6}\sqrt{1.4}} = .71$$ which is indeed less than 1.

Finally, we must ask how correlation and independence are related. In probability theory, to say that two events are (statistically) independent intuitively means that the occurrence of one event makes it neither more nor less probable that the other occurs. For example:

The event of getting a 6 the first time a die is rolled and the event of getting a 6 the second time are independent.
By contrast, the event of getting a 6 the first time a die is rolled and the event that the sum of the numbers seen on the first and second trials is 8 are dependent.

An important theorem states:

If X and Y are independent, then they are uncorrelated, that is σ_{xy} and ρ_{xy} are 0

Sampling

Up to now we have studied probability and random variables so that we can now answer the basic deductive question in statistics: What can we expect of a random sample drawn from a known population?

Random Sampling

We already have considered several examples of sampling: the poll of voters sampled from the population of all voters; the sample of light bulbs drawn from the whole production of bulbs; a sample of men's height drawn from the whole population; and a sample of two chips drawn from a bowl of chips. In cases such as these, the sample is called random if each individual in the population is equally likely to be sampled. For example, suppose that a random sample is to be drawn from the population of students in the classroom. There are several ways to actually carry out the physical process of random sampling.

1. The most graphic method is to record each person on a cardboard chip, mix all these chips in a large bowl, and then draw the sample.
2. A more practical method is to assign each person a number, and then draw a random sample of numbers. For example, suppose that a random sample of 12 students is to be drawn from a class (population) of 100 students. By counting off, each student can be assigned a different 2-digit number. Then 12 such numbers can be read out of a table of random digits as generated by the computer.

These two sampling methods are mathematically equivalent. Since the random number method is simpler to employ, it is common in practical sampling. However, the bowlful of chips is easier conceptually; consequently, in our theoretical development of random sampling, we shall often visualize drawing chips from a bowl.

We cut here matters short, since we are not so much interested in statistics theory (however, if you are, please consult the recommended text books) but me may retain the

Conclusion
We may restate the definition of simple random sampling in more mathematical terms for future reference:
A simple random sample is a sample whose n observations X1, X2, . . . , Xn are independent. The distribution of each Xi is the population distribution p(x) (with mean μ and variance σ^2).

The exception to this is sampling from a small population, without replacement. This case, which is more difficult, will not be dealt with here but should be referred to in the text books. Everywhere else, we shall assume simple random sampling.

As a first important finding: We have deduced the behaviour of a *sample* mean from knowledge of the *population.*
For example, suppose that a sample of *n* = 4 observations is drawn from the population of heights as in Figure 4-2 (b) 1. Then *X* would fluctuate around: mean $\mathbf{x,}^{-}$ and the variance $\mathbf{s^2}$

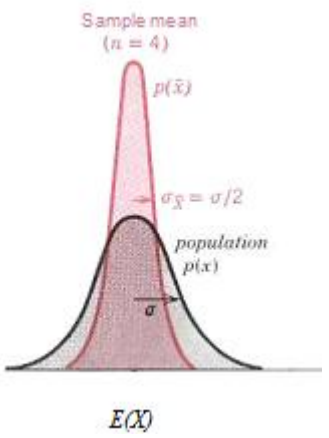
$E(\mathbf{x,}^{-}) = \mu$ E means the expected value of ..

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{4}} = \frac{\sigma}{2}$$

with a standard deviation for the sample of size 4:

The distribution of *X* is shown in Figure 5-1. It is also intuitively clear: *X* fluctuates around the same central value as an individual observation, but with less deviation because of "averaging out."
A concrete view of "averaging out" may be helpful: we might get a seven-foot man as an *individual* observation from the population, but we would be far less likely to get a seven-foot *average* in a sample of four men. This is because any seven-foot man that appears in the sample will likely be partially cancelled by a short man; or at least his effect will be diluted, because he is averaged in with other (more typical) men.
Since it is very important to distinguish between the distribution of the sample mean *X* and the population distribution, we introduce two conventions:

Figure 5-1



1. Since the population "gives birth" to the sample, we shall speak of the population distribution as the parent distribution. The distribution of X is then called a derived distribution or a sampling distribution.
2. In the diagrams in the handbook, colour is reserved for samples and sampling distributions. In contrast, parent populations are shown in grey. This convention first appears in Figure 5-1, where the distribution of the sample mean X is shown in colour, while the distribution of the parent population is shown in grey.

The Central Limit Theorem

In the preceding section, we found the mean and standard deviation of $\mathbf{x},^{-}$. Now we shall investigate the shape of its distribution.

The Distribution of $x,^{-}$ from a Normal Population

There is a very important theorem about linear combinations of normal variables, (without proof)

> If X and Y are normal, then any linear combination Z =aX + bY is also a normal random variable.

To see how this theorem will answer questions about sampling, suppose we have a parent population that is normal. Then each observation in the sample $X_1, X_2, \ldots, Xn$ has this same normal distribution.. Since the sample mean $\mathbf{x},^{-}$ is a linear combination of these normal variables, the theorem establishes that $\mathbf{x},^{-}$ is normal.

The Distribution of $x,^{-}$ from a Non-normal Population

On can show with experiments for a non-normal population; that how the distribution of the sample mean changes shape as sample size n increases. The sample mean becomes approximately normally distributed as n grows, no matter what the parent population is. This is especially remarkable for a skewed population which eventually generates the symmetric normal distribution for the sample mean. This pattern is so important that mathematicians have formulated it as:

> The central limit theorem: As the sample size n increases, the distribution of the mean $\mathbf{x},^{-}$ of a random sample taken from practically any population approaches a normal distribution (with mean μ, and standard deviation $\sigma / \sqrt{n}$).

The central limit theorem is not only remarkable, but very practical as well. For it completely specifics the distribution of $x,^{-}$ in large samples, and is therefore the key to large-sample statistical inference. In fact, in most cases when the sample size n reaches about 10 or 20, the distribution of $\mathbf{x},^{-}$ is already is practically normal. The proof of this theorem requires a very heavy mathematical background, and so we omit it.

Confidence Intervals and t-Test

So far, we considered various point estimators, except for the very beginning in chapter one. For example, we concluded that $\bar{x}$ was a good estimator of μ for populations that are approximately normal. Although on average $\bar{x}$ is on target, however, the specific sample mean $\bar{x}$ that we happen to observe is almost certainly a bit high or a bit low. Accordingly, if we want to be reasonably confident, that our inference is correct, we cannot claim that μ is precisely equal in the observed $\bar{x}$. Instead, we must construct an interval estimate or confidence interval of the form:

μ = $\bar{x}$ ± a sampling error

The crucial question is: How wide must this allowance for sampling error be? The answer, of course, will depend on how much $\bar{x}$ fluctuates (i.e., on the sampling distribution of $\bar{x}$), which we review in Figure 6-1 .
First we must decide how confident we wish to be that our interval estimate is right—that it does indeed bracket μ. It is common to choose 95% confidence; in other words, we will use a technique that will give us, in the long run, a correct interval 19 times out of 20.
To get a confidence level of 95%, we select the smallest range under the normal distribution of $\bar{x}$ that will just enclose a 95% probability. Obviously, this is the middle chunk, leaving 2.5% probability excluded in each tail. From standard tables, we find that this requires a z value of 1.96.
That is, we must go above and below the mean by 1.96 standard deviation of $\bar{x}$, as shown in Figure 6-1. The standard deviation of $\bar{x}$ (also called the standard error) is denoted by $\sigma_{\bar{x}}$, so that we may write:

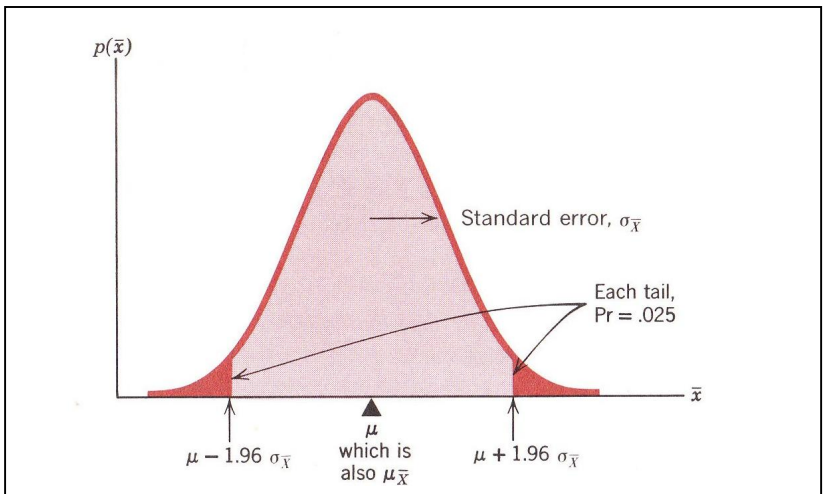


Figure 6-1 Normal distribution of the sample mean around the fixed but unknown parameter μ. 95% of the probability is contained within 1.96 standard errors

Pr ($\mu - 1.96\ \sigma_{\bar{x}} < \bar{x} < \mu + 1.96\ \sigma_{\bar{x}}$) = 95%
or turned around (Formula 6-1)
Pr ($\bar{x} - 1.96\ \sigma_{\bar{x}} < \mu < \bar{x} + 1.96\ \sigma_{\bar{x}}$) = 95%

We must be exceeding careful not to misinterpret the second part of (6-1). μ has not changed its character in the course of this algebraic manipulation. It has not become a variable but has remained a population constant. Equations (6-1) are probability statements about the random variable $\bar{x}$, or more precisely, the "random interval". It is this interval that varies and not μ.

In the previous sections, we assumed that, in constructing a confidence interval, the statistician knows the true population standard deviation. In this section, we consider the more typical case in which he does not.

Since σ is unknown, the statistician who wishes to evaluate the confidence interval (95%) must use some estimator of σ. The most obvious candidate is the sample standard deviation s (note that s, along with $\bar{x}$, always can be calculated from the sample data). Substituting s into the standard formula (remember 1-1), he estimates the 95% confidence interval for as the generalized formula:

$\mu = \bar{X} \pm z_{.025}\frac{s}{\sqrt{n}}$ where the z-value is 1.96 obtained from the table of Standard Normal Cumulative Probability

Provided that his sample is large (50 or 100), depending on the precision required), this will be an accurate enough approximation. But if the sample size is small, this substitution introduces an appreciable source of error. Therefore, if the statistician wishes to remain 95% confident, his interval estimate must be broadened. How much?
Recall that $\bar{x}$ has a normal distribution; when σ was known, we formed the standardized normal variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

By analogy we introduce “Student’s t” variable .

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

The similarity of these two variables immediately is evident. The only difference is that Z involves σ, which usually is unknown; but t involves s, which always can be calculated from the sample. The distribution of t is similar to the normal distribution. The t

distribution has a wider spread than the normal, of course, since the use of s instead of σ introduces additional uncertainty. Moreover, while there is one standard normal distribution, there is a whole family of t distributions.
Note that with small sample size, the t distribution differs substantially from the normal; but as sample size increases, it approaches the normal. The distribution of t is not tabled according to sample size n, but rather according to the divisor in s^2, which now is called "degrees of freedom" as in calculation the Variance s^2 in Formula 2-1. So for a sample size n we calculate the expected frequency and thus the confidence intervals in general terms for the population mean based on a sample:

$\mu. = \bar{x} + \bar{t}_{.025}$ (estimated standard error = $s/\sqrt{(n)}$) (Formula 6-2)

where $t_{.025}$ is the critical t value leaving 2.5% of the probability in the upper tail, with n — 1 degrees of freedom.
To sum up, we note the similarity of t estimation in (6-2) and normal estimation ($\mu. = \bar{x} + \bar{z}_{.025}$ ($\sigma/\sqrt{(n)}$). The only difference is that the observed sample value s is substituted for σ, and as a consequence the critical t value must be substituted for the critical z value.
An important practical question is: When do we use the t distribution and when do we use the normal? If σ is known, the normal distribution is appropriate; if σ is unknown, then the t distribution is appropriate — regardless of sample size. However, if the sample size is large, the normal is an accurate enough approximation of the t. So in practice, the t distribution is used only for small samples when σ is unknown and the normal is used otherwise.

Hypothesis Testing

Traditionally hypothesis testing has been treated as a separate topic in a statistics courses. It is closely related, however, to the interval estimates we just discussed in the previous chapters. Therefore, this section starts with hypothesis testing as a rewording of confidence intervals

Hypothesis Testing Using Confidence Intervals

In general, any hypothesis that lies outside the confidence interval may be judged implausible or rejected. On the other hand, any hypothesis that lies within the confidence interval may be judged plausible, or acceptable. So:

A confidence interval may be regarded as just the set of acceptable hypotheses.

Example[1]:
At a large American university in, the male and female professors were sampled independently, yielding the following annual salaries (in ten-thousands of dollars, rounded):

Men (X1)	*Women* (X2)
12, 20	9
11, 14	12
19, 17	8
16, 14	10
22, 15	16
$\bar{x}_1 = 16$	$\bar{x}_2 = 11$

These sample means give a rough estimate of the underlying population means μ_1 and μ_2. Perhaps they can be used to settle the following argument.
A husband claims that there is no difference between the salary means that is, if we denote the difference as H, he claims that: H = 0, his wife, however, claims that the difference is as large as seven thousand dollars: H=7

The calculation (cut short) of the 95% confidence interval is being used, and with the t-value for 95% = 2.16 came up with following results. The following formula is the 95% confidence interval for two means in independent samples when population variances are equal and unknown. So it translates to the Hypothesis:
$H = (\bar{x}_1 - \bar{x}_2) \pm t_{.025} * s_p \sqrt{(1/n_1 + 1/n_2)}$
...
= 5.0 ± 2.16(1.87)
= 5.0 ± 4.0

Thus, with 95% confidence, H is estimated to be between 1 and 9. Thus the claim A = 0 seems implausible, because it falls outside this confidence interval

[1] from D. A. Katz, "Faculty Salaries, Promotions, and Productivity at a Large University" American Economic Review, June 1973.

Since a 95% confidence interval is being used, it would be natural to speak of an hypothesis as being tested at a 95% confidence level. In abiding to convention, however, we use 5% -the complement of 95% - and simply call it the level of the test. Thus, we formally conclude that the hypothesis H = 0 is rejected at the 5% level. In other words, we see that there is sufficient data (a small enough sampling error) to allow us to discern (observe) a ***real*** difference. We therefore call this difference statistically discernible at the 5% level.

In summary, if a confidence interval already has been calculated, then it can be used immediately, without any further calculations, to test any hypothesis**.**

The hypothesis H = 0 in is of particular interest; since it represents no difference whatsoever, it is called the null hypothesis Ho. In rejecting it because it lies outside the confidence interval, we establish the important claim that there is indeed a difference between men's and women's income. Such a result traditionally has been called statistically significant at the 5% significance level.

There is a problem with this terminology. When the term "statistical significance" is used in this way, it simply means that enough data have been collected to establish that a difference does exist. It does not mean that the difference is necessarily important. For example, in another test based on very large samples from nearly identical populations, the 95%-confidence interval, instead of the example might be:

H = .005 ± .004

This difference is so miniscule that we could dismiss it as being of no real interest, even though it is statistically as significant as before. In other words, statistical significance is a technical term with a far different meaning than ordinary significance. Unfortunately but understandably, many people tend to confuse statistical significance with ordinary significance. To reduce the confusion, we prefer the word "discernible" to the word "significant." In conclusion, therefore, the traditional phrase "statistically significant at 5% significance level" technically means exactly the same thing as "statistically discernible at the 5% level." The latter phrase is preferable, because it is less likely to be misinterpreted.

What is the Prob-Value?

In the previous section, we developed a simple way to test any hypothesis by examining whether or not it falls within the confidence interval. Now we shall take a new perspective by concentrating on just one hypothesis— the null hypothesis Ho. We shall calculate just how much (or how little) it is supported by the data.

Prob-value (also called P-value or Probability- value) = Pr(X would be as large as the value actually observed, if Ho were true)

In general, for any hypothesis being tested, we define the Prob-value for Ho as:

The lower the Prob-value, the less likely the result is if the null hypothesis is true, and consequently the more "significant" the result is, in the sense of statistical significance.

The Prob-value is an excellent way to summarize what the data says about the credibility of Ho.

Cluster Analysis

Source: STATSOFT Statistics Textbook - STATISTICA Data Analysis Software and Services

General Purpose

The term cluster analysis (first used by Tryon, 1939) encompasses a number of different algorithms and methods for grouping objects of similar kind into respective categories. A general question facing researchers in many areas of inquiry is how to organize observed data into meaningful structures, that is, to develop taxonomies. In other words cluster analysis is an exploratory data analysis tool which aims at sorting different objects into groups in a way that the degree of association between two objects is maximal if they belong to the same group and minimal otherwise. Given the above, cluster analysis can be used to discover structures in data without providing an explanation/interpretation. In other words, cluster analysis simply discovers structures in data without explaining why they exist.

We deal with clustering in almost every aspect of daily life. For example, a group of diners sharing the same table in a restaurant may be regarded as a cluster of people. In food stores items of similar nature, such as different types of meat or vegetables are displayed in the same or nearby locations. There are countless number of examples in which clustering plays an important role. For instance, biologists have to organize the different species of animals before a meaningful description of the differences between animals is possible. According to the modern system employed in biology, man belongs to the primates, the mammals, the amniotes, the vertebrates, and the animals. Note how in this classification, the higher the level of aggregation the less similar are the members in the respective class. Man has more in common with all other primates (e.g., apes) than it does with the more "distant" members of the mammals (e.g., dogs), etc.

Statistical Significance Testing

Note that the above discussions refer to clustering algorithms and do not mention anything about statistical significance testing. In fact, cluster analysis is not as much a typical statistical test as it is a "collection" of different algorithms that "put objects into clusters according to well defined similarity rules." The point here is that, unlike many other statistical procedures, cluster analysis methods are mostly used when we do not have any a priori hypotheses, but are still in the exploratory phase of our research. In a sense, cluster analysis finds the "most significant solution possible." Therefore, statistical significance testing is really not appropriate here, even in cases when p-levels are reported (as in k-means clustering).

Area of Application

Clustering techniques have been applied to a wide variety of research problems. Hartigan [42] provides an excellent summary of the many published studies reporting the results of cluster analyses. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. In general, whenever we need to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

General Logic

The example in the General Purpose Introduction illustrates the goal of the joining or tree clustering algorithm. The purpose of this algorithm is to join together objects (e.g., animals) into successively larger clusters, using some measure of similarity or distance. A typical result of this type of clustering is the hierarchical tree.

Joining (Tree Clustering)

Hierarchical Tree

Consider a Horizontal Hierarchical Tree Plot (see graph below -Dendogram), on the left of the plot, we begin with each object in a class by itself. Now imagine that, in very small steps, we "relax" our criterion as to what is and is not unique. Put another way, we lower our threshold regarding the decision when to declare two or more objects to be members of the same cluster.

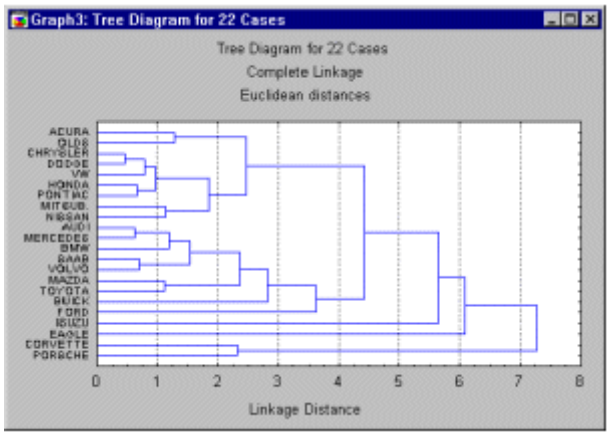


Figure 6-2 A Dendogram

As a result we link more and more objects together and aggregate (amalgamate) larger and larger clusters of increasingly dissimilar elements. Finally, in the last step, all objects are joined together. In these (Dendogram) plots, the horizontal axis denotes the linkage distance. Thus, for each node in the graph (where a new cluster is formed) we can read off the criterion distance at which the respective elements were linked together into a new single cluster. When the data contain a clear "structure" in terms of clusters of objects that are similar to each other, then this structure will often be reflected in the hierarchical tree as distinct branches. As the result of a successful analysis with the joining method, we are able to detect clusters (branches) and interpret those branches.

Distance Measures

The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items. In the previous example the rule for grouping a number of diners was whether they shared the same table or not. These distances (similarities) can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping objects. For example, if we were to cluster fast foods, we could take into account the number of calories they contain, their price, subjective ratings of taste, etc. The most straightforward way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances. If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e., as if measured with a ruler). However, the joining algorithm does not "care" whether the distances that are "fed" to it are actual real distances, or some other derived measure of distance that is more meaningful to the researcher; and it is up to the researcher to select the right method for his/her specific application.

Euclidean distance. This is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. It is computed as:

distance(x,y) = {Σi (xi - yi)2 }½

Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers). However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed. For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different. Generally, it is good practice to transform the dimensions so they have similar scales.
There are other less common distance measures to be mentioned here:

Squared Euclidean distance: distance(x,y) = Σi (xi - yi)2

City-block (Manhattan) distance: distance(x,y) = Σi |xi - yi|

Chebychev distance : distance(x,y) = Maximum|xi - yi|

Power distance:. distance(x,y) = (Σi |xi - yi|p)1/r

Percent disagreement: distance(x,y) = (Number of xi $\neq$yi)/ i

Amalgamation or Linkage Rules

At the first step, when each object represents its own cluster, the distances between those objects are defined by the chosen distance measure. However, once several objects have been linked together, how do we determine the distances between those new clusters? In other words, we need a linkage or amalgamation rule to determine when two clusters are sufficiently similar to be linked together. There are various possibilities: for example, we could link two clusters together when any two objects in the two clusters are closer together than the respective linkage distance. Put another way, we use the "nearest neighbors" across clusters to determine the distances between clusters; this method is called single linkage. This rule

produces "stringy" types of clusters, that is, clusters "chained together" by only single objects that happen to be close together. Alternatively, we may use the neighbors across clusters that are furthest away from each other; this method is called complete linkage. There are numerous other linkage rules such as these that have been proposed.

Single linkage (nearest neighbour). As described above, in this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters. This rule will, in a sense, string objects together to form clusters, and the resulting clusters tend to represent long "chains."

Complete linkage (furthest neighbour). In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors"). This method usually performs quite well in cases when the objects actually form naturally distinct "clumps." If the clusters tend to be somehow elongated or of a "chain" type nature, then this method is inappropriate.
Other Linkage Rules will only be mentioned here by name:

Unweighted pair-group average.

Weighted pair-group average.

Unweighted pair-group centroid.

Ward's method.

Two-Way Joining

Introductory Overview

Previously, we have discussed this method in terms of "objects" that are to be clustered (see Joining (Tree Clustering)). In all other types of analyses the research question of interest is usually expressed in terms of cases (observations) or variables. It turns out that the clustering of both may yield useful results. For example, imagine a study where a medical researcher has gathered data on different measures of physical fitness (variables) for a sample of heart patients (cases). The researcher may want to cluster cases (patients) to detect clusters of patients with similar syndromes. At the same time, the researcher may want to cluster variables (fitness measures) to detect clusters of measures that appear to tap similar physical abilities.

Two-Way Joining

Given the discussion in the paragraph above concerning whether to cluster cases or variables, we may wonder why not cluster both simultaneously? Two-way joining is useful in (the relatively rare) circumstances when we expect that both cases and variables will simultaneously contribute to the uncovering of meaningful patterns of clusters. For example, returning to the example above, the medical researcher may want to identify clusters of patients that are similar with regard to particular clusters of similar measures of physical fitness. The difficulty with interpreting these results may arise from the fact that the similarities between different clusters may pertain to (or be caused by) somewhat different subsets of variables. Thus, the resulting structure (clusters) is by nature not homogeneous. This may seem a bit confusing at first, and, indeed, compared to the other clustering methods described (see Joining (Tree Clustering) and k-Means Clustering), two-way joining is probably the one least commonly used. However, some researchers believe that this method offers a powerful exploratory data analysis tool

k-Means Clustering

General Logic

This method of clustering is very different from the Joining (Tree Clustering) and Two-way Joining. Suppose that you already have hypotheses concerning the number of clusters in your cases or variables. You may want to "tell" the computer to form exactly 3 clusters that are to be as distinct as possible. This is the type of research question that can be addressed by the k-means clustering algorithm. In general, the k-means method will produce exactly k different clusters of greatest possible distinction. It should be mentioned that the best number of clusters k leading to the greatest separation (distance) is not known as a priori and must be computed from the data (see Finding the Right Number of Clusters).

Example

In the physical fitness example (see Two-way Joining), the medical researchers may have a "hunch" from clinical experience that their heart patients fall basically into three different categories with regard to physical fitness. They might wonder whether this intuition can be quantified, that is, whether a k-means cluster analysis of the physical fitness measures would indeed produce the three clusters of patients as expected. If so, the means on the different measures of physical fitness for each cluster would represent a quantitative way of expressing the researchers' hypothesis or intuition (i.e., patients in cluster 1 are high on measure 1, low on measure 2, etc.).

Computations

Computationally, you may think of this method as analysis of variance (ANOVA) "in reverse." The program will start with k random clusters, and then move objects between those clusters with the goal to 1) minimize variability within clusters and 2) maximize variability between clusters. In other words, the similarity rules will apply maximally to the members of one cluster and minimally to members belonging to the rest of the clusters. This is analogous to "ANOVA in reverse" in the sense that the significance test in ANOVA evaluates the between group variability against the within-group variability when computing the significance test for the hypothesis that the means in the groups are different from each other. In k-means clustering, the program tries to move objects (e.g., cases) in and out of groups (clusters) to get the most significant ANOVA results.

Interpretation of Results

Usually, as the result of a k-means clustering analysis, we would examine the means for each cluster on each dimension to assess how distinct our k clusters are. Ideally, we would obtain very different means for most, if not all dimensions, used in the analysis. The magnitude of the F values from the analysis of variance performed on each dimension is another indication of how well the respective dimension discriminates between clusters.

Introduction to statistical regression

In the previous examples of statistical inference, we estimated the mean of a single population and we compared two population means. Finally, we might compare r population means, using analysis of variance (which we will skip here). Now we ask whether we could improve the analysis if we are able to rank the r populations numerically rather than in unordered categories.
We can use the analysis of variance to show how wheat yield depends on several different kinds inputs (like irrigation or fertilizer) If we wish to consider how yield depends on several different amounts of fertilizer, we define fertilizer application on a numerical scale. If we plot the yield Y that follows from various fertilizer applications, a scatter plot similar lo Figure 7-1 might be observed. From this scatter plot, it seems clear that fertilizer does affect yield. Moreover, it should be possible to describe how by an equation relating Y to X. Estimating an equation is, equivalent geometrically to fitting a curve through this plot.

This is called the statistical "regression" of Y on X.

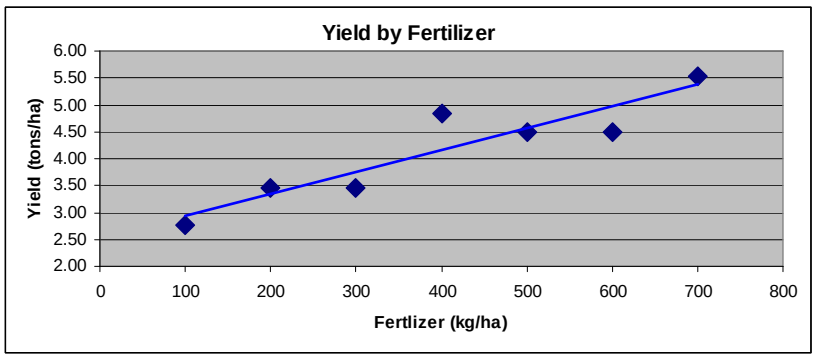
As a simple mathematical model, it will be useful as a brief and precise' description, or as a means of predicting the yield Y for a given amount of fertilizer X. Initially we restrict the discussion exclusively to how a straight line may best be fitted.

Table 7-1

X *Fertilizer (kg/ha)*	*Y* *Yield (tons/ha)*
100	2.768
200	3.460
300	3.460
400	4.844
500	4.498
600	4.498
700	5.536

Since yield depends on fertilizer, yield is called the "dependent variable" or "response variable" Y. Since fertilizer application is not depending on yield, but instead is determined independently by the experiment, we refer to it as an "independent variable" or "factor," or "regressor" X .

Figure 7-1



Example: Suppose, in a study of how wheat yield depends on fertilizer, funds are available
for only seven experimental observations. So the experimenter sets X at seven different values, taking only one observation Y in each case, as shown in Table 7-1. Graph these points, and roughly fit a line by eye in Figure 7-1. Of course it is not done by hand, but by the “Trend Line“ function of an EXCEL graph.

Fitting a line

It is time to ask, more precisely, "What is a good fit?" The answer surely is, "A fit that makes the total error small." One typical error (deviation) would be the vertical distance from the observed Y, to the fitted value Ŷi on the line, that is, (Yi - Ŷi).
We note that this error is positive when the observed Yi is above the line and negative when the observed Yi is below the line.
1. As our first tentative criterion, consider a fitted line that minimizes the sum of all these errors:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)$$

Unfortunately, this works badly. The problem is one of sign; in both cases, positive errors just offset negative errors, leaving their sum equal to zero.
There are two ways of overcoming the sign problem. The first is to minimize the sum of the absolute values of the errors:

$$\sum_{i=1}^{n}\left|Y_i-\widehat{Y}\right|_i$$

But perhaps it is not the best solution to the problem, because it pays no attention whatever to the sum of distances from the line to the observed points, that it does not force the line to be as close to the points as possible.
As a second way to overcome the sign problem, we finally propose lo minimize the sum of the squares of the errors:

$$\sum_{i=1}^{n}\left(Y_i-\widehat{Y}_i\right)^2$$

This is the famous "least squares" criterion; one of its justifications: Squaring overcomes the sign problem by making all error positive and it forces the line to be as close to the points as possible.

The reasoning her is very similar to that in the chapter “**Spread of a Distribution**” in part 1 of the Handbook

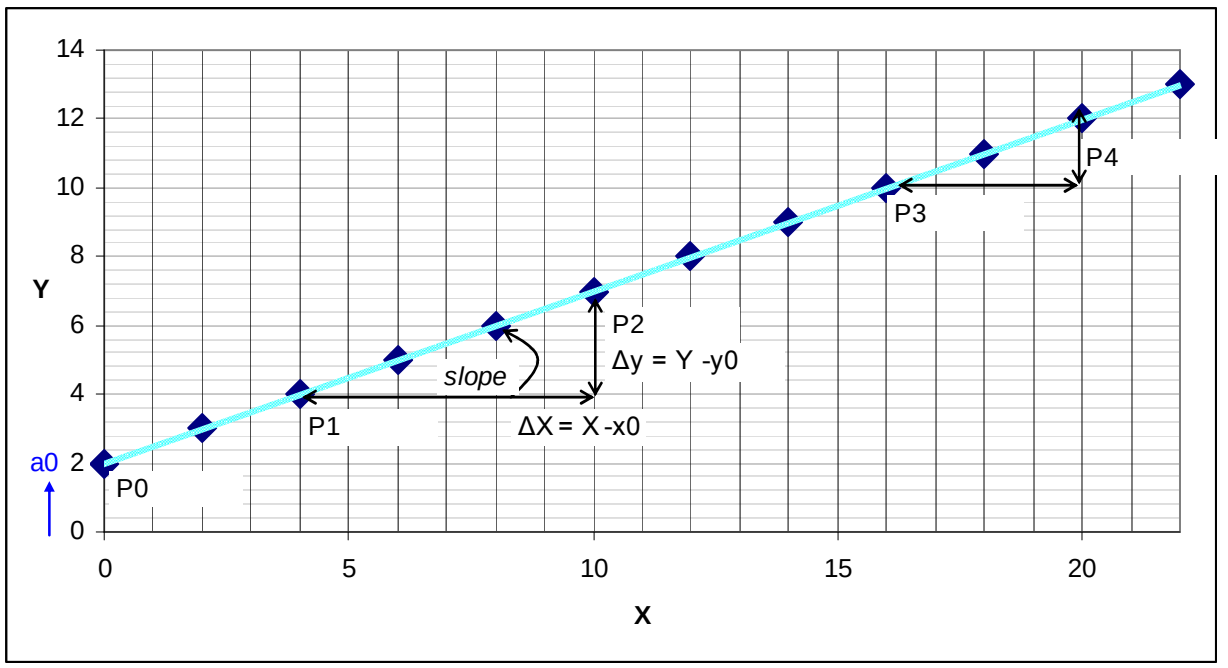
Lines and Planes; Elementary Geometry

The definitive characteristic of a straight line is that it continues forever in the same constant direction. In Figure 7-2, we make this idea precise. In moving from one point P1 to another point P2, we denote the horizontal distance by ΔX (where Δ means change, or difference), and the vertical distance by ΔY. Then the slope30 is defined as:

slope = ΔY/ ΔX

The characteristic of a straight line is that this slope remains the same everywhere: slope = ΔY/ ΔX = b (constant). For example, the slope between P3 and P4 is the same as between P1 and P2. It is now very easy to derive the equation of a line, if we know its slope *b* and any one point on the line. Suppose that the one point we know its P0 , the Y-intercept; since its coordinates, as shown in Figure 4-2 are 0 and a0 and the point is denoted *P0* (0,a0*)*. In moving to any other point *P(X,Y)*on the line, we may write: slope = ΔY/ ΔX =b= (Y-a0)/(X-0).

Figure 7-2



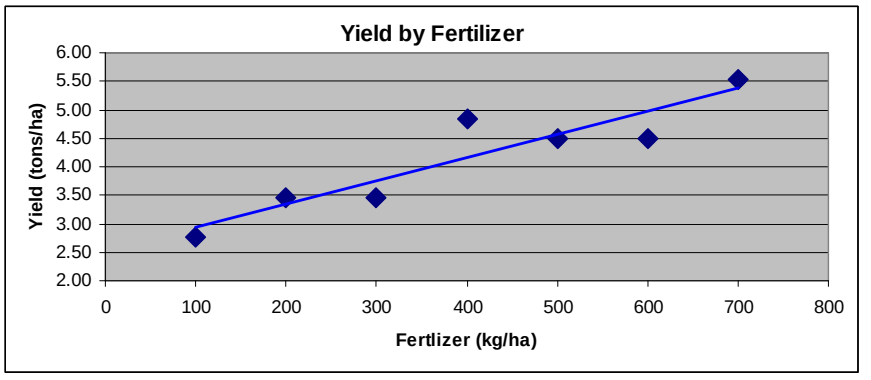
After transformation we arrive at the equation of a line Y=a0+bX, where a0 is the intercept and b the slope.

For planes in a 3-dimensional (X,Y, Z) space, we can do the same reasoning and the final equation of a plane is Y=a0+bX+ cZ, since we have two slopes b and c here. Multi-dimensional planes are mathematically constructed just correspondingly but obviously presentation in a sketch is not possible any more.

The least squares solution

The scatter of observed X and Y values from Table 7-1 is graphed again in Figure 7-3. Our objective is to fit a line: $\hat{Y}$ - a0 + bX The geometry of lines and planes, including the concepts of intercept and slope has been reviewed before The fitting of the line involves three steps:

Figure 7-3



Step 1: Translate X into deviations from its mean; that is, define a new variable x = X-$\bar{x}$. Measuring x as a deviation from $\bar{x}$ will simplify the mathematics because the sum of the new x values equals zero.

Step 2: Fit the line $\hat{Y}$ = a + bX by selecting the values for a and b that satisfy the least squares criterion; select those values of a and b that minimize $\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2$

Since each of the $\hat{Y}$ can be substituted, this means minimize: $\sum_{i=1}^{n}\left(Y_i-a-bx_i\right)^2$

With the use of calculus which is omitted here and can be reread in any of the textbooks, we arrive at

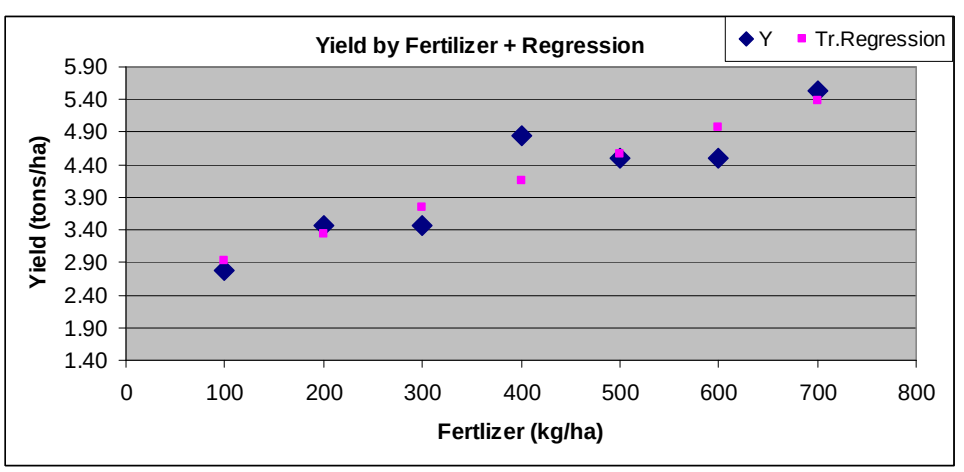
the result $a=\bar{Y}$ and $b=\dfrac{\sum_{i=1}^{n}x_iY_i}{\sum_{i=1}^{n}x_i^{\,2}}$

Step 3: If desired, the regression can now be translated back into the original frame of reference in terms of the original x values: x =(X- $\bar{x}$)

Table 7-2

X	*Y*	*xi*	*xiYi*	xi^2	*Regression*	*Tr.Regression*	$(Y- \hat{Y})$	$(Y- \hat{Y})^2$
100	2.768	-300.000	-830	90000	4.55964	2.92855	-0.16064	0.02580
200	3.460	-200.000	-692	40000	4.96741	3.33632	0.12357	0.01527
300	3.460	-100.000	-346	10000	5.37519	3.74410	-0.28421	0.08077
400	4.844	0.000	0	0	5.78296	4.15187	0.69198	0.47883
500	4.498	100.000	450	10000	6.19073	4.55964	-0.06178	0.00382
600	4.498	200.000	900	40000	6.59851	4.96741	-0.46956	0.22048
700	5.536	300.000	1661	90000	7.00628	5.37519	0.16064	0.02580
Mean		**Sums**			**Intercept and slope**		Σ	0.851
400.000	4.152	0	1142	280000	a = 4.152	2.521	s2=(1/(n-2))*Σ	0.170
					b = 0.0041	0.0041	s	0.413

Figure 7-4



Without mathematical proof, we apply the calculation in Table 7-2 using the values from Table 7-1 and obtain the regression equation as $\hat{Y} = 4.152 + 0.0041*X$ (Formula Ex 7-1) or after Step3 of the translated regression as $\hat{Y} = 2.521 + 0.0041*X$ (Formula Ex 7-2) and not surprising the scatterplot of the Regression values will be the trend line if the translated regression values would be connected (Figure 7-4)

Regression Theory

So far, our treatment of a sample of points has only involved mechanically fitting a line. Now we wish to make inferences about the parent population from which this sample was drawn. Specifically, we must consider the mathematical model that allows us to construct confidence intervals and test hypotheses.

Simplifying Assumptions

Consider again the fertilizer-yield example in the previous chapter. Suppose that the experiment could be repeated many times at a fixed level of fertilizer x. Even though fertilizer application is fixed from experiment to experiment, we would not observe exactly the same yield each time. Instead, there would be statistical fluctuation of the Y values, clustered about a central value. We can think of the many possible values of Y forming a population; the probability distribution of Y for a given x we shall call p (Y/x). Moreover, there will be a similar probability distribution for Y at any other experimental level of x. There obviously would be great problems in analyzing populations peculiar and unique in their distributions and comparing them.

To keep the problem manageable, therefore, we make several assumptions about the regularity of the populations. We assume that:
1. The probability distributions p(Yi/xi) have the same variance σ^2 for all xi.
2. The means E (Y_i) lie on a straight line, known as the true (population) regression line:
$E(Y_i) = \mu_i = \alpha + \beta x_i$ The population parameter α and β specify the line; they are to be estimated from sample information.
3. The random variables Yi are statistically independent. For example, a large value of Y1 does not tend to make Y2 large; that is Y2. is "unaffected" by Y1.

It is useful to describe the deviation of Yi from its expected value or disturbance term e_i so that the model alternatively may be written as: $Y_i = \alpha + \beta x_i + e_i$ where the error term e_i has the mean = 0 and the variance of Y that is σ^2

The Nature of the Error Term

Now let us consider in more detail the "purely random" part of Y, the error or disturbance term e_i . Where does it come from? Why doesn't a precise and exact value of Yi follow, once the value of x is given? This reasoning is important for all statistical data collection. The error may be regarded as the sum of two components:

1. Measurement error. There are various reasons why Y may be measured incorrectly. In measuring crop yield, an error may result horn sloppy harvesting or inaccurate weighing. If the example is a study of the consumption of families at various income levels, the measurement error in consumption might consist of budget and reporting inaccuracies.
2. Stochastic error occurs because of the inherent irreproducibility of biological and social phenomena. Even if there were no measurement error, continuous repetition of an experiment using exactly the same amount of fertilizer would result in different yields; these differences are unpredictable, and are called "stochastic" or "random." They may be reduced by tighter experimental control —for example, by holding constant soil conditions, amount of water, etc. But complete control is impossible —for example, seeds cannot be duplicated.

In the social sciences, controlled experiments usually are not possible. For example, an economist cannot hold U.S. national income constant for several years while he examines the effect of interest rate on investment. Since he cannot neutralize extraneous influences by holding them constant, his best alternative is to take them into account explicitly, by regressing Y on x and the extraneous factors. This is a useful technique for reducing stochastic error; it is called "multiple regressions" and is discussed later.ˆ

The Gauss-Markov Theorem

The major justification for using the least squares method to estimate a linear regression is the following:

Gauss-Markov Theorem

Within the class of linear unbiased estimators of β, the least squares estimator $\widehat{\beta}$ has minimum variance (is most efficient).
Similarly, $\widehat{\alpha}$ is the minimum variance estimator of α.

This theorem is important because it requires no assumption about the shape of the distribution of the error term. No proof will be given here, please refer to the textbooks.
It must be emphasized that the Gauss-Markov theorem is restricted; it only applies to estimators that are both linear and unbiased. It follows that there may be a nonlinear estimator that has smaller variance than the least squares estimator. For example, to estimate a population mean, the sample median is a nonlinear estimator that has smaller variance than the sample mean for certain kinds of non-normal populations as we mentioned before.

The distribution of $\widehat{\beta}$

Now we ask about the shape of the distribution of $\widehat{\beta}$. Let us add (for the first time) the strong assumption that the Yi are normal. Since $\widehat{\beta}$ is a linear combination of the Yi, it follows that $\widehat{\beta}$ also will be normal. But even without assuming that the Yi are normal, we know that, as sample size increases, the distribution of $\widehat{\beta}$ usually will approach normality. This can be justified by a generalized form of the central limit theorem. Our objective is to develop a clear intuitive picture of how this estimator varies from sample to sample.

Confidence intervals and hypothesis tests for β

Standard Error of β

Now that we have established the normality of $\widehat{\beta}$, statistical inferences about β is in order. But first we have one remaining problem: the variance σ^2 about the population line is unknown and must be estimated. A natural estimator is to use the deviations about the fitted line:

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2$$

where Ŷi is the fitted value on the estimated regression line, that is $\widehat{Y}_i = \widehat{\alpha} + \widehat{\beta}x_i$
.
s2 often is referred to as "residual variance". When s^2 is substituted for σ2 we obtain the estimated standard error $s_{\widehat{\beta}} = s/\sqrt{\sum x_i^2}$
and we can make statistical inferences.

Confidence Intervals

We can derive the 95% confidence interval for $\widehat{\beta}$ easily, arriving at a result familiar from Interval estimation in the Handbook part 1
$\beta = \widehat{\beta} \pm t_{.025}s_{\widehat{\beta}}$ which generates a 95% confidence interval for the slope after substituting the above formula,

$\beta = \hat{\beta} \pm t_{.025} s / \sqrt{\sum x_i^{\,2}}$ where the degrees of freedom for t are, as always, the same as the divisor in s^2 : d.f. = n-2.

Using a similar argument for the intercept, we could easily derive: $\alpha = \hat{\alpha} \pm t_{.025} s / \sqrt{n}$

Finally we note that $\hat{\alpha}$ and $\hat{\beta}$ are normal and so the 95% confidence interval for the mean μ_0

$\mu_0 = \hat{\mu}_0 \pm t_{.025} s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2}}$ (Formula 7-3) and the predicting a single observed Yo, once again the best estimate is the point on the estimated regression line above mean μ_0 ,in other words, the best point prediction for Yo

$\hat{Y}_0 = \hat{\alpha} + \hat{\beta} x_0 = \hat{\mu}_0$

When we try to find the interval estimate for Yo, we will face all the problems involved in the interval for the mean μ_0. And we have an additional problem because we are trying to estimate only one observed Y, rather than the more stable average of all the possible Y's. Hence, to the previous variance, we now must add the inherent variance σ^2 of an individual Y observation. So the 95% prediction interval for an individual Y observation is

$Y_0 = \hat{\mu}_0 \pm t_{.025} s \sqrt{\frac{1}{n} + \frac{x_0^2}{\sum x_i^2} + 1}$ (Formula 7-4)

This is quite an amount of formulas at a time, but the estimates of predictions for regression functions and confidence intervals is elementary for understanding and applying the regression approach in statistics

Example of Interval estimates

In the previous section, we considered the broad aspects of the model namely, the position of the whole line, remember there were several assumed populations (determined by α and β). In this section, we shall consider two narrower problems:

(a) For a given value xo, what is the interval that will predict the corresponding mean value of Yo For example, in our fertilizer problem, we may want an interval estimate of the mean yield resulting from the application of 550 kg of fertilizer. (Note that we are not deriving an interval estimate of mean yield by observing repeated applications of fertilizer; in that case, we could apply the simpler technique of estimating a population mean with the sample mean) Instead we are observing only seven different applications of fertilizer, clearly a more difficult problem.)

(b) What is the interval that will predict a single observed value of Yo (referred to as the prediction interval for an individual Yo). Again using our fertilizer example, what would we predict a single yield to be from an application 550 kg of fertilizer? This individual value clearly is less predictable than the mean value' in (a). We now consider both in detail.

We will apply the formulas to the fertilizer example and the data from Table 7-2 and start to find a 95% interval for

(a) The mean wheat yield that we would obtain if we planted many plots (μ0).

(b) The wheat yield on just one plot (Yo):

We simply start to calculate:
xo = Xo — $\bar{x}$,
= 550 - 400 = 150 and substitute it into the equation of the estimated regression line of

Formula Ex 7-1: $\hat{\mu}_0$ = Ŷ0 = 4.152 + 0.0041*150 = 4.767 This is the regression equation

(a) For an interval estimate for μ0, substitute the above value into Formula 7-3 along with s2 and $\sum x_i^2$ from Table 7-2:

$$\mu_0 = 4.767 \pm 2.571(0.413)\sqrt{\frac{1}{7} + \frac{150^2}{280000}}$$

$$\mu_0 = 4.767 \pm 1.534$$

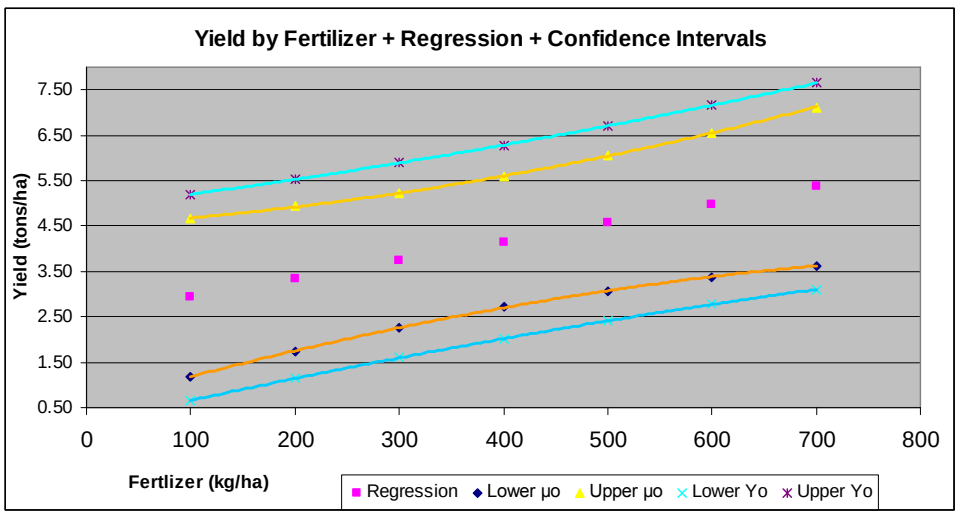
(b) For an interval estimate for Yo, yields the same calculation except for an extra 1 under the square-root sign

$$Y_0 = 4.767 \pm 2.571(0.413)\sqrt{\frac{1}{7} + \frac{150^2}{280000} + 1}$$

$$Y_0 = 4.767 \pm 2.167$$

This interval is almost 50% wider than (a), which shows how much more difficult it is to predict an individual observation than a mean.

Figure 7-5



The above calculated example is for one observation xo = 550. The relationship of prediction and confidence intervals is shown in Figure 7-5. The combined sources of error in a confidence intervals for the mean are shown in yellow, the wider, blue bands gives the prediction intervals for individual Y observations. Note how both bands expand as xo moves farther away from its central value (the mean) this reflects the fact xo^2 appears in both variances. This example gives reason for some general remarks about predictions based on statistical regression.

Dangers of extrapolation

We emphasize that, in Formulas (7-3) and (7-4), xo may be any value of x. If xo lies among the observed values x1 . . . xn, the process is called interpolation. If x0 is out beyond the observed values xl . . . xn, then the process is called extrapolation. The techniques developed in previous section may be used for extrapolation, but only with great caution, as we shall see.
There is no sharp division between safe interpolation and dangerous extrapolation. Rather, there is continually increasing danger of misinterpretation as x0 gets further and further from its central value.

Statistical Risk

We emphasized in the previous section that prediction intervals get larger as x0 moves away from the centre. This is true, even if all the assumptions underlying our mathematical model hold exactly.

Risk of Invalid Model

In practice, we must recognize that a mathematical model is never absolutely correct. Rather, it is a useful approximation. In particular, we cannot take seriously the assumption that the population means are strung out in an exactly straight line. If we consider the fertilizer example, it is likely that the true relation increases initially, but then bends down eventually as a "burning point" is approached, and the crop is overdosed. In the region of interest, from 0 to 700 kg, the relation is practically a straight line, and no great harm is done in assuming the linear model. However, if the linear model is extrapolated far beyond this region of experimentation, the result be comes meaningless. In such cases, a nonlinear model should be considered

Concluding observations

Two points warrant emphasis. First, most of the theory of this section and, in particular, the Gauss-Markov justification of least squares requires no assumption of normality of the error term. The one exception occurs when the normality assumption was required only for small sample estimation and this because of a quite general principle that small sample estimation requires a normally distributed parent population to strictly validate the t distribution. But even here, t is often a reasonably good approximation in non-normal populations.

Second, we have assumed that the independent variable x has taken on a given set of fixed values (for example, fertilizer application was set at certain specified levels). But in many cases, x cannot be controlled in this way. For example we are examining the effect of rainfall, we must recognize that x (rainfall) is a random variable that is completely outside our control.

The surprising thing is that most the findings of this section remains valid whether x is fixed or a random variable, provided that we assume that:

σ^2 (and α and β) are independent of x, amd the error term e is statistically independent of x

This greatly generalizes the application of the regression model

Multiple Regression

Introduction

Multiple regression is the extension of simple regression, to take account than one independent variable X. It is obviously the appropriate
Technique when we want to investigate the effects on Y of several variables simultaneously. Yet, even if we are interested in the effect of only one variable, it usually is wise to include the other variables influencing Y in a multiple regression analysis, for two reasons:

1. To reduce stochastic (hazard, random) error
2. Even more important, to eliminate bias that might result if we just ignored a variable that substantially affects Y.

Table 7-3

X *Fertilizer (kg/ha)*	*Y* *Yield (tons/ha)*	*Z* *Rainfall (Inches)*
100	2.768	10
200	3.460	20
300	3.460	10
400	4.844	30
500	4.498	20
600	4.498	20
700	5.536	30

Example: Suppose that the fertilizer and yield observations in our continuous were taken at seven different agricultural experiment stations across the country. If soil conditions and temperature were essentially the same in all these areas, we still might ask whether part of the fluctuation in Y (i.e., the disturbance term e) can be explained by varying levels of rainfall in different areas. A better prediction of yield may be possible if both fertilizer and rainfall are examined. The observed levels of rainfall are therefore given in Table 7-3, along with the original observations of yield and fertilizer.

The mathematical model

Yield Y now is to be regressed on the two independent variables, or "regressors": fertilizer X and rainfall Z. Let us suppose that the relation ship is of the form:

$$Y_i = \alpha + \beta x_i + \gamma z_i + e_i$$

Geometrically, this equation is a plane in the three-dimensional space**,** with the assumptions about e_i the **same as before. β** is interpreted geometrically as the slope of the plane as we move in the x-direction, keeping z constant. Similarly, γ is the slope of the plane as we move in the y-direction, keeping x constant. The least squares estimation is derived as for the simple linear regression.

The computer calculation leads to the results:

Yield = 1.944 + 0.003 Fertilizer + 0.058 Rainfall
With (Formula Ex 7-3)
Standard Errors

	0.0004	0.0107
t-values	6.532	5.401
95% CI	0.001	0.030

Let us consider some practical theoretical recommendations for the practical use of multiple regression

How many regressors should be retained?

For a Ho hypothesis (fertilizer does not improve yield, rainfall does not improve yield) the t ratios (t for rainfall = .833/.15 = 5.40)[2] for fertilizer and rainfall would lead us to reject Ho at 5% level to use the same example. We therefore should retain fertilizer and rainfall as statistically discernible variables (or, to use the traditional phrase, "statistically significant variables"); in this case, there are no problems.

But now suppose that we had weaker data (perhaps because of a smaller sample); accordingly, suppose that the standard error for rainfall Z was .55 (instead of .15). Then the t ratio would be t = .833/.55 = 1.51, which does not let us reject Ho at the 5% level. If we use this evidence to actually accept Ho (no effect of rainfall), and thus drop rainfall as a regressor, we may encounter the same difficulty that we discussed before in hypothesis testing. Since this is so important in regression analysis,

[2] We have cut short the comparison of t-values. You would have to consult a table of Student's t critical points to see that the Prob-value for a probability of <0.005 and 7 d.f. would be 3.499, so a value of 5.40 would lead to reject the Ho hypothesis, a value of 1.51 would not

let us review the argument briefly.

Although it is true that a t ratio of 1.51 for rainfall Z is statistically indiscernible, this does not prove that there is no relationship between Z and Y. It is easy to see why. We have strong biological grounds for believing that yield Y is positively related to rainfall Z. In as in (Formula Ex 7-3), this belief is confirmed by the positive coefficient y = .058. Thus our statistical evidence is consistent with our prior belief, even though it is a weaker confirmation than we would like. To actually accept the null hypothesis, and to conclude that Z does not affect Y, would be to contradict directly both the (strong) prior belief and the (weak) statistical evidence We would be reversing a prior belief, even though the statistical evidence weakly confirmed it. And this would remain true for any positive t ratio although, as t became smaller, our statistical confirmation would become weaker. Only if y is zero or negative do the statistical results contradict our prior belief.

It follows from this that, if we had strong prior grounds for believing that Z is related positively to Y, Z should not be dropped from the regression equation; instead, it should be retained, with all the pertinent information on its confidence interval, t ratio, etc.

On the other hand, what if our prior belief is that Ho is approximately true? Then the decision to drop or retain a variable would be; different. For example, a weak observed relationship (such as t = 1.51) would be in some conflict with our prior expectation of no relationship. But it is so minor a conflict that it is easily explained by chance (Prob-value <.01). Hence, resolving it in favour of our prior expectation and continuing to use Ho as a working hypothesis might be a reasonable judgment. Under these circumstances, this regressor would be dropped from the equation.

In the case of regression, there is another argument that may lead a statistician with very weak prior belief to accept Ho when the test yields a statistically indiscernible result: it keeps the model simple, and conserves degrees of freedom to strengthen tests on other regressors.

We conclude once again that classical statistical theory alone does not provide absolutely firm guidelines for accepting Ho; acceptance must In based also on extra-statistical judgment. Thus, prior belief plays a key role not only in the initial specification of which regressors should be in the equation, but also in the decision about which ones should be dropped m the light of the statistical evidence, as well as in the decision on how the model eventually will be used.

Prior belief plays a less critical role in the rejection of a hypothesis but it is by no means irrelevant. Suppose, for example, that although you be believed Y to be related to three variables, you didn't really expect it to related to a fourth; someone had just suggested that you "try on" a fourth at a 5% level. This means that if Ho (no relation) is true, there is a 5% chance of ringing a false alarm (and erroneously concluding that a relation does exist). If this is the only variable that is "tried on," then this is a risk that you can live with. However, if many similar variables are included in a multiple regression by someone who is "bag-shaking" (i.e. trying on everything in sight), then the chance of a false alarm increases dramatically. Of course, this risk can be kept small by reducing the level for each t test from 5% to 1% or less. This has led some statisticians lo suggest a 1% level with the variables just being "tried on," and a 5% level with the other variables that are expected to affect Y.

To sum up, hypothesis testing should not be done mechanically. It requires:

1. Good judgment and good prior understanding of the model being tested.

2. An understanding of the assumptions and limitations of the statistical techniques.

Interpretation of regression: "Other things being equal"

The coefficients in a linear regression model have a very simple but important interpretation, which we shall now consider.

Simple Regression Reviewed

Recall the simple regression model :

$$Y = \alpha + \beta x$$

(In this section we will ignore the error term e, since we are interested in interpreting βfor models with or without a stochastic error term.) It often is very useful to interpret β as

β = increase in *Y* if *x* is increased by one unit

For example, in the relation of wheat yield Y to fertilizer x, β is the increase in yield when fertilizer is increased one kg (called "marginal physical product" of fertilizer

To appreciate the linear model, it is useful to contrast it with a more complicated model, for example, the quadratic model:

$Y = a + \beta x + \gamma x^2$

When the marginal product of x is calculated as before, by increasing x one- unit, then: I his case, the marginal productivity of x is no longer simply the coefficient β. It also involves the coefficient γ, and the level xo. Thus, a major vantage of the linear model is that β has such a clear and direct interpretation

Multiple Regression

Consider again the multiple regression model:

$$Y = \alpha + \beta x + \gamma z$$

For example, wheat yield Y may depend on both fertilizer x and rainfall z. The interpretation of β now is:

β= the increase in Y if x is increased one unit, while z is held constant

For the general linear model:

$$Y = \alpha + \beta_1 x_1 + \beta_2 x_2 + .. + \beta_k x_k$$

It may be confirmed that the interpretation of each coefficient is similar:

β_i =the increase in Y if xi is increased one unit, while all other x variables are held constant

Factor Analysis

The basic idea of the factor analysis

In economic and social sciences one has to often to deal with indicators with a high degree of complexity, showing terms and multi relationships which cannot be expressed simply by an individual variable, whose values were determined by a question of a questionnaire or by another simple measurement.
Terms like „Creativity" , „Intelligence" , and „Qualification" or also apparently simpler terms like „Conjuncture" and „Living Standard" represent in each case for the expert a whole theory with a referred sentence of variables, which should be regarded in an appropriate representation of these terms.
In the empirical research such terms must be usually divided into many individual variables, because it is generally not possible to seize them appropriately by only one. The factor analysis follows this basic idea, goes however exactly the opposite way: Starting point of a factor analysis is a multiplicity of variables, from which one does not know a priori whether and in which way they have to do something with our terms or indicators.
With the factor analysis it is thought to be determined, whether in the examined variables there are groups of variables present, which can be identified with the terms, indicators or “background variables" like for examples to Example „ Creativity" , „Intelligence" , and „Qualification". Such background variables are called factors in the context of the factor analysis. It is the goal of each factor analysis, to reduce the high degree of complexity, represented by the multiplicity of variables. Like this it is thought to make them interpretable and manageably by making these variables represented by as few a factors as possible, which lay underneath the many heterogeneous and presumably related original variables.

An example: Study on premature infants

In n this introduction all steps of the factor analysis will be described by an example[3], which comes as a medical-psychological investigation another example in the WBT for an urban earthquake risk index.
The definition characteristic for a premature infant is not the length of the gestation t, but the birth weight of the child: Babies with a birth weight of < 1,500 g are considered as premature. For these children at the birth time different data were raised. In addition the children are observed and examined in different annual intervals in neurological and psychosocial context. It is therefore a longitudinal study. The analysis of the relationship between premature infant data on the one hand as well as neurological and psycho-social factors on the other hand became more important because of the medical development of the last decades, because at present approx. 65% of the premature infants survive, while in former times only approx. 15% did. This is due to stronger impact of intensive medicine development which probably has the price of a subsequent increased portion of neurological and psycho social disturbances in the child’s development. Apart from other questions with the investigation the group of researchers was interested whether there is a connection between certain prenatal data on the one hand and the development of intelligence on the other. To this purpose all premature infants were submitted to an investigation in its sixth year, whose results are held in 11 variables measuring the psycho-social development, like Columbia Mental Maturity Scale (http://psychology.wikia.com/wiki/Columbia_Mental_Maturity_Scale) . We will not describe in detail the 11 variables because the purpose of this section is to show how factor analysis works and not to find a solution to the briefly described problem.

The model of the factor analysis

Generally, neither the kind nor the numbers of factors are well-known in advance. For a didactical reason here the factors and their (possible) existence are communicated here, although they would be discovered as a result of the factor analysis. Thus the model of the factor analysis can be more easily described and reconstructed. Behind the 11 variables mentioned exist or (possibly) can be described the two following factors:

1. General intelligence (AI)
2. Linguistic intelligence (SI)

[3] *Veelken, Norbert* (1992): Entwicklungsprognose von Kindern mit einem Geburtsgewicht unter 1501 g. Eine regional repräsentative Studie über 371 Kinder. Habilitationsschrift Universität Hamburg.

This means that these two factors stand behind the eleven variables and it also means that they determine the variables and explain them in a scientific sense. Thus the connection between the factors and the variables - similar as with a regression analysis – can be described by a system of equations. If each of the eleven sample variables can be explained by the two Factors then an equation can be formulated for each variable, which describes this connection.
For the first variable the equation could reads as follows:

Var1 = a1 · AI + a2 · SI + Uvar1

Directly the similarity with a regression equation is noticeable: The factors AI and SI are to be regarded as explaining variables (predictors), by which the (dependent) variable *Var1* can be explained. The coefficients a1 and a2 correspond to the regression coefficients of a regression equation, and that third factor, Uvar1, corresponds to the residuals (or the errors) of an regression equation.

This error term must be taken up therefore to the equation, because generally it is not to be expected that the variable which can be explained (in this case *Var1*) completely by the remaining factors (in this case thus AI and SI is explained). Called in the factor analysis the error term (Uvar1), which is the remainder not explained by the factors which is called: “single residual factor”, The two other factors AI and SI become the “common factors” because they are used for the explanation of each variable contained in the model. This does not exclude that individual “common factors”, for different variables have only very small explanation content and thereby have only a very small influence. A small influence of a factor corresponds in an accordingly small value of the factor coefficient (a1 and/or a2).
The factors (in this example thus AI and SI) are not, as said before, known before but are determined by the factor analysis. The thought underlying the computation of factors is the following (simplifying here the rather partially quite complicated mathematical procedure):
First linear combinations of the observed variables are formed. For variables, which exhibit high correlation with another variable, it is assumed that have a common factor. By contrast variables, which correlate only weakly with one another, it is assumed that they do not have a factor in common. For the concrete different processes of estimation there are various methods of determination of the factors at the disposal.
in principle however all procedures determine the coefficients ci of the following equation, which characterizes and computes the relationship for the first factor AI.
AI = c1 · Var1 + c2 · Var2 +… + c11 · Var11

According to this equation it is formally possible that all eleven variables of the sample contribute to the explanation of the factor AI (general intelligence). The goal and the hope of the factor analysis consist however in that the factors are only determines by one part the variables. Accordingly a successful factor analysis thereby is characterized by the result that the multiplicity of the relevant variables in the sample is represented by only few factors. Nothing would be gained by factor analysis, if in the available example eleven factors were needed to characterize the relationship because then one also could work directly with the eleven variables.

Further a factor analysis can be regarded only then as successful, if the factors determined are meaningfully interpretable with regard to their content. This is one of the most difficult problems at the time of the execution of a factor analysis. This problem was covered so far thereby that already at the beginning two factors were introduced with the speaking names “General intelligence” and “Linguistic intelligence”. The factor analysis as such supplies only factors, which are called factor 1, factor 2 etc. It is the task of the analyst to interpret and them then if necessary meaningful names give these factors with regard to their content meaningful.

The four steps of a factor analysis

Usually the factor analysis is accomplished in four steps. This means however not that the factor analysis always and exclusively is accomplished in this way. Rather it is just as feasible with the StatistiXL procedure “Factor Analysis” to implement even a very complex factor analysis in only one run. On the other the numerous available options of the procedure “Factor Analysis” make it possible to evaluate the individual steps. The four usual steps of a factor analysis are the following:

Correlation Matrix

They correlation matrix is computed for all variables included into the factor analysis. From the correlation matrix can be read, which variables possibly should remain unconsidered because of their very small correlations with the other variables

Factor extraction

This step is called generally „pulling “or „extracting “of factors. There are different methods of the factor extraction and you must you indicate in the dialog fields of the procedure “Extraction”, which extraction method is to be used. Different statistical indicators, which can be defined in this step will point out whether the accepted factor model will be suitably is to represent the variables in a simple manner.

Rotation

The factors found in the second step are frequently difficult to interpret. In order to facilitate the interpretation, one makes itself the circumstance that the factors are artificial variables, which can be transformed distortion-free in such a way that they can be represented in different coordinate systems. By a suitable transformation one frequently succeeds, to point out more clearly out the connection of the factors to the observation variables and thus facilitates the interpretation of the factors.

Factor values

Although the factors can be regarded in a certain way as complex “background” variables, the substantial goal of a factor analysis can be reached in principle without ever determining concrete values of these “background”. On the other hand the goal of a factor analysis often consists to determine factors to discover variables not included into the factor analysis. Correspondingly you can use factors to explain (absent) variables, not included into the factor analysis. For these purposes you can do compute concrete factor values and if necessary store them as variable for further analysis

In the following sample file “Sol_PremInf.xls” will be used to explain the different steps of FA

Example: Correlation Matrix

Select in the “General” Folder the option “Correl/Covar” Matrix to print the correlation matrix of 11 variables.

Table 7-4 Matrix of the coefficients of correlation for the eleven sample variables

Correlation Matrix	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Var11
Var1	1.000	0.876	0.845	0.867	0.781	0.800	0.848	0.837	0.824	0.826	0.850
Var2	**0.876**	1.000	0.844	0.826	0.749	0.779	0.819	0.794	0.802	0.767	0.848
Var3	**0.845**	**0.844**	1.000	0.885	0.693	0.811	0.768	0.775	0.731	0.706	0.806
Var4	**0.867**	0.826	**0.885**	1.000	0.729	0.800	0.791	0.809	0.778	0.755	0.807
Var5	0.781	0.749	0.693	0.729	1.000	0.636	0.816	0.770	0.785	0.768	0.737
Var6	0.800	0.779	0.811	0.800	0.636	1.000	0.725	0.728	0.675	0.687	0.746
Var7	**0.848**	0.819	0.768	0.791	0.816	0.725	1.000	0.841	0.828	0.842	0.830
Var8	0.837	0.794	0.775	0.809	0.770	0.728	0.841	1.000	0.759	0.782	0.788
Var9	0.824	0.802	0.731	0.778	0.785	0.675	0.828	0.759	1.000	0.789	0.772
Var10	0.826	0.767	0.706	0.755	0.768	0.687	0.842	0.782	0.789	1.000	0.787
Var11	**0.850**	**0.848**	0.806	0.807	0.737	0.746	0.830	0.788	0.772	0.787	1.000

This correlation matrix gives a first overview of , which variables are strongly and which only weakly correlated with one another. It is about to recognize that between the variables Var1 and Var2 exists a relative strong correlation. Additionally both variables exhibit a relative to high correlation the variable Var3. On the other hand is each of the three variables only clearly more weakly also correlates to the variable Var11. If all proven coefficients of correlation would exhibit only very small absolute values, it would be a little meaningful, to continue factor analysis, since common factors exist only for such variables, which are relatively strongly correlated with one another. (The stronger correlations are marked in yellow, only the lower half of the matrix is marked like that because the matrix is symmetric).

Now we must refer to a bit more sophisticated software than we used in the WBT, like SPSS or STATS. So some of the results cannot be displayed in stastiXL but are important to understand the possibilities (and limits) of the factor analysis.

In the correlation matrix we observe numerous pairs of variables with relatively strong correlation. It is possible nevertheless that the calculated correlations appear only coincidentally in the sample, although in the population no connection between the variables exists and all coefficients of correlation have a value of zero. With the Bartlett “test on sphericity” the hypothesis can be tested, that all coefficients of correlation among the variables in the population have the value 0. The result of this test is shown in

Table 7-5

KMO- and Bartlett-Test		
Measure of Sample characteristics after Kaiser-Meyer-Olkin.		.897
Bartlett-Test on sphericity	Chi-Square	1281.153
df		55
Significance after Bartlett		.000

The test value of Bartletts test is a Chi square[4] value, which is extraordinarily high with 1281. Accordingly a significance value of 0,000 is calculated. This is to be interpreted in such a way that the hypotheses, all correlations between the eleven variables in the population are 0, can be rejected with a prob-value of 0,000. Turned around one can thus assume at least between some the eleven variables also in that Population correlations exist.

Another measure which appears here is the KMO (Kaiser-Meyer-Olkin.) value = .897
The KMO measure can take at a maximum the value 1. A value in close proximity to 1 is reached if the partial coefficients of correlation are very small. Contrary, if the KMO measure takes a small value with large partial coefficients of correlation. A small KMO value indicates that for a factor analysis the variable selection is not well chosen. The KMO measure for the eleven variables regarded in this example (here calculated with SPSS) was shown to be .897 in Table 7-5. If one accepts the evaluation scheme of the author of this indicator (Kaiser), the value with 0.897 is quite good, so that the selection of the variables for a factor-analytic model seems to be quite appropriate.

Evaluation of KMO after Kaiser [39]	
values 0,9 to 1.0	marvellous
0,8 to under 0,9	meritorious
0.7 to under 0,8	middling
0.6 to under 0,7	mediocre
0,5 to under 0,6	miserable
under 0,9 unacceptable	unacceptable

Before you finally accept the selected model, you should regard still the MSA values, which are displayed in the main diagonals the anti-image correlation matrix which is also the basis of the KMO (not displayed here). MSA is the abbreviation for Measure of Sampling Adequacy. The MSA values are in principle computed exactly like the straight descriptive KMO measure, with the difference that it refers only in each case to one variable instead of on all variables altogether: Without any presentation we just state that for the variable Var2 the MSA value is 0.925, which is to be regarded (after the evaluation of Kaiser) as marvellous. The smallest MSA value computed in the matrix amounts to 0.808 (Var9) and is still quite good. The MSA values do not offer then a cause to exclude one or more variables from the factor-analytic model.

Factor extraction

We come back to the factor analysis procedure in stastiXL. In the literature different procedures are suggested to compute the factors of a factor analysis, which is usually called factor extraction. Each of these procedures has its pro and cons. The most important of these procedures are outlined in brief. The most common, also used by most computer programs, is the procedure of the Principal Component analysis (PCA). This procedure often appears as a method on its own.

[4] The **Chi² statistic** (c^2) provides a means of testing the null hypothesis that an observed distribution has the same distribution as an expected. It is calculated as the sum of the squares of the difference between the observed and expected frequency, divided by the expected frequency *i.e.* $c^2 = S (f_i - \underline{f}_i)^2/\underline{f}_i$ where f_i is the observed frequency and $\underline{f}_i$ is the expected frequency.

PCA solves a problem similar to the problem of common factor analysis, but different enough to lead to confusion. It is no accident that common factor analysis was invented by a scientist (psychologist Charles Spearman) while PCA was invented by a statistician. PCA states and then solves a well-defined statistical problem, and except for special cases always gives a unique solution with some very nice mathematical properties. Briefly we will state similarities and differences

Similarities

PCA and FA have these assumptions in common:

- Measurement scale is interval or ratio level
- Random sample - at least 5 observations per observed variable and at least 100 observations.
- Larger sample sizes recommended for more stable estimates, 10-20 observations per observed variable
- Over sample to compensate for missing values
- Linear relationship between observed variables
- Normal distribution for each observed variable
- Each pair of observed variables has a bi-variate normal distribution
- PCA and FA are both variable reduction techniques. If communalities are large, close to 1.00, results could be similar.

PCA assumes the absence of outliers in the data. FA assumes a multivariate normal distribution when using Maximum Likelihood extraction method.

Differences

Principal Component Analysis	**Factor Analysis**
Principal Components retained account for a maximal amount of variance of observed variables	Factors account for common variance in the data
Analysis decomposes correlation matrix , using elements on the diagonals of the correlation matrix	Analysis decomposes adjusted correlation matrix. The diagonals of correlation matrix are adjusted with unique factors
Minimizes sum of squared perpendicular distance to the component axis	Estimates factors which influence responses on observed variables
Component scores are a linear combination of the observed variables weighted by eigenvectors	Observed variables are linear combinations of the underlying and unique factors

This is not to express however that the procedure of the main component analysis is superior to the remaining procedures.
With the procedure of the main component analysis linear combinations are formed of the variables. As the first main component (= factor) is that combination selected which explains the largest part of the total dispersion of all variables. The second main component is accordingly that, which explains the second largest part of the total dispersion of all variables. Formally as many main components or factors can be computed as many variables the model contains.
In the explained variance (Table 7-6) of the tabular overview of the factor extraction, there are also actually as many main components, as variables displayed.

In order to receive the following results choose the following the dialog field's selections:

Descriptive Statistics in folder "General"

Principal Components and *Extract All* in "Extraction"

Correlation Matrix and No Rotation in "Rotation"

Scree Plot in "Plots"

Table 7-6

Explained Variance (Eigenvalues)

Value	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7	Factor 8	Factor 9	Factor 10	Factor 11
Eigenvalue	8.885	0.533	0.250	0.248	0.225	0.202	0.179	0.146	0.130	0.106	0.096
% of Var.	80.769	4.845	2.276	2.257	2.047	1.839	1.625	1.327	1.181	0.961	0.873
Cum. %	80.769	85.614	87.890	90.147	92.194	94.034	95.658	96.985	98.166	99.127	100.000

The "Eigenvalue" of a factor indicates, which amount of the total dispersion of all variables of the factor model is explained by this factor. From this total dispersion the first factor (during the initial solution with altogether eleven factors) explains to 8.885 and thus 80.8% of the total dispersion. The second factor explains absolutely 0,533 or 4.9% and so on.
The column Cum to % indicates the cumulated part of the explained dispersion of the total dispersion. There is to be recognized that the first two factors already explain 86% of the total dispersion. Furthermore it is to be observed that the additional contribution of any the further factor rapidly decreases.

Determination of the number of factors

It became already clear that it would be senseless to consider as many factors in the model as there are variables. On the other hand the table 7-6 shows that the portion of the dispersions explained by the factors for individual variables with increasing factor number sinks. So the question arises, how many factors in the model are to be considered. This question cannot alone be decided on the basis of a rigid formula (You see the rather weak / flexible frame of the factor analysis). It is valid to select the number of factors by which a still sufficiently large part of the dispersions is explained, and at the same time a sufficiently large reduction of the complexity is obtained. For this reason we will select the two factors with an Eigenvalue > 0.5.

Frequently one consults the Screeplot (for the example shown in Figure 7-7). The diagram shows the factors in decreasing order of their Eigenvalues The Screeplot has the purpose and serves (to maintain in the picture) to separate the rubble, which accumulates at the foot of the mountain-slope, of those effective factors. Typically the curve of a Screeplot has the following characteristics, which can be recognized in Figure 7-7: First the curve drops very steeply, exhibits then rather soon a break, in order to drop in the further process only very slowly. As rule of thumb the recommendation is to select the number of factors at which the curve exhibits the break. On the basis the Screeplot of this example one would decide for a solution with two factors.

Figure 7-6

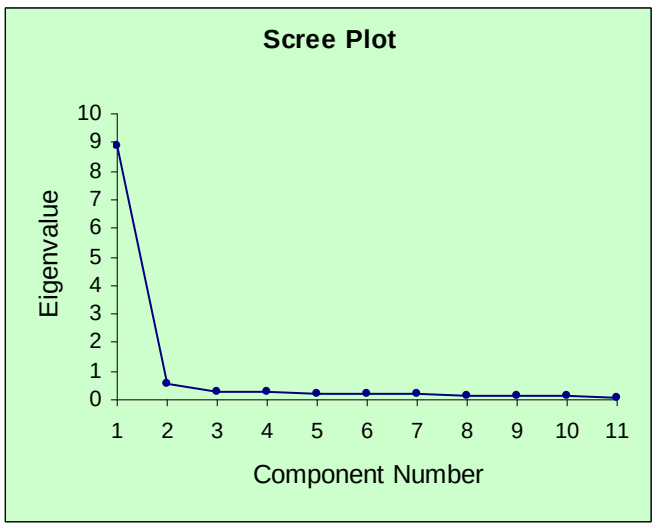


Table 7-7

Unrotated Factor Loadings		
Variable	Factor 1	Factor 2
Var1	0.947	-0.041
Var2	0.922	-0.093
Var3	0.897	-0.307
Var4	0.916	-0.200
Var5	0.855	0.329
Var6	0.848	-0.366
Var7	0.922	0.191
Var8	0.899	0.066
Var9	0.885	0.214
Var10	0.881	0.242
Var11	0.908	-0.029

Factor loadings

The encountered factors become only valuable for the analyst if their relations can be explained to the particulars variables. For this the factor matrix is used, which is shown for the available example in Table 7-7.

The factor matrix indicates the coefficients, with which the two factors enter into equation for the explanation of each of the variables of the factor model. These coefficients are mostly called factor loadings; accordingly the name of the matrix as factor load matrix. From the matrix itself we read for example for the variable Var1 that this can be described by the following equation:

Var1 = 0,947 · F1 - 0,041 · F2

From the absolute size of a factor load you can discover the impact / importance of the respective factor on the related variable. So factor 2 has a rather high impact on Var5, but only a very small one on Var1. In contrast to this Var1 is strongly “explained” by the first factor.

Different methods of factor extraction

For the factor extraction still other methods available than the one the method used so far: Principal component method.

These procedures differ in the approach in which the best model adjustment is reached:

Principal component method: This procedure is preset and was used in the example. This method is similar to principal component analysis (with which the principal component method should not be confused!) but the factor loadings are calculated as the principal component analysis solutions (eigenvectors) multiplied by the square root of the corresponding eigenvalue.

Principal Axis factor analysis: This procedure is quite similarly to the principal component method. The difference to this consists in that here squared multiple correlation coefficients are used as estimations of communalities in the diagonal of the correlation matrix in a first step. On this basis then suitable estimates of factors and the communalities are computed. These become in a second step the starting point of renewed factor estimation and new communalities etc. the iteration process continues until the communalities do not change any longer considerably.

Maximum Likelihood: With this method such parameters are estimated, for those the probability to maximize the observed

(sample) correlation matrix. It is assumed that the sample is of a parent multivariate normal distribution.
There are several other methods in other statistical packages but will be omitted here

Rotation

Purpose of the rotation

In the last step two factors were identified, which according to the criterion of their Eigenvalues have certain explanation strength. So far these factors became yet not interpreted, but were purely computational results. The factor analysis remains however without force of expression, if the meaning the factors cannot be determined. The factors must thus to be interpreted. The interpretation of the factors results from the relation, which they have to the observation variables, for which they should represent “background” variables. The relations of the factors to the individual variables can be interpreted from the factor load matrix: Large factor loads show a large, small on the other hand a small meaning of a factor for the appropriate variable on. A factor is relatively easy to interpret if some variables have a homogeneous meaning among them and have high loadings of this factor. On the contrary a meaningful interpretation is very difficult or impossible, if a factor shows relatively strong correlation to all variables of the model. Thus for example the factor loading matrix Tab 7-7 shows that the first factor has high factor loadings for all eleven variables of at least 0.8.
So this factor is hard to interpret, because it (apparently) explains many heterogeneous variables. Such a situation is not atypical for the first attempt a factor load matrix. To the easier interpretation different procedures have been developed of a rotation of the factor load matrix. The term rotation explains itself from that during the transformation the axes of the coordinate system, in that the factor loads are represented, are to be turned or rotated. This is demonstrated in the following for the two factors, their loadings represented for each value in a two-dimensional diagram (Figure 7-7)

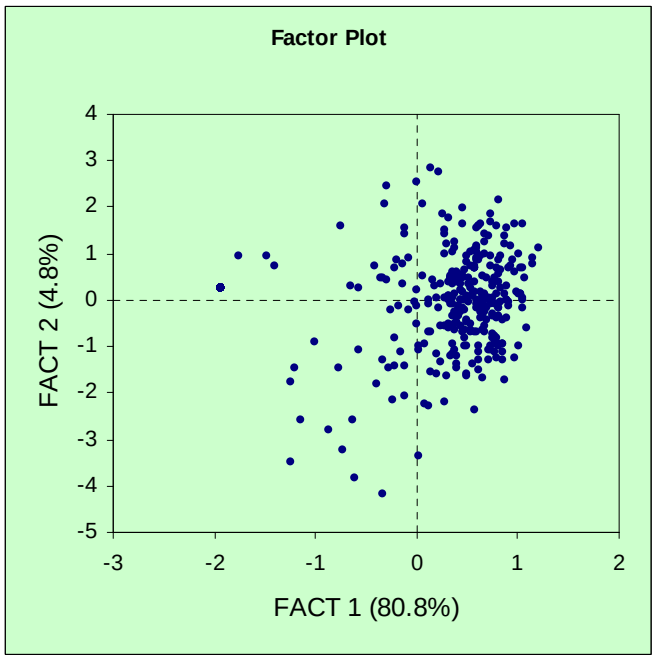


Figure 7-7

From the Figure 7-7 The graph of factor scores for Factor 1 and Factor 2 plots the case-wise factor scores and (can) include vectors which indicates the relative directions and magnitude (length of line) for the contributions of the variables to factors. (These are omitted here because of the many vectors to be plotted)

Rotation methods

Different rotation methods are available. StatistiXL has three procedures available of a orthogonal rotation. With the Varimax method the axes are rotated in such a way that the numbers of variables with high factor loadings are minimized. This is probably the most common procedure, by which above all allows to interpret the factors more easily.

Again we use an output from other software because it is not available in StatistiXL and it makes Rotation easier to understand.

If we look at the factor loading matrix before and after rotation it would look like this:

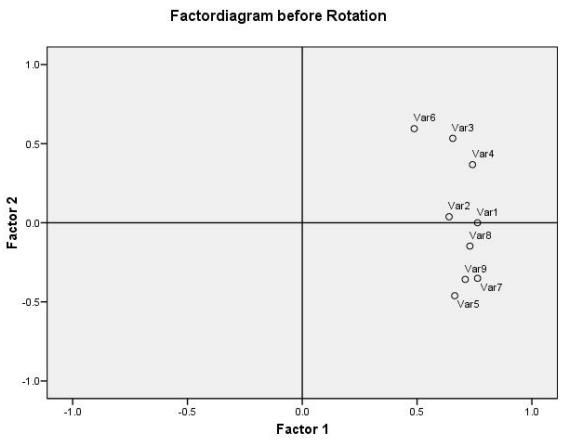


Figure 7-8

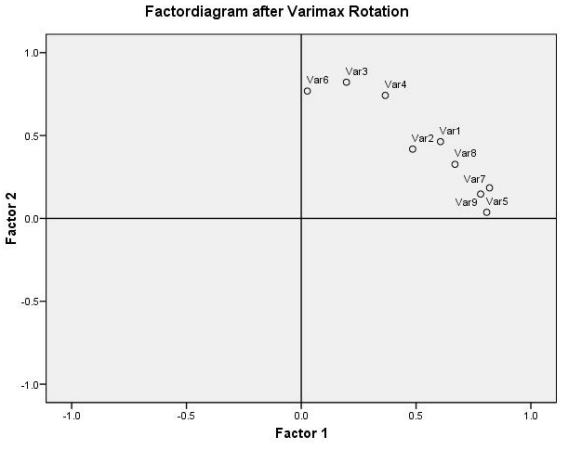


Figure 7-9

The interpretation of the factor-diagram before rotation: All variables have positive loadings for factor 1, positive and negative loadings for factor 2.The factors would be easy to interpret if all the variables would be grouped to one of the axes and if they would be in equal distance from the origin. This would mean that each variables have a high loading for the one factor and a low for the other. This is not the case.
If we could rotate the exes left by about 30 degrees, we would about reach the objective. Graphically the presentation is only feasible for two dimensions; mathematically it is possible for n coordinates.

Table 7-8

Rotated Factormatrix (Varimax)		
	Factor1	Factor2
Var7	0.81	
Var4	0.71	
Var3	0.68	
Var2	0.68	
Var8	0.59	
Var1	0.55	0.46
Var11	0.43	
Var5	0.40	
Var9		0.81
Var6		0.72
Var10		0.53

After the Varimax rotation the factor-diagram looks like Figure 7-9 and the results of the rotated factor matrix would read like the following Table 7-8. This processing of the factor coefficients facilitates the formal interpretation of the factors substantially, because a rather clear picture results (Factor loading with less than 0.4 after rotation have been cleared):

Those three variables Var6, Var9, Var10 have only for the factor 2 factor loadings > 0.4, the other variables only for the factor 1. The only exception is for Var1 which shows for both factors the loadings > 0,4 exhibits. The three variables Var6, Var9 and Var10 represent test results according to the linguistic Intelligence again. The other variables show test results to general intelligence. Factor 1 can be interpreted as the “background” variable general intelligence (AI), Factor 2 as linguistic Intelligence.

That is as far as we go in Factor Analysis: A final remark which is also interesting in the view how science follows fashion aspects, something which leads the way to theory of science, philosophy of science, and also in the sociology of scientific knowledge [41], which is far beyond the scope of this handbook:

A dubious history

If a statistical method can have an embarrassing history, factor analysis is that method. Around 1950 the reputation of factor analysis suffered from over-promotion by a few overenthusiastic partisans. In retrospect there were three things wrong with the way some people were thinking about factor analysis at that time. First, some people seemed to see factor analysis as the statistical method rather than a statistical method. Second, they were thinking in absolute terms about problems for which a heuristic approach would have been more appropriate. Third, they were thinking of overly broad sets of variables ("we want to understand all of human personality" rather than "we want to understand the nature of curiosity"). Thus in three different ways, they were attempting to stretch factor analysis farther than it was capable of going. In recent decades factor analysis seems to have found its rightful place as a family of methods which is useful for certain limited purposes…(from [40])

.

Bibliography

(Some personal remarks: Don’t take them too serious)

[1] Abelson, R.P. (1995). Statistics as Principled Argument. NJ: Lawrence Erlbaum Associates. THE DEFINITIVE Text to answer the question, "But, what does it mean?" from a social science research standpoint.

[2] Brown, F.L., Amos, J.R., & Mink, O.G. (1995). Statistical Concepts-A Basic Program. NY: HarperCollins College Publishers. A self-teaching text using the concept of programmed learning before programming had anything to do with computers. Covers all the concepts in a sequential manner, with helpful sections on Measurement, and the use of such computer programs as Minitab, SAS and SPSS. An appendix is also devoted to the use of these software applications.

[3] Baker, S. (2002). The Complete Idiot's Guide to Business Statistics IN: Alpha Books. Aside from various typos, this textbook does provide an understandable introduction to statistics as used in business settings.

[4] Best, J. (2001). Damned Lies and Statistics. Untangling Numbers from the Media, Politicians, and Activists. CA: University of California Press. If you really want to understand how statistics are abused and misused, then this is a must-read..

[5] Best, J. (2004). More Damned Lies and Statistics. How Numbers Confuse Public Issues. CA: University of California Press. Best asserts, "every piece of research contains limitations; researchers inevitably choose specific definitions, measures, designs, and analytic techniques. These choices are consequential; they shape every study's results"

[6] Bruce Bowerman, Richard O'Connell (2008), Business Statistics in Practice,, McGraw-Hill Higher Education A book for the business student with very day-to-.day related exercises

[7] Carlan, A. (1998). Everyday Math for the Numerically Challenged NY: Barnes & Nobles. Craft, J.L. (1990). Statistics and Data Analysis for Social Workers. 2nd Edition.

[8] Itasca,IL: F.E. Peacock Publishers. A straightforward, concise text about analyzing data. Diamond, I. & Jefferies, J. (2001). Beginning Statistics. An Introduction for Social Scientists. CA:Sage Publications. A good, easy-to-understand book about the basic principles of statistics.

[9] Foster, J.J. (2001). Data Analysis Using SPSS for Windows Versions 8 to 10. A Beginner's Guide. CA: Sage Publications. A great book for using SPSS. Best beginner's text. But you should have access to the software to make the most of this textbook.

[10] Freed, M.N., Ryan, J.M., & Kess, R.K. (1991). Handbook of Statistical Procedures and Their Computer Applications to Education and the Behavioral Sciences. NY: American Council on Education, Macmillan Publishing Company.

[11] Gardner, M.J., & Altman, D.G. (1989). Statistics with Confidence - Confidence Intervals and Statistical Guidelines. London:British Medical Journal. A good text on the use of confidence intervals with various statistical procedures, instead of just citing the p value in reporting research results.

[12] Gigerenzer, G. (2002). Calculated Risks. How To Know When Numbers Deceive You NY: Simon & Shuster MacMillan Co. An excellent book that introduces you to a better way of understanding statistics.

[13] Graham, A. (2003) Teach Yourself Statistics. McGrawHill Contemporary Books.

[14] Huff, D. (1954). How to Lie With Statistics. NY: W.W. Norton & Company. A classic work still worthwhile and easy to read

[15] Jaisingh.L. (2000). Statistics for the Utterly Confused. NY:McGraw Hill. It is obvious that statistics is an excellent subject for the continuous publications of books such as these.

[16] Kaplan M. & Kaplan, E. (2006). Chances Are... Adventures in Probability. NY: Penguin/Viking Group.

[17] Kranzler, G. & Moursund, J. (1995). Statistics for the Terrified. NJ: Prentice Hall. A "Read-Me-First" book for anyone who HAS TO take a statistics course.

[18] Langley, R. (1970). Practical Statistics. Dover books. Good basic text. Explains all the different significant tests, what data are required and how to interpret, with examples.

[19] Levitt, Steven D. & Dubner, Stephen J. (2005). Freakonomics. A Economist Explores the Hidden Side of Everything. NY:HarperCollins Publishers. It is actually a statistical look at modern society that no one ever bothered to do up to this point.

[20] Mlodinow, L. (2008). The Drunkard's Walk. How Randomneess Rules Our Lives NY: Vintage Books. The historical aspects of how many of the principles of probability and statistics were developed is told through biographical sketches of interesting personalities.

[21] Moore, D.S. (1991). Statistics Concepts and Controversies. 3rd Edition. NY: WH Freeman & Co. Taking a very liberal arts approach, Moore talks about statistics in the broadest sense about how it is used in a variety of disciplines. Very easy to understand if you want to understand why statistics is so important to know, from how it is used, misused, and the best ways to look at and present data, and appropriate ways to interpret statistics.

[22] Morgan, S.E., Reichert, T., & Harrison, T.J. (2002). From Numbers to Words. Reporting Statistical Results for the Social Sciences. MA: Allyn & Bacon. A slim text of gentle reminders from three young academics on the proper way to report research statistics.

[23] Myatt, M., & Ritter, S. (1997). Analysing Data. A Practical Primer Using Epi Info. Brixton Books. The best of Brixton Books series from Myatt. Does very well in explaining the statistical procedures that can be done with Epi Info, as well as explain how to report statistics in articles, and what the numbers actually mean. Good section on transforming data. Only Brixton Book worth buying.

[24] Newton, R.R., & Rudestam,K.E. (1999). Your Statistical Consultant. Answers to Your Data Analysis Questions. CA:Sage Publications. The definitive statistical answer book!

[25] Phillips, Jr., J. L. (1992). How to Think About Statistics. NY: W. H. Freeman & Co. Applying statistical concepts to the social sciences.

[26] Ramsey, F.L. & Schafer, D.W. (1997). The Statistical Sleuth: A Course in Methods of Data Analysis. CA: Duxbury Press. A really good comprehensive text on how to analyze data.

[27] Rosenthal, J.S. (2006). Struck by Lightning. The Curious World of Probabilities WASH DC: John Henry Press. Incredibly, an enjoyably readable book about probability!

[28] Salkind, N.J.(2000). Statistics for People Who (Think They) Hate Statistics. CA: Sage Publications. Yes! Finally, a statistics book written for the truly terrified.

[29] Shiffler, R.E., & Adams, A.J. (1996). Just the Basics, Please: A Quick Review of Math for Introductory Statistics. Belmont, CA: Duxbury Press.

[30] Spence, J.T., Cotton, J.W., Underwood, B.J., & Duncan, C.P. (1990). Elementary Statistics. 5TH EDITION. NJ: Princeton Hall. Simply the best elementary statistics textbook around.

[31] Tal, J. (2001). Reading Between the Numbers. Statistical Thinking in Everyday Life. NY: MacGraw Hill. GET THIS BOOK. If you are truly petrified of statistics, read this book.

[32] Urden, T.C. (2001). Statistics in Plain English. NJ: Lawrence Erlbaum Association Publishers. Though his humility is nice to see, this can easily be the best basic textbook of statistics around, not just a review text.

[33] Wallgren, A., Wallgren, B., Persson, R., Jorner, U., Haaland, J-A. (1996). Graphing Statistics & Data: Creating Better Charts. The definitive book for those who can't stand numbers as numbers and looking for a proper way to present them without misleading the reader or listener.

[34] Weaver, J.H (2000). Conquering Statistics. Numbers without the Crunch. MA: Perseus Publishing. Another for the "statistically scared" crowd.

[35] Weinbach, R.W., & Grinnell, R.M. (1991). Statistics for Social Workers. NY: Longman. (the best for simplicity in understanding how to use statistics for program evaluation).

[36] Weiss, N.A., & Hassett, M.J. (1991). Introductory Statistics. NY: Addison-Wesley Publishing Co. Basic statistics. Meant to be read sequentially. This text gives you an idea of why a basic statistics course is such a terror to those mathematically challenged.

[37] Thomas H. Wonnacott, Ronald J. Wonnacott (2007) Introductory Statistics, 5th Edition, Thomas H. Wonnacott (Author), Ronald J. Wonnacott (Author), J.Wiley & Sons, Simply the best elementary statistics textbook for today

[38] Felix Brosius (2008) SPSSS 16 Handbuch, mitp publishing house. solved with SPSS (Statistical Package for Social Science- since 1968) Excellent examples of statistical methods

[39] Kaiser, H.F. (1974): An Index of Factorial Simplicity, in: Psychometrika, Vol. 39,

[40] Richard B.Darlington, Regression and Linear Models, McGraw-Hill Companies (January 1, 1990)

[41] Paul Feyerabend and Ian Hacking, Against Method (Fourth Edition) (May 11, 2010), Verso publishing house

[42] John A. Hartigan. Clustering algorithms (Wiley, 1975, Original from the University of Michigan)