ANALYSIS OF VARIANCE AND MODELS OF CORRELATION AND REGRESSION

Tuomas Malinen

September 29, 2011

Contents

Chapter 1

Building blocks of statistical inference; Revision and introduction

1.1 Random variables

A random variable is a variable whose value is numerical outcome of a random phenomenon

- Phenomenon is called random if individual outcomes are uncertain but there is nonetheless a regular distribution of outcomes in large number of repetitions.
- The probability of random phenomenon is the proportion of times the outcome would occur in a very long series

Example 1. When you toss a coin, there are only two possible outcomes, heads or tails. Figure 4.1 shows the result of tossing a coin 5000 times twice. For each number of tosses from 1 to 5000, we have plotted the proportion of those tosses that gave a head. Trial A (red line) begins tail, head, tail, tail. The proportion of heads for Trial A starts at 0 on the first toss, rises to 0.5 when the second toss gives a head, then falls to 0.33 and 0.25 as we get more tails. Trial B (green) starts with five straigth heads, so the proportion of heads is 1 until the sixth toss.

The proportion of tosses that produces heads is quite variable at first. Trial A starts low and Trial B starts high. As we make more and more tosses, however, the proportion of heads for both trials gets close to 0.5 and stays there. If we made yet a third trial at tossing the coin a great number of times, the proportion of heads would again settle down to 0.5 in the long run. We say 0.5

is the *probability* of a head. The probability 0.5 appears as as horizontal line on the graph.

Some events are random "by definition": The outcome of a coin toss, the time between emissions of particles by a radioactive source, and sexes of the next litter of lab rats are all random. Probability theory describes random behavior. Probability can never be observed exactly. Mathematical probability is an idealization based on imagining what would happen in an indefinitely long series of trials. When examining randomness there are four *mandatory* restrictions:

- You must have a long series of independent trials, i.e. the outcome of one trial *must not* influence the outcome of any other.
- The idea of probability is empirical. We can estimate a real-world probability only by actually observing many trials.
- The proportion of an outcome often requires several hundred trials to settle down to the probability of that outcome. Psychical random devices are often too slow for this. Short runs give only rough estimates of a probability.
- All the observations must be drawn from the same parent population, i.e. from the same family of random variabless.

Random variables are denoted by capital letters near the end of the alphabet, such as X and Y. The random variables of greatest interest are outcomes such as the mean $\overline{x}$ of a random sample. When a random variable X describes a random phenomenon, the sample space S just lists the possible values of the random variable.

1.2 Probability

The French naturalist Count Buffon (1707-1788) tossed a coin 4040 times. Result: 2048 heads, or proportion 2048/4040=0.5069 for heads.

Around 1900, the English statistician Karl Pearson heroically tossed a coin 24000 times. Result: 12.012 heads, a proportion of 0.5005.

While imprisoned by the Germans during WWII, the South African statistician John Kerrich tossed a coin 10000 times. Result: 5067 heads, proportion of 0.5067.

The sample space S of a random phenomenon is the set of all possible outcomes.

Example 2. Toss a coin. There are only two possible outcomes, and the sample space is

$$S=[\text{head, tails}]$$

or more briefly, $S = [H, T]$

Example 3. Let your pencil fall blindly into table of random digits and record the value of the digit as it lands on. The possible outcomes are

$$S = [0, 1, 2, 3, 4, 5, 6, 7, 8, 9]$$

An event is an outcome or a set of outcomes of a random phenomenon, i.e. event is a subset of the sample space.

Example 4. Take the sample space of S for four tosses of a coin to be the 16 possible outcomes in the from HTHH. Then "exactly 2 heads" is an event. Call this event A. The event A expressed as a set of outcomes is

$$A = [HHTT, HTHT, HTTH, THHT, THTH, TTHH]$$

A continuous random variable X takes all values in an interval of numbers. The **probability distribution** of X is described by a density curve. The probability of any event is the area under the density curve and above the values of X that make up the event.

The cumulative distribution function F_x of any real valued random variable x is a function:

$$F_x(t) = P(x \leq t),\ t \in R^1.$$

Every cumulative distribution function has three characteristics:

- F_x is growing
- F_x is continuous to the right
- $F_x(-\infty) = 0$ and $F_x(\infty) = 1$.

When cumulative distribution function is known, the probabilities of all possible intervals $P(a < x < b)$ can be calculated. With the help of these, the probabilities of all events $x \in A$ can be calculated, if A can be formed using theoretical elementary processes within some interval. (see Moore & McCabe, p. 290-298). All continuous probability distributions assing **probability 0 to every individual outcome**. Only intervals have positive probability.

A **Density curve** is a curve that:

- is always on or above the horizontal axis and
- has area exactly 1 underneath it.

A density curve describes the overall pattern of a distribution.

Example 6. In a histogram, the *areas* of the bars represent either counts or proportions of the observations. The area of the shaded bars in figure 1.23(a) represent the students with vocabulary scores 6.0 or lower. There are 287 students, who make up the proportion 287/947=0.303 of all seventh graders in a school in question.

In figure 12.3(b) the area under the smooth curve to the left of 6.0 is shaded. Adjust the scale so that the total area under the curve is exactly 1. Areas under the curve then represent proportions of the observations. That is, *area=relative frequency*. The curve is then a *density curve*.

When exploring data, one should always:

- Plot the data, i.e. make a graph, usually a stemplot or a histrogram.
- Look for overall pattern and for striking deviations such as outliers
- Calculate an appropiate numerical summary to briefly describe center and spread.

Density curves presented above were symmetric, unimodal and bell-shaped. They are called **normal curves** and they describe *normal distributions*. All normal distributions have the same overall shape. The height of the density curve at any point x is given by

$$\frac{1}{\sigma\sqrt{2\Pi}}e^{-\frac{1}{2}\ \left(\frac{x-\mu}{\sigma}\right)^2}$$

The equation of the curve is completely determined by the mean μ and the standard deviation σ.

Because any density curve describes an assignment of probabilities, normal distributions are probability distributions. $N(\mu, \sigma)$ describes a normal distribution with mean μ and standard deviation σ. If a random variable X has the $N(\mu, \sigma)$ distribution, then the standardized variable

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

is a standard normal random variable having the distribution $N(0, 1)$.

1.3 Sampling distributions and basics of asymptotic distribution theory

The mean of a probability distribution describes the long-run average outcome. This can't be called mean $\bar{x}$. The common symbol for the **mean of a probability distribution** is μ, the Greek letter *mu*.

Select an *simple random sample* (SRS) of size n from a population, and measure a variable X on each individual in the sample. The data consist of observations on n random variables $x_1, x_2, ..., x_n$. A single x_i is a measurement on one individual selected at random from the population and therefore has the distribution of the population. The sample mean of an SRS of size n is

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + ... + x_n).$$

If the population has mean μ, then μ is the mean of each observation x_i. Therefore, by the addition rule for means of random variables (see Moore & McCabe p. 326):

$$\mu_{\bar{x}} = \frac{1}{n}(\mu_{x_1} + \mu_{x_2} + ... + \mu_{x_n}) = \frac{1}{n}(\mu + \mu + ... + \mu)$$

That is, *the mean $\bar{x}$ is the same as the mean of the population.* Thus, the sample mean $\bar{x}$ is an unbiased estimator of the unknown population mean μ.

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

Law of large numbers. Draw independent observations at random from any population with finite mean μ. Decide how accurately you would like to estimate μ. As the number of observations drawn increases, the mean $\bar{x}$ of the observed values eventually approaches the mean μ of the population as closely as you specified and then stays that close.

More formally: Lets assume that $x_1, x_2, ..., x_n, ...$ is a line of independent, identically distributed random variables ($x_n \sim i.i.d.$) with limited second moments (=variances). Let $\mu = EX_n$ and $\sigma^2 = var(x_n)$. Then

$$\bar{x_n} = \frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} \mu \ ,$$

when $n \longrightarrow \infty$.

414.pdf

Example 7. The distribution of the heights af all young women is close to the normal distribution with mean 64.5 inches and standard deviation of 2.5 inches. Suppose $\mu = 64.5$ were exactly true. Figure 4.14 shows the behavior of the mean height $\bar{x}$ on n women chosen at random from a population whose

heights follow $N(64.5, 2.5)$ distribution. The graph plots the values of $\bar{x}$ as we add women to our sample. The first woman drawn had height 64.21 inches. The second had height 64.35 inches, so for $n = 2$ the mean is

$$\bar{x} = \frac{64.21+64.35}{2} = 64.28$$

This is the second point on the line in the graph. Eventually, the mean of the observations gets close to the population mean $\mu = 64.5$ and settles down at that value.

Mean and standard deviation of a sample mean. Let $\bar{x}$ be the mean of an SRS of size n from population having mean μ and standard deviation σ. The mean and standard deviation of $\bar{x}$ are

$$\mu_{\bar{x}} = \mu$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

If the population has the $N(\mu, \sigma)$ distribution, then the sample mean $\bar{x}$ on n independent observations has the $N(\mu, \frac{\sigma}{\sqrt{n}})$ distribution.

Central Limit theorem. Draw an SRS of a size n from any population with mean μ and finite standard deviation σ. When n is large, the sampling distribution of the sample mean $\bar{x}$ is approximately normal:

$$\bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

More formally: If $x_n \sim i.i.d.$, $Ex_n \equiv \mu$ and $var(x_n) = E(x_n - \mu)^2 \equiv \sigma^2 < \infty$, then

$$y_n = \sqrt{n}\frac{\bar{x_n}-\mu}{\sigma}.$$

Example 8. Figure 5.10 shows the central limit theorem in action for a very nonnormal population. Figure 5.10 displays the density curve of a single observation of the population. The distribution is strongly rigth-skewed, and the most probable outcomes are near 0. The mean μ of this distribution is 1, and its standard deviation σ is also 1. This particular continuous distribution is called an **exponential distribution**.

In figure 5.10 the distribution of sample means from a strongly nonnormal population becomes more normal as the sample size increases. In (a) is a distribution of one observation. In (b) the distribution of average of two observations. In (c) the distribution of average for 10 observation, and in (d) the distribution of average for 25 observations.

1.4 Inference for distributions

A **parameter** is a number that decribes the **population**. A parameter is a fixed number, but in practice we do not know its value.

A **statistic** is a number that describes a **sample**. The value of a statistic is known when we have taken a sample, but it can change from sample to sample. We often use a statistic to estimate an unknown parameter.

Example 9. Sample surveys show that fewer people enjoy shopping in the past. A survey by the market research firm Yankelovich Clancy Shulman asked a nationwide random sample of 2500 adults if they agreed or disagreed that "I like buying new clothes, but shopping is often frustrating and timeconsuming." 1640, or 66% of the respondents said that they agreed. The proportion of the sample who agree is

$$p = \frac{1650}{2500} = 66\%$$

The number p=0.66 is a *statistic*. The corresponding *parameter* is the proportion (called $\hat{p}$) of all adult U.S. residents who would have said "Agree" if asked the same question. We don't know the value of the parameter p, so we use the statistic $\hat{p}$ to estimate it.

Confidence interval. A 95% confidence interval for a parameter is an interval computed from sample data by a method that has 95% probability of producing an interval containing the true value of the parameter.

Lets, for example, assume that we know that the entire population of SAT scores in the United States has a mean μ and standard deviation σ, then in repeated samples of size 400 the sample mean $\bar{x}$ has a $N(\mu, \sigma/500)$ distribution. Let say that we know that the standard deviation σ of SATM (score for mathematical reasoning ability) scores in California population is σ=100. In repeated sampling the sample mean $\bar{x}$ follows the normal distribution centered at the unknown population mean μ and having standard deviation

$$\sigma_{\bar{x}} = \frac{100}{\sqrt{500}} \doteq 4.5$$

In this case the probability is about 0.95 that $\bar{x}$ will be within 9 points (two standard deviations of $\bar{x}$ of the population mean score μ. The population mean is 461. We cannot know whether our sample is one of the 95% for which the interval $\bar{x} \pm 9$ catches μ or one of the unlucky 5%. The statement that we are 95% confident that the unknown μ lies between 452 and 470 is shorthand for saying, "We arrived at these numbers by a method that gives correct results 95% of the time".

Most confidence intervals have the form

$$\text{estimate} \pm \text{margin of error}$$

The **margin of error** shows how accurate we believe our guess is, based on the variability of the estimate. Confidence interval for sample mean $\bar{x}$ is calculated using

$$\mu \pm z * \frac{\sigma}{\sqrt{n}}$$

In other words, there is some probability C that above interval contains μ.

Example 10. Tim Kelley has been weighting himself once a week for several years. Last month his four measurements (in pounds) were

190.5 189.0 195.5 187.0

Give a 90% confidence interval for his mean weight for last month.

We treat the four measurements as an SRS of all possible measurements that Tim could have taken last month. These are estimates of μ, his true mean weigth for last month.

Examination of Tim's past data reveals that over relatively short periods of time, his weight measurements are approximately normal with a standard deviation of about 3. For our confidence interval we will use this value as the true standard deviation, that is, σ=3.

the mean of Tim's weight readings is

$$\bar{x} = \frac{190.5+189.0+195.5+187.0}{4} = 190.5$$

For 90% confidence, the formula becomes

$$\bar{x} \pm z * \frac{\sigma}{\sqrt{n}} = 190.5 \pm 1.645 \frac{3}{\sqrt{4}} = [188.0, 193.0]$$

Standard error. When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic. The standard error of the sample mean is

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}},$$

where s is a sample standard deviation. This is also sometimes called as "estimated standard error" of sample mean.

The standardized sample mean

$$z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

is the basis of the z procedures for inference about μ when σ is known. This statistic has the standard normal distribution $N(0, 1)$. If standard deviation is substituted with standard error, or estimated standard error, the statistic does not have a normal distribution.

The t distribution. If a SRS of size n is drawn from an $N(\mu, \sigma)$ population, then the *one-sample t statistic*

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

has the t *distribution* with $n-1$ degrees of freedom.

1.5 On the robustness of statistical analysis

1.5.1 Classical inference theory and robust procedures

Classical statistical inference theory presupposes a parent population (a family of random variables), whose every member contains those qualitative characteristics that the process that produced those observations is considered to have. Within this population, one can retrieve (estimate) a member that, according to some criterion, would have the best fit with the observations. Parent population should be as broad as possible, i.e. there shouldn't be any unintentional constraints.

A **response variable** measures an outcome of a study. An **explanatory variable** explains or causes changes in the response variable.

Example 10. How does drinking beer affect the level of alcohol in our blood? The legal limit for driving in most states in U.S. is 0.08%. Student colunteers at Ohio State University drank different numbers of cans of beer. Thirty minutes later, a police officer measured their blood alcohol content. Number of beers consumed is the explanatory (independent) variable and and percent of alcohol in the blood is the response (dependent) variable.

Robust procedures. A statistical inference procedure is called *robust* if the probability calculations required are insentive to violations of the assumptions made.

Example 11. Lets assume that a researcher has a one dataset on the income distribution of Finland and one dataset on the income distribution of Sweden. Both datasets include monthly observations on the incomes of households within 10 years. The Swedish dataset measures the net income of households (after taxes) and the Finnish dataset measures the net expenditures of households. Now, researcher would like to test if Sweden had the most equal income distribution within the measured era. However, it is not possible to *robustly* test this hypothesis, because the data on Sweden and Finland is from different sources. There is no one coherent parent population of random variables, and this would thus violate the assumption that random variables are drawn from a one *family of random variables*.

Theory is as important in statistical analysis as is the consistent use of statistical methods. If researcher does not have a theory to back his/her findings, obtained results might be spurious.

Example 12. Probably the most used example is the correlation between consumption of ice cream and weather temperature. It is highly likely that temperature affects on the consumption of ice cream, because it is likely that

ice cream consumption increases when temperature increases (at least to a some point). However, if researcher would have no theory he might come up with an "insane" result: Ice cream consumption raises temperature, or ice cream consumption affects on the heat of the sun, etc. Although this is an extreme example it highlights the fact that for **robust** statistical analysis, one should have some reasonable theory that is tested with statiscal analysis.

1.5.2 Power of a test and error types in hypothesis testing

In examining the usefulness of a confidence interval, we are concerned with both the level of confidence and the margin of error. The confidence level tells us how reliable the method is in repeated use. The margin of error tells us how sensitive the method is, i.e. how closely the interval pins down the parameter being estimated. The significance level, α, says how reliable the method is in repeated use. If we use 5% signifigance tests repeatedly when H_0 is true, we will be wrong (the test will reject H_0) 5% of the time, and right (the test will fail to reject H_0) 95% of the time.

The probability, computed assuming that H_0 is true, that the test statistic would take a value as extreme or more extreme than that actually observed is called the ***p*-value** of the test. The smaller the p-value, the stronger the evidence against H_0 provided by the data.

The probability that a fixed level α significance test will reject H_0 when a particular alternative value of the parameter is true is called the **power** of the test to detect that alternative.

Testing for hypothesis (H_0 vs. H_1 or H_a) is conducted with the help of probability theory. Because of this there is always a possibility of error. There are two types of errors:

- We reject H_o when it is true (**Type I error**).
- We accept H_0 when is is not true (**Type II error**).

The probability of a Type I error is the probability of rejecting H_o when it is really true. The probability of a Type II error is the probability of accepting a false null hypothesis.

The significance level α of any fixed level test is the probability of a **Type I error**. That is, α is the probability that the test will reject the null hypothesis H_0 when H_0 is in fact true.

The power of a fixed level test against a particular alternative is 1 minus the probability of a **Type II error** for that alternative.

If the probability of Type I error is denoted by α and the probability of Type II error is denoted by β the possible outcomes of testing for H_0 are:

	unknown true situation	
Result	H_0 true	H_0 false
Accept H_0	Correct decision	Type II error
	$1-\alpha$	β
Reject H_0	Type I error	Correct decision
	α	$1-\beta$

The distinction between tests of significance and tests as rules for deciding between two hypotheses lies in the reasoning that motivates the calculations. In a test of significance we focus on a single hypothesis (H_0) and a single probability (the p-value). We are thus measuring the strength of the sample evidence against H_0. Calculations of power are done to check the sensitivity of the test. If H_0 cannot be rejected, we can conclude only that there is not sufficient evidence against H_0, not that H_0 is actually true. In terms of inference, we focus on two hypotheses and give a rule for deciding between them based on the sample evidence.

Chapter 2

Analysis of variance

2.1 One-way analysis of variance

Which of the four advertising offers mailed to sample household produces the highest sales in dollars? Which of ten brands of automobile tires wears longest? How long do cancer patiens live under each of three thrapies for their cancer? In each of these settings we wish to compare several treatments. In each case the data are subject to sampling variability - if we mailed the advertising offers to another set of households, we would get different data. The question for inference is therefore posed in terms of the *mean* response. The statistical methodology for comparing several means is called **analysis of variance**, or **ANOVA**.

Two ANOVA techiques will be considered: One-way ANOVA and two-way ANOVA. One-way ANOVA is used when there is only one way to classify the populations of interest. However, often there are several ways to compare real life situations (or, in that mater, some simulations), which require two-way ANOVA.

Example 13. A gerontologist investigating various aspects of the aging process wanted to see whether staying "lean and mean", that is, being under normal body weight would lenghten life span. She randomly assigned newborn rats from a highly inbred line to one of three diets: (1) unlimited access to food, (2) 90% of the amount of food that a rat of that size would normally eat, (3) 80% on the amount of food that a rat of that size would normally eat. She maintained the rats of three diets troughout their livces and recorded their lifespans in years. Is there evidence that diet affected life span in this study? Results are reported on the table below.

Unlimited	90% diet	80% diet
2.5	2.7	3.1
3.1	3.1	2.9
2.3	2.9	3.8
1.9	3.7	3.9
2.4	3.5	4.0

2.1.1 Comparing means

To assess whether several populations all have the same mean, we compare the means of samples drawn from each population.

Example 14. A medical researcher wants to compare the effectiveness of three different treatments to lower the cholesterol of patients with high blood cholesterol levels. He assigns 60 individuals at random to the three treatments (20 to each) and records the reduction in cholesterol for each patient.

Figure 2.1: Mean serum cholesterol reduction in three groups

Figure 2.2: Large within-group variation

Figure 2.3: Small within-group variation

The *two-sample* t statistic compares the means of two populations. If the two populations are assumed to have equal but unknown standard deviations and the sample sizes are both equal to n, the t statistic is

$$t = \frac{\bar{x}-\bar{y}}{s_p\sqrt{\frac{1}{n}+\frac{1}{n}}} = \frac{\sqrt{\frac{n}{2}}(\bar{x}-\bar{y}}{s_p}$$

The square of this t statistic is

$$t^2 = \frac{\frac{n}{2}(\bar{x}-\bar{y})^2}{s_p^2}$$

If ANOVA is used to compare two populations, the ANOVA F statistic is exactly equal to this t^2.

The numerator in the t^2 statistic measures the variation **between** the groups in terms of the difference between their sample means $\bar{x}$ and $\bar{y}$. It includes a factor for the common sample size n. The denominator measures the variation **within** groups by s_p^2, the pooled estimator of the common variance. If the within-group variation is small, the same variation between the groups produces

a larger statistic and a more significant result.

The ANOVA null hypothesess is that the population means are *all* equal. The alternative is that they are not equal. To assess whether several populations all have the same mean, we compare the variation *among* the means of several groups with the variation *within* groups.

2.1.2 The ANOVA model

In statistical analysis we are usually looking for overall patterns and deviations from it:

DATA=FIT + RESIDUAL

The observations are now assumed to be drawn *randomly* from the parent population. If the observations are $x_1, x_2, ..., x_n$ we can describe this model by saying that the x_j are an SRS from the $N(\mu, \sigma)$ distribution. In the above model the x's can also be thought to be varying their population mean:

$$x_j = \mu + \epsilon_j$$

The ϵ_j's (residuals) are then an SRS from the $N(0, \sigma)$ distribution.

The model for one-way ANOVA is:

$$x_{ij} = \mu_i + \epsilon_{ij},\ i = 1, ..., I \text{ and } j = 1, ..., n_i$$

The ϵ_{ij} are assumed to be from an $N(0, \sigma)$ distribution. The standard deviation (σ) is assumed to be the same in all of the populations, but sample sizes n_i may differ.

μ_i is estimated by using the sample mean for the i:th group:

$$\bar{x_i} = \frac{1}{n} \sum_{j=1}^{n_i} x_{ij},$$

where, the residuals $\epsilon_{ij} = x_{ij} - \bar{x_i}$ reflect the variations about the sample means.

The ANOVA model assumes that the standard deviations are all equal (or that the variances are equal). If standard deviations are unequal, one can try transforming the data so that SD:s are approximately equal. Some simple tranformations include $\sqrt{x_{ij}}$ or $logx_{ij}$. These usually makes *both* group standard deviations more nearly equal and also makes the distributions of observations in each group more nearly normal.

The **assumptions** in one-way ANOVA model can be summarized as:

1. The dependent variable should be measured in interval scale, and the independent variable in nominal scale.

2. The k samples represent independent random samples drawn from k specific populations with means $\mu_1, ..., \mu_k$, where $\mu_1, ..., \mu_k$ are unknown constants.

3. Each of the k populations is normally distributed.

4. Each of the k populations has the same variance σ^2.

Because ANOVA is not extremely sensitive to unequal standard deviations, for example Moore & McCabe (2004, p. 755) do *not* recommend a formal test of equality of standard deviations before using ANOVA. Instead they give the following thumb rule:

> If the largest standard deviation is less than twice the smallest standard deviation, we can use methods based on the assumption of equal standard deviations, and our results will still be approximately correct.

Pooling gives more weigth to groups with larger sample sizes. If the sample sizes are equal, s_p^s is just the average of I sample variances.

Testing hypotheses in one-way ANOVA

The **null and alternative hypotheses** for one-way ANOVA are:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_I.$$

$$H_a : \text{not all of the } \mu \text{ are equal.}$$

In example 13 we could find the 90% and 80% diets are not different from each other in their effects on life span, but that they are significantly different from the control diet. In this case, we would accept the H_a hypotheses.

The ANOVA table

The information from analysis of variance is presented in an ANOVA table.

The colums are labelled as Source, DF, Sum of Squares, Mean Square, F value, and the Pr > F. The rows are labelled as Model, Error and Corrected Total. These are the three sources of variation in one-way ANOVA.

The Model row gives information related to the variation **among** group means (FIT).

The Error row gives information on the variation **within** groups (RESIDUAL). The term "error" is most appropriate when analysing some purely physical phenomenon, where the observations within a group differ because of a measurement error. In business and biological sciences the within-group variation is often due to the fact that not all firms or plants or people are the same. This variation is not due to errors and is best described as "residual".

The Corrected Total is labelled as "DATA", and so in ANOVA the model

DATA=FIT+RESIDUAL

becomes

total=model+residual

Each **sum of squares** is a sum of squared deviations. SS_T measures variation of the data around the overall mean, $x_{ij} - \bar{x}$. SS_G is the variation of the group means around the overall mean, $\bar{x_i} - \bar{x}$, and SS_E measures the variation of each observation around its group mean $x_{ij} - \bar{x_i}$.

Each sum of squares if associatied with the **degrees of freedom**. Because SS_T measures the variation of all observations (N) around the overall mean, its degrees of freedom are DF_T=N-1. Because SS_G measures the variation of sample means (I), its degrees of freedom are DFG=I-1. SS_E is $x_{ij} - \bar{x_i}$, and it compares N observations with I sample means, and therefore DF_E=$N - I$.

In general, it is true that

$$s_p^2 = MS_E = \frac{SS_E}{DF_E}$$

That is, the error mean square is an estimate of the within-group variance, σ^2. In the output this is the Root MS_E.

Table 2.1: Calculations in ANOVA

source	degr. of freed.	sum of squares	mean square	F
Treats	$I - 1$	$\sum n_i(x_i - \bar{x})^2$	SS_G/DF_G	MS_G/MS_E
Error	$N - I$	$\sum(n_i - 1)s_i^2$	SS_E/DF_E	
Total	$N - 1$	$\sum(x_{ij} - \bar{x})^2$	SS_T/DF_T	

The computational formulas for sum of squares are:

$$SS_{Total} = \sum_i \sum_j X_{ij}^2 - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N}$$

$$SS_{Group} = \sum_i \left[\frac{\left(\sum_j X_{ij}\right)^2}{n_i}\right] - \frac{\left(\sum_i \sum_j X_{ij}\right)^2}{N}$$

$$SS_{Error} = \sum_i \left[\sum_j X^{2_{ij}} - \frac{\left(\sum_j X_{ij}\right)^2}{n_i}\right]$$

To obtain **mean squares**, divide the corresponding sum of squares by its degrees of freedom.

- $DF_T = N - 1$
- $DF_G = I - 1$
- $DF_E = N - I$.

The F test

If H_0 is true, there are no differences among the group means. The ratio MS_G/MS_E (MSG/MSE in table below) is approximately 1 if H_0 is true and tends to be larger if H_a is true.

Solution to example 13. The hypotheses are

$$H_0 : \mu_U = \mu_{90\%} = \mu_{80\%}$$
$$H_a : \text{At least one pair of the } \mu_i\text{'s are not equal.}$$

Source	SS	df	MS	F	critical value
Groups	3.15	2	1.575	7.76	3.89
Error	2.44	12	0.203		
Total	5.60	14			

Since $7.76 > 3.89$, the **mean squares diets** is significantly bigger than the **error mean squares** indicating at least one pair of the diets aren't equal. Conclusion at this point is only that *some* of the diets result in different life spans (i.e. reject H_0). To find out where the differences lie we must compare the means.

2.1.3 Comparing the Means

Contrasts

The ANOVA F test answers to the question: Are the differences among observed group means statistically significant? This just tells us that the group means are not all the same. Plotting and inspecting means gives us an indication of where the differences are. In the ideal situation, some specific questions regarding comparisons among the means are posed before the data is collected.

A **contrast** expresses an effect in the population as a combination of population means. Contrast is estimated by forming corresponding **sample contrast** by using sample means in place of population means. Under the ANOVA assumptions, a sample contrast is a linear combination of independent variables and has, therefore, a normal distribution. Inference is based on t statistics.

Bonferroni's and Duncan's tests

The Bonferroni t test statistics is

owaf.pdf

contrasts.pdf

$$\text{Bonferroni } t = \frac{\bar{x_{i.}} - \bar{x_{j.}}}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_i}\right)}},$$

where i and j represent any two means to be separated and the degrees of freedom is $N - k$. Bonferroni t tests are usually run at a smaller α level than the global F test of the ANOVA table, to manage the probability of type I error. If all Bonferroni tests are performed at significance level α, then the overall probability of at least one Type I error (α') is larger than α and its value is usually unknown. It can be shown that with three tests at the α level, α' is at most $1 - (1 - \alpha)^3$. If, for example, $\alpha = 0.05$ (5%) and three Bonferroni tests are conducted, the probability of making at least one Type I error becomes at most $1 - (1 - 0.05)^3 = 0.143$. Using $\alpha = 0.05$ for ten comparisons $\alpha' = 0.401$. So the overall probability of Type I error increases rapidly in comparison with increases in number of means to be separated.

Because of this, most Bonferroni tests are conducted at an α level lower than the global F test. A general rule is to determine an *experimentwise* acceptable upper boundary for the probability of Type I error, say b, and divide this probability by the actual number of comparisons run, to determine the α level for each comparisons. If we wished to experimentwise α' to be 0.05 in example , we would use $0.05/3 = 0.017$ as the α level for each Bonferroni t test.

A second method used to separate means in fixed treatment ANOVA is called the **Duncan's multiple range test**. This test uses the rank orders of the sample means to determine the *shortetst significant range* or SSR_p. Any two means that differ by more than this value are considered significantly different. Test assumes *that all samples have the same size.* The test protocol is as follows:

1. Linearly order the k sample means from smallest to largest.
2. Calculate the shortest significant range by

$$SSR_p = r_p \sqrt{\frac{MSE}{n}},$$

   where

   - r_p is a critical value of the test (from tests own table),
   - MSE is the error mean square from the ANOVA table,
   - n is the common sample size,
   - and v is the degrees of freedom for the MSE.

3. For any subset of p sample means $2 \leq p \leq k$, compare the appropriate SSR_p. If the range of the means under consideration is greater than the SSR_p, the population means are considered significantly different and denoted with different superscript letters.

To use the DMRT to separate means based on *unequal sample sizes* requires an adjustment (C. P. Kramer 1956):

$$SSR'_p = r_p \sqrt{MSE},$$

where r_p is from the table of tests critical values and MSE from the ANOVA table. The test statistic becomes

$$|\bar{x_i} - \bar{x_j}| \sqrt{\frac{2n_i n_j}{n_i + n_j}}.$$

Model II ANOVA

All that has been presented above has been constructed for completely randomized design with fixed effects, i.e. the model I ANOVA. In the model I ANOVA we are interested on treatments "at hand", and the results can be reproduced with the same treatments. In model II ANOVA, or random-effects ANOVA, we are only interested on the difference between treatments, and the results usually change, because the treatments may change between studies.

Example 15. An endocrinologist studying genetic and environmental effects on insulin production of pancreatic tissue, raised five litters of experimental mice. At age 2 months he sacrificed the mice, dissected out pancreatic tissue and treated the tissue specimens with glucose solution. The amount of insulin released from these specimens was recorded in pg/ml. Are there significant differences in insulin release among the litters? Following table gives the amounts of litter.

	1	2	3	4	5
	9	2	3	4	8
	7	6	5	10	10
	5	7	9	9	12
	5	11	10	8	13
	3	5	6	10	11
$T_{i\cdot}$	29	31	33	41	54
$\bar{x_{i\cdot}}$	5.8	6.2	6.6	8.2	10.8

$$SS_T = \sum_i \sum_j x_{ij}^2 - \frac{T^2}{N} = 220.24$$

$$SS_G = \sum_i \frac{T_i^2}{n_i} - \frac{T^2}{N} = 83.84$$

$$SS_E = SS_T - SS_G = 136.40$$

Source of variation	SS	df	MS	F	c.v.
Among litters	83.84	4	20.96	3.07	2.87
Error	136.40	20	6.82		
Total	220.24	24			

Since $3.07 > 2.87$, we reject H_0 and accept H_a. Thus there seems to be a significant variability among the litters.

There is also a model III type ANOVA, which combines random- and fixed-effects treatments. However, this method is beyond the scope of this course.

2.1.4 Power of tests in ANOVA

When planning a study using ANOVA, it is important to perform power calculations to check that the sample sizes are adequate to detect differences among means that are judged to be important. Power calculations also help evaluate and interpret the results of studies in which H_0 was not rejected.

Procedures for power calculations in ANOVA:

1. Specify

   - an alternative H_a that you consider important; that is, values for the true population means $\mu_1, ..., \mu_I$;
   - sample sizes $n_1, ..., n_I$; usually these will all be equal to the common value n;
   - a level of significance α, which is usually equal to 0.05; and
   - a guess at the standard deviation σ.

2. Find the degrees of freedom DFG=$I-1$ and DFE=$N-I$ and the critical value that will lead to rejection of H_0. This value, which we denote by $F*$, is the upper α critical value for the F distribution.

3. Calculate the **noncentrality parameter**

   $$\lambda = \frac{\sum n_i(\mu_i - \bar{\mu})^2}{\sigma^2},$$

   where $\bar{\mu}$ is a weighted average of the group means,

   $$\bar{\mu} = \sum w_i \mu_i$$

   and the weights are proportional to the sample sizes,

   $$w_i = \frac{n_i}{\sum n_i} = \frac{n_i}{N}$$

4. Find the power, that is, the probability that the observed F is greater than $F*$. Under H_a, the F statistic has a distribution known as the **noncentral F distribution**.

If the n_i are all equal to the common value n then $\bar{\mu}$ is the ordinary average of the μ_i and

$$\lambda = \frac{n\sum(\mu_i - \bar{\mu})^2}{2}$$

If the means are all equal (the ANOVA H_0), then $\lambda = 0$. The noncentrality parameter measures how unequal the given set of means is. Large λ points to an alternative far from H_0, and we expect the ANOVA F test to have high power.

2.2 The Two-Way Analysis of Variance

Two-way ANOVA compares the means of populations that are classified in two ways or the mean responses in two-factor experiments. In the two-way ANOVA model, there are two factors, each with its own number of levels, comparing to one-way ANOVA where we had only one categorical variable.

Example 16. In an experiment on the influence of dietary minerals on blood pressure, rats receive diets prepared with varying amounts of calcium and varying amounts of magnesium, but with all other ingredients of the diets the same. Calcium and magnesium are the two factors in this experiment. As is

common in such experimetns, high, normal, and low values for each of the two minerals were selected for study. So there are three levels for each of the factors and a total of nine diets, or treatments. The following table summarizes the factors and levels for this experiment.

	Calcium		
Magnesium	L	M	H
L	1	2	3
M	4	5	6
H	7	8	9

For example, Diet 2 contains magnesium at its low level combined with the normal (medium) level of calcium. Each diet is fed to 9 rats, giving a total of 81 rats used in this experiment. The response variable is the blood pressure of a rat after some time on the diet.

Example 17. A textile researcher is interested in how four different colors of dye affect the durability of fabrics. Because the effects of the dyes may be different for different types of cloth, he applies each dye to five different kinds of cloth. The two factors in this experiment are dyes with four levels and cloth types with five levels. Six fabric specimens are dyed for each of the 20 dye-cloth combinations, and all 120 specimens are tested for durability.

If one or more specific dye-cloth combinations produced expectionally bad or exceptionally good durability measures, the experiment should discover this combined effect.

Effect of the dyes that may differ for different types of cloth are represented in the FIT part of the a two-way model as **interactions**. In contrast, the average values for dyes and cloths are represented as **main effects**. The two-way model represents FIT for each of the two factors and interaction. One-way desings that vary a single factor and hold other factors fixed cannot discover interactions.

2.2.1 The two-way ANOVA model

Advantages of two-way ANOVA:

1. It is more efficient to study factors simultaneously rather than separately.
2. We can reduce the residual variation in a model by including a second factor thought to influence the response.
3. We can investigate interactions between factors.

When discussing two-way models in general, we will use the labels A and B for the two factors. In two-way desing every level of A appears in combination with every level of B, so that $I \times J$ groups are compared. The sample size of level i of factor A and level j of factor B is n_{ij}.

The total number of observations is

$$N = \sum n_{ij}.$$

Assumptions for two-way ANOVA:

We have independent SRSs of size n_{ij} from each $I \times J$ normal populations. The population means μ_{ij} may differ, but all populations have the same standard deviation σ. The μ_{ij} and σ are unknown parameters.

Let x_{ijk} represent the kth observations from the population having factor A at level i and factor B at level j. The statistical model is

$$x_{ijk} = \mu_{ij} + \epsilon_{ijk}$$

for $i = 1, ..., I$ and $j = 1, ..., n_{ij}$, where the deviations ε_{ijk} are from an $N(0, \sigma)$ distribution.

The FIT part of the model is the means μ_{ij}, and the RESIDUAL part is the deviations ϵ_{ijk} of the individual observations from their group means. Estimation of a population mean μ_{ij} is done by using the sample mean of the observations in the samples from this group:

$$\bar{x_{ij}} = \frac{1}{n_{ij}} \sum_k x_{ijk}$$

The k below $\sum$ means that we sum the n_{ij} observations that belong to the (i, j)th sample.

The RESIDUAL part part of the models contains the unknown σ. σ^2 is estimated by calculating the sample variances for each SRS and pooling these:

$$s_p^2 = \frac{\sum(n_{ij}-1)s_{ij}^2}{\sum(n_{ij}-1)}$$

Just like in one-way ANOVA, the numerator in this fraction is SSE and the denominator is DFE (DFE=$N - IJ$).

2.2.2 The ANOVA table

Source	DF	SS	MS	F
A	$I-1$	SS_A	SS_A/DF_A	MS_A/MS_E
B	$J-1$	SS_B	SS_B/DF_B	MS_B/MS_E
AB	$(I-1)(J-1)$	SS_{AB}	SS_{AB}/DF_{AB}	MS_{AB}/MS_E
Error	N-IJ	SS_E	SS_E/DF_E	
Total	$N-1$	SS_T	SS_T/DF_T	

Next we go through two special cases (from many possible model extensions).

2.2.3 Randomized complete block design ANOVA

Example 18. Each of six garter snakes, *Thamnophis radix*, was observed during the presentation of petri dishes containing solutions of different chemical stimuli. The number of tongue flicks during a 5-minute interval of exposure was recorded. The three petri dishes were presented to each snake in random order.

	Stimulus		
Snake	Fish mucus	worm mucus	dH_2O
1	13	18	8
2	9	19	12
3	17	12	10
4	10	16	11
5	13	17	12
6	11	14	12

Each of the 18 data points can be looked at *two* ways. $x_{11} = 13$ is part of column of numbers that represent response to the first stimulus (fish mucus). That x value is also related to $x_{21} = 18$ and $x_{31} = 8$ because they are all values recorded from the same snake. They represent a block of data that is analogous to a pair of data points in a paired t test. If the snakes are randomly chosen from a larger group of garter snakes, we are not interested in differentiating among the snakes according to their response and have "blocked" the design *only* to remove some of the variability among snakes in their response to olfactory stimuli.

The null hypothesis would be $H_0 : \mu_{fm} = \mu_{wm} = \mu_{dH_2O}$ with the alternative hypothesis H_a: At least on pair of μ_i's not equal.

Randomized means that each treatment is assigned randomly within blocks and **complete** implies that each treatment is used exactly once within each block.

The randomized complete block model

Model assumptions:

1. Each observation constitutes a random, independent sample from a population with mean μ_{ij}. There are $k \times b$ of these populations sampled.
2. Each of the $k \times b$ populations is normally distributed with the same variance.
3. The treatment and block effects are *additive*, i.e. there is no interaction between blocks and treatments.

The model is:

$$x_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij},$$

where

- μ is the overall mean, τ is the effect due to the fact that the experimental unit received the i:th treatment

$$\tau_i = \mu_{i.} - \mu,$$

- β_j is the effect due to the fact that the experimental unit was in the j:th block

$$\beta_j = \mu_{.j} - \mu,$$

- ϵ_{ij} is the residual

$$\epsilon_{ij} = x_{ij} + \mu_{ij},$$

where $_{ij}$ is the mean for the combination of the ith treatment and the jth block.

Sum of Squares identity

$$SS_{Total} = \sum_i \sum_j (x_{ij} - \bar{x})^2$$

with df=$v = bk - 1$. The computational formula is:

$$SS_{Total} = \sum_j \sum_j x_{ij}^2 - \frac{T_{..}^2}{N}$$

The computational formula for sum of squares of treatmes:

$$SS_{treat} = \sum_i \sum_j (\bar{x_{i.}} - \bar{x_{..}})^2 = \sum_i \frac{T_{i.}^2}{b} - \frac{T_{..}^2}{N},$$

with df=$k - 1$.

The computational formula for the sum of squares of blocks:

$$SS_{blocks} = \sum_i \sum_j (\bar{x_{.j}} - \bar{x_{..}})^2 = \sum_j \frac{T_{.j}^2}{k} - \frac{T_{..}^2}{N}$$

with df=$b - 1$.

For the errors the sum of squares is

$$SS_{Error} = \sum_i \sum_j (x_{ij} - \bar{x_{i.}} - \bar{x_{.j}} + \bar{x_{..}})^2 = SS_{Total} - (SS_{Treat} + SS_{Block})$$

with df=$(k-1)(b-1)$.

The ANOVA table

The sum of squares divided by the appropriate degrees of freedom generate *mean squares* than can be used to test hypotheses about means:

$$MS_{treat} = \frac{SS_{Treat}}{k-1}$$

$$MS_{Blocks} = \frac{SS_{Blocks}}{b-1}$$

$$MSE = \frac{SS_{Error}}{(k-1)(b-1)}$$

Expected mean squares

To test $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, the ratio used is

$$\frac{MS_{Treat}}{MSE} = F_{(k-1),(k-1)(b-1)}.$$

The value of the test statistics is 1, if H_0 seems to be true.

Mean separation

In two-way ANOVA the Bonferroni's test is as in one-way ANOVA

$$\text{Bonferroni } t = \frac{\bar{x_{i.}} - \bar{x_{j.}}}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_i}\right)}},$$

but in Duncan's test the denominator in shortest significant range has changed

$$SSR_p = r_p \sqrt{\frac{MSE}{b}}.$$

2.2.4 Factorial desing two-way ANOVA

Example 19. In an attempt to find the most effective methods for training companion dogs for the physically challenged, an experiment was conducted to compare three different training regimes in combination with three different reward systems. All the animals in the study were Labrador retrievers and were 6 to 8 months old at the start of the experiment. Individual dogs were assigned to a combination of training regime and reward system randomly. At the end of a 10-week training period the dogs were given a standardized test to measure their ability to function as companion dogs for the visually impaired. The results of the test were:

	training regime		
Reward	I	II	II
Praise	45	51	52
	69	50	18
	53	62	25
	51	68	32
Tangible	54	53	51
	72	63	59
	69	67	47
	66	70	42
Praise and tangible	91	69	66
	87	73	68
	89	77	70
	91	74	64

The behavioral scientist running this study would like to know which training regime, if any, is best and which reward system is best. Interactions between training regimes and reward systems are also of interest. These can be positive

(synergy) or negative (inference).

The experiment is called **factorial** because we are interested in two different factors' effect's on the measured response. It is **completely random** because individuals are assigned to a combination of training regime and reward system randomly. These combinations are called **cells**. The effects are *fixed* because we are interested in only training regimes I, II, and III, and reward systems praise, tangible, and praise with tangible. Neither the training regimes nor the reward systems are chosen at random from a larger group of interest. Thus, this is a Model I design for both factors.

The factorial desing two-way model

Each observation in the dataset can be thought as the sum of various effects. For $i = 1, ..., a$, $j = 1, ..., b$, $k = 1, ..., n$.

Model assumptions:

1. The observations in each *cell* constitute an independent random sample of size n from a population with mean μ_{ij}.
2. Each of the populations represented by the cell samples is normally distributed and has the same variance, σ^2.

The model is:

$$x_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk},$$

where

- μ is the overall mean,
- α_i is the effect due to the fact that the experimental unit receiced the ith level of factor A

$$\alpha_i = \mu_{i..} - \mu,$$

- β_j is the effect due to the fact that the experimental unit received the jth level of factor B

$$\beta_j = \mu_{.j.} - \mu$$

- $(\alpha\beta)_{ij}$ is the effect of the *interaction* between the ith level of factor A and the jth level of factor B

$$(\alpha\beta)_{ij} = \mu_{ij.} - \mu_{i..} - \mu_{.j.} + \mu,$$

- ϵ_{ijk} is the residual

$$\epsilon_{ijk} = x_{ijk} - \mu_{ij}$$

Sums of squares

Computational formulas are:

$$SS_{Total} = \sum_i \sum_j \sum_k x_{ijk}^2 - \frac{T_{...}^2}{abn}$$

$$SS_A = \sum_i \left(\frac{T_{i..}^2}{bn} \right) - \frac{T_{...}^2}{abn}$$

$$SS_B = \sum_j \left(\frac{T_{.j.}^2}{an} \right) - \frac{T_{...}^2}{abn}$$

$$SS_{Cells} = \sum_i \sum_j \left(\frac{T_{ij.}^2}{n} \right) - \frac{T_{...}^2}{abn}$$

These calculations are highly tedicious and are always done with computer programs. The degrees of freedom associated with the sums of squares are:

$$SS_{Total} : v = abn - 1 \text{ or } N - 1$$

$$SS_{Cells} : c = ab - 1$$

$$SS_A : v = a - 1$$

$$SS_B : v = b - 1$$

$$SS_{A \times B} : v = (a-1)(b-1)$$

$$SS_{Error} : v = ab(n-1)$$

Test for interaction between factors

The hypotheses are

$$H_0 : (\alpha\beta)_{ij} = 0 \ \forall \ i, j$$

$$H_a : (\alpha\beta)_{ij} \neq 0 \text{ for some } i, j$$

The appropriate test for null hypothesis is

$$F_{A \times B} = \frac{MS_{A \times B}}{MS_E},$$

with degrees of freedom

$$v_1 = (a-1)(b-1),\ v_2 = ab(n-1).$$

As with all ANOVA F tests, if the numerator is significantly larger than the denominator, the H_0 will be rejected.

If the null hypothesis of **no interactions** is accepted, the analysis is continued with the help of two tests.

- Test whether there are differences in means for A-factor treatments:

$$H_0 : \mu_{i..}\text{'s are equal } (\alpha_i = 0 \ \forall \ i)$$

$$H_a : \text{At least on pair of } \mu_{i..}\text{'s is not equal,}$$

using the test statistic

$$F_A = \frac{MS_A}{MS_E}$$

with degrees of freedom

$$v_1 = a - 1,\ v_2 = ab(n-1).$$

- Test whether there are differences in means for B-factor treatmens similarly.

If the null hypothesis is *rejected*, the F_A and F_B tests are omitted. Because the levels of Factor A do not behave consistently across the levels of Factor B, and vice versa, the best combination of A factor and B factor should be used. The various cell means can be separated with DMRT of the form

$$SSR_p = r_p\sqrt{\frac{MS_E}{n}}.$$

2.3 Introduction to nonparametric methods in ANOVA

Many of the methods used for inference about the means of quantative response variables assume that the variables in question have normal distribution in the populations or populations from which we draw our data. In practice, of course, no distribution is exactly normal. Fortunately, usual methods for inference about population means are quite robust. That is, the results of inference are not very sensitive to moderate lack of normality, especially when samples are reasonably large. What to do if plotting the data suggests that the data are clearly not normal, especially when we have only few observations? The basic options are:

1. If there are extreme **outliers** in a small data set, any inference method may be suspect. An outlier is an observation that may not come from the same population as the other observations. To decide what to do, you must find the cause of the outlier. Equipment failure that produced a bad measurement, for example, entitles you to remove the outlier. If the outlier appears to be "real data", it is risky to draw any conclusions from just few observations.

2. Sometimes it is possible to **transform** the data so that their distributions is more nearly normal.

3. In some settings, **other standard distributions** replace the normal distributions as models for overall pattern in the population.

4. Finally, there are **nonparametric methods** that do not reguire the normal distribution of the population.

2.3.1 The Kruskal-Wallis test

A nonparametric analog to Model I one-way ANOVA

Example 20. Preliminary observations on North Stradbroke Island indicated that the gastropod *Austrocochlea obtusa* preferred the zone just below the mean tide line. In an experiment to test this, *A. obtusa* were collected marked, and placed either 7.5m above this zone (Upper Shore), 7.5m below this zone (Lower Shore), or back in the original area (Control). After two tidal cycles, the snails were recaptured. The distance each had moved (in cm) from where it had been placed was recorded. Is there significant difference among the median distances moved by the three groups?

Upper shore	Control	Lower shore
59	19	5
71	32	13
75	55	15
85	62	41
95	82	46
148	114	51
170	144	60
276		106
347		200

Ranking the data.

Upper shore	Rank	Control	Rank	Lower shore	Rank
59	10	19	4	5	1
71	13	32	5	13	2
75	14	55	9	15	3
85	16	62	12	41	6
95	17	82	15	46	7
148	21	114	19	51	8
170	22	144	20	60	11
276	24			106	18
347	25			200	23
Sum	162	Sum	84	Sum	79

If the three populations are identical, then the rankings should be randomly allocated to each of the samples. There is no reason to expect any one population to have a large number of high or low ranks. The average rank in each group should be roughly the same. In this example there is a grand total of N=25 observations. The average, or expected, rank is the mean (median) value of the numbers from 1 to 25, which is $\frac{N+1}{2} = \frac{25+1}{2} = 13$. Under the assumption of no difference between groups, we expect the average rank within each group to be roughly the overall average. Since there are 9 observations in the Upper Shore sample, its average is $\frac{162}{9} = 18$. Similarly the average for the Control is $\frac{84}{7} = 12$ and for the Lower Shore $\frac{79}{9} = 8.78$. The averages are different from 13, but are they significantly different? The test statistics for this situation is based on these differences, as we describe below.

Assumptions:

1. Independent random samples of sizes $n_1, ..., n_k$ are drawn from k continuous populations.
2. The null hypothesis is that all k populations are identical, or equivalently, that all the samples are drawn from the same population.

We like to test whether all k populations have the same location. Only one pair of hypotheses is possible:

H_0: All k populations have the same median,

H_a: At least one of the populations has a median different from the others.

Test statistic and decision rule

Rank all the observations without regard to the sample they come from. Use midranks for tied values.

- Let N denote the total number of measurements in the k samples.
- Let R^i denote the sum of ranks associated with the ith sample.
- The *grand mean* or *average rank* is $\frac{N+1}{2}$
- The *sample mean rank* for the ith sample is $\frac{R_i}{n_i}$.

The test statistics H measures the dispersion of the sample mean ranks from the average rank:

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2.$$

If the null hypothesis is true, this deviation should be close to 0 since the average rank should be close to each sample mean rank. If one or more of the groups has a sample mean rank that differs from the average rank then H will be positive. The larger H is, the less likely that H_0 is true.

Paired comparisons

Once either test is done, the are two possible, mutually exclusionary states that have realized:

1. We are unable to reject H_0: Based on the available data we cannot detect any differences among the k population means, and the analysis is complete.
2. We were able to reject H_0: Based on the available data we conclude there are differences among the k population means, and the analysis continues.

Following test is the nonparametric analog of the Bonferroni's t test. Hypotheses are:

H_0: The means of the ith and jth populations are the same.

H_a: The means of the ith and jth populations are different.

Choose an overall significance level α'. As with the Bonferroni's test the larger the k is, the larger α' must be in order to run the comparisons at a reasonable level of significance, α. But when α' is larger, we run a greater risk of rejecting a null hypothesis inappropriately (Type I error). Therefore, in the design of the experiment, the number of treatments k under investigation should be kept to the minimum of *those tests of real interest to the researcher.*

The test statistic for a two-tailed comparisons of the ith and jth treatments is based on the difference in mean ranks for the two samples:

$$z_{ij} = \frac{\left|\frac{R_i}{n_i} - \frac{R_j}{n_j}\right|}{\sqrt{\frac{N(N+1}{12}\left(\frac{1}{n_i}+\frac{1}{n_j}\right)}},$$

where N is the total number of observations. This test statistics has an approximate standard normal distribution, and $z_{1-\frac{\alpha}{2}}$ is the critical value. Thus, H_0 is rejected if $z_{ij} \geq z_{1-\frac{\alpha}{2}}$.

2.3.2 The Friedman k-sample test: Matched data

This is a nonparametric analog to Randomized complete block desing ANOVA. There are four assumptions for the Friedman test:

1. The data consists b mutually independent blocks of samples from k random variables. The data is presented in b rows and k columns.
2. The random variable x_{ij} comes from the ith block and is associated with treatment j.
3. The k treatments within each block are assigned randomly. Block are assumed independent.
4. The observations (in blocks) may be ranked according to the variable interest.

There is only one hypothesis:

H_0: All k treatmens have identical effects.

Test uses corrected sum of squares

$$S = \sum_{j=1}^{k}\left[R_j - \frac{b(k+1)}{2}\right]^2.$$

If H_0 is accepted, then all rank sums of treatments should be close to the expected value of $\frac{b(k+1)}{2}$ and S should be close to 0. If H_0 is false, S larger.

The Friedman test statistic is:

$$T = \frac{12}{bk(k+1)}\sum_{J=1}^{k}\left[R_j - \frac{b(k+1)}{2}\right]^2,$$

which should follow χ^2_{k-1} distribution.

In **paired comparisons** the hypotheses

H_0: The effects of the ith and jth treatments are the same;

H_a: The effect of the treatments are different.

The test statistic for a two-tailed comparisons is

$$z_{ij} = \frac{|R_i - R_j|}{\sqrt{\frac{bk(k+1)}{6}}} \ .$$

The critical value is $z_{1-\frac{\alpha}{2}}$. H_0 is rejected if $z_{ij} \geq z_{1-\frac{\alpha}{2}}$.

It should be noted that statistical analysis on variables with **clearly non-normal** distributions is totally its own topic, and this section has just been an introduction on this challenging topic. In general, nonparametric methods should only be used when there is *no other* way to analyze the data. In that case additional reading is also reguired.

More on this topic:
Hollander & Douglas (1999): *Nonparametric Statistical Methods.* John Wiley and Sons.
Gutiérrez-Peña, E., & Walker, S.G. (2005). Statistical Decision Problems and Bayesian Nonparametric Methods. *International Statistical Review*, 3, 309-330.
Rubin, D.B. (1981). The Bayesian boostrap. *The Annals of Statistics*, 9, 130-134.

Chapter 3

Correlation analysis and models of regression

3.1 Correlation analysis

Two variables measured on the same individuals (statistical units) are **associated** if some values of one variable tend to occur more often with some values of the second variable than with other variables of that variable.

The *correlation* measures the direction and strength of the linear relationship between two quantative variables. If we have data on variables x and y for n individuals, the means and standard deviations of the variables are $\bar{x}$ and s_x for the x-values, and $\bar{y}$ and s_y for the y-values. The correlation between x and y is

$$r = \frac{1}{n-1} \sum \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

3.1.1 Pearson product-moment correlation coefficient

Pearsons product-moment correlation coefficient is usually just called **correlation coefficient**.

Consider the standardized normal deviates for x and y

$$\frac{x_i - \bar{x}}{s_x} \text{ and } \frac{y_i - \bar{y}}{s_y}.$$

When corresponding deviates are multipled together and summed, we get the **index of association**

$$\sum \left(\frac{x_i - \bar{y}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

The association has the following characteristics:

1. If large x's are associated with large y's and small x's with small y's, then both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will have the same sign and their product will be positive, and we say that there is a **positive correlation** between x and y.

2. For all pairs of x's and y's, if large x's are associated with small y's and vice versa, the index will be negative because $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will have opposite signs. Thus we'll say that there is a **negative correlation** between x and y.

If we divide the index of association by $n-1$ (degrees of freedom) we obtain the Pearson product-moment correlation coefficient:

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{(n-1)s_x s_y} = \frac{\sum xy - \frac{(\sum n)(\sum y)}{n}}{\sqrt{\left[\sum_x^2 - \frac{(\sum x)^2}{n}\right]\left[\sum y^2 - \frac{(\sum)^2}{n}\right]}} = \frac{SS_{xy}}{\sqrt{SS_x SS_y}},$$

whose values have the range $-1 \leq r \leq 1$. The r is also an *estimate* of the parameter ρ defined by

$$\rho = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

The standard error of the correlation coefficient is

$$s_r = \sqrt{\frac{1-r^2}{n-2}}, \text{ with degrees of freedom } n-2.$$

Using this we can develop a test of hypothesis for ρ:

$$H_0 : \rho = 0$$

$$H_a : \rho \neq 0$$

with test statistic

$$t = \frac{r-0}{s_r}.$$

Example 21. A malacologist interested in the morphology of West Indian chitons, *Chiton olivaceous*, measured the lenght (anterior-posterior) and width of the eight overlapping plates composing the shell of 10 of these animals. She would like to know is the length associated with the width of the overlapping plates.Her data:

Animal	Lenght (cm)	Width (cm)
1	10.7	5.8
2	11.0	6.0
3	9.5	5.0
4	11.1	6.0
5	10.3	5.3
6	10.7	5.8
7	9.9	5.2
8	10.6	5.7
9	10.0	5.3
10	12.0	6.3

The correlation coefficient is

$$r = \frac{599.31 - \frac{(105.8)(56.4)}{10}}{\sqrt{\left[1123.9 - \frac{(105.8)^2}{10}\right]\left[319.68 - \frac{(56.4)^2}{10}\right]}} = \frac{2.598}{\sqrt{(4.536)(1.584)}} = 0.969$$

The standard error of the correlation coefficient is

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-(0.969)^2}{10-2}} = 0.087.$$

So the test statistic is

$$t = \frac{r-0}{s_r} = \frac{0.969-0}{0.087} = 11.14$$

The critical value from t distribution for v=8 with α=0.05 are $\pm$2.306. Thus there seems to be *strong linear correlation* between length and width of chiton shells.

Determination of a 95% confidence interval for ρ is complicated by the fact that only when $\rho = 0$ can r considered to come from an approximate normal distribution. For other values than $\rho = 0$ Fisher's Z transformation must be employed.

$$Z = tanh^{-1}r = 0.5ln\left(\frac{1+r}{1-r}\right).$$

This is usually done with the help of a statistical program. The 95% confidence interval for for ρ in example 20 is [0.8625, 0.9927]. This again indicates a strong co-relationship between the two morphological measurements.

In the future the malacologist could save time and energy by measuring only one of the dimensions, either length or width, because these variables behave in a coordinated and highly predicteble fashion.

3.1.2 Correlation analysis based on ranks

Data for a correlation analysis can also consists of a bivariate random sample of paired observations of size n, $(x_1, y_1), ..., (x_n, y_n)$. If each pair is assumed to come from a continuous population, there should be no tied x's or y's. However, with the nonparametric tests described in this secdion it is sufficient if the x and y observations can be ranked from lowest to highest.

Kendall's measure of correlation, τ

The Kendall correlation coefficient depends on a direct comparison of the n observations (x_i, y_i) with each other. Two observations, for example (190, 186) and (182, 185), are called **concordant** if both members of one pair are large than the corresponding members of the other pair (here $190 > 182$ and $186 < 185$). A pair of observations, such as (180, 188) and (182, 185) are called **discordant** if a number in the first pair is larger than the corresponding number in the second pair ($188 > 185$), while the other number in the first pair is smaller than the corresponding number in the secodn pair ($180 < 182$). Pairs with at least one *tie* between respective members are neither concordant nor discordant.

Let C denote the number of concordant pairs of observations, D the number of discordant pairs, and E the number of ties. The Kendall correlation coefficient is defined in terms of the difference between C and D divided by the total number of comparisons,

$$\tau = \frac{C-D}{\frac{n(n-1)}{2}} = \frac{2(C-D)}{n(n-1)}.$$

The τ is a statistic, not a parameter. Some notices about τ:

- If all $\frac{n(n-1)}{2}$ comparisons are concordant (a "perfect" positive correlation), then $\tau = +1$.
- If all $\frac{n(n-1)}{2}$ comparisons are discordant, then $\tau = -1$.
- In other cases $-1 < \tau < +1$.
- Ties are not counted in τ since they do not constitute evidence for either a positive or negative correlation.

Test for independence or zero correlation is conducted using hypotheses:

- $H_0 : \tau = 0$ (or "x and y are independent").
- $H_a : \tau \neq 0$ (or "x and y are not independent").

Because exact distribution is difficult to tabulate, this test is usually conducted with the help of statistical programs.

Spearman's coefficient, r

On of the most common correlation coefficients that appear in the literature is **Spearman's rank correlation coefficient**, r_s. The idea is to rank the x and y observations separetely and compute the Pearson correlation coefficient on the *ranks* rather than on the original data. the value r_s is usually different from the value of Pearson's r calculated on the *original data*, but for large sample sizes the two values are usually relatively close. However, if there are no ties among the x's and y's, then r_s can be computed much more simply than Pearson's r.

Spearman's rank correlation coefficient (assuming no ties):

$$r_s = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2-1}$$

where $d_i = r_{x_i} - r_{y^i}$ is the difference in the rank of x_i and y_i.

3.1.3 Properties of correlation

Allthough correlation is fairly simple analysis tool there are several things that need to be taken into account when interpreting the results of correlation coefficient:

- Correlation makes no use of the distinction between explanatory and response variables.
- Correlation requires that both variables be quantative, so that it makes sense to do the arithmetic indicated by the formula for r.
- Because r uses the standardized values fo the observations, it does not change when we change the units of measurement of x, y, or both.

- Positive r indicates positive association between the variables, and negative r indicates negative association.
- The correlation is always a number between -1 and 1. The values near 0 indicate a very weak linear relationship.
- Correlation measures the strength of only the linear relationship between two variables.
- Like the mean and standard deviation, the correlation is not resistant: r is strongly affected by a few outlying observations.

It should also be remembered that **correlation is not a complete description of two-variable data**, even when the relationship between the variables is linear. You should report the means and standard deviations of both x and y along with the correlation. Conclusions based on correlations alone may require rethinking in the light of a more complete description of the data.

3.1.4 Inference for correlation

The correlation coefficient is a measure of the strength and direction of the linear association between two variables. Correlation does not require an explanatory-response relationship between the variables. We can consider the sample correlation r as an estimate of the correlation in the population and base inference about the population correlation r.

The correlation between variables x and y when they are measured for every member of population is the **population correlation**. Let ρ be the population correlation. When $\rho = 0$ there is no linear association in the population. In the case where the two variables x and y are both normally distributed (or **jointly normal**), the condition $\rho = 0$ is equivalent to the statement that x and y are independent. That is, there is no association of any kind between x and y.

Example 21. Pearsons sample correlation between body density and the log of the skinfold measures is r= -0.849. The size of the sample is n=92. The t statistic for testing the null hypothesis that the population correlation is zero is

$$\begin{gathered} t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \\ = \frac{-0.849\sqrt{92-2}}{\sqrt{1-(-0.849)^2}} \\ =\text{-}15.5. \end{gathered}$$

Using n-2=90 we have $P<0.0001$.

3.2 Joint distributions

Two variables x and y belong to a same *probability space* if they describe the same *phenomenom*. For example, the whitecell count of a cancer patient (X) and remainig life time of the patient (Y) belong to a same probability space. When variables belong to a same probability space they are said to have a **joint probability distribution** that describes their relationship.

If X and Y are discrete random variables, the function f(x,y), which gives the probability that $X = x$ and $Y = y$ for each pair of values (x,y) within the range of values of X and Y, is called the joint probability distribution of X and Y.

Example 22. Let's assume that a field, which soil is homogenous, is divided on different areas (n) which has been fertilized with different amounts $x_1, ..., x_n$ of the same fertilizer. The observed crop yields are $y_1, ..., y_n$. Random variables x_n and y_n now clearly belong to a same probability space.

Lets now assume that we have a *two dimensional distribution* where all density functions are normal curves. If the two variables are X and Y, then the two dimensional distribution function has the form

$$f_{X,Y} = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-q(x,y)}, \tag{3.1}$$

where

$$q(x,y) = \frac{1}{2(1-\rho^2)} \left[(\frac{x-\mu_1}{\sigma_1})^2 + (\frac{y-\mu_2}{\sigma_2}) - 2\rho(\frac{x-\mu_1}{\sigma_1})(\frac{y-\mu_2}{\sigma_2}) \right]$$

In this case

$$X \sim N(\mu_1, \sigma_1^2) \text{ and } Y \sim N(\mu_2, \sigma_2^2).$$

It is also easy to show that

$$\sigma_{12} = cov(X,Y) = E(X-\mu_1)(Y-\mu_2) = \rho\sigma_1\sigma_2$$

where parameter ρ represents the correlation coefficient

$$corr(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = \rho$$

Thus we have a distribution family (represented in equation 3.1) with the parameters $\mu_1, \mu_2, \sigma_1, \sigma_2$ and ρ (with constraints $\sigma_1, \sigma_2 > 0, |\rho| \leq 1$). Instead of correlation ρ, covariance can also be used as parameter

$$\sigma_{12} = cov(X,Y) = \rho\sigma_1\sigma_2$$

Lets now assume that we have a two dimensional random variable Z which is presented in a vector form as $Z = (X \quad Y)'$. The parameters in the distribution 3.1 are usually grouped to a *vector of expected values*

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = E \begin{pmatrix} X \\ Y \end{pmatrix} = EZ$$

and to a *covariance matrix*

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} = E \begin{pmatrix} X - \mu_1 \\ Y - \mu_2 \end{pmatrix} (X - \mu_1 \quad Y - \mu_2)$$

Now the distribution of Z is

$$Z = (X \quad Y)' \sim N_2(\mu, \Sigma)$$

The shape of the distribution function (equation 1) reveals that when

$$\rho = 0 \Longleftrightarrow X \perp\!\!\!\perp Y,$$

that is, when then population correlation is zero, variables X and Y in vector Z are said to be independent.

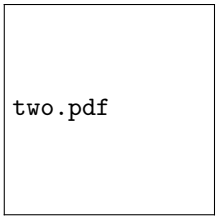
If we form a conditional density function $f_{Y|X=x}(y)$, i.e. density function with the *condition* that Y is defined by changes in values of X, on equation 3.1, we have

$$f_{Y|X=x}(y) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(y-\mu_Y(x))^2} \tag{3.2}$$

where

$$\mu_Y(x) = E(Y|X = x).$$

The conditional distribution function (**??**) is clearly a one-dimensional normal distribution function, which is determined by variance (σ^2) and the term $(y - \mu_Y(x))^2$.



3.3 Simple linear regression model

3.3.1 Least-squares regression

Correlation measures the direction and strength of the linear (straight-line) relationship between two quantative variables. If a scatterplot shows a linear relationship, we could summarize this overall pattern by drawing a line on the scatterplot.

A *regression line* is a straight line that describes how a response variable y changes as explanatory variable x changes. Regression line is often used to **predict** the value of y for a given value of x. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

Example 23. How do children grow. The pattern of growth varies from child to child, so we can best understand the general pattern by following the average height of a number of children. Table 2.7 presents the mean heights of a group of children in Kalama, an Egyptian village that was the site of a study of nutrition in developing countries. The data were obtained by measuring the heights of 161 children from the village each month from 18 to 29 months of age.

Figure 3.1 below is a scatterplot of the data in table 2.7. Age is the explanatory variable, which we plot on the x axis. The plot shows a strong positive linear association with no outliers. The correlation is r=0.994, so a line drawn trough the points will describe these data very well.

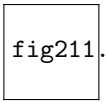


Figure 3.1: KalamaI

Figure 3.2: KalamaII

3.3.2 Fitting a line to the data

Fitting a line to data means drawing a line that comes close as possible to the points. The equation of a line fitted to the data gives a compact description of the dependence of the response variable on the explanatory variable. It is a mathematical model for the straight-line relationship.

Suppose that y is a response variable and x is a explanatory variable. A straight line relating y to x has an equation of the form

$$y = a + bx$$

In this equation, b is the **slope**, the amount by which y changes when x increases by one unit. The number a is the **intercept**, the value of y when $x = 0$.

Any straight line describing the Kalama data has the form

$$\text{height}=a + b \times age.$$

In figure 3.2 we have drawn the regression line with the equation

$$\text{height}=64.93+0.635\times age.$$

The figure 3.2 shows that this line fits the data well. The slope b=0.635 tells us that the height of Kalama children increases by about 0.6 centimeter for each month of age. The slope b of a line $y = a + bx$ is the *rate of change* in the response y as the explanatory variable x changes. The slope of a regression line is an important numerical description of the relationship between the two variables. The intercept a=64.93 cm, would be the mean height at birth if the straight-line pattern of growth were true starting at birth. Children don't grow at fixed rate from birth, so the intercept a is not important in our situation except as part of the equation of the line.

We can use a regression line to **predict** the response y for a specific value of x. The accuracy of predictions from a regression line depends on how much scatter about the line the data shows. In the Kalama example, the data points are all very close to the line, so we are confident that our prediction is accurate. If the data show a linear pattern with considerable spread, we may use a regression line to summarize the pattern but we will put less confidence in predictions based on the line (figure 3.3).

Extrapolation is the use of a regression line for prediction far outside the range of values of the explanatory variable that you used to obtain the line. Such predictions are often not accurate.

In Kalama data birth (age 0) lies outside the range of data. Thus the estimate on mean height at birth (64.93 cm) is inaccurate. This is also a question of parent population. In Kalama example your parent population is clearly restricted to Kalama and you cannot use it to predict the height of children in Cairo, much less in Tokyo.

3.3.3 Method of least-squares

Diffent people might draw different lines by eye on a scatterplot. This holds especially in cases when data points are widely scattered. No line will pass exactly through all the points, but we want one that is as close as possible. We will use the line to predict y from x, so we want one that a line that is as close as possible to the points in the *vertical* direction. That's because the prediction errors we make are errors in y, which is the vertical direction in the scatterplot (in this case). If we predict 85.25 centimeters for the mean height at age 32 months and the actual mean turns out to be 84 centimeters, our error is

error=observed height - predicted height=84 - 85.25= -1.25 centimeters.

We want a regression line that makes these prediction errors as small as possible. Figure 3.4 illustrates the idea. For clarity, the plot shows only three of the points from figure 3.2, along with the line, on an expanded scale.

The **least-suares regression line of** y **on** x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

If we have n observations on two variables x and y as

$$(x_1, y_1), ..., (x_n, y_n)$$

and we draw a line $y = a + bx$ through the scatterplot of these observations, the line predicts the value of y corresponding to x_i as $\hat{y} = a + bx_i$.

The method of least-squares chooses the line that makes the sum of the squares of these errors as small as possible. To find this line, we must find the values fo the intercept a and the slope b that minimize

$$S(a, b) = \sum (error)^2 = \sum (y_i - a - bx_i)^2$$

for given observations x_i and y_i. Function S is minimized by taking partial derivatives with respect to a and b:

$$\begin{aligned} \frac{\partial S}{\partial a} &= 2na - 2\sum y_i + 2\sum x_i b \\ \frac{\partial S}{\partial b} &= 2\sum x_i^2 b - 2\sum x_i a \end{aligned}$$

Kalamaesim.pdf

Figure 3.3: KalamaIII

Figure 3.4: KalamaIV

and setting these partial derivatives to zero and multiplying them by 0.5, from which we get

$$\begin{array}{c} na - \sum y_i + \sum x_i b = 0 \\ \sum x_i^2 b - \sum x_i y_i + \sum x_i a = 0. \end{array}$$

Thus, the parameter estimators are:

$$\begin{array}{c} a = \bar{y} - b\bar{x} \\ b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} \end{array}$$

We have data on an explanatory variable x and a response variable y for n individuals. The means and standard deviations of the sample data are $\bar{x}$ and s_x for x and $\bar{y}$ and s_y for y, and the correlation between x and y is r. The **equation of the least-squares regression line of y on x is**

$$\hat{y} = a + bx$$

with slope

$$b = r\frac{s_y}{s_x}$$

and intercept

$$a = \bar{y} - b\bar{x}.$$

Example 24. From table 2.7 the mean and standard deviations of the 12 ages are

$$\bar{x}\text{=23.5 months and } \bar{s_x}\text{=3.606 months}$$

The mean and standard deviation óf the 12 heights are

$$\bar{y}\text{=79.85 cm and } s_y\text{=2.302 cm.}$$

The correlation between height and age is r=0.9944. Thus

$$b = 0.9944\frac{2.302}{3.606}\text{=0.6348 cm per month}$$

and intercept

$$a\text{=79.85-0.6348}\times\text{23.5 = 64.932 cm}$$

The equation of the least-squares line is

$$\hat{y}\text{=64.932+0.6348}x.$$

3.3.4 Interpreting the regression line

The slope b=0.6348 centimeters per month in example 24 is the rate of change in the mean height as age increases. The unit "centimeters per month" come from the units of y (centimeters) and x (months). Although the correlation does not change when we change the units of measurement, the equation of the least-squares line does change. The slope in centimeters per month is 2.54 times as large as the slope in iches per month, because there are 2.54 centimeters in an inch.

The expression $b = r\frac{s_y}{s_x}$ for the slope says that along the regression line, **a change of one standard deviation in x corresponds to a change of r standard deviations in y**. When the variables are perfectly correlated (r=1 or r=-1), the change in the predicted response $\hat{y}$ is the same (in standard deviation units) as the change in x. Otherwise, because $-1 \leq r \leq 1$ the change in $\hat{y}$ is less than the change in x. As the correlation grows less strong, the prediction $\hat{y}$ moves in less in response to changes in x.

The least-squares regression line always passes trough the point $(\bar{x}, \bar{y})$ on the graph of y against x.

3.4 Inference for regression

A simple linear regression studies the relationship between a response variable y and a single explanatory variable x. We expect that different values of x will produce different mean responses. Figure 5 illustrates the statistical model for a comparison of the blood pressure in two groups of experimental subjects, one group taking a calcium supplement and the other a placebo.

Figure 3.5: Calcium and placebo

In linear regression the explanatory variable x can have many different values. We can think of the values of x as defining different **subpopulations**, one for each possible value of x. Each subpopulation consists of all individuals in the population having the same value of x. If you, for example, give x=1500 milligrams of calcium to everyone, then these are subjects "who actually received 1500 milligrams of calcium". In this case, the subpopulation would be the same as the sample.

The statistical model for simple linear regression assumes that for each value of x the observed values of the response variable y are normally distributed about a mean that depends on x. We use μ_y to represent these means. We are interested in how the many means μ_y changes as x changes. In general the means μ_y can change according to any sort of pattern as x changes. In simple

linear regression we assume that they all lie on a line when plotted againts x. The equation of the line is

$$\mu_y = \beta_0 + \beta_1 x.$$

This is the **population regression line**; it describes how the mean response changes with x. The model assumes that this variation, measured by the standard deviation σ, is the same for all values of x.

3.4.1 Simple linear model

Example 25. In example 22 it was assumed that we had a field, which soil is homogenous, and which is divided on different areas (n) which have been fertilized with different amounts $x_1, ..., x_n$ of the same fertilizer. The observed crop yields are $y_1, ..., y_n$. We would like to build a model that would describe the effect of fertilization of the crop yields.

Lets assume that observed yields can be intrepeted as observed values of independent random variables $Y_1, ..., Y_n$. We are interested on the expected values of these random variables, namely $E(Y_i) \equiv \mu_i$, which depend on the prechosen and non-random amounts of fertilizer x_i ($i = 1, ..., n$). We assume that this relationship is linear $\mu_i = \beta_1 + \beta_2 x_i$. We also assume that the amount of fertilizer does not affect on the (unknown) variances of the random variables Y_i, i.e. $var(Y_i) = \sigma^2$ for all $i = 1, ..., n$.

The model is

$$Y_i = \beta_1 + \beta_2 x_i + \epsilon_i, \quad i = 1, ..., n, \tag{3.3}$$

where $\epsilon_1, ..., \epsilon_n$ are independent, identically distributed random variables with $E(\epsilon_i) = 0$ and $var(\epsilon_i) = \sigma^2$. As with ANOVA these can be interpreted as *random fluctuations*, which is not explained by the amount of fertilizers.

Equation 3.3 is called a **simple linear regression model**. A general linear model assumes that variables are normally distributed and that they have a joint normal distribution, i.e. they have a *multinormal* distribution. Because amount of fertilizers is interpreted as non-random, the joint probability distribution is formed by the assumption

$$Y_i \sim N(\beta_1 + \beta_2 x_i, \sigma^2)$$

fig102.pdf

Figure 3.6:

Although it is reasonable to assume that $\beta_1 > 0$, we model parameter space on more common way $\beta_1, \beta_2 \in \mathbb{R}, \sigma^2 > 0$. Usually it is also assumed that

$$\epsilon_i \sim N(0, \sigma^2)$$

In general, given n observations on the explanatory variable x and the response variable y

$$(x_1, y_1), ..., (x_n, y_n)$$

the **statistical model for simple linear regression** states that the observed response y_i when the explanatory variable takes the value x_i is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Here $\beta_0 + \beta_1 x_i$ is the mean response when $x = x_i$.

Because the means μ_y lie on the line $\mu_y = \beta_0 + \beta_1 x$, they are all determined by β_0 and β_1. Once we have the estimates of β_0 and β_1, the linear relationship determines the estimates of μ_y for all values of x. Linear regression allows us to do inference not only for subpopulation for which we have data but also for the corresponding to x's not present in the data.

3.4.2 Estimating the regression parameters

The predicted value of y for a given value $x*$ of the x is the point on the least-squares line $\hat{y} = b_0 + b_1 x*$. This is an unbiased estimator of the mean response μ_y when $x = x*$. The **residuals** are

$$e_i\text{=observed response - predicted response}$$

$$= y_i - \hat{y}_i$$

$$= y_i - b_0 - b_1 x_i.$$

The residuals correspond to the model deviations ϵ_i. The e_i sum to 0, and the ϵ_i come from a population with mean 0.

The remaining parameter to be estimated is σ, which measures the variance of y about the population regression line. Because this parameter is the standard deviation of the model deviations, we use the residuals to estimate it:

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

3.4.3 Assumptions in linear regression models

Let

$$y_n = \beta' X_n + \epsilon_n$$

be a linear regression model, where β is a vector of parameters to be estimated and X_n a vector of explanatory variables. Following assumption must be valid in order to make the estimation of the above model feasible in such a way that the estimates are subject to statistical inference:

1. $E(\epsilon_n | x_n) \equiv 0$
2. $cov((\epsilon_1, ..., \epsilon_n)') = \sigma^2 I$
3. $\{\epsilon_n\} \perp\!\!\!\perp \{X_n\}$
4. $\epsilon_n \sim NID(0, \sigma^2)$


autocorrelationI.pdf


autocorrelationII.pdf


heteroscedasticity.pdf

3.4.4 Confidence intervals and tests of significance

In regression analysis the basic idea about variation in the data is similar to that in the ANOVA. That is

$$\text{DATA=FIT+RESIDUAL}$$

This can also be expressed as

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

If we square each of the three deviations above and then sum over all n observations, it is algebraic fact that the sums of squares add:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

This equation can be rewrited as

$$SS_T = SS_M + SS_E$$

Where T, M, and E stand for total, model, and residual (or unexplained variation). Estimate on the variance about the population regression line is:

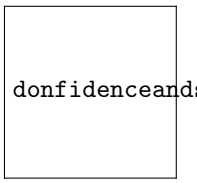
$$MS_E = s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

With the help of these we can calculate the *standard errors for estimated regression coefficients* or SE's. The standard error of the slope b_1 of the least-squares regression line is

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}.$$

The standard error of the intercept b_0 is

$$SE_{b_0} = s\sqrt{\frac{1}{n} + \frac{\bar{x^2}}{\sum (x_i - \bar{x})^2}}$$



The fraction of variation in y explained by x can be calculated with the help of sums of squares, namely

$$r^2 = \frac{SS_M}{SS_T} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}.$$

This is because SS_T is the total variation in y and SS_M is the variation due to the regression of y on x.

3.4.5 The ANOVA table

Source	DF	SS	MS	F
Model	p	$\sum(\hat{y}_i - \bar{y})^2$	SS_M/DF_M	MS_M/MS_E
Error	$n-p-1$	$\sum(y_i - \hat{y}_i)^2$	SS_E/DF_E	
Total	$n-1$	$\sum(y_i - \bar{y})^2$	SS_T/DF_T	

3.5 Multiple regression

3.5.1 Inference for multiple regression

In the multiple regression setting, the response variable y depends on not one but n explanatory variables. These explanatory variables will be denoted by $x_1, x_2, ..., x_n$. The mean response is a linear function of the explanatory variables:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n.$$

This expression is the **population regression equation**. We do not observe the mean response because the observed values of y vary about their means. We can think of subpopulations of responses, each correponding to a particular set of values for *all* of the explanatory variables $x_1, ..., x_n$. In each subpopulation, y varies normally with a mean given by the population regression equation.

3.5.2 Multiple linear regression model

Statistical unit	Response variable; y	Explanatory variables; $x_1, ..., x_p$
1	y_1	x_{11},...,x_{1p}
.	.	.
.	.	.
.	.	.
n	y_n	$x_{n1}, ..., x_{np}$

If $Y_1, ..., Y_n$ are random variables, whose observed values are $y_1, ..., y_n$ like in equation 3.3, the *general linear model* can be defined as

$$Y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, i = 1, .., n \tag{3.4}$$

from which the equation 3.3 is a special case where $P = 2$ and $x_{i1} = 1 \ \forall$ i. In equation 3.4 the explanatory variables (x_{ij}) on the right side are *non-random* or *fixed* numbers, $\beta_1, ..., \beta_p$ are unknown parameters, and ϵ is a non-observable random variable that is related to statistical unit i. ϵ_i describes the part of the variation in the response variable that the explanatory variables or their linear combination $\beta_1 x_{i1} + ... + \beta_p x_{ip}$ cannot explain. Linear combination $\beta_1 x_{i1} + ... + \beta_p x_{ip}$ is called the *structure* or *systematic part* of the model. Linearity means that systematic part of the model is a *linear function* of the parameters $\beta_1, ..., \beta_p$ and that the residual becomes added to the systematic part *additively*. In order to make the model 4 functional under the assumption of statistical inference the joint distribution function and parameter space must be determined. This is done by incorporating the assumptions 1-4 presented in section 3.3 in to model 4. Assuming NID (normal, identical distribution) on residuals transforms model 4 to a *multiple linear regression model*.

The multiple linear regression model is a special case of the general linear model in equation 3.3.

$$Y_i = \beta_1 + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, i = 1, ..., n. \tag{3.5}$$

In this case is assumed that variables in the model are an SRS from a normally distributed population.

multiplelinear.pdf

Example 26. Lets consider a study where factors that affect on the short-run growth of trees were studied. In the long run the growth of trees is determined by rainfall, soil, and many other factors. In short-run it can be argued that the single most important factor that determines the level of growth hormone in a tree is temperature (x). To study this, growth of a single pine seedling were observed in two hour intervals for 10 days. The height of the seedling was measured with a measuring rod and a automatic camera, with the accuracy of the measurement being 0.1 millimeters.

A simple linear model was fitted on the data:

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t,\ \epsilon_t \sim NID(0, \sigma^2)$$

where

$$Y_t \text{ is the growth rate of the pine during time period } t$$
$$X_t \text{ is the temperature during time period } t.$$

After regression the parameter estimate of β_2 was $\hat{\beta}_2 = 0.053$ and the "explanatory power" of the model was $R^2_{X,Y} = 0.32$.

However, because the temperature usually changes rather slowly, the X_{t-1} might also effect on the value of the explanatory variable on time t. If this is not taken into account in the model, estimated effects of imminent temperature changes (on time t) on the growth rate may be biased. Thus if the changes in the temperature affect on the growth rate of the pine after two hours, X_{t-1} does have an effect also on Y_t. On the other hand, temperature changes for over 4 hours or more can also affect on the growth rate. Therefore, the model should be something like

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + \beta_5 X_{t-3} + \epsilon_t.$$

Estimation of above model gives $\hat{\beta}_2 = 0.021$ and $R^2_{X,Y} = 0.51$. The fit of the model is considerably better.

Even if discarding the additional explanatory variables would not lead to *biased* parameter estimates, there might be other reasons why one would want to include "additional" explanatory variables on the model. One goal could be the attemp to build a quantative model that had the *greatest* possible explanatory power on the variations in y. Other reasons might include:

- urge to describe and test the dependencies (or interactions) between explanatory variables x and response variable y,
- forecasting the values of y with the help of some background measures x_n,
- and finding out the *suitable* values of x when one wants to *guide* y to a certain level.

3.5.3 Estimation of multiple regression parameters

Let

$$b_0, b_1, ..., b_n$$

denote the estimators of the parameters

$$\beta_0, \beta_1, ..., \beta_n.$$

For the ith observation the predicted response is

$$\hat{y_i} = b_0 + b_1 x_{i1} + ... + b_p x_{ip}$$

The ith residual, the difference between the observed and predicted response, is therefore

$$\begin{aligned} e_i &= \text{observed response - predicted response} \\ &= y_i - \hat{y_i} \\ &= y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip} \end{aligned}$$

The **method of ordinary least squares** chooses the values of the b's that make the sum of the squares of the residuals as small as possible. In other words, the parameter estimates $b_0, b_1, ..., b_p$ minimize the quantity

$$S(b_0, b_1, ..., b_p) = \sum (y_i - b_0 - b_1 x_{i1} - ... - b_p x_{ip}).$$

This is done the same way as in the case of simple linear regression model. First we take partial derivatives on S with respect ot parameter estimates:

$$\frac{\partial S}{\partial b_j} = 2 \sum e_i \frac{\partial e_i}{\partial b_j} = -2 \sum e_i \frac{y_i}{\partial b_j} = -2 \sum e_i x_{ij}$$

Thus the parameter estimates $b_0, ..., b_p$ satisfies p simultaneous equations

$$\sum \hat{e_i} x_{ij} = 0, \, \forall \; j = 0, 1, ..., p$$

In matrix notation $X' \hat{e} = 0$. When we insert $\hat{e} = y - Xb$ in to this, we get

$$X'(y - Xb) = 0.$$

Solving for b gives:

$$b_{OLS} = (X'X)^{-1} X'y$$

The parameter σ^2 measures the variability of the responses about the population regression equation. As in the case of simple linear regression, we estimate σ^2 by an average of the squared residuals. The estimator is

$$\begin{aligned} s^2 &= \frac{\sum e_i^2}{n-p-1} \\ &= \frac{\sum (y_i - \hat{y_i})^2}{n-p-1} \end{aligned}$$

The quantity $n - p - 1$ is the **degrees of freedom** associated with s^2. The degrees of freedom equal the sample size n minus $p + 1$, the number of β's we must estimate to fit the model. In the simple linear regression case there is just one explanatory variable, and $p = 1$ and the degrees of freedom are $n - 2$. To estimate σ we use

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

3.5.4 Confidence intervals and significance tests for regression coefficients


confidencebetaj.pdf

3.5.5 ANOVA table for multiple regression

Source	DF	SS	MS	F
Model	p	$\sum(\hat{y}_i - \bar{y})^2$	SS_M/DF_M	MS_M/MS_E
Error	$n-p-1$	$\sum(y_i - \hat{y}_i)^2$	SS_E/DF_E	
Total	$n-1$	$\sum(y_i - \bar{y})^2$	SS_T/DF_T	

In multiple regression analysis the hypotheses tested are

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$

$$H_a : \text{At least one of the } \beta_j \text{ is not 0.}$$

For simple linear regression we noted that the square of the sample correlation could be written as the ratio of SS_M to SS_T and could be interpreted as the proportion of variation in y explained by x. A similar statistic is routinely calculated for multiple regression. The statistic

$$R^2 = \frac{SS_M}{SS_T} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

is called **the squared multiple correlation**. Often R^2 is multiplied by 100 and expressed as a percent. The square root of R^2 is called **multiple correlation coefficient**, and it measures the correlation between the observations y_i and the predicted values $\hat{y}_i$.

3.6 Notes

3.6.1 Identification

Lets assume that we have a model

$$Y = \beta_1 + \beta_2 X + \epsilon, \quad \epsilon \sim N(0.\sigma^2) \tag{3.6}$$

but there would be *interactions* between quantative variables y and x, which describe the random variables X and Y, i.e there would be interactions on $x \longrightarrow y$ and also on $y \longrightarrow x$. In this case the assumption 3 ($\{\epsilon_n\} \perp\!\!\!\perp \{X_n\}$) would be unrealistic, because the ϵ was assumed to be part of the random variable Y. Because the Y would affect on X, X could not be independent on ϵ.

This can demonstrated by modelling the both interactions with linear regression models:

$$Y = \beta_1 + \beta_2 X + \epsilon, N(0, \sigma^2) \tag{3.7}$$

$$X = \gamma_1 + \gamma_2 Y + \epsilon_*, N(0, \sigma_*^2 \tag{3.8}$$

Even if "errors" ϵ and ϵ_* were assumed independent, the distribution of variables X and Y would still be some two-dimensional normal distribution in above model. In all two-dimensional distributions only there are only 5 *parameters* that can be estimated. However, there are already *six* parameters in the above models. Thus it does not contain enough *information* so that all six parameters could could be estimated.

This can also be shown by denoting that correlation coefficient $\rho = corr(X, Y)$ in a *symmetric* measure of dependence. So, if there are causal relationships in *both* directions, correlation cannot be used to interpret the relationships.

Another way to describe this identification problem would be to consider the *moments* of linear regression model:

$$Y = \beta_1 + \beta_2 X + \epsilon, \quad \epsilon \perp\!\!\!\perp X, \quad \epsilon \sim N(0, \sigma^2), \tag{3.9}$$

where our parameter space is

$$B = \{(\beta_1 \quad \beta_2 \quad \sigma^2)' | \sigma^2 \geq 0\} \subset \Re^3.$$

We can estimate our model (3.9) by using *moments* of random variables. These are expected values, variance, covariance, etc. From model (3.9) we can see that

$$\begin{cases} E(Y) = \beta_1 + \beta_2 E(X) \\ cov(X, Y) = \beta_2 var(X) \\ var(Y) = \beta_2^2 var(X) + \sigma^2 \end{cases}$$

Now we have three "unknown" parameters β_1, β_2 and σ^2, which can be solved from the three equations above:

$$\begin{cases} \beta_2 = \frac{cov(X,Y)}{var(X)} \\ \beta_1 = E(Y) - \beta_2 E(X) = E(Y) - \frac{cov(X,Y)}{var(X)} E(X) \\ \sigma^2 = var(Y) - \beta_2^2 var(X) = var(Y) \left(1 - \frac{cov(X,Y)^2}{var(X)var(Y)}\right). \end{cases}$$

The moments that are present in these expressions can be estimated with "sample moments":

$$\begin{aligned} &\widehat{E(X)} = \bar{X} = \tfrac{1}{n} \textstyle\sum_{i=1}^n X_i, \\ &\widehat{E(Y)} = \bar{Y} = \tfrac{1}{n} \textstyle\sum_{i=1}^n Y_i, \\ &\widehat{var(X)} = S_X^2 = \tfrac{1}{n-1} \textstyle\sum_{i=1}^n (X_i - \bar{X})^2, \\ &\widehat{var(Y)} = S_Y^2 = \tfrac{1}{n-1} \textstyle\sum_{i=1}^n (Y_i - \bar{Y})^2, \\ &\widehat{cov(X,Y)} = S_{X,Y} = \tfrac{1}{n-1} \textstyle\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X}). \end{aligned}$$

By writing correlation coefficient as

$$R_{X,Y} = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} = \frac{\sum_{i=1}^n (Y_i - \bar{Y}(X_i - \bar{X})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

our parameter estimates become

$$\begin{aligned} &\hat{\beta_2} = S_{X,Y} = R_{X,Y} \tfrac{S_Y}{S_X} \\ &\hat{\beta_1} = \bar{Y} - \beta_2 \bar{X} = \bar{Y} - R_{X,Y} \tfrac{S_Y}{S_X} \bar{X} \\ &\hat{\sigma^2} = S_Y^2 (1 - R_{X,Y}^2). \end{aligned}$$

However, if the assumption $\{\epsilon_n\} \perp\!\!\!\perp \{X_n\}$ is broken, we have *simultaneous equation* model like the one presented in models 3.7 and 3.8. In this case our parameter space would be $B = \{(\beta_1 \quad \beta_2 \quad \gamma_1 \quad \gamma_2 \quad \sigma^2 \quad \sigma_*^2)' | \sigma^2, \sigma_* \geq 0\} \subset \Re^6$. Now there are six parameters, but we only have 5 sample moments (5 dimensional parameter space) in the joint normal distribution of two random variables. Therefore we do not have *enough* information in the model (3.6) so that all the parameters in the model could be estimated.

Example 27. Consider the annual development of prices and wages. It can be argued that wages affect prices and prices affect wages. Therefore there is a clear two-way causal relationship and regression analysis cannot be used to analyze this dependence.

In this case there would be a need to find some instrumental variable Z on X, that would be highly correlated with X, but $EZ\epsilon \equiv 0$. This instrumental variable would provide enough information for the identification of all the parameters in equation 6. The estimation would now be conducted with instrumental variables estimator:

$$b_{IVE} = (Z'X)^{-1}Z'Y$$

However, this method is beyond the scope of this course. Additional reading:
- Pearl, J. *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2000.
- Greene, William, H. *Econometric Analysis*, Prentice Hall, 2000/2003.

Lets now assume that variables x and y are related to each other through some "deterministic" physical law. In these situations there is usually a need to prepare for considerable measurement errors when measuring the variables x and y. Lets concider the setup where these *observations* would be described by random variables X and Y, where

$$\begin{gathered} y = \beta_1 + \beta_2 x \\ Y = y + \epsilon \\ X = x + \epsilon_* \end{gathered}$$

Then a model that would describe these *observations* would be of the form

$$Y = \beta_1 + \beta_2 X + \kappa, \tag{3.10}$$

$\kappa = \epsilon - \beta_2 \epsilon_* \sim N(0, \sigma^2 + \beta_2^2 \sigma_*^2)$

In this case the assumption $\kappa \perp\!\!\!\perp X$ is clearly unrealistic, because ϵ_* is included in both κ and X. If x could be measured without error, i.e. $\sigma_*^2 = 0$, there would, of course, be no problem. Thus **the variable that is measured with greater error should be the response variable.** Alternatively could try to find some instrumental variable Z.

3.6.2 Linearity in parameters

Linearity in regression models means that regression function is linear with respect to *parameters* β, not with respect to values of x. This means that, if the

relationship between variables x and y can be *linearized* with some tranformation, even some *non-linear relations* between variables can be examined using regression analysis.

Example 28. Does the GDP level of a nation affect on its population level? In development economics, some theories argue that too high population could lead to a *poverty trap*, where economy of cannot grow because providing the population requires a "lions share" of the capital. The dataset includes data on real gross domestic product per capita (GDP) (with the base year 1996), and population rate (thousands) in 188 countries in 2000.

The figures show that although the fit for data is not so good, the common behaviour of variables can be described with two-dimensional normal distribution, where the regression function of Y with respect to X would be linear. So, random variables $population = e^X$ and $GDP = e^Y$ could be used to describe the quantative variables GDP and population. Now we have

$$\begin{aligned} E(GDP|population = w) &= E(e^Y|log(population) = logw) \\ &= E(e^{\beta_1+\beta_2 X+\epsilon}|X = logw) \\ &= e^{\beta_1+\beta_2 logw} E(e^\epsilon|X = logw) \\ &= e^{\beta_1} w^{\beta_2} E(e^\epsilon) \\ &= \beta_1^* w^{\beta_2} \end{aligned}$$

where $\beta_1^* = e^{\beta_1} E(e^\epsilon)$. Now, when $\beta_2 < 0$, the regression function (the third function above) does somewhat resemble the original values of GDP and population in the scatterplot above.

However, this model might suffer from *under identification*, because the assumption that GDP level does not affect on the level of population may be unrealistic. Some theories in demography argue that the population growth slows as country gets more richer. Thus, there would be a two-way causal relationship between the variables and assumption of probabilistic independence between the two variables would be invalid.

The point in the above example was to show that one should always pay enough attention on the "initial transformation" of variables before conducting a *regression analysis*. Otherwise the estimated model could produce extremely poor fit on the data and obtained parameter estimates could be *biased*.

3.7 Introduction to the analysis of time series

A series y_t in which the values of the dependent (response) variable y is observed successively on time periods t, is called a *time series*. Usually observations are collected with *equal* intervals so that the time index gets some whole numbered values, for example between $t = 1, ..., n$.

The general linear regression model has some limitations when it is used to desribe time series data.

$$y_t = \beta' X_t + \epsilon_t \tag{3.11}$$

$$\text{where } \epsilon_t \perp\!\!\!\perp X_t \text{ and } E\epsilon_t \equiv 0.$$

We, of course, have to know that the causality is of the form $X_t \longrightarrow y_t$. In addition, we usually are forced to accept some facts, when we want to model the relationships between time series in realistic way:

- The explanatory variable's X_t effect on the response variable may be lagging.
- Usually, it is not possible to construct such a linear combination $\beta' X_t$ that $y_t - \beta' X_t$ would have a totally irregular variation with respect to time.

Thus, in many cases, the "static" regression model (3.11) has to be replaced with its "dynamic" form

$$y_t = \beta_0' X_t + \beta_1' X_{t-1} + \beta_2' X_{t-2} + \beta_3' X_{t-3} + ... + \epsilon_t \tag{3.12}$$

$$\epsilon_t \perp\!\!\!\perp X_t,\ E\epsilon_t \equiv 0, \text{ but } E(\epsilon_t \epsilon_{t'}) \neq 0 \text{ when } t \neq t'.$$

More complicated models are needed if there are interactions between y_t and X_t.

3.7.1 Expectations and stationarity of a series

If we concider a sequence of T independent and identically distributed (i.i.d.) variables ϵ_t,

$$\{\epsilon_1, \epsilon_2, ..., \epsilon_T\}$$

with

$$\epsilon_t \sim N(0, \sigma^2), \text{ and } E\epsilon_t = 0$$

this is called a *Gaussian white noise process.* In this kind of time series model, there is no systematic variation.

Consider a *realization* (a observed sample) of some random variable Y_t:

$$\{y_1, y_2, ..., y_T\}. \tag{3.13}$$

The expexted value of this series might be viewed as the probability limit of the ensemble average:

$$E(Y_t) = plim_{I\to\infty}(1/I)\sum_{i=1}^{I} Y_t^{(i)}.$$

If, for example, process $\{Y_t\}_{t=-\infty}^{\infty}$ presents the sum of a constant μ plus a Gaussian white noice process $\{\epsilon_t\}_{t=-\infty}^{\infty}$, it is

$$Y_t = \mu + \epsilon_t. \tag{3.14}$$

The mean of this series is

$$E(Y_t) = \mu + E(\epsilon_t) = \mu.$$

If Y_t is a time trend plus gaussian white noise,

$$Y_t = \beta t + \epsilon_t,$$

its mean is

$$E(Y_t) = \beta t$$

The expected value of Y_t with respect to it lagged values

$$\gamma_{jt} = E(Y_t - \mu_t)(Y_{t-j} - \mu_{t-j}),$$

with some number of j observations, is called *autocovariance* of a process (3.13). For example for process (3.14) the autocovariances are

$$\gamma_{jt} = E(Y_t - \mu)(Y_{t-j} - \mu) = E(\epsilon_t \epsilon_{t-j}) = 0 \text{ for } j \neq 0.$$

If the mean μ and the autocovariances γ_{jt} do not depend on time t

$$E(Y_t = \mu), \forall\ t$$
$$E(Y_t - \mu)(Y_{t-j} - \mu) = \gamma_j \ \forall\ t \text{ and } j$$

then the process Y_t is said to be *weakly stationary*. Other way to describe stationarity is to consider the values $Y_t, ..., Y_{t+p}$ as a *joint distribution functions* of random variable Y_t with respect to time $t_1, ..., t_p$. Let $F_{Y_{t_1},...,Y_{t_p}}$ represent a *cumulative distribution function* of the joint distributions of Y_t. Now, if

$$F_{Y_{t_1},...,Y_{t_p}} = F_{Y_{t_{1+\tau}},...,Y_{t_{p+\tau}}} \ \forall\ p, \tau, t_1, ..., t_p$$

process is said to be *strongly stationary*. In more general terms a process is said to be stationary if it has no systematic variation, for example time trends. White noice process described above is a *strongly* stationary process.

3.7.2 MA and AR processes

A process

$$Y_t = \mu + \epsilon_t + \theta\epsilon_{t-1}, \quad E(\epsilon_t) = 0 \quad and \quad E(\epsilon_t^2) = \sigma^2, \tag{3.15}$$

where μ and θ could be any constants, is called *first order moving average process* denoted as $MA(1)$. Equation (3.15) is called as moving average because Y_t is constructed from a weighted sum, or average, of two most recent values of ϵ.

The expected value of (3.15) is

$$E(Y_t) = E(\mu + \epsilon_t + \theta\epsilon_{t-1}) = \mu + E(\epsilon_t) + \theta E(\epsilon_{t-1}) = \mu,$$

where μ represents a constant term which is *anticipated* to be the mean of the process. The variance of Y_t is

$$\begin{aligned} E(Y_t - \mu)^2 &= E(\epsilon_t + \theta\epsilon_{t-1})^2 \\ &= E(\epsilon_t^2 + 2\theta\epsilon_t\epsilon_{t-1} + \theta^2\epsilon_{t-1}^2) \\ &= \sigma^2 + 0 + \theta^2\sigma^2 \\ &= (1 + \theta^2)\sigma^2 \end{aligned} \tag{3.16}$$

A *qth order moving average process* has a form

$$Y_t = \mu + \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q} \tag{3.17}$$

$$E(\epsilon_t) = 0,\ E(\epsilon_t^2) = \sigma^2$$

where $(\theta_1, ..., \theta_q)$ could be any real numbers. The mean of the process (3.17) is again μ, and the variance

$$E(Y_t - \mu)^2 = (1 + \theta_1^2 + \theta_2^2 + ... + \theta_q^2)\sigma^2.$$

A process that satisfies the difference equation

$$Y_t = c + \phi Y_{t-1} + \epsilon_t \tag{3.18}$$

$$E(\epsilon_t) = 0,\ E(\epsilon_t^2) = \sigma^2$$

is called a *first-order autoregression* or $AR(1)$. When $|\phi| < 1$ the process (8) has a solution

$$\begin{aligned} Y_t &= (c + \epsilon_t) + \phi(c + \epsilon_{t-1}) + \phi^2(c + \epsilon_{t-2}) + ... \\ &= [c/1 - \phi] + \epsilon_t + \phi\epsilon_{t-1} + ... \end{aligned} \tag{3.19}$$

Thus, the mean of a stationary $AR(1)$ process is

$$\begin{aligned} E(Y_t) &= [c/(1 - \phi)] + 0 + 0... \\ &\Rightarrow \mu = c/(1 - \phi) \end{aligned} \tag{3.20}$$

Similar to MA processes, the *pth-order autoregression*, $AR(p)$, satisfies

$$Y_t = c + \phi_t Y_{t-1} + \phi_2 Y_{t-2} + ... + \phi_p Y_{t-p} + \epsilon_t \tag{3.21}$$

$$E(\epsilon_t) = 0,\ E(\epsilon_t^2) = \sigma^2.$$

Figure 3.7: Hamilton 1994

.

3.7.3 ARX models

Mixed average moving average, or $ARMA$, processes are processes that include both the autoregressive and the moving average terms:

$$Y_t = c+\phi_1 Y_{t-1}+\phi_2 Y_{t-2}+...+\phi_p Y_{t-p}+\epsilon_t+\theta_1\epsilon_{t-1}+\theta_2\epsilon_{t-2}+...+\theta_q\epsilon_{t-q} \quad (3.22)$$

Estimation of the parameters in combined ($ARMA$) time series models (3.22) is clearly more complicated than in the case of normal linear regression and these models are beyond the scope this course. However, there are one type of time series models whose estimation is quite straightforward. These are called ARX models.

Lets assume that factors X_t would define the *target* level y_t* of the response variable according to

$$y_t* = \beta' X_t,\ t = 1, ..., n,$$

but the process of y_t would include some "inertia" according to

$$y_t - y_{t-1} = \alpha(y_t * -y_{t-1}) + \epsilon_t, \quad \epsilon_t \perp\!\!\!\perp X_t, \quad 0 < \alpha < 1.$$

Thus, we would have a model

$$y_t = (1-\alpha)y_{t-1} + \alpha\beta' X_t + \epsilon_t \quad (3.23)$$

$$\epsilon_t \sim NID(0, \sigma^2), \quad \epsilon_t \perp\!\!\!\perp X_t, \quad t = 2, ..., n$$

Model (3.23) is a simplest formulation of difference model, and it can also be interpreted as

$$y_t = \mu + (1-\alpha)y_{t-1} + \epsilon_t \quad (3.24)$$

where the exogenous variables X_t would only affect linearly on the level ($\frac{\mu}{\alpha}$) to where y_t is *stabilizing*. The model is

$$\alpha(L)y_t = \mu + \sum_{i=1}^{m} w_i(L)x_{it} + \epsilon_t. \quad (3.25)$$

$$\begin{aligned}
&\{\epsilon_t\} \perp\!\!\!\perp \{x_{it}, ..., x_{mt}\}, \quad \epsilon_t \sim NID(0, \sigma^2)\\
&\alpha(L) = 1 - \alpha_1 L - ... - \alpha_p L^p\\
&w_i(L) = w_{i0} - w_{i1} - ... - w_{ir_i} L^{r_i},
\end{aligned}$$

Where (L) denotes lag operator, i.e. $(L)y_t = y_{t-1}$, $(L^2)y_t = y_{t-2}$, etc.. The parameters in model (3.25) can be estimated utilizing the OLS (ordinary least squares) techniques.

The model (3.25) includes one restrictive assumption on the evolution of y_t. That is, the exogenous factors x_t can only affect *linearly* on the location of "equilibrium" of the system, but *not* on to the structure of the system. This way of reasoning might be suitable in some applications, but certainly not in general. The "solution" of the model (15) is

$$y_t = \frac{\mu}{\alpha(1)} + \sum_{i=1}^{m} \frac{w_i(L)}{\alpha(L)} + \frac{1}{\alpha(L)}\epsilon_t,$$

which would mean that the effect of *every* explanatory variable on y_t should the of the *same* form on long lags. This is not very realistic assumption in many cases.

Example 29. On example 26 factors that affect on the short-run growth of trees were studied. We had a simple linear model that was fitted on the data:

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t,\ \epsilon_t \sim NID(0, \sigma^2)$$

where

Y_t is the growth rate of the pine during time period t
X_t is the temperature during time period t.

and a "expanded" model:

$$Y_t = \beta_1 + \beta_2 X_t + \beta_3 X_{t-1} + \beta_4 X_{t-2} + \beta_5 X_{t-3} + \epsilon_t.$$

In addition to this we might also think that the growth in previous periods might affect on the growth rate. For example the production of growth hormone might depend on the production of growth hormone in the previous period. If we would like to take this into account, the model should be transformed to

$$Y_t = \mu + \phi_1 Y_{t-1} + \beta_1 X_t + \beta_2 X_{t-1} + \beta_3 X_{t-2} + \beta_4 X_{t-3} + \epsilon_t.$$

So, we are now assuming that the production of growth hormone *stabilizes* to a certain level that is determined by the temperature.

3.8 Introduction to the analysis of panel data

A *panel* data set or *longitudinal* data consists of several, usually short time series (index $t = 1, ..., T$) for several cross-sectional units (index $i = 1, ..., n$). i can be country, a municipality, a firm, tree, field and so on. Therefore, the observations on y can be collected to a single vector, for example:

$$\begin{array}{l} Y_i = (y_{i,1} \quad ... \quad y_{i,Ti})' \quad i = 1, ..., m \\ Y = (Y_1' \quad ... \quad Y_m')'. \end{array}$$

Example 30. In examples 26 and 29 a study where factors that affect on the short-run growth of trees were examined. Our response variable was the growth rate of a *single* pine seedling. If we had observations on 30 pine seedlings measured in the same time intervals, our original model

$$Y_t = \beta_1 + \beta_2 X_t + \epsilon_t, \quad \epsilon_t \sim NID(0, \sigma^2)$$

would transform to

$$\begin{array}{ll} Y_{it} = \beta_1 + \beta_2 X_{it} + \epsilon_{it}, & \epsilon_{it} \sim NID(0, \sigma^2) \\ & i = 1, ..., 30. \end{array} \tag{3.26}$$

Now we have a panel data, i.e. time series of observations from 30 different pine seedlings.

The use of panel data posses several major advantages over conventional cross-sectional or time-series data. Panel data usually give the researcher a large number of data points, which increases the degrees of freedom and thus improves the efficiency of estimates. Panel data also allows researcher to analyze a number of important questions that cannot be addressed using cross-sectional or time-series data sets.

Example 31 (Hsiao 2003). Suppose that a cross-sectional sample of married women is found to have an average yearly labor-force participation rate of 50 percent. At one extreme this might be interpreted as implying that each woman in a homogenous population has a 50 percent change of being in the labor force in any given year, while at the other extreme it might imply that 50 percent of the women in a heterogenous population always work and 50 percent never work. In the first case, each women would be expected to spend half of her married life in the labor force and half out of the labor force, and job turnover would be expected to be frequent, with an average job duration of two years. In the second case, there is no turnover, and current information about work status is a perfect predictor of future work status. To discriminate between these two models, we need to utilize individual labor-force histories to estimate the probability of participation in different subintervals of the life cycle. This is possible only if we have sequential observations for a number of individuals.

3.8.1 Issues involved in the use of panel data

Heterogeneity bias

The power of panel data derives from their theoretical ability to isolate the effects of specific actions, treatments, policies, etc.. This is based on the assumption that the data under analysis are generated from controlled experiments in which the outcomes are random variables with probability distribution that is a smooth function of the various variables describing the conditions of the experiment. If the available data is generated from **simple controlled experiments, standard statistical methods can be applied**.

However, if the observations are not from controlled experiments, different statistical units, or individuals, may be subject to the influences of different factors. If important factors peculiar to given individual are left out, the assumption that the variable y is generated by a parametric probability distribution function $P(y|\theta)$, where θ is an m-dimensional real vector, identical to all individuals at all times, may be unrealistic. Ignoring the individual, or time-specific effects that are not captured by the included explanatory variables can lead to heterogeneity of parameters in the model specification. Ignoring this heterogeneity could lead to meaningless parameter estimates. For example, lets

consider a simple linear regression model:

$$y_{it} = \alpha_i + \beta_i x_{it} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma^2), \quad i = 1, ..., N, \quad t = 1, ..., T. \tag{3.27}$$

The parameters α_i and β_i may stay constant over time, but may be different for different cross-sectional units. If we estimate a *normal* linear regression using all NT observations, we have a model

$$y_{it} = \alpha + \beta x_{it} + \epsilon_{it}, \qquad \epsilon_{it} \sim N(0, \sigma^2). \tag{3.28}$$

Now, consider two realizations of observations:

1. Heterogenous intercepts ($\alpha_i \neq \alpha_j$), homogenous slopes ($\beta_i = \beta_j$). Figures 1.1, 1.2, and 1.3 demonstrate the variety of circumstances that may arise. Solid lines serve the same purpose for the least squares regression of model (3). Obviously, pooled regression should not be used. The direction of the bias cannot be identified *a priori* either.

2. Heterogenous intercepts and slopes ($\alpha_i \neq \alpha_j$, $\beta_i \neq \beta_j$). In figures 1.4 and 1.5 circled numbers signify the individuals whose regression have been included in the analysis. Point scatters are not shown. In the example presented in figure 1.4, pooling would lead to nonsensical parameter estimate, because it would just tell the average of coefficients that are greatly different in different units. In the figure 1.5 pooling would results to curvilinear relationship, which results to false inference.

There will be similar biases, if the intercepts and slopes vary through time, even if they would be identical for all individuals during some time period.

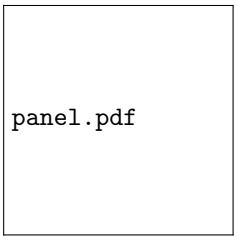


Figure 3.8: Biases in panel data (Hsiao 2003, p. 10)

Selectivity bias

Another frequent bias in cross-section and panel data models is that the sample may not be drawn randomly from the parent population.

Example 32 (Hsiao 2003). New Jersey negative income tax experiment excluded all families in the geographic areas of the experiment who had incomes above 1.5 times the officially defined poverty level. Lets assume that in the population the relationship between earnings (y) and exogenous variables (x), including education, intelligence, etc., is

$$y_i = \beta' x_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2).$$

If the participants of an experiment are restricted to have earnings less than L, the selection criterion for families considered for inclusion in the experiment can be stated as

$$y_i = \beta' x_i + \epsilon_i \leq L, \text{ included},$$
$$y_i = \beta' x_i + \epsilon_i > L, \text{ excluded}.$$

For simplicity we assume that the values of exogenous variables, except for education, are the same for each observation. In figure 1.6 upward-sloping solid line indicates the average relation between education and earnings in our population, and the dots represent the distribution of earnings around this mean for selected values of education. All individuals whose earnings would be above given level L (horizontal line) would be eliminated from the experiment. When estimating the effect of education on earnings, we would observe only the points below the line (circled). This would lead to underestimation of the effect of education on earnings. More importantly, explanatory variable would now be measured with considerable error ($x_i + \epsilon_*$), which would lead to *under-identification* of parameters (see section 3.6.1 on regression).

3.8.2 Simple regression with variable intercepts

In panel data models, the conditional expectation of y give x is examined using the linear regression

$$y_{it} = \alpha_i + \beta X'_{it} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma^2), \quad \epsilon_{it} \perp\!\!\!\perp X_{it}. \tag{3.29}$$

where β is a $K \times 1$ vector of coefficient (excluding intercept). For estimation the properties of the intercept must be specified. There are three possibilities:

1. Constant: If $\alpha_i = \alpha \forall i$ we can use ordinary least squares to estimate equation (4).
2. Random effects: If α_i is a stochastic (random) variable that is independent of x_i, OLS is not efficient. We note the "random individual effect" as u_i.
3. Fixed effects: If α_i is a "group mean" and if $E(\alpha_i x_i) \neq 0$, OLS will produce biased parameter estimates.

In principle, the most important question when dealing with panel models, is the assumptions made on the intercept. As presented above, the intercept, or individual effect, can basically have three different forms. If we have rejected the

overall homogeneity hypothesis, but the specification of a model seems proper, a simple way to take account the heterogeneity across individuals and/or through time is to use variable-intercept models. These models assume that, conditional on the observed explanatory variables, the effects of all excluded (omitted) variables are driven by individual time-invariant and individual time-varying variables (in some cases there is also a need to take the period individual-invariant effects into account, but as they make the estimation of models clearly more complicated we discard these effects).

The individual time-invariant variables are the same for given statistical unit through time, but vary across statistical units (e.g. gender, ability, and socio-economic background variables). The individual time-varying variables vary across cross-sectional units at given point in time and also through time (the "common" residual). In general, two different estimators are used: fixed-effects or random effects estimators.

In the case of **fixed effects** estimator we have a model

$$\begin{aligned} &y_{it} = \alpha_i + \beta' X_{it} + \epsilon_{it}, \qquad \epsilon_{it} \sim N(0, \sigma^2), \quad \epsilon_{it} \perp\!\!\!\perp X_{it} \\ &i = 1, ..., N, \quad t = 1, ..., T. \end{aligned}$$

β' is a $1 \times K$ vector of constants and α_i is a 1×1 scalar constant representing the effects of those variables specific to the ith individual in more or less the same fashion over time. OLS estimator of fixed effects is called the **least-squares dummy variables** estimator. LSDV estimator removes the individual effects effects, usually by assuming $\sum_{i=1}^{N} \alpha_i = 0$. This way the individual effects α_i represent the deviation of the ith individual from the common mean, and they are "eliminated" from the estimation.

Random effects estimator is

$$y_{it} = \alpha + \beta X'_{it} + u_i + \epsilon_{it}, \quad u_i \sim N(0, \sigma_u^2), \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2), \tag{3.30}$$

where following assumptions must also hold:

$$Eu_i = E\epsilon_{it} \equiv 0 \tag{3.31}$$

$$Eu_i u_j = \begin{cases} \sigma_u^2 & \text{if} i = j \\ 0 & \text{if} i \neq j, \end{cases} \tag{3.32}$$

$$E\epsilon_{it}\epsilon_{jt} = \begin{cases} \sigma_\epsilon^2 & \text{if} i = j, t = s \\ 0 & \text{otherwise,} \end{cases} \tag{3.33}$$

and

$$u_i \perp\!\!\!\perp X_{it}, \tag{3.34}$$

$$\epsilon_{it} \perp\!\!\!\perp X_{it}, \tag{3.35}$$

$$\epsilon_{it} \perp\!\!\!\perp u_i. \tag{3.36}$$

In the case of random effects, OLS is no longer BLUE estimator, i.e. *best linear unbiased* estimator. Estimation in the case of random effects must be conducted with *generalized least squares* estimator, or GLS.

3.8.3 Notes

Random or fixed-effects?

The obvious problem in the random effects estimator is that it is biased if the assumption $u_i \perp\!\!\!\perp X_{it}$ is broken (which is not the case in fixed-effects estimator because the individual effects, α_is, are "eliminated" before estimation). This is problematic because the individual effects are not known and their estimation is generally impossible. The assumption about the independence of individual effects and explanatory variables can be tested with the help of *Hausman's specification test*. There you compare the parameter estimates and error variances between random effects and fixed-effects estimations.

The problem of choosing between random and fixed-effects is somewhat problematic concerning the modern statistical inference theory. Cheng Hsiao (2003, p.43), has concluded the issue the following way:

> The situation to which a model applies and the inferences based on it are the deciding factors in determining whether we should treat effects as random or fixed. When inferences are going to be confined to the effects in the model, the effects are more appropriately considered fixed. When inferences will be made about a population of effects from which those in the data are considered to be a random sample, then the effects should be considered random.

However, this is only one view. Some think that there is no real difference between the inferences in these two estimators as long as the sample is randomly drawn (Mundlak 1978). What makes the situation more complicated is the fact that the fixed-effects and random effects estimators do not always produce similar results even when the assumption $u_i \perp\!\!\!\perp X_{it}$ holds. In general, if the inference is not clearly restricted to the sample, one should compare the estimates of fixed- and random coefficient models, for example with Hausman's test, and use random effects estimator if it seems appropriate according to theory and tests.

Stationarity

It is usually assumed that panel data consists on many statistical units, but only on few observation through time. So, generally N is assumed to be large and T small. If time series are stationary and estimator is consistent and efficient in the case where T is fixed, it is also consistent and efficient when both N and T are large.

Chapter 4

Sampling

The idea of *sampling* is to study a part in order to gain information about the whole. Data are often produced by sampling a population of people or things. For example opinion polls report the views of the entire country based on interviews with a sample of about 1000 people. Goverment reports on employment and unemployment are produced from a monthly sample of about 60,000 household. The quality of manufactured items is monitored by inspecting small samples each hour of ech shift.

In all of these examples, the expense of examining every item in the population makes the sampling a practical necessity. Timeliness is another reason for preferring a sample to a **census**, which is an attempt to contact every individual in the entire population.

An **observational study** observes individuals and measures variables of interest but does not attempt to influence the responses. An **experiment** deliberately imposes some treatment on individuals in order to observe their responses.

Statistical inference produces answers to specific questions, along with a statement of how confident we can be that the answer is correct. The conclusions of statistical inference are usually intended to apply beyond the individuals actually studied. Successful statistical inference usually requires **production of data** intended to answer the specific questions posed.

[Yrjön esimerkki taululle tai ehkei vielä tähän]

4.1 Design of experiments

The individuals on which the experiment is done are the **experimental units**. When the units are human beings, the are called **subjects**. A specific experimental condition applied to the units is called a **treatment**.

Because the purpose of an experiment is to reveal the response of one variable to changes in other variables, the distinction between explanatory and response

variable is important. The explanatory variables are often called **factors**. Many experiments study the joint effects of several factors. In such a experiment, each treatment is formed by combining a specific value (often called a **level** of each of the factors.

Example 33. What are the effects of repeated exposure to an advertising message? The answer may depend both on the length of the ad and on how often it is repeated. An experiment investigated this question using undergraduate students as *subjects*. Some subjects saw a 30-second commercial; others, a 90-second version. The same commercial was shown either 1, 3, or 5 time during the program.

This experiment has two *factors*: length of the commercial, with 2 levels, and repetition, with 3 levels. The combinations of one level of each factor form 6 *treatments*. Figure 3.2 shows the layout of the treatments. After viewing, all of the subjects answered questions about their recall of the ad, their attitude toward the camera, and their intention to purchase it. These are the *response* variables.

Laboratory experiments in science and engineering often have a simple design with only a single treatment, which is applied to all of the experimental units. The design of such an experiment can be outlined as

$$\text{Treatment} \longrightarrow \text{Observed response.}$$

Example 34. "Gastric freezing" is a clever treatment for ulcers in the upper intestine. The patient swallows a deflated balloon with tubes attached, then a refrigerated liquid is pumped through the balloon for an hour. The idea is that cooling the stomach will reduce its production of acid and so relieve ulcers. An experiment reported in the *Journal of American Medical Association* showed that gastric freezing did reduce acid production and relieve ulcer pain. The treatment was safe and easy and was sidely used for several years. The design of the experiment was

$$\text{Gastric freezing} \longrightarrow \text{observe pain relief}$$

However, the gastric freezing experiment was poorly designed. The patients' response may have been due to the **placebo effect**. This may be due to trust in the doctor and expectations of a cure or simply to the fact that medical conditions often improve without treatment. The response to a dummy treatment is the placebo effect.

Later, a test where patients were divided in to two groups, one receiving a placebo and other the gastric freezing treatment. 34% of the 82 patients in the treatment group improved, but so did 38% of the 78 patients in the placebo group.

A **lurking variable** is a variable that is not among the explanatory or response variables, but may still influence the interpretation of relationships among those variables. A placebo effect' is a lurking variable. In general, the desing of a study is **biased** if it systematically favors certain outcomes.

4.1.1 Randomization

The **design of an experiment** first describes the response variable of variables, the factors (explanatory variables), and the layout of the treatments, with comparison as the leading principle. Figure 3.2 illustrates this aspect of the design of a study of response to advertising. The second aspect of design is the rule used to assign the experimental units to the treatments. Comparison of the effect of several is valid only when all treatments are applied to similar groups of experimental units. If one corn variety is planted on more fertile groud, or if one cancer drug is goven to more seriously ill patients, comparisons among treatments are meaningles. Systematic differences among the groups of experimental units in a comparative experiment cause bias.

Experimenters often attempt to match groups by elaborate balancing acts. Medical researchers, for example, try to match the patients in a "new drug" experimental group and a "standard drug" control group by age sex, physical condition, smoker or not, and so on. Matching is often helpful but not adequate, because there are too many lurking variables that might affect the outcome.

> The statistician's remedy is to rely on chance to make an assignment that does not depend on any characteristics of the experimental units and that does not rely on the judgment of the experimenter in any way.

The use of chance to divide experimental units into groups is called **randomization**.

Example 35. Does talking on a hands-free cell phone distract drivers? Undergraduate students "drove" in a high-fidelity drving simulator equipped with a hands-free cell phone. The car ahead brakes: how quickly does the subject respond? Twenty students (the control group) simply drove. Another 20 (the experimental group) talked on the cell phone while driving.

This experiment has a single factor (cell phone use) with two levels. The researchers must divide the 40 student subjects into two groups of 20. To do this in a completely unbiased fashion, put the names of the 40 students in a hat, mix them up, and draw 20. These students form the experimental group and the remaining 20 make up the control group. Figure 3.3 presents the desing of the experiment.

The logic behind the design presented in figure 3.3 is

- Randomization produces two groups of subjects that we expect to be similar in all respects before the treatments are applied.

- Comparative design helps ensure that influences other than the cell phone operate equally on both groups.
- Thus, differences in average brake reaction time must be due either to talking on the cell phone or the play of chance in the random assignment of subjects to the two groups.

Principles of experimental design

1. **Control** the effects of lurking variables on the response, most simply by comparing two or more treatments.
2. **Randomize**, i.e. use impersonal chance to assign experimental units to treatments.
3. **Repeat** each treatment on many units to reduce chance variation in the results.

The aim in statistical analysis is to find a difference in the response that is so large that it is unlikely to happen just because of chance variation. We use the laws of probability to learn if the treatmen effects are larger than we would expect to see if only chance were operating. An observed effect so large that it would rarely occur buy chance is called **statistically significant**.

4.1.2 How to randomize

The idea of randomization is to assign subjects to treatments by *drawing names from a hat*. If one cannot use computer programs on randomization, it can be done with the help of *table of random digits*.

A **table of random digits** is a list of the digits 0,1,2,3,4,5,6,7,8,9 that has the properties

1. The digit in any position in the list has the same chance of being any one of 0,1,2,3,4,5,6,7,8,9.
2. The digits in different positions are independent in the sense that the value of one has no influence on the value of any other.

When all experimental units are allocated at random among all treatments, the experimental desing is **completely randomized**. Completely randomized designs can compare any number of treatments. The treatments can be formed by levels of a singly factor or by more than one factor.

Example 36. In the cell phone experiment, we must divide 40 students at random into two groups of 20 students each.

Step 1: Label. Give each student a numerical label, using as few digits as possible. Two digits are needed to label 40 students, so we must use labels

01, 02, 03, ... , 39, 40

It is also correct to use labels 00 and 39 or some other choice of 40 two-digit labels.

Step 2: Table. Start anywhere in the table of random digits and read two-digit groups. Suppose we begin at some line, which is

69051 64817 87174 09517 84534 06489 87201 97245

The first 10 two-digit groups in this line are

69 05 16 48 17 87 17 40 95 17

Each of these two-digit groups is a label. The labels 00 and 41 to 99 are not used in this example, so we ignore them. The first 20 labels between 01 and 40 that we encounter in the table choose students for the experimental group. Of the first 10 labels in our line, we ignore four because they are too high (over 40). The others are 05, 16, 17, 17, 40, and 17. The students labelled 05, 16, and 40 go into the experimental group. Ignore the second and third 17s because student is alreayd in the group. Run your finger across our line (and continue to the following lines) until you have chose 20 students. These students form the experimental group. The remaining 20 are the control group.

4.1.3 Some cautions about experimentation

The logic of randomized comparative experiment depends on our ability to treat all the experimental units identically in every way except for the actual treatments being compared. Good experiment thus requires careful attention to details. For example, the subjects in the second gastric freezing experiment all got the same medical attention during the study. Moreover, the study was **double-blind**, i.e. neither the subject themselves nor the medical personnel who worked with them knew which treatment any subject had received.

In general, many experiments have some weaknesses in detail. The environment of an experiment can influence the outcomes in unexpected ways. Probably the most serious weakness of experiments is **lack of realism**. The subjects or treatments or setting of an experiment may not realistically duplicate the conditions we really want to study.

Example 37. How do layoffs at a workplace affect the workers who remain on the job? Psychologists asked student subjects to proofread text for extra course credit, then "let go" some of the workers (who were actually accomplices of the experimenters). Some subjects were told that those let go had performed poorly (treatment 1). Others were told that no all could be kept and that it was just luck that they were kept and others let go(treatment 2). We can't be sure that the reactions of the students are the same as those of workers who survive layoff in which other workers lose their jobs. Many behavioral science experiments use student subjects in a campus setting. Do the conclusion apply to the real world?

Most experimenters want to generalize their conclusions to some setting wider than of the actual experiment. Statistical analysis of an experiment cannot tell us how far the results will generalize to other settings.

4.2 Sampling design

The entire group of individuals that we want information about is called the **population**. A **sample** is a part of the population that we can actually examine in order to gather information.

Population is defined in terms of our desire for knowledge. The **design** of a sample survey refers to method used to choose the sample from the population. Poor sample designs can produce misleading conclusions.

Example 38. The american Family Association (AFA) is a conservative group that claims to stand for "traditional family values". It regularly posts online poll questions on its Web site - just click on a response to take part. Because the respondents are people who visit this site, the poll results always support AFA's positions. Well, almost always. In 2004, AFA's online poll asked about the heated issue of allowing same-sex marriage. Soon, email lists and social- network sites favored mostly by young liberals pointed to the AFA poll. Almost 850,000 people responded, and 60% of them favored legalization os same-sex marriage. AFA claimed that homosexual rights groups had skewed its poll.

Online polls are now everywhere - some sites will even provide help in conducting your own online polls. As the AFA poll illustrates, you can't trust the results. People who take the trouble to respond to an open invitation are not representative of the entire adult population. That's true of regular visitors to AFA's site,, of the activists who made a special effort to vote in the marriage poll, and of the people who bother to resopnd to write-in, call-in, or online polls in general.

In example 38 the sample was selected in a manner that quaranteed that it would not be representative of the entire population. These sampling schemes display *bias*, or systematic error, in favoring some parts of the population over others. Usually, **voluntary response samples** are biased because people with strong opinions, especially negative opinions, are most likely to respond.

4.2.1 Stratified samples

A **simple random sample (SRS)** of size n consists of n individuals from the population chosen in such a way that every set of n individuals has an equal chance to be in the sample actually selected. Each treatment group in a completely randomized experimental design is an SRS drawn from the available experimental units. An SRS is selected by labeling all the individuals in the population and using software or a table of random digits to select a sample of the desired size.

SRS is an example of a **probability sample**. A probability sample is a sample chose by chance. We must know what samples are possible and what chance, or probability, each possible sample has. As SRS gives a equal chance to each member in the population to be selected, it is sometimes reasonable to sample important groups within the population separately and then combine sample. To select this kind of **stratified sample**, first divide the population into groups of similar individuals, called **strata**. The choose a separate SRS in each stratum and cobine these SRS's to form the full sample.

Example 39. A dentist is suspected of defrauding insurance companies by describing some dental procedures incorrectly on claim forms and overcharging for them. An investigation begins by examining a sample of his bills for the past three years. Because there are five suspicious procedures, the investigators take a stratified sample. That is, they randomly select bills for each of the five procedures separately.

4.2.2 Multistage samples

Another common means of restricting random selection is to choose the sample in stages. This is common practice for national samples of households or people. For example, data on employment and unemployment are gathered by the government's Current Population Survey, which conducts interviews in about 60,000 households each month. Clearly, sending interviewers to the widely scattered households in an SRS would be too costly, and the government wants data broken down by states and large cities. Therefore **multistage sampling design** is used. The sampling design is roughly as follows:

Stage 1 Divide the United States into 2007 geographical areas called Primary Sampling Units, or PSUs. PSUs do not cross stage lines. Select a sample of 754 PSUs. This sample includes the 428 PSUs with the largest population and a stratified sample of 326 of the others

Stage 2 Dividive each PSU selected into smaller areas called "blocks". Stratify the blocks using ethnic and other information and take a stratified sample of the blocks in each PSU.

Stage 3 Sort the housing units in each block into clusters of four nearby units. Interview the households in a probability sample of these clusters.

Cautions about sample surveys

Random selection eliminates bias in the choise of a sample from a list of the population. Sample surveys of large human populations require much more than a good smapling desing. **Undercoverage** occurs when some groups in the population are left out of the process of choosing the sample. **Nonresponse** occurs when an individual chose for the sample cannot be contacted or does not cooperate. The behavior of the respondent or of the interviewer can also cause **response bias** in sample results. The **wording of questions** is the most important influence on the answers given to a sample survey.

Example 40. In response to the question "Are you heterosexual, homosexual, or bisexual?" in a social science research survey, one women answered, "It's

just me and my husband, so bisexual". This is a classic example of trusting too much on people's knowledge on different terms.

How do American's feel about goverment help for the poor? Only 13% think we are spending to much on "assistance to the poor", but 44% think we are too much on "welfare". How do the Scots feel about the movement to become independent from England? Wekk, 51% would vote for "independece for Scotland", but only 34% support "an independent" Scotland separate from the United Kingdom". It seems that "assistance to the poor" and in"independence" are nice, hopeful words. "Welfare" and "separate" are negative words.

4.3 On statistical inference

The basic fact in sampling is that the value of a statistic varies in repeated random sampling. This is called **sampling variability**. Random samples eliminate *bias* from the act of choosing a sample, but they can still be wrong because the *variability* that results when we choose at random. If the variation when we take repeat samples from the same population is too great, we can't trust the results of any one sample.

The problem is eased by the fact that if we take lots of random samples of the same size from the same population, the variation from sample to sample will follow a predictable pattern. Basicly, all of statistical inference is based on one idea. That is, to see **how trustworthy a procedure is if we repeat it many times**.

As mentioned above, sampling variability is not *fatal*. If we took many samples, we would end up to some distribution. The procedure is as follows

- Take a large number of samples from the same population.
- Calculate the sample proportion $\hat{p}$ for each sample.
- Make a histrogram of the values $\hat{p}$.
- Examine the distribution displayer in the histogram for shape, center, spread and outliers and other deviations.

In practice, however, it is often too expensive to take many samples from a large population.

Example 41. Suppose that in fact 60% of the population is U.S. find clothes shopping time-consuming and frustrating (see example 9). Then the true value of the parameter we want to estimate is $p = 0.6$. We can now imitate the population by a table of random digits, with each entry standing for a person.

Six of the ten digits (say 0 to 5) stand for people who find shopping frustrating. The remaining four digits (6 to 9) stand for those who do not. Because all digits in a random number table are equally likely, this assignment produces a population proportion of frustrated shoppers equal to $p = 0.6$. We then imitate an SRS of 100 people from the population by taking 100 consecutive digits from table of random numbers. The statistic $\hat{p}$ is the proportion of 0s to 5s in the

sample. Here are the first 100 entries from table of random digits:

19223	95034	05756	28713	96409	12531	42544	82853
73676	47150	99400	01927	27754	42648	82425	36290
45467	71709	77558	000095				

There are 64 digits between 0 and 5, so $\hat{p} = 64/100 = 0.64$. A second SRS based on the second 100 entries in table of random digits would give a result of $\hat{p} = 0.55$ (check from the table presented in appendix). That is sampling variability.

4.3.1 Sampling distributions

The **sampling distribution** of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Example 42. Figure 3.9 illustrates the process of choosing many samples and finding the sample proportion $\hat{p}$ for each one. The histogram at the right of the figure shows the distribution of the values of p from 1000 separate SRSs of 100 drawn from a population with $p = 0.6$.

Figure 3.10 is parallel to figure 3.9. It shows the process of choosing 1000 SRS, each size of 2500 from a population in which the true proportion is $p = 0.6$. The 1000 values of $\hat{p}$ from these samples form the histrogram at ther right of the figure. Both figures are drawn on the same scale. These histograms display the *sampling distribution* of the statistic $\hat{p}$ for two sample sizes.

We can generalize the distributions in figures 3.9 and 3.10 as follows:

- **Shape**: The histrograms look normal with the 1000 values for samples of size 2500 are very close to normal. Figure 3.11 presents a normal quintile plot of the values of $\hat{p}$ for our sample of size 100. The normal curves drawn through the histrograms describe the overall shape quite well.
- **Center**: In both cases the values of the sample proportion $\hat{p}$ vary from sample to sample, but the values are centered at 0.6, which is the true population parameter. Therefore, we can conclude that $\hat{p}$ has no *bias* as an estimator of p.
- **Spread**: The values of $\hat{p}$ from sample of size 2500 are much less spread out than the values from sample of size 100.

4.3.2 Bias and variability

Bias concerns the center of the sampling distribution. A statistic used to estimate a parameter is **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter being estimated.

The **variability of a statistic** is described by the spread of its sampling distribution. This spread is determined by the sampling desing and the sample size n. Statistic from larger probability samples have smaller spreads.

To reduce bias, we use random sampling. When we start with a list of the entire population, simple random sampling produces unbiased estimates, i.e. the values of a statistic computed from an SRS netiher consistently overestimate nor consistently underestimate the value of the population parameter.

To reduce the variability of a statistic from an SRS, use larger sample. You can make the variability as small as you want by taking a large enough sample. Results from a sample survey usually come with a **margin of error** that set bounds on the size of the sample statistics. Thus, it is smaller for larger samples.

The size of the population is usually not a critical factor in analysis, if the sample is really a *random sample*. Moore and McCabe give this following "thumb rule":

> The variability of a statistic from a random sample does not depend on the size of the population , as long as the population is at least 100 times larger than the sample.

It should be remembered that many times *drawing a random sample* reguires the use of sophisticated probabilistic methods. Therefore all that has been presented in these lectures on sampling should be treated only as "raw guidelines".

Additional reading:
Cochran, William (1977): *Sampling techniques.* John Wiley & Sons.
Thomson, Steven (2004): *Sampling.* John Wiley & Sons.

REFERENCES:

Glover, Thomas and Mitchell, Kevin (2002). *An Introduction to Biostatistics.* Waveland Press Inc.

Hamilton, James D. (1994). *Time Series Analysis.* Princeton University Press.

Hsiao, Cheng (2003): *Analysis of Panel Data.* Cambridge University Press.

Moore, David S. and McCabe, George P. (2006). *Introduction to the Practice of Statistics* W.H. Freeman & Company.

Mundlak, Y. (1978): On the pooling of time-series and cross-section data. *Econometrica*, 46, 69-85.

Tuominen, Pekka (2007). *Todennäköisyyslaskenta I.* Limes Ry.