# V. COMPARISON OF MEANS

*Salim Daya, MB, MSc,*

*Departments of Obstetrics and Gynaecology, and Clinical Epidemiology and Biostatistics, McMaster University*

## ABSTRACT

*The t test is used commonly by researchers to test hypotheses involving data that are normally distributed and can be summarized by means and standard deviations. The paired t test is used to compare two paired observations on the same individual, or on matched individuals. The unpaired t test is used to compare the means of two independent samples. Although these tests are now easily performed using statistical programmes available for personal computers, it is important to understand the concepts involved in the calculations. Misuse of the test will often lead to false conclusions.*

## RÉSUMÉ

*Le test t est utilisé couramment par les chercheurs pour tester des hypothèses sur des données normalement distribuées pouvant être résumées au moyen de moyennes et d'écarts types. Le test t apparié permet de comparer deux observations portant sur le même individu ou les observations portant sur deux individus assortis. Le test t non apparié est utilisé pour comparer les moyennes de deux échantillons indépendants. Bien que ces tests soient maintenant d'utilisation simple grâce aux programmes de statistiques des ordinateurs individuels, il est important de comprendre les notions qui interviennent dans les calculs. L'utilisation erronée de ces tests aboutit souvent à des conclusions fausses.*

J SOGC 1995;17:571-9

## KEY WORDS

*t test, paired t test, unpaired t test, hypothesis testing, statistics, mean, standard deviation, standard error of the mean, variance.*

The availability of user-friendly statistical programmes for personal computers has made data analysis and hypothesis testing much easier. The purpose of this paper is to highlight the theoretical framework for testing of means, so that the report generated from such programmes can be understood and used appropriately to make inferences. In a previous installment in this series on Understanding Science, the discussion focused on the importance of confidence intervals, which emphasize that the mean observed in a study is an estimate of the true (but unknown) population mean. The sample mean and its standard error are used to construct the confidence interval, which represents the likely values for the population mean, within the predetermined interval; the wider the interval, the more likely the population mean will be included within the interval.

▼ ▼ ▼

This paper describes an alternate, but related approach to determine whether the sample mean is consistent with the population mean. This approach, which is known as significance testing, is based on the use of either the normal distribution, or the t distribution. The tests can also be used to compare the means from two different samples, by following similar rules for testing. The formulae used have been modified to accomodate additional assumptions which are made in the two sample situation.

Significance testing is not as informative as using the confidence interval approach, but is closely connected. For example, if the confidence level chosen is $1-\alpha$, the corresponding significance level is $\alpha$. Thus, if $\alpha = 0.05$, then the confidence level is $1-0.05 = 0.95$, and represents the 95 percent confidence interval. Similarly, the significance level 0.05 represents a five percent chance of rejecting the hypothesis being tested in the study.

## TESTING A POPULATION MEAN

### A) Known Standard Deviation

As discussed previously in this series, a normally distributed random variable (x), with a mean $\mu$ and standard deviation $\sigma$, can be transformed to the standard normal (z) distribution, with mean 0 and standard deviation 1, using the following transformation:

$$z = \frac{x - \mu}{\sigma}$$

This formula refers to the sampling distribution of individual observations. In contrast, the sampling distribution of the mean ($\bar{X}$) requires transformation using the standard error of the mean. According to the central limit theorem, the mean of this sampling distribution is still $\mu$, but the standard deviation is $\sigma/\sqrt{n}$ (i.e. the standard error of the mean). The formula can be rewritten:

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

Thus, if the true population mean of a distribution is $\mu$, and if the mean in a random sample taken from the population is observed to be $\bar{X}$, then using z tables it is expected that the value calculated for z will fall within the interval –1.96 to + 1.96 about 95 percent of the time (i.e. the area under the standard normal distribution curve between these limits of z is 0.95; recall that the area under the curve between –1 and +1 is 0.683, which is the probability of the value of z falling within one standard deviation of the mean). Whenever a study is conducted in which continuous data are gathered, one is interested in testing the hypothesis that the observed mean ($\bar{X}$) is the true mean $\mu$. To test this hypothesis, the significance level chosen is the probability that z will fall outside the limits governed by this level. For example, the selection of the conventionally accepted five percent significance level corresponds to the probability that the value calculated for z using the observed mean, will fall outside the interval –1.96 to + 1.96, if the observed mean is the true mean.

**Example**

Let us assume that systolic blood pressure is a normally distributed variable, with a known mean of 118 mm Hg and standard deviation of 10 mm Hg, in a population of healthy women in the third trimester of pregnancy.

Suppose a group of 25 women seen in an antenatal clinic was observed to have a mean systolic blood pressure of 121 mm Hg. How likely is this observation to have occurred by chance? The null hypothesis ($H_0$) is that the true mean systolic BP is 118 mm Hg. Thus, if $H_0$ is true, then the probability that the observed mean falls outside the interval ± 1.96 by chance is only 0.05.

$$z = \frac{121 - 118}{10/\sqrt{25}} = 1.5$$

This value of z lies within the limits chosen, indicating that the probability that the observed mean systolic BP (i.e. 122 mm Hg) occurred by chance is > 0.05. Therefore, there is insufficient evidence to reject the null hypothesis that the true mean is 118 mm Hg.

### B) Unknown Standard Deviation and the t Distribution

The approach discussed above is based on the understanding that the true value of $\sigma$ for the population is known (or at least there is a good approximation to this value). The estimates of $\sigma$ for the variables being measured are obtained from random samples. However, because there is much variability between and within subjects, the larger the sample studied,

the closer the estimate gets to the true (but unknown) value of $\sigma$. Often, in hypothesis testing the mean value of the normal distribution ($\mu$) of a particular variable is known, but the variance ($\sigma^2$) is not. Consequently, the above method of using the standard normal deviate cannot be used because the standard error of the mean (i.e. $\sigma/\sqrt{n}$) cannot be calculated. The only estimate of $\sigma$ available is the standard deviation (s) obtained from the sample of subjects being studied. Thus, by replacing $\sigma$ with s in the formula for the standardized normal deviate, one obtains the statistic t instead of the normal deviate z:

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

As previously mentioned in the formula, when n is large, s is a good approximation to $\sigma$. Therefore, as n gets larger, the sampling distribution of t is expected to become closer to that of z (i.e. close to a standard normal distribution with mean = 0 and standard deviation = 1). In contrast, when n is small, s may differ considerably from $\sigma$ causing t to have more random variability than z.

The t distribution was derived by W.S. Gossett (1876-1937) and published under the pseudonym 'student' to whom its derivation is erroneously credited. It is also known as the student's t distribution. This distribution belongs to a family of distributions called the t distributions, which are distinguished by the "degrees of freedom" (df). In the present context, t has a distribution on n – 1 degrees of freedom (i.e. one less than the sample size). As the df rises, the distribution tends to become more like the standard normal distribution, to which it is identical when df = $\infty$.

The term df refers to the number of independent measurements or observations in a quantity (in this case $\sigma$). For example, in choosing the three angles of a triangle, the third angle can be determined automatically, after having chosen two of the angles. In this scenario, there are two df, because one can choose two of the angles without restriction, but the third angle is defined, because the sum of the angles must be 180°. Similarly, in a sample of n measurements, one df is lost because $\bar{X}$ is used to estimate $\mu$. This leaves n – 1 independent values to calculate s. Tables are used to determine the probabilities for each t statistic at a particular df. From such a table, which is shown in the appendix, it can be verified that when df is infinity, the five percent significance level is 1.96, which is identical to that for the z distribution.

Because the individual observations are used to estimate $\sigma$, the df becomes smaller as the sample size is reduced. The penalty for such reduction in sample size is a reduction in the precision of the estimate. For example, with only one df (i.e. sample size of 2), the critical value at the five percent significance level is 12.71 instead of 1.96 (when $\sigma$ is known, and the z distribution can be used). The larger the critical value, the more difficult it is to reject the null hypothesis. Fortunately, by the time one gets to 10 df, the loss is not so great (e.g. the critical value is 2.23 versus 1.96; a reduction of only 14%, considering the distance from 10 to infinity). In general, for sample sizes $\geq$ 30, t is a good approximation to z. Also, the t distribution is valid only if the variable being studied has a normal distribution.

Calculation of the confidence interval is also performed using the t distribution, instead of the normal distribution. For example, to calculate the 95 percent confidence interval around $\mu$, the following formula is used:

$$\bar{X} \pm t_{df,\,0.05}\frac{s}{\sqrt{n}}$$

Here, $t_{df,\,0.05}$ represents the five percent point of the t distribution on n – 1 degrees of freedom; this value can be read from the intersection of the 0.05 column and the appropriate degrees of freedom row from the t table. As with hypothesis testing, the effect of this critical value on the confidence interval is large for small sample sizes.

**Example**

One may want to test the hypothesis that the mean gestational sac diameter, measured using transvaginal ultrasonography at seven weeks' gestation, in infertile women who conceived with *in vitro* fertilization (IVF) treatment, is equal to the known standard value (found to be 18.5 mm in a previous study).

The following sac diameter data (in mms) were obtained from 25 such women.

| | | | | |
|---|---|---|---|---|
| 21 | 17 | 16 | 20 | 25 |
| 23 | 15 | 22 | 27 | 24 |
| 26 | 21 | 17 | 17 | 22 |
| 18 | 24 | 22 | 19 | 17 |
| 19 | 15 | 25 | 20 | 15 |

The hypothesis to be tested can be stated in the null ($H_0$) and alternate ($H_A$) formats as follows:

$H_0$: $\mu = 18.5$ mm

$H_A$: $\mu \neq 18.5$ mm

It is assumed that the variable "gestational sac diameter" is normally distributed with an unknown $\sigma$. Therefore, the data have to be used to estimate $\sigma$, so that the t distribution can be used to test the hypothesis. The data set has $n = 25$, and $\bar{X} = 507/25 = 20.3$ mm.

To calculate the t statistic, the standard error of the mean ($s\sqrt{n}$) has to be determined.

$$S^2 = \Sigma(x-\bar{x})^2$$
$$= \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$
$$= 10603 - \frac{257049}{25}$$
$$= 321.04$$
$$\therefore S = \sqrt{321.04}$$
$$\rightarrow SEM = \frac{s}{\sqrt{n}} = \frac{\sqrt{321.04}}{\sqrt{25}} = 3.584$$
$$Thus\ t = \frac{20.3-18.5}{3.584} = 0.502$$

The degrees of freedom $(n - 1) = 25 - 1 = 24$.

From the t table, the probability values close to the calculated t value are shown below, together with the exact value obtained using a computer programme.

| | | | |
|---|---|---|---|
| t | 0.531 | 0.390 | 0.502 |
| p | 0.6 | 0.7 | 0.62 |

The t distribution with 24 df, $\alpha = 0.05$, and a two-tailed test to detect a difference greater or less than 18.5 mms, has a critical value from the t table of 1.711, indicating that the null hypothesis can only be rejected if the value of the t statistic calculated is either less than –1.711 or more than +1.711 (because a two-sided test was selected). In our example, the t statistic obtained (0.502) does not fall outside these limits. Therefore, based on the observed data, $H_0$ is not rejected. Stated differently, the precise significance level ($p = 0.62$) indicates that there is insufficient evidence (at the conventional 5% significance level) to reject $H_0$, because it is highly likely (i.e. 62% chance) that the observed mean value (20.3 mms) occurred by chance. Thus, there is no evidence that the mean gestational sac diameter at seven weeks' gestation in IVF pregnancies is different from the value in the normal population.

## COMPARISON OF TWO MEANS

In many studies, the researcher has to compare the means of variables measured in two groups of subjects, to determine if there are any differences, so that appropriate inferences can be made. Unlike the design discussed in the previous section, in which the emphasis was on comparing the mean in the sample with a standard mean, this section deals with two study designs, in which the **difference** between the means observed in the two groups is tested.

In the *paired* (or matched) design, the two groups are of equal size, because the individual subjects of one sample are paired with members of another group, or the individual subjects in one group are subjected to a measurement on two separate occasions (e.g. before and after an intervention). In the latter approach, which is commonly used, the subject serves as his/her own control, and the investigator is trying to determine whether the intervention makes any difference. In the *unpaired* (or independent) design, the two groups are totally separate and independent from each other.

### A) Paired Comparisons

In certain situations, a paired design is an extremely useful and powerful method for detecting differences. It is well known that biological measurements exhibit wide variation in subjects, so that any differences between subjects, in response to the intervention, may become overshadowed. A method to control for variability within subjects is to use the subjects as their own controls, and measure the change observed with the intervention. In this way, differences can be detected more easily.

The data obtained in a paired design can be analyzed using the paired t test, in which the difference between the two measurements for each subject (e.g. before and after an intervention) is treated as a single measurement. These differences are independent of each other, and if normally distributed, can be analyzed using the t test method previously described for a single mean. This approach is valid because the difference between two means is equal to the mean of the differences. (This statement can be verified in the example that follows). Under the null hypothesis, the difference arises from a distribution of differences with a mean of zero. However, the standard deviation of the differences is not equal to the difference in standard deviations between the two measurements. Therefore, the individual subject differences

have to be used to calculate the amount of variability in the distribution of the mean of differences.

If the differences are normally distributed, then the paired t test can be calculated as follows:

$$t = \frac{\bar{d}}{s/\sqrt{n}}$$

Here, $\bar{d}$ is the mean difference between the two groups.

$$i.e.\ \ \bar{d} = \frac{\Sigma d}{n}$$

and the standard deviation of the difference is:

$$s_d = \sqrt{\frac{\Sigma (d-\bar{d})^2}{n-1}}$$

**Example**

A study was conducted in ten women who had symptomatic uterine fibroids, and were treated with a gonadotrophin releasing hormone agonist (GnRHa) for 12 weeks. Uterine volume changes were documented with magnetic resonance imaging, which was performed prior to and at the end of treatment. The following results were obtained.

| PATIENT NUMBER | UTERINE VOLUME ($cm^3$) | | |
|---|---|---|---|
| | Pre-treatment | Post-treatment | Difference (d) |
| 1 | 526 | 433 | 93 |
| 2 | 549 | 478 | 71 |
| 3 | 739 | 496 | 243 |
| 4 | 608 | 497 | 111 |
| 5 | 566 | 427 | 139 |
| 6 | 631 | 462 | 169 |
| 7 | 698 | 511 | 187 |
| 8 | 687 | 443 | 244 |
| 9 | 623 | 388 | 235 |
| 10 | 709 | 538 | 171 |
| Mean | 633.6 | 467.3 | 166.3 |
| Standard deviation | 72.933 | 45.260 | 62.677 |

The hypotheses being tested are:

$H_0$: $\delta = 0$ (the mean difference in uterine volumes in the population is zero)

$H_A$: $\delta \neq 0$ (the mean difference is not zero)

Here, $\delta$ is the parameter for the mean difference in the population.

From the data, it can be seen that treatment produced a mean reduction in uterine volume of 166.3 $cm^3$. Note that the difference in means (633.6 – 407.3 = 166.3) is the same as the mean of the differences. Using the paired t test to test the hypothesis:

$$t = \frac{\bar{d}}{s/\sqrt{n}} = \frac{166.3}{62.677/\sqrt{10}} = 8.39$$

The critical value from the table, for a two-tailed t test with 9 df (i.e. n – 1) at the 0.05 level, is 2.262 which is much lower than the value calculated from the data (8.39). Therefore, there is sufficient evidence to reject the null hypothesis, in favour of the alternate hypothesis, that the mean difference is not zero, i.e. there is a significant reduction in uterine volume in response to treatment. The exact probability value using a computer programme is $p < 0.001$.

### B) Unpaired Comparisons

When observations are made in two separate or independent groups, the appropriate test to compare the means of these groups is the two-sample t test. Although in theory, the z distribution can be used, particularly for sample sizes greater than 30, in practice, it is rarely used because the population standard deviations for each group are not known.

Several assumptions have to be made before the two-sample t test can be used. The observations in both groups should be normally distributed, although this requirement of normality becomes less critical for large samples (i.e. $\geq 30$). The variances should be equal in the two groups, based on the premise that the two groups come from the same population in which, the distribution of the variable being studied in the two groups should have the same mean and variance. Equality of variances should be determined (using standard procedures available in statistical computer programmes) before performing the t test, although for equal sample sizes, the need for equal variances can be ignored, without having a major effect on the significance level of the test. In other words, the t test is robust for unequal variances if the sample sizes are equal. Finally, the observations should occur independently, so that the observations in one group do not influence the observations in the other group. This requirement cannot be tested formally, but can be assured by appropriate study designs.

▼ ▼ ▼

**1. Equal sample sizes**

When the sample size is the same in both groups, a test confirming that both groups have equal variances is not required. Thus, whenever possible, investigators should ensure the accrual of equal numbers of subjects into each group.

Under the null hypothesis, it is presumed that the difference in the means of the two groups has arisen from a distribution of differences with an overall mean of zero. The t test is used to test the null hypothesis that there is no **difference** between the means.

i.e. $H_o$: $\mu_1 - \mu_2 = 0$

In this way, the t test is different from the z test which is used to test the hypothesis that the two means are equal (i.e. $H_o$: $\mu_1 = \mu_2$). Although this distinction may appear to be a question of semantics, it is worth re-emphasizing that the t test in this situation is focused on the distribution of **differences** between the two groups.

Also, because the standard deviation of the population is not known, it has to be estimated from the data. A strategy to determine the standard deviation of this distribution of differences between means would be simply to assume that the error of the difference is the sum of the errors of the two estimated means. This approach is partially correct, but for mathematical reasons, the standard errors of the mean should first be squared before being added:

$$SE_{difference} = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

Here, $S_1^2$ and $S_2^2$ are the variances for the two groups of measurements. Because the sample sizes are equal (i.e. $n_1 = n_2$), the equation can be simplified to:

$$SE_{difference} = \sqrt{\frac{S_1^2 + S_2^2}{n}}$$

The t test is then performed using the same format shown previously, but modifying the equation in the following way:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{S_1^2 + S_2^2}{n}}}$$

The degrees of freedom are $(n_1 + n_2 - 2)$, because one df each is lost in calculating $\bar{X}_1$ and $\bar{X}_2$. The variance ($S^2$) for each group is calculated as previously described.

**Example**

In the example used before, let us assume that the 20 patients with uterine fibroids were randomly allocated to treatment with either GnRHa or placebo. Twelve weeks later, MRI examination produced the following data for uterine volume:

| | UTERINE VOLUME AFTER TREATMENT ($cm^3$) | |
|---|---|---|
| | Placebo | GnRHa |
| | 512 | 433 |
| | 544 | 478 |
| | 641 | 496 |
| | 615 | 497 |
| | 604 | 427 |
| | 692 | 462 |
| | 642 | 511 |
| | 561 | 443 |
| | 647 | 388 |
| | 539 | 538 |
| Mean | 599.7 | 467.3 |
| Standard deviation | 58.195 | 45.260 |

The hypotheses being tested are:

$H_o$: $\mu_1 - \mu_2 = 0$

$H_A$: $\mu_1 - \mu_2 \neq 0$

From the data, it can be seen that the group treated with GnRHa had a lower mean uterine volume than the group given placebo.

The test of the null hypothesis using the unpaired t test is as follows:

$$t = \frac{599.7 - 467.3}{\sqrt{\frac{(58.195)^2 + (45.260)^2}{20}}}$$

$$= 8.032$$

The critical value from the table, for a two-tailed t test with 18 df (i.e. $n_1 + n_2 - 2$) at the 0.05 level, is 2.101, which is much lower than the value calculated from the data (i.e. 8.032). Therefore, there is sufficient evidence to reject the null hypothesis, in favour of the alternate hypothesis, that the difference between the two means is not zero. i.e. GnRHa treatment causes a significant reduction in uterine volume. (The exact probability value is $p < 0.001$).

**2. Unequal sample sizes with equal variances**

If there are unequal sample sizes in the two groups but the variances are similar, the formula becomes a little more complex. Before the t test is performed, the

▼ ▼ ▼

standard error of the mean has to be obtained from the sample variance. However, because there are two groups, there will be two separate estimates of this statistic. A simple solution would be to take the mean of the two variances, but this is not an appropriate strategy because it does not take into account the fact that variances are influenced by the sample size. It is reasonable to presume that the standard deviation from the larger of the two groups is a better estimate of the population value. Therefore, when deriving the overall **pooled** estimate of the variance for the two groups, it would be appropriate to place more emphasis (i.e. weight) on the estimate from the larger group than on the estimate from the smaller group. Consequently, the sample size must be included in the pooled estimate of the variance.

Mathematically, it can be shown that the following equation is appropriate to calculate the pooled variance for two groups, with sizes $n_1$ and $n_2$, and variances $S_1^2$ and $S_2^2$.

$$Pooled\ S^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1) + (n_2-1)}$$

This pooled estimate falls somewhere between the separate values of the variance for each group. The standard error of the difference in the means of the two groups (i.e. $\bar{X}_1 - \bar{X}_2$) can now estimated by:

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\left[S^2\left(\frac{1}{n_1}\right) + \left(\frac{1}{n_2}\right)\right]}$$

To test the null hypothesis that the population means are equal (i.e. $\mu_1 = \mu_2$), the t test can be performed using the following formula:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{SE(\bar{X}_1 - \bar{X}_2)}$$

which can be re-written as:

$$t = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2} x \left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$$

This statistic follows the t distribution on $(n_1 + n_2 - 2)$ degrees of freedom. From the formula, it can be seen that the precision of the comparison between the means of the two groups is increased by higher values of $n_1$ and $n_2$, because the standard error of $\bar{X}_1 - \bar{X}_2$ is lowered. Thus, the larger the sample size, the better the precision of the difference observed between the two groups.

The two sample (unpaired) t test cannot be used to compare paired data because systematic differences between pairs would not be eliminated, but would form part of the variance used in the denominator of the t statistic equation. This approach is incorrect and would lead to a less sensitive analysis.

**Example**

In our example of patients with uterine fibroids, let us assume that the allocation to GnRHa or placebo was performed with a 2:1 randomization scheme. Thus, for every two patients who received GnRHa, one received placebo. The results of the MRI examination 12 weeks after commencing treatment are as follows:

| | UTERINE VOLUME AFTER TREATMENT($cm^3$) | | |
|---|---|---|---|
| | Placebo | GnRHa | |
| | 512 | 433 | 447 |
| | 544 | 478 | 412 |
| | 641 | 496 | 422 |
| | 615 | 497 | 439 |
| | 604 | 427 | 465 |
| | 692 | 462 | 375 |
| | 642 | 511 | 580 |
| | 561 | 443 | 405 |
| | 647 | 388 | 457 |
| | 539 | 538 | 470 |
| Mean | 599.7 | 452.3 | |
| Standard deviation | 58.195 | 41.618 | |

The hypothesis being tested are:

$H_O$: $\mu_1 - \mu_2 = 0$

$H_A$: $\mu_1 - \mu_2 \neq 0$

From the data, it can be seen again that GnRHa treatment resulted in a lower mean uterine volume than placebo.

The test of the null hypothesis using the unpaired t test for unequal sample sizes is as follows:

$$t = \frac{599.7 - 452.3}{\sqrt{\frac{(10-1)58.195^2 + (20-1)41.618^2}{(10+20-2)} x \left[\frac{1}{10}+\frac{1}{20}\right]}}$$

$$= 7.999$$

The critical value from the table, for a two-tailed t test with 28 df (i.e. $n_1 + n_2 - 2$) at the 0.05 level is 2.048, which is much lower than the value calculated from the

data (i.e. 7.999). Therefore, the evidence is strong enough to reject the null hypothesis, and to conclude that GnRHa treatment is efficacious in reducing uterine volumes in patients with fibroids (the exact p value is < 0.001).

**3. Unequal variances**

In situations where the groups are of unequal sizes, the variances may also differ considerably and cannot be pooled, because this approach would violate the assumption of equality of variances. Several options can be used to deal with this problem.

First, the scale of measurement can be transformed (e.g. by using the logarithm or some other transformation of the original measurement) until a scale is found which produces similar variances, but different means. However, if the original means are not too different, it will be difficult to find a transformation that reduces significantly the disparity between the variances. Second, the degrees of freedom for the t test can be reduced by using a formula known as the Satterthwaite correction. This approach increases the magnitude of the t statistic required to reject the null hypothesis (that the means are different). Consequently, a larger difference between means is required to conclude that the difference is statistically significant. A third option is to use the separate variances, instead of a pooled variance for the two groups. The standard error of the difference between the two means would be calculated as follows:

$$SE(\bar{X}_1 - \bar{X}_2) = \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}$$

Then, the significance test of the null hypothesis would be based on the d distribution, which is an approximation to the t distribution.

$$d = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)}}$$

The critical value for any particular probability level depends on $S_2^2 / S_2^2$, $n_1$, and $n_2$ and can be obtained from appropriate tables.

## SUMMARY

The t distribution is a probability distribution, which is similar to the standard normal (z) distribution. It is used to test hypotheses involving continuous data, which can be summarized using means and standard deviations. To determine whether a mean value observed in a study is different from a standard mean, or from a mean in another group, it is important to know:

a) the magnitude of the difference in means,
b) the variability in the observations made in the group(s), and
c) the sample size(s).

In general terms, the critical value for the t test is the ratio of the observed difference in means to the standard error of the mean, which is derived from the standard deviation in the group, divided by the square root of the sample size. The probability value for the t statistic thus calculated is determined from the t table, using the appropriate degrees of freedom, which is a function of the sample size.

Although the statistical calculations may seem intimidating, they do not need to be performed manually, because statistical programmes are now available for personal computers to simplify the task. However, it is important to understand the results of such computer-assisted statistical calculations. Before comparing means, it is necessary to test for equality of variances; the result of t tests are usually provided for both equal and unequal variances. If the sample sizes are equal, then the results of the t test for equal variances should be used. If the sample sizes are not equal, then the results of the formal test for equality of variances should be consulted. If this test is not significant (i.e. there is no significant difference in the variances), then the t test for equal variances should be used. If the equality of variances test is significant (i.e. the variances are not similar), then the results of the t test for unequal variances should be used. Depending on the computer programme used, it is quite likely that the appropriate adjustments for the degrees of freedom will have been made for the latter situation. Alternatively, the data could be transformed first, and the test for equality of variances repeated to determine whether the t test is performed better on the transformed data.

The t test is one of the most commonly used statistical tests in clinical research. Misuse of the test will often produce false conclusions. The details in this paper are presented in an effort to illustrate the principles of the various types of t test, and to identify appropriate situations in which they can be used.

## APPENDIX

CRITICAL VALUES FOR t DISTRIBUTION

| Degrees of freedom (df) | Two sided probability value (area in two tails) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0.9 | 0.8 | 0.7 | 0.6 | 0.5 | 0.4 | 0.3 | 0.2 | 0.1 | 0.05 | 0.02 | 0.01 | 0.001 |
| 1 | 0.158 | 0.325 | 0.510 | 0.727 | 1.000 | 1.376 | 1.963 | 3.078 | 6.314 | 12.706 | 31.821 | 63.657 | 636.619 |
| 2 | 0.142 | 0.289 | 0.445 | 0.617 | 0.816 | 1.061 | 1.386 | 1.886 | 2.920 | 4.303 | 6.965 | 9.925 | 31.598 |
| 3 | 0.137 | 0.277 | 0.424 | 0.584 | 0.765 | 0.978 | 1.250 | 1.638 | 2.353 | 3.182 | 4.541 | 5.841 | 12.924 |
| 4 | 0.134 | 0.271 | 0.414 | 0.569 | 0.741 | 0.941 | 1.190 | 1.533 | 2.132 | 2.776 | 3.747 | 4.604 | 8.610 |
| 5 | 0.132 | 0.267 | 0.408 | 0.559 | 0.727 | 0.920 | 1.156 | 1.476 | 2.015 | 2.571 | 3.365 | 4.032 | 6.869 |
| 6 | 0.131 | 0.265 | 0.404 | 0.553 | 0.718 | 0.906 | 1.134 | 1.440 | 1.943 | 2.447 | 3.143 | 3.707 | 5.959 |
| 7 | 0.130 | 0.263 | 0.402 | 0.549 | 0.711 | 0.896 | 1.119 | 1.415 | 1.895 | 2.365 | 2.998 | 3.499 | 5.408 |
| 8 | 0.130 | 0.262 | 0.399 | 0.546 | 0.706 | 0.889 | 1.108 | 1.397 | 1.860 | 2.306 | 2.896 | 3.355 | 5.041 |
| 9 | 0.129 | 0.261 | 0.398 | 0.543 | 0.703 | 0.883 | 1.100 | 1.383 | 1.833 | 2.262 | 2.821 | 3.250 | 4.781 |
| 10 | 0.129 | 0.260 | 0.397 | 0.542 | 0.700 | 0.879 | 1.093 | 1.372 | 1.812 | 2.228 | 2.764 | 3.169 | 4.587 |
| 11 | 0.129 | 0.260 | 0.396 | 0.540 | 0.697 | 0.876 | 1.088 | 1.363 | 1.796 | 2.201 | 2.718 | 3.106 | 4.437 |
| 12 | 0.128 | 0.259 | 0.395 | 0.539 | 0.695 | 0.873 | 1.083 | 1.356 | 1.782 | 2.179 | 2.681 | 3.055 | 4.318 |
| 13 | 0.128 | 0.259 | 0.394 | 0.538 | 0.694 | 0.870 | 1.079 | 1.350 | 1.771 | 2.160 | 2.650 | 3.012 | 4.221 |
| 14 | 0.128 | 0.258 | 0.393 | 0.537 | 0.692 | 0.868 | 1.076 | 1.345 | 1.761 | 2.145 | 2.624 | 2.977 | 4.140 |
| 15 | 0.128 | 0.258 | 0.393 | 0.536 | 0.691 | 0.866 | 1.074 | 1.341 | 1.753 | 2.131 | 2.602 | 2.947 | 4.073 |
| 16 | 0.128 | 0.258 | 0.392 | 0.535 | 0.690 | 0.865 | 1.071 | 1.337 | 1.746 | 2.120 | 2.583 | 2.921 | 4.015 |
| 17 | 0.128 | 0.257 | 0.392 | 0.534 | 0.689 | 0.863 | 1.069 | 1.333 | 1.740 | 2.110 | 2.567 | 2.898 | 3.965 |
| 18 | 0.127 | 0.257 | 0.392 | 0.534 | 0.688 | 0.862 | 1.067 | 1.330 | 1.734 | 2.101 | 2.552 | 2.878 | 3.922 |
| 19 | 0.127 | 0.257 | 0.391 | 0.533 | 0.688 | 0.861 | 1.066 | 1.328 | 1.729 | 2.093 | 2.539 | 2.861 | 3.883 |
| 20 | 0.127 | 0.257 | 0.391 | 0.533 | 0.687 | 0.860 | 1.064 | 1.325 | 1.725 | 2.086 | 2.528 | 2.845 | 3.850 |
| 21 | 0.127 | 0.257 | 0.391 | 0.532 | 0.686 | 0.859 | 1.063 | 1.323 | 1.721 | 2.080 | 2.518 | 2.831 | 3.819 |
| 22 | 0.127 | 0.256 | 0.390 | 0.532 | 0.686 | 0.858 | 1.061 | 1.321 | 1.717 | 2.074 | 2.508 | 2.819 | 3.792 |
| 23 | 0.127 | 0.256 | 0.390 | 0.532 | 0.685 | 0.858 | 1.060 | 1.319 | 1.714 | 2.069 | 2.500 | 2.807 | 3.767 |
| 24 | 0.127 | 0.256 | 0.390 | 0.531 | 0.685 | 0.857 | 1.059 | 1.318 | 1.711 | 2.064 | 2.492 | 2.797 | 3.745 |
| 25 | 0.127 | 0.256 | 0.390 | 0.531 | 0.684 | 0.856 | 1.058 | 1.316 | 1.708 | 2.060 | 2.485 | 2.787 | 3.725 |
| 26 | 0.127 | 0.256 | 0.390 | 0.531 | 0.684 | 0.856 | 1.058 | 1.315 | 1.706 | 2.056 | 2.479 | 2.779 | 3.707 |
| 27 | 0.127 | 0.256 | 0.389 | 0.531 | 0.684 | 0.855 | 1.057 | 1.314 | 1.703 | 2.052 | 2.473 | 2.771 | 3.690 |
| 28 | 0.127 | 0.256 | 0.389 | 0.530 | 0.683 | 0.855 | 1.056 | 1.313 | 1.701 | 2.048 | 2.467 | 2.763 | 3.674 |
| 29 | 0.127 | 0.256 | 0.389 | 0.530 | 0.683 | 0.854 | 1.055 | 1.311 | 1.699 | 2.045 | 2.462 | 2.756 | 3.659 |
| 30 | 0.127 | 0.256 | 0.389 | 0.530 | 0.683 | 0.854 | 1.055 | 1.310 | 1.697 | 2.042 | 2.457 | 2.750 | 3.646 |
| 40 | 0.126 | 0.255 | 0.388 | 0.529 | 0.681 | 0.851 | 1.050 | 1.303 | 1.684 | 2.021 | 2.423 | 2.704 | 3.551 |
| 60 | 0.126 | 0.254 | 0.387 | 0.527 | 0.679 | 0.848 | 1.046 | 1.296 | 1.671 | 2.000 | 2.390 | 2.660 | 3.460 |
| 120 | 0.126 | 0.254 | 0.386 | 0.526 | 0.677 | 0.845 | 1.041 | 1.289 | 1.658 | 1.980 | 2.358 | 2.617 | 3.373 |
| ∞ | 0.126 | 0.253 | 0.385 | 0.524 | 0.674 | 0.842 | 1.036 | 1.282 | 1.645 | 1.960 | 2.326 | 2.576 | 3.291 |

For infinite degrees of freedom, the t distribution is identical to the normal distribution.

## SUGGESTED READING

The following books were used in preparing this manuscript.


1. Armitage P, Berry G. Statistical Methods in Medical Research. Oxford Scientific Publications, 1994.
2. Remington RD, Schork MA. Statistics with Applications to the Biological and Health Sciences. Prentice-Hall Inc. 1970.
3. Dawson-Saunders B, Trapp RG. Basic and Clinical Biostatistics. Appleton and Lange, 1990.
4. Norman GR, Streiner DL. Biostatistics. The Bare Essentials. Mosby-Year Book Inc. 1994.