Institute for the Advancement of University Learning
&
Department of Statistics

Descriptive Statistics for Research
(Hilary Term, 2002)

Lecture 8: Univariate and Multivariate Tests

We finish this series of lectures by presenting a collection of commonly used statistical procedures.

(I.) Tests for One Variable

In this section, we are interested in the population properties of only one variable (at a time) rather than in the relationship between two or more variables.

(I.a.) Continuous Variable

(I.a.i.) Tests for the Mean

If we are interested in investigating the properties of one continuous variable, then, often, we will be most interested in the variable's (population) mean. In Lecture 7, we saw an example of a test of two competing hypotheses about a variable's mean, assuming that the population variance of the variable is known. Recall that this test also assumes that the sample mean has an approximately normal distribution; in other words, the test assumes either that the underlying distribution of the variable is normal or that the sample size is large. However, the assumption of known variance is not often fulfilled in practical applications. Therefore, we consider a similar test of two competing hypothesis about the mean for the case where σ^2 is not known.

(1.) Normality assumed, variance unknown

Suppose that $\mathbf{X} = \{X_1, X_2, \cdots X_n\}$ is a sample of n observations of a random variable that is distributed as $N(\mu, \sigma^2), \sigma^2$ unknown. To test

$$H_0 : \mu = \mu_0 \text{ vs } H_1,$$

where H_1 is a one - or two - sided alternative, the optimum test statistic (in the Neyman-Pearson sense) is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}},$$

where S is the square root of the unbiased version of the sample variance. Under H_0, the statistic T has a Student's t distribution with n-1 degrees of freedom. When testing H_0 against $H_1: \mu \neq \mu_0$ (a two-sided alternative), large values of $|T|$ (i.e. large negative *and* positive values of T) correspond to the rejection region. If the alternative hypothesis is $H_1: \mu > \mu_0$ (or the other possibility for a one-sided alternative, $H_1: \mu < \mu_0$), then large positive (negative) values of T define the critical region.

Example: Consider the measurements of heights (in mm) of the ramus bone for a sample of 20 boys aged 8½ years. The data are:

45.3, 46.8, 47.0, 47.3, 47.5, 47.6, 47.7, 48.5, 48.8, 48.9, 49.2, 50.0, 50.4, 50.8, 51.4, 51.7, 52.8, 53.0, 53.2, 54.6

We assume that these heights come from a normally distributed population, as linear biological measures often do (we could investigate this assumption using graphs or a formal goodness of fit test). However, we do not know the population variance.

Suppose that we are interested in testing the null hypothesis that the population mean is equal to 50mm against the alternative that it is not. That is, we want to test

$$H_0: \mu = 50 \text{ vs } H_1: \mu \neq 50$$

at the $\alpha = 0.05$ level of significance.

A: For these data, we have that $\bar{x} = 49.63$, s=2.54 and $s.e.(\overline{X})$ = 0.568. The observed value of T is t^*= -0.664. We can compare this value with a Student's t distribution with n-1 = 19 degrees of freedom (the null distribution).

Note that, since the alternative is two-sided, the critical region for this test will consist of two parts, each with $\alpha/2$ of probability. This region consists of values of T below -2.09 and above 2.09 and appears as the shaded areas in Figure 1. These values were obtained using statistical tables for the Student's t distribution; alternatively, these values could have been obtained using a statistical software package.

Since t^* does not fall into the critical region, we conclude that there is insufficient evidence to reject H_0 at a significance level of 0.05. The p-value

for this test, which is the area below t^* plus the area above $-t^*$ (= 0.664), is 0.516. Note that, if our alternative had been one-sided (e.g. $\mu < 50$ or $\mu > 50$), then

a. our *p*-value would have been half that *p*-value (i.e. 0.258).

b. the critical region would have included only points below -1.73 or above 1.73, as these are the values which accumulate 0.05 and 0.95 of probability under a t_{19} distribution, respectively.

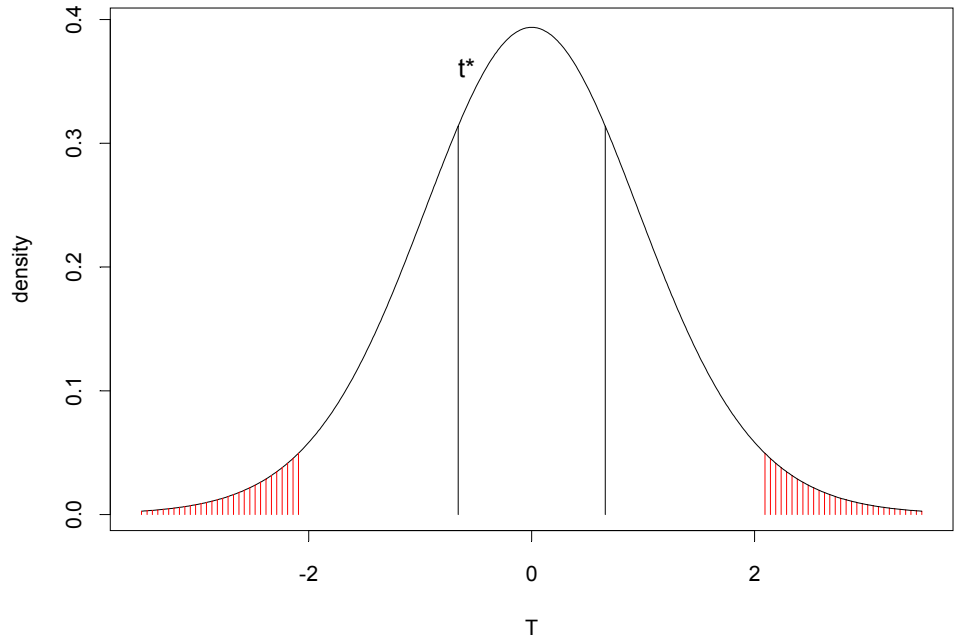

***Figure 1**: The observed value of T and the critical region for α=0.05*

(2.) Normality not assumed (nonparametric tests)

In situations where the assumption of normality is not satisfied, we may try transforming the variable in an effort to make its distribution approximately normal so that we could still use the above tests. If we are unable to induce normality via a transformation, we could instead use a nonparametric test. The nonparametric analogue of the test presented above is a variation of the ***Wilcoxon rank sign*** test and tests the null hypothesis that the *median* of a random variable equals a particular value. This test is implemented in most statistical software packages, and appears, for example, in the nonparametric submenu in SPSS.

(I.b.) Categorical Variable (the Chi – squared Test)

Suppose now that the one variable whose properties we hope to investigate is categorical, rather than continuous. Further, suppose that this variable has *K* (mutually exclusive)

categories (i.e. each unit in the underlying population falls into exactly one of *K* categories) and that we have a random sample of *n* units. In this case, we are not interested in the variable's mean (obviously, since the mean is not defined for ordinal and nominal variables). Instead, we are interested in testing the null hypothesis that the proportions of the population in the *K* categories are $\pi_1,...,\pi_K$, respectively, where the π_i sum to one. Often, we will be interested in the particular hypothesis that the π_is are equal (i.e. that each π_i equals *1/K*). However, it should be noted that the test presented below is quite general and does not simply apply to testing the equality of category proportions for a single categorical variable.

If we let π_i represent the hypothesised proportion of units falling into the i^{th} category, we would expect $E_i = n\pi_i$ occurrences of the objects in the *i*th category if the null hypothesis were true. Letting O_i denote the observed number of units (out of *n*) in the i^{th} category, the following test statistic measures how far the hypothesised (expected) data is from the observed data. Intuitively, if this value is large, there is evidence against the null hypothesis, whereas if this value is small, the information contained in the sample provides little evidence against the null. The test statistic is

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

which, under the null hypothesis, has a chi-squared distribution with *K-1* degrees of freedom. Therefore, the rejection region for this test consists of all values that are greater than the $(1-\alpha)$ quantile of the χ^2_{K-1} distribution, for an appropriate choice of α. The quantiles of the chi-squared distribution (with the appropriate number of degrees of freedom) can be easily calculated in the majority of statistical software packages or found in chi-squared tables in most statistics texts. Before proceeding with an example, we should note that this test should only be used for large samples; for guidelines on what constitutes a sufficiently large sample, see the following section.

> **Example:** We wish to test whether there is a seasonal effect of murders in a particular U.S. state. We observe 1361 crimes in total, 334 of which were in spring, 372 in summer, 327 in autumn, and 328 in winter. If we let π_i be the proportions of crimes in each of the four seasons, then our hypothesis is
>
> $$H_0 : \pi_1 = \pi_2 = \pi_3 = \pi_4 = 0.25$$
> $$H_A : \text{at least one proportion is different}$$
>
> We produce the following table to help us to see how the test works.

	Spring	Summer	Autumn	Winter

Observed	334	372	327	328
Expected	340.25	340.25	340.25	340.25
(Obs-Exp)²/Exp	0.11	2.96	0.52	0.44

The observed value of the test statistic is 4.03, with 3 degrees of freedom. The rejection region, at the 0.05 level of significance, is values of the chi-squared statistic that are greater than $\chi^2_{0.05,3} = 7.82$. Hence, we would not reject the null hypothesis; that is, we conclude that there is insufficient evidence of a seasonal effect on the murder rate.

(II.) Two Variables: Testing Relationships Between Variables

(II.a.) The General Statistical Model

Often in applied statistics, we are interested in investigating the relationship between two (or more) variables; in other words, we might want to see either (i) how the variables influence each other or (ii) how one (or some) of the variables is (are) influenced by other variables. In the former case, both variables are treated equally. This is not true in the latter type of investigation, in which case the variables are treated asymmetrically; in this case, we will first need to specify a ***statistical model*** for these variables.

A statistical model describes the relationship between one or more ***response*** variables (usually denoted by ***Y***, and sometimes called the ***dependent*** variable) and one or more ***explanatory*** variables (usually denoted by ***X***, and known variously as the ***independent*** variables, ***predictor*** variables, or ***covariates***). It is generally assumed that the *X*s are measured without any experimental or measurement error; therefore, any uncertainty in our model will concern only the response variable. We write a statistical model, in its most general form, as

$$Y = f(X;\beta) + \varepsilon,$$

where ε denotes a random error, *f* is a function describing how *Y* changes in relation to *X*, and β represents the parameters of the model. The function *f* is typically assumed known. The error term, ε, includes both random measurement error and possible aspects of *Y* that are not adequately explained by the model (i.e. by the function of the *X*s). We assume that the expected value of the errors is zero.

It is usually too ambitious to model all aspects of the response variable, and thus it is common practice to analyse how the *mean* of *Y* changes in relation to changes in *X*. The expected value of *Y*, given a particular value of *X*, is written as

$$E[Y|X = x] = f(x;\beta).$$

Note that the error term does not appear in this equation because we assume that the expected value of ε is 0.

Notice that we have not restricted this scheme to a particular type of variable. Both Y and X can be nominal, ordinal, or numeric. The type of variables used in the model, as well as the assumptions placed on the error component, determines which method of analysis to use. In the remainder of the lecture, we outline some of these methods of analysis. Before doing so, however, we digress briefly to discuss the concept of causality.

(II.b.) Causality

The idea of ***causality*** is central to scientific research. For instance, we might be interested in the relationship between interest rates and inflation or between smoking and lung cancer. The researcher usually states a *causal* hypothesis, which is asymmetric in the sense that one set of variables influence the others, but not vice versa. Often, a causal hypothesis of this type is the motivation driving statistical investigations of how a variable is influenced by (one or more) other variables. Clearly, in science (and in statistics), we can never *prove* that one variable is the cause of another since we cannot directly observe their relationship; all we can do is 'disprove' hypotheses by showing that empirical evidence contradicts them.

Some scientists (and statisticians) believe that a relationship between variables must satisfy the following three criteria to be considered causal:

1. the variables must be associated;
2. there must be an appropriate time ordering, with cause preceding effect;
3. alternative explanations have to be eliminated.

Statistical methods are used in order to measure the extent of the association between two or more sets of variables (condition 1). Conditions 2 and 3 pertain more to the design of the study than to the examination of empirical evidence, and it is usually at this stage of an experiment/study that the attempt to establish causality comes unstuck. For instance, failure to control for ***confounding*** factors will nullify any attempt at demonstrating causality because condition 3 will not be satisfied. This particular problem with establishing causality can arise when either the design of the study or the choice of variables for inclusion in the subsequent analysis introduces a bias of some sort. Examples of the former and latter include, respectively, ***self-selection bias*** and ***omitted variable bias***, of which Simpson's paradox, discussed below, is a special case.

(II.c.) Some Common Statistical Procedures

In the remainder of this lecture, we will present some commonly used statistical procedures for examining and testing relationships between (usually two) variables. Each subsection deals with a particular combination of variable types: for example, continuous response with categorical predictor, and so on.

(II.c.i.) Categorical Variables Only

(II.c.i.1.) Testing Hypothesised Proportions

Binary Response with Categorical Predictor

Suppose we have K populations from which we have taken samples of size n_1, n_2, …, and n_K, respectively. In addition, suppose that each of the K populations can be further divided into a 'success' category and a 'failure' category. Alternatively, we could describe this situation by saying that we have a categorical predictor variable with K levels and a binary (1=success, 0=failure) response variable. We are normally interested in testing whether the true population proportions of success in the K groups are equal to some hypothesised values $\pi_1,...,\pi_K$, where each π_i is between 0 and 1. Frequently, the researcher may be interested in testing whether the probability of success is the same in each of the populations, in which case the K hypothesised proportions would all be equal. However, this is not the only null hypothesis that can be investigated using the following test.

If the null hypothesis were true, then we would expect $E_i = n_i\pi_i$ successes in the i^{th} sample for i = 1, …, K. Let O_i denote the observed number of successes in the i^{th} sample. To use these observed and expected frequencies to test the general null hypothesis specified above against an appropriate alternative, we employ the chi-squared test statistic defined in the previous section since it measures, in some sense, the distance between the observed and expected category frequencies. However, in this situation, the test statistic has a chi-squared distribution with K, rather than K-1, degrees of freedom under the null hypothesis: the rejection criteria for this test is values of the calculated χ^2 statistic greater than the $(1-\alpha)$ quantile of the χ^2_K distribution, for an appropriate choice of α. The intuitive reasoning behind the difference in the degrees of freedom is that, in the previous example, we constrained the total number of observations to be n, whereas in this case we allow each cell to have any number of observations, with no constraint on the total.

Before proceeding, we should note that the chi-squared statistic can also be used to test the goodness of fit of distributions (as was shown at the end of Lecture 7). To do so, we divide the observations into groups and count how many of the observed values fall into each. Then, based on our hypothesised distribution, we calculate how many of the observations we would have expected to fall into each group, calculate the chi-squared statistic, and use it to test whether our hypothesised distribution appears to be a reasonable one.

(II.c.i.2.) Testing for Independence

We now consider how to statistically analyse whether or not two or more categorical variables are independent of each other.

(II.c.i.2.A.) Two Variables

One way of representing the relationship between two categorical variables is via a ***two-way table***. Let one of the variables have r levels and the other have c levels. Then, a two–way table is a table with r rows and c columns, with each cell containing the observed

number of objects falling into that crossed category. Let N_{ij} denote the number of observations in the i^{th} row and j^{th} column, i=1,…,r, j=1,…,c, C_j denote the column totals, R_i denote the row totals, and n the total sample size. Then, the general form of a two-way table is

	Column						
Row	**1**	**2**	**…**	**j**	**…**	**c**	**Total**
1	N_{11}	N_{12}	…	N_{1j}	…	N_{1c}	R_1
2	N_{21}	N_{22}	…	N_{2j}	…	N_{2c}	R_2
:	:	:		:		:	:
I	N_{i1}	N_{i2}	…	N_{ij}	…	N_{ic}	R_i
:	:	:		:		:	:
R	N_{r1}	N_{r2}	…	N_{rj}	…	N_{rc}	R_r
Total	C_1	C_2	…	C_j	…	C_c	n

CHI-SQUARED TEST

We are interested in testing whether the row and column variables are independent: the null hypothesis is that there is no relationship between the row and column classifications. There are two possible scenarios that might lead us to test this hypothesis. In the first, we have a sample of n units drawn from a population. We believe that each unit in the sample can be classified according to two categorical variables. In this case, the variables have a symmetric relationship; this is not the case in the second scenario, in which there are c populations, and samples of sizes C_1,…,C_c are drawn from each. Each unit is then classified according to a categorical variable with r possible levels.

The difference between these two situations lies in how the data is collected. In the first case, the researcher sets the sample size, n, and then classifies each unit into one of the rc cells. In the second case, the column totals are the sample sizes selected at the design stage. The first situation is known as ***multinomial sampling***, and the second as ***product multinomial sampling***. Although these are two completely separate scenarios, the chi-squared testing procedure described below is the same for both.

The statistic that tests the null hypothesis in an r x c table compares the observed counts with the expected counts, the latter being calculated under the assumption that the null hypothesis is true. The expected count in the ij^{th} cell of the table is given by

$$E_{ij} = \frac{R_i C_j}{n},$$

and we again use the chi – squared test statistic

$$\chi^2 = \sum_{ij} \frac{(N_{ij} - E_{ij})^2}{E_{ij`}}.$$

This statistic has a chi-squared distribution on (*r*-1)(*c*-1) degrees of freedom. Therefore, we reject the null hypothesis that the rows and columns are independent if the observed value of the chi-squared statistic is greater than $\chi^2_{\alpha,(r-1)(c-1)}$, the $(1-\alpha)$ quantile of the chi-squared distribution with (*r*-1)(*c*-1) degrees of freedom.

Note that these results are based on approximations, and thus there are certain important assumptions that need to be satisfied in order to use this test. To apply this theory to real data examples, we must assume that we have a *large enough* sample. This tends to be satisfied if:

- The smallest expected count is 1 or more.
- At least 80% of cells have an expected count of 5 or more.

If this is not the case, then we should consider using an alternative testing procedure, such as the one presented in the following section.

Example: Each of 250 job applicants at a large firm was classified in two ways: (1) by whether or not they got a job offer and (2) by their ethnic group. The data are as follows:

	ETHNICITY		
GOT JOB OFFER?	**Anglo**	**African Am.**	**Hispanic**
Yes	24	13	18
No	124	39	32

We are interested in testing whether receiving a job offer is independent of the ethnicity of the applicant. In doing so, we find the chi-squared test statistic to be 8.869, which we compare to a χ^2 distribution with $(3-1)\times(2-1)=2$ degrees of freedom. The rejection region, at the 0.05 level of significance, corresponds to values greater than 5.99: hence we reject the null hypothesis that receiving a job offer is independent of ethnicity. We conclude that there is some degree of racial discrimination when offering a job in this particular company.

FISHER'S EXACT TEST

As noted above, the chi-squared test for independence is only useful when we have a 'large enough' sample. Fisher's Exact Test is a probability test for independence in *r*x*c* (two-way) tables, applicable when the conditions described above are not satisfied. The test determines the exact probability of obtaining the observed result or one more extreme if the two variables are independent and the marginal (row and column) totals are fixed.

Most statistical software packages implement Fisher's Exact Test for 2x2 tables; however, Splus and R allow up to 10 levels per variable. As an example, we might wish to know whether a particular drug and a biochemical reaction in the body are independent of each other. We observe the following data

	AS Yes	**AS No**
R24H Yes	0	2
R24H No	23	15

which has expected values

	AS Yes	**AS No**
R24H Yes	1.2	0.8
R24H No	21.8	16.2

Fisher's Exact Test should be used to test the hypothesis of independence in this case, as we have expected values less than 1. For this particular example, the test returns a *p*-value of 0.17436, implying that the null hypothesis of independence cannot be rejected at the 5% level.

(II.c.i.2.B.) More Than Two Variables

If we have a data set with more than two categorical variables, there are a number of possible techniques available to us for investigating whether those variables are independent of each other. The simplest approach is to collapse over some of the variables so that a two-way contingency table is obtained, and carry out the analysis as described above. This is not always a good idea, however, because of a phenomenon known as ***Simpson's paradox***.

Simpson's paradox (which, technically, is not a paradox at all) refers to the situation in which an overall conclusion regarding the relationship between two or more variables is 'contradicted' when another variable of importance is added to the analysis. Consider the following example.

> **Example:** In 1972 a survey of the electoral roll was carried out in Wichkham, near Newcastle upon Tyne. The study was primarily concerned with thyroid and heart disease and their relationship with smoking. In 1994, a follow-up study was conducted. The following results are for two age groups of females. Each table shows the twenty-year survival status of smokers and non-smokers.

Age 55-64

	Dead	**Alive**
Smoker	51 (44%)	64 (56%)
Non-Smoker	40 (33%)	81 (67%)

Age 65-74

	Dead	**Alive**
Smoker	29 (80%)	7 (20%)
Non-Smoker	101 (78%)	28 (22%)

For each table, it appears that a larger proportion of smokers have died than non-smokers. Now observe what happens when the tables are combined (i.e. when we collapse over age).

Combined

	Dead	**Alive**
Smoker	80 (53%)	71 (47%)
Non-Smoker	141 (56%)	109 (44%)

It now appears that smokers have a lower death rate. The reason for this 'paradox' is that most smokers have died before reaching the older age classes, and so the higher (absolute) number of deaths of non-smokers in the older age classes has obscured the result.

Rather than collapsing over (potentially) important variables, it is generally much better to use a ***log-linear analysis*** to investigate independence (or, identically, dependence) between more than two categorical variables. Log-linear models are similar to linear models except that the response is a count rather than a continuous variable. These types of model are a specific type of a broader model class known as ***Generalized Linear Models***, which allow us to analyse the relationships between non-normal and non-continuous variables. Whereas log-linear models give equal status to all the categorical variables in the model, some other generalized linear models for analysing the relationships between two or more categorical variables allow asymmetric (i.e. response-predictor) relationships between the variables.

(II.c.ii.) Continuous Response with a Categorical Predictor

We now consider cases in which the response variable is continuous and the predictor variable is categorical. In general, we will be interested in seeing how the predictor variable affects the mean of the continuous variable; in other words, we will compare the mean of the continuous variable across the populations defined by the categorical predictor variable. Throughout the majority of this section, we will assume that the samples (from each of the populations defined by the predictor variable) come from *normally* distributed populations with unknown mean and variance.

(II.c.ii.1.) Binary Predictor

(II.c.ii.1.A) Comparing the Means of Two Populations

PAIRED TWO-SAMPLE *T*-TEST

In certain situations we might be interested in comparing the effect of a particular treatment on *pairs* of observations. These pairs can either come from the same individual measured before and after the treatment (***self-pairing***) or from pairs of similar individuals (e.g. pairs of patients of the same sex, age, etc.) given different treatments. Pairing is used in an attempt to make comparisons between treatments more accurate. It does this by making members of any pair as similar as possible in all areas except treatment category. Thus, any difference we do see can be attributed (in theory) to treatment effects.

Denote the paired sample as (X_i, Y_i), $i = 1, \ldots, n$. We assume that each pair element is individually normally distributed with means μ_X and μ_Y, respectively, and unknown variances. The null hypothesis is $\mu_X - \mu_Y = \mu_0$, against a one-sided or two-sided alternative. Note that the value μ_0 is the hypothesised mean (population) difference between treatments. Often, the researcher will be interested in testing the hypothesis that $\mu_0 = 0$ (the hypothesis that there is no difference between treatments).

Note that paired data are statistically dependent, and thus violate the assumption independence. In order to remove this dependency, we take pairwise differences, and use these differences as the data sample. Thus, let $D_i = X_i - Y_i$ (i.e. the difference for the *i*-th pair). The statistic for this test is then

$$T_P = \frac{\overline{D} - \mu_0}{s.e.(\overline{D})}$$

where $\overline{D}$ is the mean of the differences and its estimated standard error is given by

$$s.e.(\overline{D}) = \sqrt{\frac{\sum_{i=1}^{n}(D_i - \overline{D})^2}{n\,(n-1)}}\,.$$

The null distribution of T_P is a Student's *t* distribution with *n-1* degrees of freedom. Calculation of the rejection region for this test is identical to that described previously (in Lecture 7), and details are omitted.

Following lecture 5, it is relatively easy to construct confidence intervals (CIs) for the difference of means. A $100(1-\alpha)\%$ CI for the mean difference in treatment is given by

$$\overline{D} \pm t_{1-\alpha/2}^{(n-1)}\, s.e.(\overline{D})\,.$$

If the sample size is large, the approximate validity of the test (and the confidence interval) follows from the Central Limit Theorem (even if the underlying distribution is not normal); if n is small and the true distribution of the differences is not normal, then the stated probability levels will be incorrect.

Example: Consider data collected on eight patients, each of whom had one eye affected with glaucoma and one not. The measurements are the corneal thickness (in microns) of both eyes. The null hypothesis is that having glaucoma makes no difference in the corneal thickness of eyes. Suppose we believe that glaucoma decreases the thickness of the cornea; thus, we are in a position to specify a lower – tailed test. The hypotheses can be written as

$$H_0: \mu_D = 0 \text{ vs } H_1: \mu_D < 0 .$$

The data are:

Affected:	488	478	480	426	440	410	458	460
Not affected:	484	478	492	444	436	398	464	476
Differences:	4	0	-12	-18	4	12	-6	-16

We calculate the mean difference to be $\bar{d} = -4$, and hence $t^* =$ -1.053. There are 7 degrees of freedom, giving a two-sided p-value of 0.3274. Therefore, the (one-sided) p-value of interest is 0.1637. This indicates that we cannot reject H_0 at the 0.05 level of significance.

A 95% confidence interval for mean corneal thickness difference is [-12.98, 4.98]. Recall that this interval contains those values of the statistic for which we would not reject the null hypothesis. Therefore, as this interval contains our hypothesized value (0), we cannot reject the hypothesis of no difference in mean corneal thickness.

Even though the mean value of the differences is –4, we are unable to reject the null hypothesis. If the sample size were larger (i.e. we had more information with which to work), we might have been able to find a significant difference.

INDEPENDENT TWO - SAMPLE *T*-TEST

Suppose we have two independent samples of sizes n_1 and n_2. Further, suppose each sample is from normally distributed random variables X_1 and X_2. Consider a null hypothesis regarding the difference in the variables means such as $H_0 : \mu_{X_1} - \mu_{X_2} = \mu_0$. As before, μ_0 is often 0 in practice. For this test, we use the statistic

$$T_I = \frac{(\overline{X}_1 - \overline{X}_2) - \mu_0}{s.e.(\overline{X}_1 - \overline{X}_2)}.$$

We must distinguish between situations where the samples share a common variance and situations where they do not. The formulae for the standard error of the difference of the means, as well as the degrees of freedom for the null distribution of $T_{I,}$ are different depending which of these situations holds.

(With Equal Variance)
Let us first assume that both samples have equal (unknown) variances. In this case, the estimated value of the quantity in the denominator of T is

$$s.e.(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{(n_1 + n_2)[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{n_1\, n_2\, (n_1 + n_2 - 2)}}$$

where s_1^2 and s_2^2 are the unbiased estimates for the variances in each sample. The null distribution of T_I is a Student's t with $(n_1 + n_2 - 2)$ degrees of freedom.

Example: Consider the following data obtained from two lots of chicks, between 28 and 84 days old, treated with two different diets (one a high protein diet and the other a low protein diet). The response variable is the gain in weight (in grams) over that period of time.

It is reasonable to assume that weight gain is normally distributed. The sample sizes are 12 and 7 for the high and low protein diets, respectively. We would like to test the hypothesis of no difference in the mean weight gain for the two diets against the general (two-sided) alternative.

The data are:

High protein diet: 134, 146, 104, 119, 124, 161, 107, 83, 113, 129, 97, 123
Low protein diet: 70, 118, 101, 85, 107, 132, 94

The observed means are 120g and 101g; the observed variances are 457 and 425, which does not indicate gross deviations from the assumption of homogeneity (equal variance). We calculate the degrees of freedom to be 17. The critical region, for $\alpha = 0.05$, consists of the values of T_I which are larger than 2.11 or smaller than -2.11. The observed value of the statistic is

$$t_I^* = \frac{120 - 101}{10.04} = 1.89,$$

which does not fall in the critical region. The *p*-value is 0.075, indicating that there is a probability of about 1 in 12 of observing, by chance, a result as large or larger than the observed value of T_1 under H_0. Therefore, we do not reject the null hypothesis. Note that if our alternative hypothesis had been one-sided (e.g. chicks on a high protein diet have greater mean weight gain than those on a low protein diet), then the critical region for $\alpha = 0.05$ would have been those values of T_1 larger than 1.74, and our *p*-value would have been 0.0375. Thus, in the case of the one-sided alternative, we would reject the null hypothesis in favour of the alternative at the $\alpha = 0.05$ level of significance.

(With Unequal Variances)

If the variances of the two samples are different, we need to use a different estimated standard error of the mean and to calculate the degrees of freedom using an approximation (due to Satterthwaite). In this situation, the estimated standard error of the mean is

$$s.e.(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}},$$

and the degrees of freedom approximation equation is given by

$$df = \frac{\left(n_1 + n_2 - 2\right)^2}{\frac{s_1^2}{n_1(n_1 - 1)} + \frac{s_1^2}{n_1(n_1 - 1)}}.$$

Note that this number might be non-integer.

Example: Consider data on two methods for obtaining the concentration of a chemical component in a vat. The first technique is a standard method (precise, but slow and expensive), and the second is quick and cheap, but erratic. Clearly, the variances of the two populations will be different.

Suppose we are interested in comparing the means of the concentrations produced by these two procedures in order to examine whether or not the quick method systematically over - or under - estimates the true concentration. The hypotheses are

$$H_0 : \mu_s = \mu_Q \text{ vs } H_1 : \mu_s \neq \mu_Q.$$

The data are as follows:

Standard method: 25, 24, 25, 26

Quick method:23, 18, 22, 28, 17, 25, 19, 19

The means and variances of the two methods are (25, 21) and (0.67, 17.71), respectively. Using the formula above, the (approximate) degrees of freedom are 7.988. With these degrees of freedom and $\alpha = 0.05$, the critical region of this test consists of the values of T_1 smaller than -2.306 or larger than 2.306. The observed value of the T_1 statistic is $t^* = 2.5923$, indicating a significant (at a 5% significance level) difference between the mean concentrations produced by the two methods. The *p*-value is 0.0321, which is twice the area to the right of t^*, under the corresponding Student's *t* distribution, since the alternative hypothesis is two-sided.

For this example, if we had used the T_1 statistic (assuming equal variances) we would not have rejected the null hypothesis. As a general rule, whenever there is doubt it is better to avoid the assumption of equality of variances. Another possibility is to transform the original values of the variables (for instance, using logarithms or square root) to see if that makes the variances homogeneous.

Lastly, we should note that, whether or not the assumption of equal variance is made, the independent two-sample *t*-test assumes that the two underlying populations have normal distributions. If this assumption does not appear to be satisfied (as evidenced by informal and formal tests using the two samples), then we should consider using a nonparametric test. The nonparametric analogue of the two-sample *t*-test is called the ***Mann - Whitney test*** or the ***Wilcoxon rank sum test***; this test will be discussed in greater detail later in this lecture.

(II.c.ii.1.B) Comparing the Variances of Two Populations (The F-test)

We have seen that one particular form of the two-sample *t*-test assumes (among other things) homogeneity of variances. We are able to test the assumption of equality of variance by noting the following result. If two independent random variables, X_1 and X_2, each have chi–squared distributions on υ_1 and υ_2 degrees of freedom, respectively, then the random variable

$$F = \frac{X_1 / \upsilon_1}{X_2 / \upsilon_2}$$

has an ***F-distribution*** with parameters υ_1 and υ_2, denoted $F(\upsilon_1, \upsilon_2)$.

It can be shown that, if $X_1,...,X_n$ is a random sample from a $N(\mu, \sigma^2)$ population,

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},$$

where S^2 denotes the unbiased sample variance. Therefore, to test the hypothesis

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1$$

against a one – or two – sided alternative, we use the statistic

$$F = \frac{S_1^2}{S_2^2},$$

where S_i^2 is the unbiased sample variance estimate for sample *i*, *i*=1, 2. From the result stated above, the null distribution of the test is $F(n_1 - 1, n_2 - 1)$, and this can be used to construct a rejection region appropriate to the form of the alternative hypothesis. Most statistical packages will perform this test (for example, `var.test` in R). Note that this result relies on the assumption that each sample is normally distributed and independent.

(II.c.ii.2.) Polytomous Predictor

Suppose that we want to see how the mean of a continuous variable changes across more than two groups (i.e. across the levels defined by a polytomous predictor variable).

ANOVA Approach

It is possible to generalise the analysis for comparing the means from two independent, normally distributed populations. The technique is known as *Analysis of Variance (ANOVA)*. This name can be misleading since the main interest is in comparing two or more population means. However, the technique compares these means by examining certain components of variance, and it is from this that it derives its name.

We use ANOVA in the following situation. Suppose we have random samples (of potentially varying sizes) from each of *k* normally distributed populations. We write the null hypothesis (of homogeneity of means) as

$$H_0 : \mu_1 = \mu_2 = \cdots \mu_k$$

This hypothesis is tested against the *general* alternative that at least one of the population means is different from the others. Note that the preceding statement defines several alternative hypotheses. For instance, suppose that we are analysing 3 samples. It may be that only one of the means is different: the remaining two means might be equal and form a homogeneous group. Additionally, we would also want to reject H_0 if the three means were all different from each other. If we do not reject the null hypothesis, we simply conclude that there are no differences between the means. If we do reject the null, we must

still explore the possible patterns that led to this rejection. This exploration is known as *post-hoc analysis,* and it will be discussed later.

The basic idea underlying ANOVA is as follows. An important theorem in mathematical statistics states that the *total variance* of the *pooled* sample can be divided into two components:

- the *within groups variance* and
- the *between groups variance*.

The within groups variance is simply the sum of the variances calculated for each individual group. The between groups variance is the variance obtained using the means of each group as data. The within groups variance represents the internal variability of the groups; the between groups variance measures how well separated these groups are. If the groups are well separated from each other, the ratio of between group variance to within group variance should be large. In order to decide how large the value of this ratio should be in order to be considered significant, we use the F distribution. The degrees of freedom are $(k-1)$ and $(n-k)$, where n is the total sample size.

Note that the two-sample t-test (with equal variance) is a particular case of ANOVA (with k=2). Thus, we would expect that the assumptions of normality and homogeneity of variance (required by the t-test) are also required for comparing k independent samples. This is the case, and, as before, we should test that these assumptions are satisfied.

> **Example:** Thirty magazines were ranked by the educational level of their readers. The thirty ranked magazines were then divided into 3 equal sized groups, and three magazines were randomly selected from each group. Six advertisements were randomly selected from each of the nine selected magazines. The chosen magazines were:

- **Group 1** Highest educational level.
    1. Scientific American
    2. Fortune
    3. The New Yorker;
- **Group 2** Medium educational level.
    4. Sports Illustrated
    5. Newsweek
    6. People;
- **Group 3** Lowest educational level.
    7. National Enquirer
    8. Grit
    9. True Confessions;

> For each advertisement, the following variables were recorded:

WDS = number of words in advertisement copy
SEN = number of sentences in advertising copy
SYL3 = number of 3+ syllable words in advertising copy
MAG = magazine (1 through 9 as above)
GRP = educational level (as above)

A Kolmogorov-Smirnov test confirms that the data do not contradict the assumption of normality for each of the variables WDS and SEN stratified by MAG and GRP. We show the results of the ANOVA procedure for WDS (the response variable) using GRP as the factor (predictor variable).

```
Variable       WDS        num of words in ad
By Variable  GROUP         educational group

                                  Analysis of Variance

                        Sum of      Mean             F         F
Source            DF    Squares     Squares       Ratio      Prob

Between Groups     2    10141.81   5070.91       1.1762     .3167
Within Groups     51    219866.77   4311.11
Total             53    230008.59


Levene's Test for Homogeneity of Variances

    Statistic     df1     df2        2-tail Sig.
      2.2422       2      51           .117
```

We do not reject the hypothesis of homogeneity of variances. The *p*-value indicates support for the null hypothesis that the three groups have the same mean number of words per ad. The following error bar graph illustrates this result.

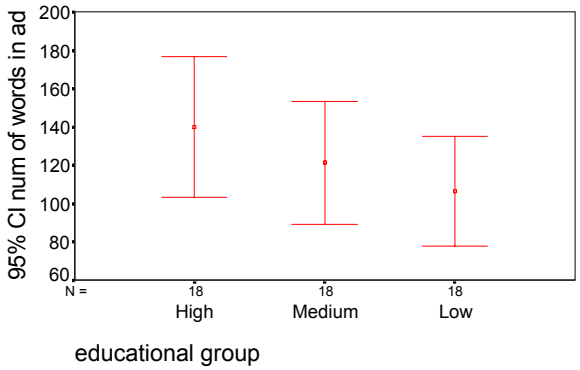

Figure 2: *Comparison of three groups*

Nonparametric Approach

When using the ANOVA procedure, we assume that the data from each group follow a normal distribution and that the groups we are comparing have homogeneous variances. If the variances are not all equal, then the conclusions

about the group means drawn from ANOVA analysis may not be valid since the observed ANOVA *p*-value will be smaller than the one we would have obtained if the assumption of equal variance were satisfied. This means that ANOVA will yield an *anti-conservative p*-value (that is, an increase in the probability of Type I error) if the homogeneity of variances assumption is not satisfied. Therefore, it is important to test, either formally or informally, that the homogeneity of variance assumption is satisfied, as has been stated before.

If this assumption does not appear to be satisfied, transforming the data (perhaps using a logarithmic transformation) can sometimes help, as we will see in the Post-hoc analysis section below. In addition to homogenising the variances, using a transformation can also sometimes induce approximate normality in the data.

However, if we cannot find a transformation that appears to homogenise the variance or normalise the variables, then we should consider using a nonparametric test. As mentioned previously, nonparametric tests do not require specific assumptions regarding the distribution of the underlying population (e.g. normality). As we have already mentioned, the nonparametric analogue of the independent two-sample *t*-test is the Mann-Whitney test. The nonparametric equivalent of ANOVA is the ***Kruskal-Wallis test***, which is a generalisation of the Mann-Whitney test for greater than two groups. For both the general test and its two-sample version, the null hypothesis is that the **medians** are equal, against the general alternative that at least one differs from the others. Note that it makes sense to compare the medians, rather than the means, because if the data are skewed, as they probably would be if we are using a nonparametric test, then the value of the mean will be artificially inflated (or deflated). The Mann-Whitney and the Kruskal-Wallis tests test this null hypothesis by transforming the data into *pooled ranks* (i.e., they start by assigning the rank 1 to the smallest observation in the pooled sample, and so on) and then calculating a test statistic from these ranks. Lastly, both tests appear in the nonparametric submenu of SPSS and as individual functions in Splus and R

> **Example:** The following histogram shows the distribution of wealth (measured in billions[•] of dollars) of a sample of U.S. billionaires, compiled by Fortune magazine. The figure also displays the normal curve corresponding to the mean and standard deviation of the data. Clearly, the assumption of normality is not reasonable.

[•] American billion (i.e. one thousand million)

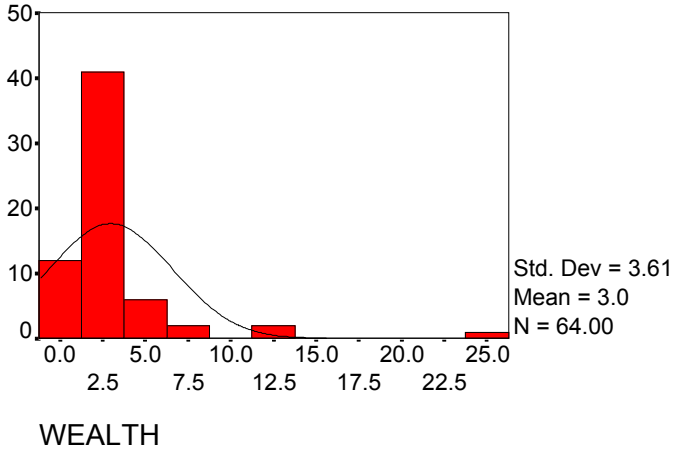

Figure 3: *Histogram and normal curve for wealth data in the USA*

In addition to the U.S., four other regions were sampled, including Europe and Asia. The distribution of the data is quite skewed, and a Kolmogorov-Smirnov test rejects the null hypothesis of normality for the wealth variable in each region.

In order to compare the medians of the five regions, we use the Kruskal-Wallis test. Doing so yields the following results:

```
   WEALTH
by REGION0   region


      Mean Rank    Cases

       129.53       38   REGION0 = 1   Asia
       108.59       80   REGION0 = 2   Europe
       119.66       22   REGION0 = 3   Middle East
       115.19       29   REGION0 = 4   Other
       119.98       64   REGION0 = 5   USA

                   233   Total

                                  Corrected for ties
Chi-Square      D.F.  Sign.    Chi-Square      D.F.       Sign.
 2.7365          4    .6028     2.7441          4         .6015
```

The Kruskal-Wallis test uses a chi-square approximation to the null distribution, and this is the reason for the name appearing in the results. If we have ties in the data (as we do here), then we should use the *p*-value that has been corrected for ties. In this case, it is 0.6015, indicating that there is insufficient evidence to reject the hypothesis of homogeneity of median wealth between the five regions.

Post – Hoc Analysis

Whether we use ANOVA or the Kruskal-Wallis test to do so, when we investigate how the centre of the continuous variable changes across the *k* groups, our initial null hypothesis is that the population means (or medians) are all homogenous. If we fail to reject this

hypothesis, the analysis ends there. If we do reject the initial null hypothesis, then we will have to establish the reason(s) for doing so. For example, it may be that only one group mean differs from the rest or that there is a particular pattern in which the groups appear to be separated.

One way of finding significant differences between the means is to make all possible pairwise comparisons (i.e. test if each pair of means is equal). Note that we can use either of the two-sample *t*-tests to make these pairwise comparisons if we assume that the populations are normal; if not (i.e. if we used the Kruskal-Wallis test to test the initial null hypothesis), then we should use the Mann-Whitney test to make these pairwise comparisons. In either case, making these pairwise comparisons leads to the problem of *multiple comparisons*, as we saw in Lecture 7. As we saw then, one way to ensure an overall significance level of α is to use the Bonferroni correction, whereby each individual test (comparison) is conducted at the

$$\alpha^* = \frac{\alpha}{\binom{k}{2}}$$

level of significance. Note that $\binom{k}{2} = \frac{k(k-1)}{2}$ is the number of possible pairwise comparisons between k groups. There are other techniques for adjusting the significance levels used for multiple comparisons. Statistical software such as SPSS give several possibilities, including some which correct for unequal population variances. We will only concentrate on the Bonferroni correction in these notes.

As a final note, it should be pointed out that another way of investigating how the means of the groups differ is by using ***contrasts*** (or ***planned comparisons***). These are typically specified before data analysis (and sometimes before data collection) and reflect comparisons of interest to the researcher.

> **Example:** We return to the magazine advertisement data. We have already established that there is no difference in mean word length between adverts appearing in magazines of different educational levels (low, medium, high). We now consider the relationship between word length and the factor *magazine,* which has 9 levels. The following graph shows the mean word lengths and their 95% confidence intervals across the 9 magazines.

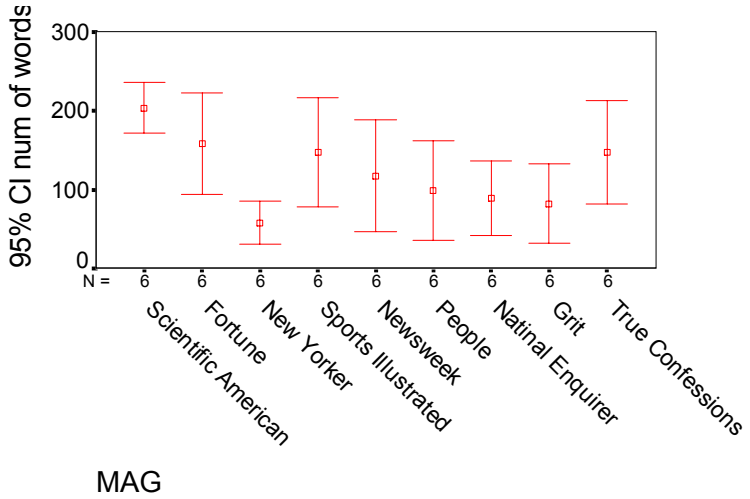

Figure 4: *Comparing nine means*

Note that there is evidence of non-homogeneous variances among the groups. The ANOVA table is as follows:

```
Variable     WDS        num of words in ad
By Variable  MAG

                                  Analysis of Variance

                   Sum of          Mean             F       F
Source     D.F.    Squares        Squares          Ratio   Prob.
Between     8    99217.2593     12402.1574        4.2671  0.0007
Within     45   130791.3333      2906.4741
Total      53   230008.5926


Levene Test for Homogeneity of Variances

    Statistic     df1     df2       2-tail Sig.
      2.3837       8      45          0.031
```

It appears that the null hypothesis of homogeneity of means is rejected (p-value = 0.0007), and we can conclude that at least one of the 9 mean word lengths differs from the others.

However, note that the p-value for the test for homogeneity of variances is quite small, indicating probable departures from the null hypothesis. Since our main interest is in the differences between the means, this test is relevant only in ensuring that the assumptions of ANOVA are satisfied, and thus that the p-value obtained from testing the means is correct. However, as noted above, the p-value produced by the ANOVA procedure is not correct if the homogeneity of variance assumption is not satisfied. We should, therefore, attempt to remedy this situation.

As mentioned earlier, one potential solution of this problem is to transform the data so that the groups have similar variances. Usually, taking natural

logarithms of the data is an adequate transformation to produce homogeneous variances (note that if the data include 0s or negative values, we cannot use the logarithmic transformation. If this is the case, we simply add a constant to the data values so as to make them positive, and then use the logarithms of these new values).

The following error bar graph shows the means, and their confidence intervals, for the natural logarithm of word length.

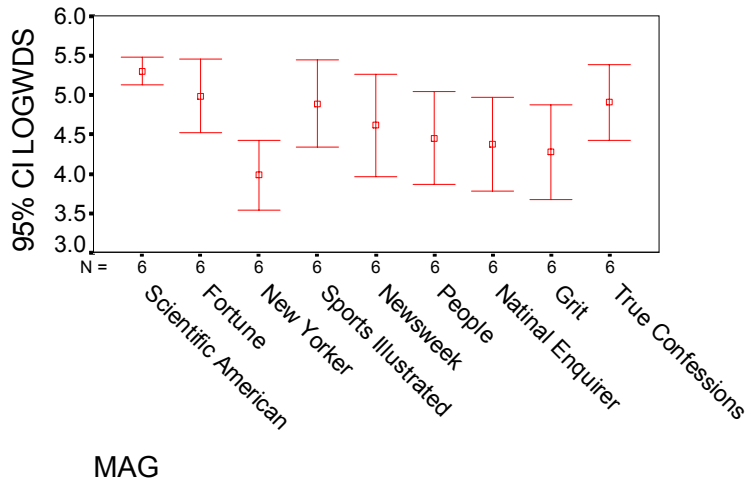

Figure 5: *Comparisons of means using a logarithmic transformation*

Note that, perhaps with the exception of *Scientific American*, the variances seem to be of a comparable magnitude for all magazines. The formal tests are as follows.

```
Variable  LOGWDS
By Variable  MAG

                                Analysis of Variance

                 Sum of          Mean            F        F
Source   D.F.    Squares         Squares         Ratio    Prob.
Between   8       8.1901          1.0238        4.1274   0.0009
Within   45      11.1619          0.2480
Total    53      19.3519

Levene's Test for Homogeneity of Variances

    Statistic     df1     df2       2-tail Sig.
      1.0169       8      45          0.437
```

It appears that the logarithmic transformation has achieved its desired result: we do not reject the null hypothesis of homogeneity of variances (p-val. = 0.437). Note that our previous conclusion regarding mean word lengths still holds: at least one magazine has a different mean word length than the others (p-val. = 0.0009).

In order to investigate the patterns in the means, we must perform a *post-hoc* analysis using the Bonferroni correction to adjust for multiple comparisons. Doing so shows that only 3 pairwise differences are significant. They are (New Yorker, Fortune), (New Yorker, Scientific American) and (Scientific American, Grit). Note that because of the small sample sizes in our example (each magazine contributes only 6 adverts), the standard errors of the means (not shown) are quite large. Therefore, we cannot establish a significant difference even between pairs of magazines that apparently differ in mean word length by a large margin. For instance, there is no significant difference in the mean word lengths for the New Yorker (53.88) magazine and True Confessions (135.88), even though the latter has a mean almost 2.5 times that of the former. This is an example of how small sample sizes increase the probability of a Type II error.

(II.c.iii.) Continuous Response with a Continuous Predictor

The most common form of statistical analysis for the case in which both the response and the (one) predictor are continuous is known as ***simple linear regression***. Linear regression is a parametric procedure, as the response variable is assumed to follow a normal distribution. A complete description of linear regression is beyond the scope of this course, and details are omitted. Note, however, that the IAUL and Department of Statistics do run an introductory modelling course in which such techniques are described fully.

In simple linear regression, we are able to test for a linear relationship between the two variables. Linear regression models the (linear) relationship between the response (y) and predictor (x) as

$$E(y \mid x) = \alpha + \beta x ,$$

where α denotes the ***intercept*** parameter and β the ***slope*** parameter. Statistical techniques such as ***maximum likelihood***, or numerical techniques such as ***least squares***, are used to estimate these parameters and provide their standard errors. By testing the hypothesis that β =0, we are testing whether or not y is linearly related to x. Note that the test statistic is particularly straightforward when we assume the response is normally distributed.

Linear regression is performed by all statistical software packages. Most give the estimate of the regression parameters and their standard errors, the test statistic (t) for testing the hypothesis of no relationship between y and x, and the p-value of this test.

(II.c.iv.) Categorical Response with a Continuous Predictor

(II.c.iv.1.) Grouped or Ungrouped Binary Response

When exploring the relationship between a categorical response variable and a continuous predictor variable, the simplest scenarios are when the response variable can only take the values 0 ('failure') or 1 ('success'), in which case the data are said to be ***ungrouped***, or when the response variable counts the number of 'successes' out of the number of 'trials,' in which case the data are ***grouped***. For both situations, a commonly used analysis technique is ***logistic regression***.

Logistic regression, like log-linear models and linear regression, is a special case of a generalized linear model (GLM). As we mentioned previously, GLMs allow us to model the relationship between a (potentially) non-normally distributed and (potentially) non-numerical response variable and various predictors. In the case of logistic regression, we can see that the response has a Bernoulli (or binomial) distribution.

The relationship between the response and the predictor is modelled in a similar fashion to that of linear regression, except here we relate a *function of* the expected value of Y to the Xs, as in

$$f[E(y \mid x)] = \alpha + \beta\, x\,.$$

For logistic regression, the function f is the logistic function (hence the name), which is defined as

$$f(x) = \ln\left(\frac{x}{1-x}\right),$$

for values of x between 0 and 1. Note that, in this case, the expected value of y is the probability of success, p. Logistic regression therefore models the relationship between the log of the odds ratio of success and the predictor(s).

Most statistics packages implement logistic regression. Typically, the output from such a test includes estimates of the regression parameters, α and β, their approximate standard errors, and an approximate test statistic and p-value.

(II.c.iv.2.) Polytomous Response

If the polytomous response variable is nominal categorical (no intrinsic ordering of categories), the modelling procedure is a relatively straightforward extension of logistic regression. Specifically, we can assume the response is a realization of a *multinomial* random variable, and use ***multinomial logistic regression*** to explore the relationship between the expected response and the predictor(s). Most statistical software packages have this facility. For example, SPSS, from version 9.0 onwards, implements multinomial logistic regression via the NOMREG macro.

If the polytomous response is ordinal (for example, the response has categories low, medium, and high), there are two options open to us. One is to ignore the ordering and simply use a multinomial regression. However, we lose information in doing so. The second option is to account for this ordering in the modelling procedure. An example of such a technique is ***proportional odds logistic regression (POLR)***. Several statistical software packages, including SAS, SPlus, R, Stata, and SPSS implement this facility.

Useful references on this material can be found in Agresti (1996) and Johnson and Albert (1999).

(II.d.) More Than One Predictor Variable

We are not limited, when using techniques such as linear and logistic regression, to only one predictor variable. Indeed, we often wish to examine the relationship between a response and several predictors. One common reason we might wish to do so is to control for confounding variables, as we saw in our example on Simpson's paradox.

The extension to more than one predictor, in terms of modelling and computation, is quite straightforward. For example, if we wish to examine the relationship between Y and three predictors X_1, X_2, and X_3 via a linear regression, we can write the model as

$$E[Y \mid x_1, x_2, x_3] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 .$$

Each parameter $(\alpha, \beta_1, \beta_2, \beta_3)$ is then estimated, and standard errors are obtained for each. Hypothesis tests regarding the overall linear relationship between the response and all the predictors can be constructed via an *F*-test. Additionally, each marginal linear relationship (i.e. each relationship between *Y* and one predictor variable) can be tested.

As a final point, we note that the interpretation of model parameters when there is more than one predictor differs slightly from that when there is only one predictor. In the single predictor case, we interpret the slope parameter, β, as the increase in (a function of) the expected value of Y per unit increase in X. In the multiple predictor case, we interpret the slope parameter for the i^{th} predictor as the increase in (a function of) the expected value of Y per unit increase in X_i, conditional on the values of the remaining predictors (i.e. holding all other predictors *constant*).

(III.) References

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*. Wiley, New York.

Appleton, D.R., French, J.M., and Vanderpump, M.P. (1996). "Ignoring a covariate: An example of Simpson's Paradox." *American Statistician*, 50, pp. 340-341.

Johnson, V E. and Albert, J.H. (1999). *Ordinal Data Modelling*. Springer, New York.

MCB (I-2000), *KNJ* (III-2001), *JMcB* (III-2001)