*Clinical Research Methods*

# Hypothesis testing (Part II): Inference from means

ABHAYA INDRAYAN, PIYUSH GUPTA

## INTRODUCTION

The focus of this article is on quantitative data that are generally summarized in terms of mean. Mean is a statistical tool that depends on the pattern of distribution of values in the target population. Thus, forms such as Gaussian distribution are especially important to draw inference from sample means.

The mean of any variable differs from sample-to-sample due to inter-individual variability. This has been discussed previously.[1] If the mean level of free thyroxine ($T_4$) is 0.62 ng/dl in a sample of 18 children with thalassaemia major, can it be concluded that it is lower than the minimal normal 0.7 ng/dl? How can one be confi-dent that another sample of such children will not give a mean higher than the lower limit of normal? If the mean $T_4$ level in a sample of thalassaemic boys is 0.65 ng/dl and in a sample of girls 0.56 ng/dl, can it be concluded that the $T_4$ level is affected by the gender of the children? If these children are divided into different groups by growth pattern (normal, slightly retarded, moderately retarded and severely retarded) and a difference of thyroid function parameters is observed, can it be confidently stated that this difference would persist in repeated samples? Or, how can we conclude that this difference is genuinely present in such subjects in the target population and is not a chance occurrence in the sample? A host of questions arise pertaining to the uncertainties inherent in means of samples. Most of these can be satisfactorily answered by the application of appropriate statistical methods discussed in this article.

## COMPARISON OF MEANS IN ONE AND TWO GROUPS UNDER GAUSSIAN CONDITIONS: STUDENT'S *t*-TEST

### *Comparison with a pre-specified mean*

***Example 1.* Suppose the interest is in finding out whether a random sample of 10 patients with chronic diarrhoea have the same average haemoglobin (Hb) level (say, 13.8 g/dl with SD = 1.672) as normally seen in healthy people (say 14.6 g/dl) in the area. In this example, the sample mean is lower than the normal. This could occur if the sample constitutes people with a lower Hb level. How can you conclude with reasonable confidence on the basis of this sample that patients of chronic diarrhoea have a lower Hb level than the average?**

There is only one sample in this example and the comparison is with a known average in healthy people. It is a one-sample problem though the comparison is of two means, one found in the sample and the other known for the healthy population.

The answer to this problem naturally depends on the magnitude of the difference between the sample mean and the known mean in healthy people. In the example, this difference is 13.8–14.6= –0.8 g/dl. The larger the difference, the greater is the chance that the patients actually have a lower Hb level. This magnitude of difference is assessed relative to the expected variation in means from sample-to-sample. The latter is measured by the standard error (SE) of mean, which is $\sigma/\sqrt{n}$. The standard deviation (SD) $\sigma$ would be rarely known and is replaced by its estimate *s*. This replacement changes Gaussian distribution to Student's *t*. Thus the criterion for this set-up is

$$\text{Student's } t\text{-test: } t_{n-1} = \frac{\bar{x}-\mu_o}{s/\sqrt{n}}$$

where $\mu_o$ is the value of the mean under the null hypothesis $H_o$. As explained in our first article,[1] *P*-value is the probability of *t*-value as much as in the sample or more extreme in favour of $H_1$. The exact *P*-value for a particular value of *t* is provided by most standard statistical packages. Alternatively, standard probability tables can be consulted to check whether or not *P* is less than the threshold such as 0.05 or 0.01. The distribution of *t* depends on the degrees of freedom (df). For *t* in the above equation, df=*n*–1.

**For the above example,**

$$t_9 = \frac{13.8-14.6}{1.6720/\sqrt{10}} = -1.51$$

**A statistical package gives $P(t<-1.51)=0.0827$. This is more than 0.05. Thus, the chance is more than 5% that the null hypothesis, i.e. $H_o = 14.6$ g/dl, is true. The difference between the sample mean 13.8 g/dl and the population mean 14.6 g/dl is not statistically significant. The sample mean 13.8 g/dl could have arisen due to sampling fluctuation when the sample is from a population with a mean of 14.6 g/dl. Thus, the plausibility of the population mean being 14.6 g/dl is not adequately ruled out. The sample fails to provide sufficient evidence against this null hypothesis.**

### *Comparison of means of two samples*

Consider a situation where two samples are available. These could be from two groups such as men and women, suffering from disease A and disease B, of age 20–39 years and age >40 years, or from one group before and another after treatment. The latter is called a paired sample set-up. The erythrocytic sedimentation rate measured twice by two different methods in the same group of subjects also exemplifies pairing. Pairing also occurs when the subjects in the two groups are one-to-one matched such as in some case–control studies. The procedure to calculate the SE of difference in case of paired samples is different from that in unpaired

University College of Medical Sciences, Dilshad Garden, Delhi 110095
ABHAYA INDRAYAN Department of Biostatistics and Medical Informatics
PIYUSH GUPTA Department of Paediatrics

Correspondence to ABHAYA INDRAYAN

samples. Following the same argument as in the case of a one-sample set-up, the general form of the criterion in the two-sample set-up is

$$\text{Student's } t \text{ (two-sample)} = \frac{\text{mean difference}}{\text{estimated SE of difference}}$$

We explain the procedure below with the help of examples.

*Paired sample set-up*

*Example 2:* Given below are the serum albumin levels (g/dl) in six randomly chosen patients of dengue haemorrhagic fever before and after treatment. Can it be concluded that the mean albumin level after treatment is different from the mean before treatment?

| | | | | | | |
|---|---|---|---|---|---|---|
| Before treatment | 4.8 | 4.1 | 5.3 | 3.9 | 4.5 | 3.8 |
| After treatment | 5.2 | 4.9 | 5.2 | 4.8 | 4.6 | 4.4 |

In this case,
differences, $d_i$: 0.4 0.8 −0.1 0.9 0.1 0.6
mean difference, $\bar{d}=0.45$, and SD of differences, $s_d=0.3937$

Also SE of difference, $s_d/\sqrt{n}=0.1607$. We need to find the chance of getting a mean difference of 0.45 or larger when the actual difference in the target population is zero. If this chance is exceedingly small, say less than 0.05, it can be concluded that the mean difference is not zero. Now,

$$t_5 = \frac{0.45}{0.1607} = 2.80$$

Since there is no assertion in this case that the albumin level after treatment will increase or decrease, the alternative hypothesis is $H_1: \mu_1 \neq \mu_2$. For this $H_1$, two-tailed probability $P(|t|>2.80)$ is needed. For $n-1=5$ df, this is $P=0.019$ from a statistical software. Since $P$ value is less than 0.05, the null hypothesis is rejected. It can be deduced with reasonable assurance that the mean albumin level after treatment is different from the mean before treatment.

*Unpaired samples set-up*

*Example 3:* Suppose the idea of measuring serum albumin level surfaced later and the levels before treatment were not available in those very patients where post-treatment levels could be obtained. However, albumin levels before therapy were obtained later in another sample of six newly admitted patients. Consider the same data as given in *Example 2* but now the observations belong to 12 different patients in place of pairs of observations on 6 patients. Now,
mean albumin level in the before-treatment group,
»$x_1=4.40$, SD $s_1=0.5797$,
mean albumin level in the after-treatment group,
»$x_2=4.85$, SD $s_2=0.3209$

$$\text{df}=n_1+n_2-2=6+6-2=10.$$

This gives: $s^2_p=0.2195$ and $t_{10}=-1.66$, where $s^2_p$ is the pooled variance.

The probability of getting these samples or more extreme in favour of $H_1$ when $H_0$ is true, is given by $P(|t|>1.66)$. From a statistical package, $P$-value is 0.1272, which is large. Thus, the null hypothesis of equality of means is plausible and cannot be rejected. The evidence is not strong enough to conclude that the mean albumin level after treatment is any different from the mean before treatment. Note that the same values in a paired set-up give different results from that in an unpaired set-up.

*Cross-over design*

The cross-over design economizes on the subjects because the same subject is used for trial twice. Comparison is within subjects and therefore more precise. We illustrate a simple method to analyse data from cross-over experiments.

*Example 4:* Consider a trial of $n=16$ asthma patients who were randomly divided into two equal groups of 8 each. The first group received treatment A (say, formoterol) then treatment B (say, salbutamol), while the second group received treatment B and then treatment A. We abbreviate them as trA and trB. An adequate wash-out period was provided before switching treatment so that there was no carry-over effect. The response variable is forced expiratory volume in one second ($FEV_1$). The data obtained are given below.

*Group I: AB sequence*

| Subject no. | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| $FEV_1$ (L/min) | | | | | | | | |
| Period 1 (trA) | 1.28 | 1.26 | 1.60 | 1.45 | 1.32 | 1.20 | 1.18 | 1.31 |
| Period 2 (trB) | 1.25 | 1.27 | 1.47 | 1.38 | 1.31 | 1.18 | 1.20 | 1.27 |

*Group II: BA sequence*

| Subject no. | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
|---|---|---|---|---|---|---|---|---|
| $FEV_1$ (L/min) | | | | | | | | |
| Period 1 (trA) | 1.27 | 1.49 | 1.05 | 1.38 | 1.43 | 1.31 | 1.25 | 1.20 |
| Period 2 (trB) | 1.30 | 1.57 | 1.17 | 1.36 | 1.49 | 1.38 | 1.45 | 1.20 |

*Step 1: Test for group effect.* In this example, the groups identify the sequence and the group effect is the same as the sequence effect. If the sequence is not affecting the values, the mean difference between trA and trB should be the same in the two groups. Calculate the mean (trA−trB) and their SD in the two groups separately and then compare these for equality by the usual two sample $t$-test as described for unpaired samples. In this example, $t_{14}=-1.13$. This is not statistically significant ($P>0.05$). If the sequence effect is present, its reasons should be ascertained and the trial done again after eliminating those causes. This is not the case in this example.

*Step 2: Test for carry-over effect.* If a positive carry-over effect is present, the values for Period 2 should be consistently higher than for Period 1 in both the groups. To test for its presence, calculate the period differences for this example.

Period differences in Group I (Period 1−Period 2)
+0.03 −0.01 +0.13 +0.07 +0.01 +0.02 −0.02 +0.04
Period differences in Group II (Period 1−Period 2)
−0.03 −0.08 −0.12 +0.02 −0.06 −0.07 −0.20 0.00

If no carry-over effect is present, the mean of these 16 differences should be close to zero. The test of the null hypothesis of no carry-over effect is done by considering the differences in the two groups together as one sample. In this example, the Student's $t$ for paired-sample is $t_{15}=-0.86$. One-tail $P$-value for $t=-0.86$ at 15 df (from probability tables), is greater than 0.05. Thus, there is no definite evidence to conclude that a carry-over effect is present.

***Step 3: Test for treatment effect.*** **Cross-over design is not a good strategy when carry-over effect is present. If group and carry-over effects are not present, the two groups can be considered together as one. Then the paired *t*-test can be applied to the joint sample. In our example, $t_{15}=3.35$. From the probability table for 15 df, this gives $P<0.01$. Thus, the treatment difference is statistically highly significant.**

Data from cross-over trials can be analysed more meticulously by using the analysis of variance (ANOVA) method.[2]

## COMPARISON OF MEANS IN THREE OR MORE GROUPS UNDER GAUSSIAN CONDITIONS: ANOVA *F*-TEST

Consider a trial on a new once-a-day hypertensive drug with three different dosages and a control. The objective is to find whether different dosages have a differential effect on diastolic blood pressure. There are four groups in this trial and the comparison index is the mean reduction in diastolic pressure.

The generic method used for comparing means in three or more groups is called analysis of variance (ANOVA). The name comes from the fact that the total variance in all the groups combined is broken down into components such as within-groups variance and between-groups variance. Between-groups variance is the systematic variation occurring due to group differentials. The residual left after this extraction is considered a random component arising due to intrinsic biological variability between individuals, which is the within-groups variance. If genuine group differentials are present, then the between-groups variance should be large relative to the within-groups variance. Thus, the ratio of these two components of variance can be used as a criterion to find whether or not the group means are different. Between-groups variance is kept in the numerator and within-groups in the denominator. The test criterion now used is called *F*-test (after the name of the statistician Fisher) or variance ratio.

### *One-way ANOVA*

Consider a study in which plasma amino acid (PAA) ratio for lysine is calculated in healthy children and in children with Grades I, II and III malnutrition. This ratio is the difference in PAA concentration in blood before and after the meal, expressed as percentage of the amino acid requirement. There are four groups in this study. The set-up is called one-way since no further classification of subjects, say by age or gender, is sought in this case. Groups define a factor, in this case grade of malnutrition. The 'response' is a quantitative variable. PAA is a ratio but can still be considered to follow a Gaussian pattern within each group. When other factors are properly controlled, the difference in PAA ratio among subjects would either be due to the degree of malnutrition or due to intrinsic inter-individual variation in the subjects of different groups (Fig. 1). The former is the between-groups variation and the latter is the within-groups variation.

Note that part of the within-groups variation can be due to factors such as heredity, age, gender, height and weight of the children but these are assumed as being under control and disregarded in this set-up.

If group differences are not really present, the between-groups variance and the within-groups variance would both arise due to intrinsic variation alone and both will be nearly equal. Since the between-groups variance is in the numerator, a value of *F* substantially more than unity implies that between-groups variation is

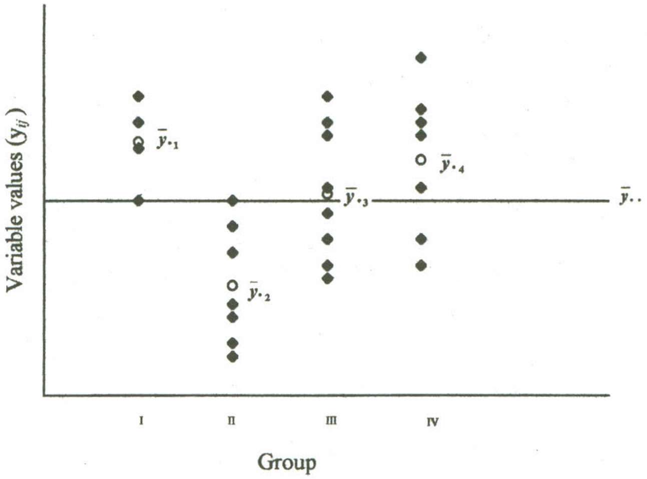

FIG 1. A graphical display of within-groups and between-groups variance (within groups is the difference of individual points with the respective group-mean, and between-groups is the difference of group-means with the overall mean)

- The *t*-test is based on the magnitude of difference and its variation. If the interest is in the proportion of subjects showing a rise, and not in the magnitude of rise, then use the methods described for proportion.[3]
- The same difference may be statistically significant in a paired set-up but not significant in the unpaired set-up. This can occur if the difference in the paired set-up is fairly consistent, with each patient serving as his own control. In the unpaired set-up, the inter-individual variation can be large. A paired set-up with 'before' and 'after' measurements is usually desirable but in case it is not feasible, close matching of controls with cases simulates pairing.
- Student's *t* is valid only when *means* follow a Gaussian pattern. When *n* is large, the pattern is nearly always Gaussian due to the central limit theorem. When *n* is small (say <30), the *t*-test is valid only if the underlying distribution is Gaussian. If the underlying distribution is far from Gaussian and *n* is small, then non-parametric (e.g. Wilcoxon) tests are used. Some of these are discussed in this article.
- The samples have to be random. Statistical inference is not valid for non-random samples.
- The Student's *t*-test and other means-based tests can also be carried out on geometric means after taking logarithm of the values. However, the conclusions will be applicable to log-values and not to the values themselves. Geometric means are used for multiplicative variables such as antibody titre.
- The *t*-test is for averages and not for individual values. The conclusion is valid only for the groups. Individuals can behave in a very unpredictable manner, though they are likely to follow a trend.

large relative to the within-groups variation. This is an indication that the groups are indeed different with respect to the mean of the variable under study. The criterion *F* is calculated assuming the null hypothesis of equality of means in different groups is true. The larger the value of *F*, smaller becomes the probability that $H_o$ is true. Standard statistical packages provide corresponding *P*-value for calculated values of *F* and df, so that a decision can be made instantly. Cut-off values of *F* for levels such as $\alpha$=0.05 and different dfs can also be obtained from standard probability tables.

If *n* is small, the ANOVA procedure is valid only when the pattern of distribution of the variable is Gaussian. If the pattern is far removed from Gaussian, use non-parametric tests.

### *Two-way ANOVA*

Consider a clinical trial in which three doses (including a placebo) of a drug are given to a group of anaemic boys and girls to assess the rise in haematocrit (Hct) level. It is suspected that the effective dose may be different for boys and girls. This differential response is called interaction. In this example, interaction is likely between the drug dose and gender.

*Two-factor design.* The objective of the trial in the above example is to find the effect of dose, gender and their interaction on the response (such as percentage rise in Hct level) in anaemic children following administration of iron tablets. This is called a two-way ANOVA situation as there are two factors involved, i.e. the dose and gender. Note that there are three dose-groups of boys in this trial and another three dose-groups of girls. The researcher may like to have *n*=10 subjects in each of these six groups, making a total of 60 subjects. To minimize the role of other factors causing variation, these 60 subjects should be as homogeneous as possible with respect to all those characteristics (for example, body weight, type and amount of diet consumed, etc.) that might influence the response. Once the eligible subjects are identified, 30 boys and 30 girls need to be randomly allocated to the three dose levels. Such allocation increases the confidence of asserting that any difference now occurring is mostly, if not exclusively, due to the factors under study, namely dose of the drug and gender of the subjects in this example.

As in the case of a one-way ANOVA, within- and between-group variances are obtained. Criterion *F* is calculated separately for each of the two factors and for their interaction. The *P*-value is obtained as usual corresponding to the calculated value of *F*. A separate decision for factor 1, factor 2 and the interaction is made regarding their statistical significance. When the interaction is not significant, the factors are called additive.

*Repeated measures.* In many medical situations, as in the case of administering an anaesthetic agent, it is necessary to monitor a subject by repeatedly observing vital signs such as heart rate and blood pressure at specified intervals. In this case, each subject can be regarded as level of a factor and a two-way ANOVA can be done. The second factor will be the group, for example, patients receiving two or three different anaesthetic agents. For more complex designs, the method of analysis changes.

Special caution is required in the case of repeated measures. ANOVA for repeated measures requires not only uniform variance but also uniform co-variance (or correlation) between each pair of repeated measures. If *n* is large, small differences are negligible but the differences could be unduly large in some cases. Those measurements close in time may be highly correlated as compared to those widely separated in time. The correlation between heart rate (HR) after 1 minute and 5 minutes after anaesthesia would be higher and the HR after 1 minute and 30 minutes poorly correlated. If the correlations are really unequal, adjustments may be needed, especially if *n* is small.

- A problem in the comparison of three or more groups by criterion *F* is that its significance indicates only that a difference exists. It does not tell exactly which group or groups are different. Further analysis, called multiple comparison, is required to identify the groups that have a different mean. This is discussed below.
- ANOVA is a procedure based on means. Any means-based procedure is severely disturbed when outliers are present. Prior to using ANOVA, ensure that there are no outliers in your data. If there are any, examine if they can be excluded without affecting the conclusion.
- Random sampling is required, as always, for validity of conclusions from ANOVA *F*-test.
- Important assumptions for validity of ANOVA are (i) Gaussian pattern, (ii) homoscedasticity, and (iii) independence.

  The assumption of Gaussian pattern is not a strong requirement. ANOVA *F*-test is quite robust to minor departures from *Gaussian pattern*. When the pattern is really far from Gaussian, particularly when *n* is small, it is advisable to use non-parametric methods.

  *F*-test also requires that the variance in different groups is nearly the same. This property is called *homoscedasticity* and can be checked by the Hartley's $F_{max}$ test.[4] If the pattern is not Gaussian, Hartley's test cannot be used. Transformation of the data, such as logarithm (ln *y*), square ($y^2$) and square root ($\sqrt{y}$) are tried in such a case.

  The assumption of *independence* of observations is the most serious requirement for validity of ANOVA *F*. This is violated particularly in cases where serial observations are taken and the value of an observation depends on what it was at the preceding time. A different set of methods, called *hierarchical* or *repeated measures*, are generally applied for analysis of such data.
- Extension of ANOVA to three or more factors is straight, with similar main effects and interactions as in two-way ANOVA. However, there would now be several two-factor interactions, three-factor interactions, etc. Their details are beyond the scope of this article. Interested readers may consult Lindman.[5]
- It is desirable that higher-order interactions are tested before lower-order ones, since it is difficult to attach a meaning to the lower-order interactions when higher-order interactions are present.

## MULTIPLE COMPARISONS: BONFERRONI AND TUKEY TESTS

Once the overall significance is indicated by the $F$-test, the next step is to identify the groups that are different from one or more of the others. This requires pair-wise comparisons. If there are four groups, the comparisons are between group 1 and group 2, group 1 and group 3, group 1 and group 4, group 2 and group 3, group 2 and group 4, and group 3 and group 4. These are a total of six comparisons called multiple comparisons. Means in two groups are generally compared by Student's $t$-test. However, repeated application of this test at, say, 5% level of significance on the same data, increases the total probability of Type I error to an unacceptable level. If there are 15 tests, each done at 5% level, then the overall (experiment-wise) Type I error could be as high as $1-(1-0.05)^{15}= 0.54$. Compare this with the desired 0.05. To keep the probability of Type I error within a specified limit such as 0.05, many procedures for multiple comparisons are available. Each of these is generally known by the name of the scientist who first proposed that test. Among them are Bonferroni, Tukey, Scheffe, Newman–Keul, Duncan and Dunnett.[6] The last is used specifically when each group is to be compared with the control only.

The Bonferroni and Tukey procedures are commonly used in medical and health literature and, in our opinion, are also the most suitable ones. These procedures test for differences in multiple groups and at the same time, ensure that the probability of Type I error does not exceed the desired level $\alpha$.

### *Bonferroni procedure*

In this procedure, each comparison is done by using Student's $t$-test but a difference is considered significant only if the corresponding $P$-value is less than $\alpha/H$ where $H$ is the number of comparisons. If there are four groups and all pair-wise comparisons are required, then $H$=6. A difference would be considered significant at 5% level if $P<0.05/6$, i.e. if $P<0.0083$. This procedure is efficient when the number of comparisons, i.e. $H$ is small.

### *Tukey test*

This is best suited when the interest is in all pair-wise comparisons. The procedure works in a slightly different manner. We have not provided the details.

Many statistical packages would perform Tukey test or other multiple comparison procedures at the specified level of significance. They will also indicate which groups, if any, are significantly different from others. Tukey test for multiple comparisons may sometimes give results at variance with the results of the $F$-test. It is possible that the $F$-test is significant but none of the pair-wise comparisons is significant. Conversely, the $F$-test may not show significance but comparison for a specific pair may still be signi-ficant. This happens because both require a Gaussian pattern but the underlying distribution may not be exactly Gaussian. They behave differently for a departure from this pattern. The problem may arise more frequently for a small $n$ than for a large $n$ because a large $n$ is an insulation against violation of the Gaussian pattern in most cases.

## NON-PARAMETRIC TESTS FOR NON-GAUSSIAN CONDITIONS

We continue with the set-up where the response is quantitative, practically continuous, and the interest is in comparing two or more groups. If $n$ is small and the underlying distribution of the response variable does not follow a Gaussian pattern, non-parametric methods, also called distribution-free methods, are needed for comparison. This can happen when we study a quantitative variable in a restricted class of subjects, for example, studying duration of labour at the time of childbirth, blood glucose level in diabetics, haemoglobin level in anaemics, etc. The distribution pattern here can be highly skewed.

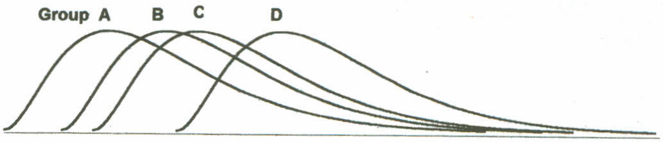

FIG 2. Distribution in four groups different in location only

The term non-parametric method implies a method that is not for any specific parameter. Student's $t$-test, for example, is a parametric method because it is concerned with a parameter, namely, the mean. In the case of non-parametric methods, the hypothesis is concerned with the pattern of the distribution as in the goodness-of-fit test, or with some characteristic of the distribution of the variable such as randomness and trend. More commonly, though, the interest would be in location of the distribution without specifying the parameter. This is illustrated in Fig. 2, where the distributions are identical but are different with respect to location only. Location shift is clear and there is no need to mention mean or median or any such parameter in this case.

When Gaussian conditions are present, performance of non-parametric tests is not as good as Student's $t$-test or ANOVA $F$-test.

### *Comparison of two groups: Wilcoxon tests*

As in the case of Student's $t$-test, the comparison of two groups can be done in two types of situations—paired and unpaired. Non-parametric methods for these situations are different.

#### SITUATION 1—PAIRED DATA

*Sign test:* Sign test is one option in case of paired data. The only information it utilizes is the direction of difference, whether negative or positive, within pairs. Under the null hypothesis that there is no difference, the negative sign is as likely as the positive sign. Thus, $H_0$: $\pi=½$, where $\pi$ is the probability of, say, the positive sign. The test can be carried out by calculating the binomial probability under this $H_0$.

Sign test is considered inadeqaute because it ignores the magnitude of difference. However, there are situations where only sign is important and the magnitude can indeed be ignored. This can happen, for example, in a behavioural problem where a judgement on 'greater than' or 'less than' between pairs of performances can be easily made but not about the magnitude of difference. In an iron supplementation programme for antenatal women, the interest may lie in responders who show an increase in the Hb level in excess of say, 0.5 g/dl and not in the actual amount of increase. However, such considerations convert the quantitative data to qualities, and thus compromise the power of the procedure to detect a difference.

*Wilcoxon signed-rank test for matched pairs.* If the magnitude as well as the direction of the differences is important, the Wilcoxon signed-rank test is a more powerful test. This test gives more weight to a pair that shows a large difference relative to a pair that shows a small difference. The test criterion $Ws$ is based on the ranked value of the differences after excluding ties.[7] The test is not applicable when $n<5$. Gaussian approximation can be used when $n\geq 20$.

SITUATION 2—INDEPENDENT SAMPLES

The two independent samples counterpart of the Wilcoxon test is called the Wilcoxon rank-sum test.[7] The rank-sum test for two independent samples will not give any statistical significance at 5% level if either $n_1$ or $n_2$ is less than 4. When any of these sample sizes is 10 or more, the Gaussian approximation can be invoked.

### *Comparison of three or more groups—Kruskal–Wallis test (non-Gaussian distribution, small n)*

Consider an example of cholesterol level in isolated diastolic hypertensives, isolated systolic hypertensives, 'clear' hypertensives and control. All subjects are adult women of medium build and the groups are matched for age. They all belong to the same socio-ethnic group. Thus, the factors that may affect cholesterol level are controlled to a large extent. It is expected that the pattern of distribution of cholesterol level in different hypertension groups would be the same but not Gaussian. One or more group may have measurements higher or lower than the others. The objective is to find whether or not the differences between groups are statistically significant. Since the underlying distributions are not Gaussian and if, in addition, the number of subjects in different groups is also small, the conventional ANOVA cannot be used.

The non-parametric Kruskal–Wallis test is the right method for such a set-up. Interested readers may consult Hollander and Wolfe[7] for the exact methodology. The non-parametric method for two-way tables is called Friedman's test. This is also available in Hollander and Wolfe.[7]

Various tests used for comparing quantitative data under different conditions are summarized in Table I. This will help in selecting a proper test for the type of data and the problem in hand.

TABLE I. Statistical procedures for test of hypothesis on means or locations

| Set-up | Conditions | Main criterion |
|---|---|---|
| Comparison of two groups | *Paired* | |
| | —Gaussian | Student's $t$ |
| | —non-Gaussian | |
| | $5\leq n\leq 19$ | Wilcoxon signed-rank $W_s$ |
| | $20\leq n\leq 29$ | $W_s$ referred to Gaussian Z |
| | $n\geq 30$ | Student's $t$ |
| | *Unpaired* | |
| | —Gaussian | Student's $t$ |
| | | Pooled variance when $\sigma_1^2\approx\sigma_2^2$ |
| | | Separate variance when $\sigma_1^2\neq\sigma_2^2$ |
| | —non-Gaussian | |
| | $n_1$, $n_2$ between (4,9) | Wilcoxon rank-sum $W_R$ |
| | $n_1$, $n_2$ between (10,29) | $W_R$ referred to Gaussian Z |
| | $n_1$, $n_2\geq 30$ | Student's $t$ |
| | *Cross-over design* | |
| | —Gaussian | Student's $t$ (paired) |
| | —non-Gaussian | Not discussed |
| Comparison of three or more groups | *One-way layout* | |
| | —Gaussian | ANOVA $F$ |
| | —non-Gaussian | |
| | $n\leq 5$ | Kruskal–Wallis $H$ |
| | $n\geq 6$ | $H$ referred to Chi-square |
| | *Two-way layout* | |
| | —Gaussian | ANOVA $F$ |
| | —non-Gaussian | |
| | $J\leq 3$, $K\leq 13$ | Friedman $S$ |
| | Larger $J$, $K$ | $S$ referred to Chi-square |
| Multiple comparisons | —Gaussian | Tukey $Q$ |
| | —non-Gaussian | Not discussed |

## TESTING FOR THE PRESENCE OF MEDICALLY IMPORTANT DIFFERENCE IN MEANS

The null hypotheses discussed so far are for no difference. When this $H_0$ is rejected, the only conclusion reached is that a difference is present. No inference can be deduced from the magnitude of difference. The difference could be so small that it has no clinical implication or could be large enough to be medically important. This uncertainty is tackled by setting up an $H_0$ that specifies the magnitude of difference. Consider the following examples:

1. When can a new antihypertensive drug be considered clinically effective, i.e. if it results in an average decrease in diastolic blood pressure of at least 2 mmHg, 5 mmHg, 8 mmHg or 10 mmHg?
2. An iron supplementation programme in adolescent women is organized. After intake of the supplement for 30 days, the mean Hb level rises from 13.6 g/dl to 13.8 g/dl. Is this average gain of 0.2 g/dl sufficient to justify the programme? What gain can be considered enough to justify the expenditure and efforts in running the programme—0.5 g/dl, 1 g/dl or more?
3. The normal intraocular pressure (IOP) in healthy subjects when measured by applanation tonometry is 15.8 (SD 2.5) mmHg. In glaucoma, it is elevated. In a group of 60 patients with primary open-angle glaucoma, suppose the average IOP was 22.7 (SD 4.5) mmHg. After treatment with a new beta-adrenergic blocker, it came down to 19.5 (SD 3.7) mmHg. This reduction is statistically significant but the reduced level in the treatment group is still higher than the level in healthy subjects. Is this reduction still clinically important? What kind of difference from the normal level of 15.8 mmHg is clinically tolerable—0.5 mmHg, 1 mmHg, 2 mmHg or higher?

In all such problems, the clinician needs to decide the minimum acceptable or tolerable difference to justify intervention. A method similar to those already discussed can be adapted to test whether or not such a medically important difference is present.

### *Equivalence tests*

We are now in a position to discuss what are called equivalence tests in pharmaceutical literature. The primary aim of these tests is to disprove a null hypothesis that two means, or any other summary measure, differ by a clinically important amount. Equivalence tests are designed to demonstrate that no important difference exists between a new and the current regimen. They can also be used to demonstrate stability of a regimen over time, equivalence of two routes of dosage, and equipotency.

Equivalence can be demonstrated either in the form of 'at least as good as the present standard' or as 'neither better nor worse than the present standard'. The former is called clinical equivalence and the latter is called bioequivalence. In the case of clinical equivalence, the alternative hypothesis is one-sided, while in the case of bioequivalence it is two-sided. The latter can help to demonstrate that the dose of the drug delivered by the new route is neither higher nor lower than that delivered by the standard route. This is a typical quality control equipotency or bioequivalence goal.

An equivalence study may have two independent groups, paired groups or a cross-over design. If the outcome is a quantitative variable, the null hypothesis is $H_0: \mu_1-\mu_2=\mu_0$ but can also be $H_0: \mu_1-\mu_2=0$. Sometimes the interest is in the ratio of means instead of the difference. In that case, $H_0: \mu_1/\mu_2=\Delta$ where $\Delta$ could also be one. After logarithm, this reduces essentially to the former case. These hypotheses can be tested by the methods already described.

## THE NATURE OF STATISTICAL INFERENCE

Let us re-emphasize that nearly all information in health and medicine is empirical in nature because it is gathered from samples, which by themselves are a big source of uncertainty. Sample size also plays a dominant role in statistical inference. Inference from a test of hypothesis procedure can be drawn with minimal chance of error when *n* is large. However, a side-effect of a large *n* is that a very small difference can become statistically significant. This difference may or may not be medically significant. Therefore, caution is needed in drawing conclusions from statistical significance. Some of these are explained below:

1. *Whether or not a statistically significant result has any medical significance*
   If a drug significantly increases the cure rate of a particular disease from 70% to 73%, the question of medical relevance is whether this rise of 3% is worth it. One has to consider the price of the drug, efforts in procuring it, inconvenience in ingesting it, complying with the drug schedule and possible side-effects prior to recommending it for use. On the basis of statistical significance alone, this drug could not be recommended.

2. *Whether or not a plausible medical reason is available for the observed difference*
   Consider a random sample of 24 men and 15 women patients of leukaemia. Suppose 4 men and 7 women survive for 5 years. The difference in their survival rate is statistically significant since *P* is found to be less than 0.05 for one-sided $H_1$. However, no worthwhile reason may be available for this difference in their survival rate. Note that when the level of significance is 5%, there is a 1 in 20 chance that false significance is obtained. On the other hand, there might be factors, so far unknown, that could account for such a difference and this indeed could be real. For example, an inborn resistance in women, which leads to their higher life expectancy, can be an explanation. Statistical significance without proper medical explanation is rarely useful. However, a medical explanation may not be immediately available and may emerge later.

3. *Whether or not the P-value obtained is sufficiently small*
   While the convention is to use a threshold of 0.05 to label a *P*-value small or large, this is not uniformly applicable in all cases. In cases where the consequence of accepting $H_1$ can be grave, a *P*-value of 0.01 or less should be used. On the other hand, an inflated threshold such as 0.10 can also be used as in behavioural research.

4. *Whether or not multiple statistical tests are used on the same group of subjects*
   Procedures mentioned in this article are applicable to only one variable at a time. If you measure arterial blood gas $HCO_3$, $PCO_2$ and $PO_2$ in asthma patients and observe statistically significant ($P<0.05$) alteration in all three parameters individually, then any joint composite conclusion on all three of them should not be drawn. Multivariate methods are required when the variables are to be considered simultaneously. The results obtained in multivariate set-ups are not necessarily the same as those obtained by multiple tests on individual variables.

A distinction has to be made between significant, real and important differences. We have already said that a very large *n* can make a medically unimportant difference statistically significant. A statistically significant difference is very likely to be real though there is a small chance that it is not. On the other hand, if *n* is small a real difference may not be statistically significant. Similarly, a large and medically important difference can also be statistically not significant if *n* is not sufficiently large. A real difference, if it is small, such as 3 mg/dl in average total plasma cholesterol between a treatment response in men and women, can be medically unimportant or of no prognostic consequence. Also, small and large values together can produce a middling kind of mean. Similarly, a large variation between individuals can mask the difference between two or more groups. Averages can be deceptive, and there is always a need to be cautious while interpreting them.

## REFERENCES

1 Indrayan A, Gupta P. Sampling techniques, confidence intervals and sample size. *Natl Med Journal India* 2000;**13:**29–36.

2 Everiti SB. *Statistical methods for medical investigations.* London:Edwin Arnold, 1994:77–91.

3 Indrayan A, Gupta P. Hypothesis testing (Part I). *Natl Med J India* 2000;**13:**86–93.

4 Berenson ML, Levine DM, Goldstein M. *Intermediate statistical methods and applications: A computer package approach.* Englewoold Cliffs, New Jersey:Prentice Hall, 1983.

5 Lindman HR. *Analysis of variance in experimental design.* New York:Springer-Verlag, 1992.

6 Miller RG. *Simultaneous statistical inference.* New York:Springer-Verlag, 1981.

7 Hollander M, Wolfe DA. *Nonparametric statistical methods.* New York:John Wiley, 1973.