

16. Statistical Issues

In conducting an inquiry one major hurdle most people face is how to handle the statistics. With the availability of powerful software which is also increasingly becoming user friendly the temptation will be to tap out commands for statistical tests which may be inappropriate to the question the researcher is asking. In previous chapters suggestions have been made about tests for different situations. In this chapter the purpose is to provide a commentary about the commonly employed tests, and at the same time introduce the reader to the popular software EPI-INFO jointly put out by the World Health Organization and the Centres for Disease Control, Atlanta.

What is EPI-INFO

EPI-INFO does a number of things. For a detailed description the reader is referred to the manual which accompanies the software. For our purpose we would consider its three main parts, amongst many. These are:

EPED, a word processing package for typing the questionnaire or any other data collection form. As a word processor it can also be used for writing the report. The resulting file gets copied onto a floppy disc, and always has a suffix .QES. With a little experience the researcher can use the questionnaire so prepared also for data collection without much modification.

ENTER programme is for entering the data. In this mode the computer reads the .QES file, and presents a blank questionnaire on the screen. The operator is then prompted to enter the appropriate data for each variable. The entered data gets saved on a floppy, and at the end the operator has a file with a suffix .REC.

ANALYSIS programme is for doing the analysis. In this mode the .REC file is first read, after which the operator is able to perform the basic analysis, including simple graphics like bar charts, histograms, pie charts, scatter plots etc. for exploring the data.

Table 16.1 Programs menu of EPI – INFO

Programs
EPED word processor
ENTER data
ANALYSIS of data
CHECK customize entry
IMPORT files
EXPORT files
MERGE files
STATCALC calculator
CSAMPLE analyze surveys
EPITABLE calculator
EPINUT anthropometry

The reader would now be guided through an investigation using EPI-INFO whilst providing a commentary on various statistical issues that arise. The research question on which the investigation is based is the following:

Can Women's Organizations bring about Health Development?

To answer this question a study was designed in Kandy, Sri Lanka where a Women's Development Centre (WDC) has been established, and had been operating in 12 communities at the time of the study. Four communities were randomly selected. Mothers with a child less than five years old were eligible for inclusion, and 100 mothers were enrolled from the 4 study communities. For each selected community an adjoining community was selected such that the WDC had not yet commenced activities there. From amongst these adjacent communities 100 mothers were enrolled to form a comparison group. (The variable "group" distinguishes between the study and comparison groups).

(Please refer to Appendix 12.1 – Approach to Data Analysis). Having identified the outcome, explanatory and confounding variables the next task is that of summarizing the values of the key variables in a comprehensible manner. The summary measures commonly employed are mean, median, and mode. They all give an idea of the central tendency. But just knowing the central tendency is not enough. One also needs to know the variability in order to get a complete picture. Variability is a measure of how far away from the centre the values occur. In a Normal distribution, the mean gives the central value, and variance is the average deviation squared. The square root of the variance is the standard deviation. The terminology sounds daunting, but the advantage of knowing the mean and standard deviation in a Normal distribution is that approximately two-thirds of the values will fall within the range $\text{mean} \pm 1 \text{ sd}$, and 95% of them would be in the range $\text{mean} \pm 2\text{sd}$. No summary can be more succinct. Moreover, this also forms the basis for significance testing. But before we do so there is

still one small matter to resolve. Significance testing for a difference between two means (for example, in our study the mean ages of the mothers in the study and comparison group) assumes a Normal distribution. One way of checking is by visual inspection of a bar chart. If the distribution were to be skewed then one can get round it by logarithmic transformation of the data. Treating the data in that way helps to pull in the long right hand tail, and extend the left hand tail.

Table 16.2 Commands available under Analysis

General	Var manip	Graphics
READ	SELECT	PIE
CLOSE	SORT	BAR
VARIABLES	DEFINE	HISTOGRAM
ROUTE	LET	LINE
LIST	IF	SCATTER
FREQ	RECODE	
TABLES	COMBINE	
DESCRIBE		
MEANS		
REGRESS		
SUMFREQ		
SUMTABLES		
MATCH		

Significance testing

Statisticians use the term "significance" differently from its dictionary meaning. In the statistical sciences "significance" of an observed result means that it is "probably real and not by chance". When the results of any study are derived three possibilities arise viz.

- The results could have arisen by chance.
- The results are so because of bias.
- The results are influenced by a confounding variable.

Appropriate statistical tests are resorted to for seeking an answer to the first possibility viz. "Could the results occur by chance?" With regard to the second possibility, **bias** is an issue of study design. Its origins lie in the manner of selection of subjects for the study or in the way information has been collected from the study subjects. Bias is carefully looked for in the design stage of the study. One can do very little about bias once the study is completed and the data analysis is in progress. The third

possibility, **confounding**, is an issue of an alternative explanation of the results, and is to be dealt with both at the time of designing a study and during data analysis.

Coming back to significance testing, by convention the dividing line is at a significance level of 0.05. A value of $P < 0.05$ means that the observed results could arise by chance only 1 in 20 times. However, one must resist the temptation that $P < 0.04$ means firmly established, and $P < 0.06$ is to be ignored.

In selecting a test of significance one has to take account of two features of the data viz.

- (i). How was the data collected? For example were the study and comparison groups recruited independently, or was there any matching done? If two populations are being compared the samples may have been drawn independently or they are paired. If any matching or pairing was done during sampling then it should be allowed for in significance testing. One performs a paired t ' test as compared to the t ' test.
- (ii). Some tests for continuous data assume a Normal distribution. These are the parametric tests like z score, t ' test, standard error of the mean, and so on. Others do not assume a Normal distribution. These are the non-parametric tests. They are based on the principle of replacing the actual data values by their ranks, and ordering the variables to be analysed by rank. The Mann-Whitney test is the non-parametric equivalent of the t ' test; McNemar's test (for proportions) and the Signed Rank test (for ordinal data) are equivalents of paired t ' test; and the Kruskal-Wallis test is the equivalent of analysis of variance in the parametric tests.

In the Sri Lankan study which we are using as demonstration of these principles, we may wish to ensure that the study and comparison groups were similar (i. e. matched) on certain key variables like mothers' ages, the number of children they had borne, maternal literacy, maternal education, and uptake of health services as judged by attendance at antenatal care, breastfeeding, and immunization. Since the samples were drawn independently (not paired), significance testing for the former would be by means of the t ' test for numeric variables, and by the Mann-Whitney test for the categorical variables.

One tailed or two tailed test?

These terms often cause confusion. Significance testing depends on the question one is asking. For example, "Is the new treatment better or worse than the placebo?" requires a two-tailed test. On the other hand "Is the treatment better than the placebo?" is answered by 1-tailed test. The 1-tailed significance result is typically half of the 2-tailed result.

When we use a test of significance to compare two groups we usually start with the null hypothesis that there is no difference between the study populations from which the data were obtained. If the null hypothesis is not true then the alternative hypothesis must be true viz. there is a difference. Neither the null hypothesis nor the alternative one specify a direction, for example, that the study population is

doing better than the comparison group or vice versa. Thus, in most instances we are using two-sided tests. If one-sided test is to be used the direction must be specified in advance. In general, a one-sided test is appropriate when a large difference in one direction would lead to the same action as no action at all.

Significance testing or Confidence Interval

One shortcoming of significance testing is that there occurs a tendency to process the data in a way where all that matters is whether a test result is significant or not. If there is a difference, however small, a significance test result can be arrived at by taking a large enough sample. Such a game with numbers can have no place in scientific research. The investigator is seeking after the truth, and what really matters is whether the study has been able to detect a clinically important difference between two population means, and if so, how large the difference is likely to be. The confidence interval gives a range for the difference between the two means. The investigator can now decide whether the difference is of practical significance. Thus confidence interval is a desirable statistic because it gives a clear indication of the size of the difference between two means. A very large confidence interval indicates large variability, and therefore a large sample size is called for to reduce the variability. The current practice is to report the confidence interval along with the test statistics like mean, odds ratio, relative risk, and so on. (See Appendix 16.2)

Demonstrating relationships between variables

It is often the case that the researcher wishes to demonstrate a relationship between two or more variables e.g. between height and body weight; between maternal weight and the size of the baby at birth; between the method of feeding and episodes of diarrhoea, and so on. If both the variables are continuous a preliminary scatter plot is helpful in making a visual assessment of a relationship. The scatterplot would show whether the points are randomly scattered (no relation), clustered about a straight line (linear relationship), or a curve (curvilinear relationship). Log transformation is often helpful in changing a curvilinear relationship to one approximating a straight line by transforming the values on either the X or Y-axes to logarithms. The statistical measure of relationship between variables is the product-moment correlation coefficient, and is denoted by the letter "r". (Pearson's Correlation coefficient). It takes a value between -1 to +1. The nearer the value is to zero the less strong is the relationship. A negative sign means an inverse relationship. The square of "r" is the proportion of variability in the value of one variable, which is explained by its linear relation with the other variable. Thus if the value of "r" is 0.7, one can surmise that about 50% of the variability in one variable is explained by its linear relationship with the other.

If the two variables under study are not continuous but categorical the rank correlation coefficient is used instead. It also gives the "r" statistic (Spearman's correlation coefficient) with similar meaning as in the case of the product-moment correlation coefficient for continuous variables.

Once a scatter plot is obtained one may wish to fit a regression line in order to get a more vivid picture of the association between the two variables. The terminology now changes somewhat. One of the variables is labelled as the outcome (or response or dependent) variable, and the other is called the explanatory (or predictor or independent) variable. Which one is called which depends on the research question. For example, if the study is about the effect of maternal weight gain in pregnancy on the size of the newborn, the weight of the baby at birth is the outcome (or response or dependent) variable and maternal weight is the explanatory (or predictor or independent) variable.

Fitting a regression line also helps to obtain the algebraic statement of the relationship like $y = \alpha + \beta x$. One must bear in mind that in trying to fit a regression line to a scatter plot three assumptions are being made viz.

- (i). The relationship is linear.
- (ii). The spread of the data about the line is the same at all levels of the independent (or predictor or explanatory) variable.
- (iii). The deviations of the data values from the line are normally distributed.

These assumptions can be checked by making what is called a "diagnostic plot of the residuals". This is the difference between the observed values of the response variable values and the values as fitted on the regression line.

Before leaving the subject of correlation, it should be stressed that 'r' measures only the linear relationship. Insignificant 'r' means that some other relationship exists, hence the importance of viewing the scatter plots for prior exploration to get a feel of the spread of the data.

Significance Testing of Categorical Variables

One is often comparing two populations on categorical variables like immunized/not immunized; ill/well; affected/not affected and so on. There are tests for significance of difference between two proportions similar to those for difference between two means. When actual counts are being compared instead of proportions, one ends up with a 2 x 2 table. This is a powerful tool and some of its uses have been described in connection with sample size calculation, the odds ratio, and diagnostic tests. Here the 2 x 2 table is being considered for the Chi square (χ^2) test which is one of the most commonly employed tests of significance. EPI – INFO calculates the χ^2 value as well as the odds ratio each time a 2 x 2 cross tabulation is performed.

The Chi square test has two restrictions:

(i). Either the total sample size should be greater than 40

or

(ii). If the size is between 20 and 40, the expected value in a cell should not be less than 5.

If these conditions are not met the Fisher's exact test is resorted to, which most computer software would perform unprompted.

Multiple Regression Analysis

We have so far considered tests of significance for association between two variables. In practice, however, one outcome may be a result of several variables. Multiple regression analysis is used to tease out the effect of several predictor variables on the outcome variable. Different types of analysis are needed for handling different types of outcome variables (e.g. numerical, or categorical), for allowing for interaction between variables, as well as for dealing with confounding. Several powerful packages are available for the purpose. The problem is not with performing the analysis, but ending up with an inappropriate type of analysis leading to an erroneous inference.

Appendix 16.1.

Commonly used statistical tests

Name of Test	Test statistic	Parametric (P) Or Non-parametric (NP)	Purpose
' <i>t</i> ' test for independent samples	' <i>t</i> '	P	To test the difference between the means of two independent groups
' <i>t</i> ' test for paired samples	' <i>t</i> '	P	To test the difference between the means of two paired (matched) groups
ANOVA	<i>F</i>	P	To test the difference in the means of three or more independent groups
Mann-Whitney U test	<i>U</i>	NP	To test the difference in the categorical scores obtained between two groups
Wilcoxon Signed Rank Test	<i>Z</i>	NP	To test the difference in the categorical scores between two paired (matched) groups
Kruskal-Wallis Test	<i>H</i>	NP	To test the difference in the categorical scores between three or more independent groups
Chi-square Test	χ^2	NP	To test the difference in proportions in two groups
Pearson's product-Moment Correlation	<i>r</i>	P	To test for association between two numeric variables
Spearman's Correlation coefficient	<i>r</i>	NP	To test for association between two categorical variables

Appendix 16.2

Confidence Interval

Confidence Interval (CI) is based on the idea that the same study carried out on different samples of subjects would not yield identical results but would be spread round the true (but unknown results). CI estimates this variation. It may be interpreted as:

- 1). If several samples are drawn from the same population and the confidence interval of the result calculated then 95% of such intervals will contain the true value. and
- 2). If only one sample has been drawn then C.I. gives the range within which we can be 95% certain that the population value lies.

Standard Error of a mean = Standard deviation / $\sqrt{\text{number of subjects}}$.

(5% confidence Interval of the mean = Mean \pm 1.96 \times standard error of the mean.

Standard error of a proportion = $\sqrt{p \times (1 - p) / n}$

If p =40% (i.e.0.4) out of n= 60, then SE = $\sqrt{0.4 \times 0.6 / 60}$ 0.063 =6.3%

95% C.I. = 40% \pm 1.96 \times 6.3% = 27.6% to 52.4%

2. Confidence Interval of Relative Risk

If r1 and r2 events are observed among n1 and n2 subjects then the observed proportions are p1 = r1/n1 and p2 = r2/n2.

The relative risk is p1/p2.

The standard error of the log normal of RR is given by the formula $\sqrt{1/r1 + 1/r2 - 1/n1 - 1/n2}$

Example: In a randomised controlled trial comparing a new treatment in 125 patients (n1) 15 (r1) showed full recovery (p1= 15/125 =12%).

By comparison with the old treatment 30 patients (r2) out of 120 (n2) showed full recovery (p2=30/120 = 25%).

Then RR = p1/p2 = 12/25 = 0.48 = 48%.

The normal logarithm of 48% is -0.734.

The standard error (SE) of the RR by the above formula is $\sqrt{1/15 + 1/30 - 1/125 - 1/120}$ = 0.289

95% confidence interval of the normal logarithm of the RR = -0.734 \pm (1.96 \times 0.289)
= -1.301 to -0.167

Then the Confidence Interval of the RR = $e^{-1.301}$ to $e^{-0.167}$ = 0.272 to 0.846 = 27.2% to 84.6%

3. When the outcome is a difference between two means (e. g. mean blood pressure reduction in two groups)

The standard error of difference between two means is given by the formula

$$\sqrt{\frac{(n_1-1)Sd_1^2 + (n_2-1)Sd_2^2}{n_1 + n_2 - 2}} \times (1/n_1 + 1/n_2)$$

where Sd_1 and Sd_2 are the standard deviations of the two means, and n_1 and n_2 are the number of subjects in the two groups.