International Journal of Scientific Engineering and Applied Science (IJSEAS) – Volume-4, Issue-9, September 2018
ISSN: 2395-3470
www.ijseas.com

# Statistical Analysis of Data

**Dr.S.Narayanan[1], Dr.C.Sabarigirinathan[2], Dr.K.Vinayagavel[3], Dr.D.Deepiha[4], Dr.S.BalaSiddharth[5], Dr.M.Saravanapriya[6]**

**[1](PhD student of Pacific University of Higher Education and Research (PAHER), Udaipur, Rajasthan)**

**[2](Professor and HOD, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)**

**[3](Professor, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)**

**[4&6](Post Graduate Student, Dept of Prosthodontics, Tamilnadu Govt Dental College & Hospital, Chennai, India)**

**[5](MBBS – Private, Chennai, India)**

## Abstract

Statistical analysis is a component of data analytics. Statistical analysis involves collecting and scritinizing every data sample in a set of items from which samples can be drawn.Statistically significant is the likelihood that a relationship between two or more variables is caused by something other than chance. Statistical hypothesis testing is used to determine whether the result of a data set is statistically significant. This article explains the procedure of performing various statistical operations.

## Introduction

In most research studies, the information collected represents only a sample from the population of interest (target population). Drawing conclusions about the population, whether it is a simple descriptive study or a randomized controlled trial, depends on statistical analysis of the data. This manual is intended to assist in the preparation of a research proposal, and not with data analysis. However, since the choice of the design has a direct impact on the analysis of the data, it is important to have an idea of the type of analysis anticipated when designing the study. Therefore, we will briefly review the important aspects of statistical analysis.

## Basis for statistical analysis

The fundamental principles of probability theory (briefly reviewed in Chapter 7) are used in statistical inference. All the inferences are based on three primary entities: the population (U) that is of interest, the set of characteristics (variables) of the units of this population (V), and the probability distribution (P) of these characteristics in the population.

### *The population* (U)

The population is a collection of units of observation that are of interest, and is the target of the investigation. For example, in determining the effectiveness of a particular drug for a disease, the population would consist of all possible patients with this disease. In determining the prevalence of the incidence of HIV infection among commercial sex workers in a country, the population would consist of all the commercial sex workers in the community. The 'population' here is synonymous with the 'target population' identified in Chapter 7.

It is essential, in any research study, to identify the population clearly and precisely. The success of the investigation will depend to a large extent on the identification of the population of interest. Often, the population of interest is not observable, and a smaller population is identified as the subject of investigation. For example, in clinical trials, some patients are excluded for various reasons prior to randomization, and the studied population is therefore somewhat different from the target population. This distinction should be clear at the beginning of the study, but also at the time of data analysis and interpretation, so that the inferences drawn from the study will be valid.

### *The variables* (V)

Once the population is identified, we should clearly define what characteristics of the units of this population (subjects of the study) we are planning to investigate. For example, in the case of the HIV study above, one needs to define HIV (reliable and valid method of identifying HIV in people), and what other characteristics of the people (e.g. age, sex, education, etc.) one intends to study. Clear and precise definitions and methods for measuring these characteristics (a simple observation, a laboratory measurement, or a battery of tests using a questionnaire) are essential for the success of the research study.

The variables are characterized in many ways; for statistical considerations, the variables are usually classified as discrete or continuous. Discrete variables are those in which only a small number of values is possible (e.g. sex: male, female), incidence of a disease (yes, no)). Continuous variables are those which, theoretically, can take any value within a specified range of minimum and maximum value (e.g. age, blood pressure). There are some variables that are discrete in nature, but the number of categories make them similar to continuous variables, and these are considered as continuous in most statistical calculations (e.g. number of years of schooling, number of people in a household).

### *The probability distribution* (P)

The most crucial link between the population and its characteristics, which allows us to draw inferences on the population based on sample observations, depends on this probability distribution. The probability distribution is a way to enumerate the different values the variable can have, and how frequently each value appears in the population. The actual frequency distribution is approximated to a theoretical curve that is used as the probability distribution.

Common examples of probability distributions are the binomial, Poisson and normal. Most statistical analyses in health research use one of these three common probability distributions. For example, the incidence of a relatively common illness may be

approximated by a binomial distribution, whereas the incidence of a rare condition (e.g. number of deaths from motor vehicle accidents) may be considered to have a Poisson distribution. Distributions of continuous variables (blood pressure, heart rate) are often considered to be normally distributed.

Probability distributions are characterized by 'parameters': quantities that allow us to calculate probabilities of various events concerning the variable, or that allow us to determine the value of probability for a particular value. For example, the binomial distribution has two parameters: n and π. The binomial distribution occurs when a fixed number (n) of subjects is observed, the characteristic is dichotomous in nature (only two possible values), and each subject has the same probability (π) of having one value and (1-π) of the other value. The statistical inference then involves finding out the value of π in the population, based on an observation of a carefully selected sample.

The normal distribution, on the other hand, is a mathematical curve represented by two quantities, μ and σ. The former represents the mean of the values of the variables, and the latter, the standard deviation. (Definitions in section 8.3.3.)

The type of statistical analysis done depends very much on the design of the study: in particular, whether the study was descriptive, and what sampling design was used to draw the sample from the population.

## Descriptive studies

In descriptive studies, the object is to estimate the values of the parameters of the probability distribution, or a function of these parameters. Based on what was observed in the sample, an estimate (best guess) of the values in the population is made, and a measure of the accuracy of this estimate is obtained. The measure of accuracy is based on what is known as the sampling distribution of the estimate.

### Accuracy of estimates

When a descriptive study is conducted and an estimate (E) of a parameter is obtained from the study, we need to know how this value, E would change if we took another sample. The distribution of values of E over different repetitions of sampling (under identical conditions to the ones we have already employed) is known as the sampling distribution of E. The sampling distribution can be empirically determined by actually repeating the process. Clearly, this is both difficult and unwarranted. It is possible to get an approximate idea of the sampling distribution, purely based on sampling theory. Once the sampling distribution is obtained, we can answer questions such as 'how close is my estimate likely to be to the true value of the parameter?' Obviously, we cannot get a 100% certain answer to this question, because we have only observed a sample. However, based on the sampling distribution, we can state with a certain amount of confidence (e.g. 95% sure) that it will be within ±x of the true value. This interval is known as the confidence interval. The greater the confidence in the statement, the larger is the value of x (wider interval). As we see below for specific examples, it is also known that the width of the interval for the same amount of confidence will decrease with an increase in sample size. Intuitively, the more information we have (large n) the more confident we are (smaller width of interval, or larger confidence for same interval).

## Estimation of parameters of the binomial distribution

When the study deals with a dichotomous event (such as incidence of a disease), the objective is to obtain an estimate for the probability of the event (incidence rate) occurring in the population. Based on the binomial probability distribution, it has been shown that the best estimate is the sample proportion, p (number of events in the sample/ sample size, n).
In order to assess how accurate this estimate is (how close is p to the true value, π), we need to know how much variability is expected in p in repeated samples using the same design (sampling distribution of p). It has been shown that, for p, the distribution is approximately normal, with mean p and standard deviation, and $s = \pi(1-\pi)/n$ (s is known as the standard error of p). Using the properties of the normal distribution, we can then say that the true value of π is within p ± 1.96s, with 95% confidence.

### Example 1

In a study to determine the prevalence of HIV infection among commercial sex workers (CSW), a sample of 150 CSW was tested, and 42 were found to be positive for HIV. The estimate for HIV prevalence was therefore 28%, with a standard error of 3.67%. The 95% confidence interval for HIV prevalence among CSW in this community is therefore 28±1.96*3.67 = (20.82%, 35.18%), i.e. based on this survey, we can state with 95% confidence that the true prevalence could be as low as 21%, or as high as 35%.

Notice that, in Chapter 7, we discussed many parameters or functions of parameters from binomial distributions when we discussed the incidence and prevalence and the risk ratios. The RR and OR that we get from cohort and case-control studies are the estimates of true risk ratios in the population from which the study samples were selected. To complete the picture, therefore, we need to calculate the sampling distributions of these estimates. In most cases, the sampling distributions are assumed to be approximately normally distributed (a statistically acceptable result if the sample size is large and the sampling is done using probability methods), so that we need only to calculate the standard error of these estimates to construct confidence intervals. Most computer programs that calculate relative risks or odds ratios will also report their standard errors, and in some cases the confidence intervals.

### Estimation of parameters for normal distribution

For a variable, X that has a normal distribution, we would need to know the mean μ and the standard deviation, σ. The best of these parameters is the sample mean $\bar{x}$ (arithmetic average of all the observations in the sample) and the sample standard deviation,

$$s = \sqrt{\sum_i (x_i - \bar{x})^2/(n-1)}$$

[The Normal distribution has the property that it is a symmetric probability distribution, the centre of the distribution is μ.
Also, the mean ± 1.96 (standard deviation) contains 95% of the values of the variable (i.e. the probability that the variable has values within this interval is 95%).]

Another reason that the normal distribution is commonly used in statistical inference is that most sample functions (sample mean, risk ratios, correlation coefficient, etc.) have the normal distribution as the sampling distribution, if the sample size is sufficiently large.

Most of the inferences in health research involve only inferences on the mean value. The sample mean has a normal distribution with mean m and standard deviation (standard error of the mean), s/√n. Thus, the 95% confidence interval for the population mean, m is therefore:

$$\bar{x} \pm 1.96 s/\sqrt{n}$$

Or, more simply, sample mean ± 2* (standard error of mean). For a more detailed description of common estimation problems and formulae for confidence intervals, see Kleinbaum, Kupper and Morgenstern, or Glantz.

## Analytical studies

In contrast to descriptive studies, analytical studies involve the testing of hypothesis in addition to description of the population. The study will have formulated research hypotheses, and on the basis of the observations in the research study, we need to draw conclusions as to the validity of these hypotheses. The inference is therefore a two-step process: estimate the parameters of the relevant probability distributions; test hypotheses (also known as testing of significance) involving these parameters.

### Statistical tests of hypotheses

A test of hypothesis has several steps:

#### Step 0. Identify the null hypothesis

This is a re-statement of the research hypothesis in the 'null' form, i.e. 'no effect of treatment', 'no difference in survival rates', 'no difference in prevalence rates', 'relative risk is one', etc. The null hypothesis is often stated with the research objectives. The null hypothesis should be 'testable', i.e. it should be possible to identify which parameters need to be estimated, and it should be possible to estimate the parameter, its standard error and the sampling distribution, given the study design.

#### Step 1. Determine the levels, α and β of errors acceptable in the inference

Since the inference is based on a sample of the population, one will never be absolutely sure if the hypothesis is true or not in the population. The decision is a dichotomous one: to accept the null hypothesis H, or to reject $H_0$. Two types of errors in inference are possible. The type I error (α) is the probability of falsely rejecting the null hypothesis, and the type II error (β) is the probability of falsely accepting the null hypothesis. These are summarized in the table below:

| | 'Truth' (in the population) | |
|---|---|---|
| Decision (based on sample results) | $H_0$ is true | $H_0$ is false |
| Accept $H_0$ | No error | Type II or β |
| Reject $H_0$ | Type I or α | No error |

Notice that the aim of the research study is to minimize both α and β; however, they work in opposite directions. If we decrease one, the other tends to increase. The researcher often designs the study to achieve a desired level for α, and minimize β for this situation. The statistical testing of hypothesis, therefore, is often done with a choice of α and the best statistical test available that will minimize β. The choice of α and β is made after determining the consequences of each of the errors, and is fixed at the time of design.

Step 2. Determine the best statistical test for the stated null hypothesis

This depends on the design, the type of variables, and the type of the probability distribution of the variable. For example, suppose that the null hypothesis is that the prevalence rates of a disease among two population groups are the same, and simple random samples have been obtained from the two population groups independently (design). The variable is the disease, which is a (discrete) dichotomous variable, and the sample size is fixed. Therefore, a binomial distribution is the probability distribution under consideration, and the prevalence rate is the
parameter of the distribution, which is estimated by the sample prevalence rates. These have approximately normal
distributions (sampling distribution). Therefore, a z-test or chisquare ($\chi^2$) test (see below) is the most appropriate.

Step 3. Perform the statistical test

This involves calculating the appropriate test statistic (the z or $\chi^2$) and comparing the computed value with its theoretical distribution. If the observed value is outside the limits in which the probability is $<\alpha$ for the sampling distribution, the null hypothesis is rejected.

Step 4. Calculate the power of the test

If the null hypothesis is not rejected, i.e. the computed value of the test statistic is within the limits for the α, then the statistical power of the test (1-β) should be computed for some acceptable minimum departure from the null hypothesis. If the power is too low, one would recommend that the study be repeated with a larger sample size. If the power is acceptable, one accepts the null hypothesis.

Sometimes, instead of deciding on 'acceptance' or 'rejection' of $H_0$, the test statistic is compared with the sampling distribution, and the value of α at which the test would reject the $H_0$ is calculated. This is called the P-value for the test.
In the above example, if the computed value of z were less than -1.96, or greater than 1.96, or equivalently, if the $\chi^2$ value were above 3.84, one would reject the null hypothesis, with α = 0.05.

It should also be noted that rejecting a null hypothesis does not necessarily mean that the effect or difference (departure from the null hypothesis) is 'clinically' significant. The differences may be trivial in terms of practical usefulness, and yet statistically significant, if the sample size is large. For example, an odds ratio of 1.1 can be statistically significant at 5% level of significance, if the sample size is very large (say 100 000), but one would not worry too much about an increase in relative risk by such a small amount. (Of course, it depends on the particular disease, and the smallest difference that makes a significant impact is often called the minimally acceptable difference, and is used in calculating the sample size when designing the study; see Chapter 5)
When we reject a null hypothesis, we usually accept an alternative hypothesis, $H_1$, which in most cases is the opposite of $H_0$. For example, if $H_0$ = the means of two populations are equal, then $H_1$ = the two means are not equal. This type of alternative hypothesis is called a two-sided alternative. When the mean of one population is too large or too small compared with the other, we reject the null hypothesis. There may be cases in which we are interested only in detecting whether the difference is on one side of the hypothesis (e.g. does the drug improve the survival rate?) In this case, the testing can be one-sided, and the $H_0$ rejected when the difference is too large and showing the benefit of the drug, but not if the difference is too large and showing that the drug is detrimental. Obviously, since we reject $H_0$ only half the time, the type I error is reduced; equivalently, for the same type I error, $H_0$ is rejected more often, increasing the power of the test. The decision to use a one-sided or two-sided test should be made in advance (before data collection), and should be based on solid scientific reasoning, lest the comparison be biased.

## Some common statistical tests of hypotheses

### Comparison of two proportions (z-test; $\chi^2$ test)

A common test of significance in epidemiological studies involves the comparison of two proportions. Examples include the comparison of incidence rates (in cohort studies) and the comparison of prevalence rates (in case-control or cross-sectional studies).
Comparison of proportions involves the testing of a null hypothesis of the form $H_0 : \pi_1 = \pi_2$, where $\pi_1$ and $\pi_2$ are the probabilities of an event in two independent populations. The common design involves a simple random sample of subjects, taken from the two populations independently, or using some form of matching (e.g. paired observations, such as matched case-control studies with exact matching on age). The event or characteristic, such as the incidence or prevalence of a disease, exposure to a risk factor, belonging to a particular race, etc., is either dichotomous, or is made dichotomous by grouping all the events not of interest into one group (e.g. in a multiracial country such as Canada, the interest may be to compare the white population with the rest). The probability distribution assumed is binomial.

The test of the hypothesis is based on the observed proportions, $p_1$ and $p_2$ in the two samples. If $H_0$ is true, one would expect $(p_1 - p_2)$ to be zero. The sampling distribution of $(p_1 - p_2)$ is approximately normal, with mean $(p_1 - p_2)$ and standard deviation (standard error of the difference) given by the formula:

$$\sqrt{[p_1(1-p_1)/n_1]+[p_2(1-p_2)/n_2]}$$

Therefore the test statistic,

$$z = (p_1 - p_2)/\sqrt{p_1(1-p_1)/n_1]+[p_2(1-p_2)/n_2]}$$

has a normal distribution, with mean 0 and standard deviation 1, if the $H_0$ is true. Under the null hypothesis, $\pi_1 = \pi_2 = \pi$; therefore, the standard error is:
$\sqrt{[p_1(1-p_1)(1/n + 1/n_2)}$ and is estimated by

$$\sqrt{p(1-p)(1/n_1+1/n_2)}$$

where $p = (n_1p_1 + n_2p_2)/(n_1 + n_2)$.

Equivalently, $\chi^2 = z^2$ has a chi-square distribution with one degree of freedom. The statistical test therefore, is to calculate z or $\chi^2$ and compare with the appropriate distribution. For example, if a = 0.05, the cut point for z is ± 1.96, and the cut point for $\chi^2$ is 3.84. Notice that $\chi^2$ can also be calculated in a simple way from the two-way table, as illustrated below:

OBSERVED FREQUENCIES

| | | Population 1 | Population 2 | Total |
|---|---|---|---|---|
| Event | Yes | $O_{11}$ | $O_{21}$ | a |
| | No | O12 = n1-O11 | O22 = n2-O21 | b |
| | Total | $n_1$ | $n_2$ | $n = n_1+n_2 = a+b$ |

If $H_0$ is true, based on the sample sizes, we have the following:

EXPECTED FREQUENCIES

| | | Population 1 | Population 2 | |
|---|---|---|---|---|
| Event | Yes | E11 = a*n1/n | E21 = a*n2/n | a |
| | No | E12 = n1-E11 = b*n1/n | E22 = n2-E21 = b*n2/n | b |
| | Total | $n_1$ | $n_2$ | n |

$c^2 = \sum(O-E)^2 / E$

Where the summation is over the four cells of the 2x2 table, O = observed frequency and E = expected frequency.

## Example 2

In a cohort study of low birth weight, 250 women of Chinese origin and 150 women of Indian origin were followed throughout their pregnancies for various risk factors for low birth weight (birth weight less than 2500 grams). Twelve Chinese women and 18 Indian women gave birth to infants weighing less than 2500 grams. The research question was whether the incidence of low birth weight was higher among the Indian women.

Variable: low birth weight (dichotomous: yes/no).

Parameter of the binomial distribution π = incidence rate.

Null hypothesis: $\pi_C = \pi_I$: type I error = 0.05.

Data:

| | Chinese | Indian |
|---|---|---|
| Low birth weight | 12 | 18 |
| Normal birth weight | 238 | 132 |

Estimates of incidence rates:

Chinese: $p_C = 12/250 = 4.8\%$

Indian: $p_I = 18/150 = 12\%$.

Test procedure:

(a) $z = (4.8 - 12)/ \sqrt{(4.8 * 95.2/250) + (12 * 88/150)}$

$= (-7.2/2.98) = -2.42$ cut point for $z = \pm 1.96$

Since the calculated z is less than -1.96, we reject the null hypothesis, and conclude that the incidence rates are different in the two populations. [The difference in the incidence rates is statistically significant (P<0.05).]

(b) Expected frequencies for the four cells of the above table are:

| | Chinese | Indian |
|---|---|---|
| Low birth weight | 18.75 | 11.25 |
| Normal birth weight 2 | 31.25 | 138.75 |

$\chi^2 = [(-6.75)^2/18.75 + (6.75)^2/11.25 + (6.75)^2/231.25 + (-6.75)^2/138.75)] = 7.01$, which is larger than the chi-square, with one degree of freedom cut point for a 5% tail area. Therefore, we reject the null hypothesis at the 5% level of significance, and conclude that the two incidence rates are different.

Comparison of incidence in cohort studies, or prevalence in casecontrol studies

A specific example of the comparison of incidence is a cohort study. In such a case, the index of comparison might be the relative risk, rather than the risk difference as above. The null hypothesis, $I_1 = I_2$ may be re-stated to the null hypothesis, RR = 1. We need to find the sampling distribution of the sample risk ratio, rr, in order to test this hypothesis. Since it is a ratio, the function, ln(rr) is assumed to have a normal distribution with mean zero. Based on this, a test of significance involves the computation of the standard error of ln(rr), using the test statistics, z = ln(rr) / s.e.(ln(rr)) as above. In practice, however, the test of significance is done on the hypothesis of equal incidence rates, and the chi-square test is appropriate. For further discussion, see Kleinbaum, Kupper and Morganstern.

Comparison of two proportions when the samples are matched

When the two samples are matched, especially in a one-to-one matching, the resulting observations are not statistically independent. Therefore, the standard error of the difference will involve a covariance term. Moreover, the difference may not have a normal distribution. Therefore, the z test for two independent samples is no longer valid. There are statistical tests that take into account the dependency between the samples. One test in particular, the McNemar's chi-square, is worth mentioning. Suppose the two samples are matched one-to-one, so that we have n pairs of observation (total 2n observations). McNemar's test involves separating these n pairs into concordant (both members of the pair have the event, or neither has) and discordant (one member of the pair has the event, the other does not). Characterizing the event as + and the non-event as -, the four categories of observations and their frequencies are summarized below:

+/+ a; +/- b; -/+ c; -/- d

The two concordant groups, +/+ and -/-, are discarded, as they do not provide information on the null hypothesis of equal probability in the two populations. If the null hypothesis were true, one would expect the discordant pairs +/- and -/+ to have equal frequencies, so the expected numbers in these groups are (b+c)/2.
A chi-square based on these sets of two observed frequencies (b,c) and two expected frequencies [(b+c)/2, (b+c)/2] follows a chisquared distribution with one degree of freedom, if the null hypothesis is true.
The McNemar's $\chi^2 = (r-c)^2 / (r+c)$, and should be compared with a chi-square distribution with one degree of freedom for the test of significance of the null hypothesis. If the numbers of pairs are small, a continuity correction is often applied to this formula:

$$\chi^2 = (r-c-1)^2 / (r+c)$$

Example 3

In a case-control study of nasopharyngeal carcinoma (NPC), 200 cases of NPC were matched to 200 control subjects (patients from the same hospital admitted with other conditions, matched for age, sex and race of the patient). One of the risk factors considered in the study was exposure to Epstein-Barr virus (EBV). The following table summarizes the results for the 200 pairs of subjects in the study, with respect to this risk factor:

| No. of pairs | EBV exposure | |
|---|---|---|
| | Among cases | Among controls |
| 45 | + | + |
| 28 | - | + |
| 56 | + | - |
| 71 | - | - |

The null hypothesis is that there is no association between the exposure and the disease, which translates, as the frequencies of the two discordant pairs are equal.

Discarding the 'tied' pairs (+,+), (-,-), the discrepant pairs have frequencies of 28 and 56 respectively. The McNemar's chi-square is therefore $(28-56)^2 / (28+56) = 10.6$. If we choose the type I error ($\alpha$) = 0.05, the cut point for $\chi^2$ is 3.84, and we reject the null hypothesis of no association between the exposure (EBV) and the disease (NPC).

Comparison of two proportions when the sample size is small

All the above tests are based on normal approximation of the test statistics, which depend on the sample size being large. The requirement is that np is more than 5 (the expected frequency in each cell in the contingency table is above 5). When the sample size is too small to have this requirement, the normal approximation may be incorrect. Sometimes a continuity correction is applied to the chisquare, although this is not widely accepted. A test that does not use the normal approximation, the Fisher exact test, is used in such situations. See Glantz for further details.

Comparison of two means (independent samples)

When the variable of interest is a continuous one, the relevant probability distribution is the normal distribution. In such cases, the null hypothesis often takes the form, $H_0$: $m_1 = m_2$, where $m_1$ and $m_2$ are the means of the variable in the two populations, respectively. The test of the hypothesis follows the same steps as in the case of testing the difference in proportions, except that the parameter of interest is the difference in means, $\bar{x}$. The best estimate of the population mean, m, is the sample mean. Therefore, to test the null hypothesis, we compute the standardized difference in means.
In the case of the two samples being obtained independently (for example, in a clinical trial where the patients have been randomly allocated to two groups, or in an unmatched case-control study), this value has a normal distribution, with mean 0 and standard deviation 1, if the null hypothesis is true. Here, we are assuming that the standard deviations in the two populations, $\sigma_1$, $\sigma_2$ are known. In practice, however, we seldom know these quantities and they have to be estimated by their respective sample standard deviations. Commonly, it is assumed that the two populations have the same standard deviations, and a pooled estimate of the common standard deviation, s, is used in the calculations:

$$s = \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)}$$

Then the standardized difference,

$$t = (\bar{x}_1 - \bar{x}_2)/s\sqrt{(1/n_1 + 1/n_2)}$$

has a student's t-distribution with $(n_1+n_2-2)$ degrees of freedom, if the null hypothesis is true. Therefore, the test would be to compare the computed value of t with table values for the appropriate t-distribution for the chosen α.
Example 4

From a study of the incidence of low birth weight (birth weights of 2500 grams or less) among various ethnic groups in Malaysia, the average birth weights, along with the standard deviations, are given below:

| Ethnic group | Mean | N | Std. deviation |
|---|---|---|---|
| Malay | 2816.71 | 458 | 597.52 |

| | | | |
|---|---|---|---|
| Chinese | 2692.05 | 156 | 577.95 |
| Indian | 2914.26 | 135 | 538.52 |
| Other | 2776.99 | 136 | 548.69 |
| Total | 2803.51 | 885 | 580.81 |

We want to test the null hypothesis that the average birth weight for Malay children is the same as that for Indian children. The test statistic is computed as below. The pooled standard deviation is:

$$s = \sqrt{[(457*597.52^2) + (134*538.52^2)]/134 +457)}$$

$$= 584.66 \quad t = (2816.71-2914.26)/\{584.66[(1/458)+(1/135)]\}$$

$$= -17.40$$

This should have a t-distribution with 591 degrees of freedom, if the null hypothesis is true. The t-distribution is approximately the same as the normal distribution when the sample size is large (more than 50). Thus, the cut point for a 5% level of significance would be ± 1.96, and since the calculated t is outside these limits, we would reject the null hypothesis and conclude that the two groups have different average birth weights.
[Note: Comparison of more than two groups would require more advanced statistical tests such as the Analysis of Variance and F-tests, which are beyond the scope of this manual. Refer to other statistical texts, such as Glantz, for details.**]**

**Comparison of two means (paired samples)**

As in the case of the McNemar's test, when the two samples are not independent (usually paired due to matching), a similar t-test can be computed. The procedure involves the computation of differences in the outcome variable between the two members of the pair, and calculating the mean and the standard error of these differences. The ratio, t = (mean difference / standard error of difference), then follows a t-distribution with (n-1) d.f., where n is the number of pairs, when the null hypothesis is

*References and further reading*

Fleiss J.L. Statistical methods for rates and proportions. New York, John Wiley and Sons, 1981.

Glantz S.A. Primer of biostatistics, 4 ed. Singapore, McGraw Hill, 1997.

Kahn H.A., Sempos C.T. Statistical methods in epidemiology. Oxford, Oxford University Press, 1989.

Kelsey J.L., Thompson W.D., Evans A.S. Methods in observational epidemiology. Oxford, Oxford University Press, 1986.

Kleinbaum D.G., Kupper L.L., Morganstern H. Epidemiologic research: principles and quantitative methods. New York, Van Nostrand Reinhold, 1982.

Lillienfeld A.M., Lillienfeld D.E. Foundations of epidemiology, 2 ed. Oxford, Oxford University Press, 1980.

Schlesselman J.J. Case control studies. Oxford, Oxford University Press, 1982.