CHAPTER 22

Comparison of Two Groups: t-Tests and Nonparametric Tests

Contents

Basic Concepts 337
- Introduction 337
- Paired t-Test 338
- Sample Size for Paired Test 346
- Unpaired t-Test 346
- Unequal Variances 349
- Sample Size for Unpaired Test 350
- Conclusion 350
- Nonparametric or Distribution Free Tests 351
  - *The Wilcoxon-Signed Rank Test* *351*
  - *The Sign Test* *353*
  - *The Mann—Whitney U Test* *354*

Advanced Concepts 356
- Comparing Two Coefficients of Variation 356
- The Paired t-Test Implies an Additive Model 356
- Confidence Limits for Medians 358
- Ranking Transforms 359
- The Meaning of the Mann—Whitney Test 359

Appendix 360
References 361

BASIC CONCEPTS

Introduction

One of the most frequent questions asked is if the means of two groups are different enough that they are unlikely to have come from the same population. Did the blood sugar decrease more with drug A than drug B? To evaluate this question with continuous data, we usually use the t-test for comparing the means of two groups. It is the prototype for almost all other statistical inferences, and brings into play most of the considerations involved in making these inferences. The t-test is a version of the more general analysis of variance (ANOVA) restricted to two groups.

Biostatistics for Medical and Biomedical Practitioners
http://dx.doi.org/10.1016/B978-0-12-802387-7.00022-6

There are two types of t-tests. In one, data are collected in pairs of subjects and the set of differences between each pair is tested to determine if the variation is consistent with the null hypothesis that the set of differences comes from a population with a mean difference of zero, that is, the two groups are not different. This is termed the paired t-test. Its counterpart is where two different groups are compared and the question asked is if the two means could have come from the same or different populations.

Paired t-Test

Paired comparisons are frequent. For example, a blood sample is divided into half and placed in two tubes: one is a control and the other has some chemical added, with the question being whether the chemical causes a change in the concentrations of the substance of interest. Blood from several subjects is examined, each time in pairs. Another type of paired experiment might have one pair of rat littermates from several different litters, with one of each pair being given a standard diet and the other member of the pair being given the same diet with a food additive to determine if the additive affects growth. A third experiment might be to study a group of hypertensive people before and after a given dose of a drug to determine if it lowers blood pressure.

In order to determine if the experimental group differs from the control group or if the drug lowers pressure, examine the *differences* between each pair of data values. For example, a study of the biological value of raw (R) versus roasted (P) peanuts as judged by the weight gain of rat littermates (in grams) produced the data of Table 22.1 (Mitchell et al., 1936).

For each pair of rats the difference D in weight gain is calculated, giving the data in column 3. Now ask: "If there is no average difference between the weight gains on the

Table 22.1 Weight gain of paired littermates fed either raw or roasted peanuts in their diet

Raw peanuts R	Roasted peanuts P	Difference D
61	55	6
60	54	6
56	47	9
63	59	4
56	51	5
63	61	2
59	57	2
56	54	2
44	63	−19
61	58	3
		$\Sigma X_i = 20$ $\overline{X_D} = 2$

two diets, how likely is it that there would be a difference of as much as 2 g?" That is, Ho: $\mu_D = 0$.

If it is very likely, then we would not consider that roasting affected the nutritional value of peanuts, but if it is an unlikely difference, then we might want to consider that roasting affected their nutritional value. Assess the probability of the null hypothesis by determining how many standard deviations from the mean that difference represents. If the difference is many standard deviations from the mean, then there is reason to reject the null hypothesis. To do the required calculations, calculate the mean and standard deviation of the differences, $\overline{X_D} = 2$, $\sum(X_i - \overline{X})^2 = 536$. Therefore, $s^2 = 59.56$, $s = 7.72$, and $s_{\overline{X}} = 7.72/\sqrt{10} = 2.44$. Then relate the difference to the standard error to determine the probability of observing that difference if the true population difference is zero.

$$t = \frac{2 - 0}{2.44} = 0.82.\ \mathrm{P} = 0.43.$$

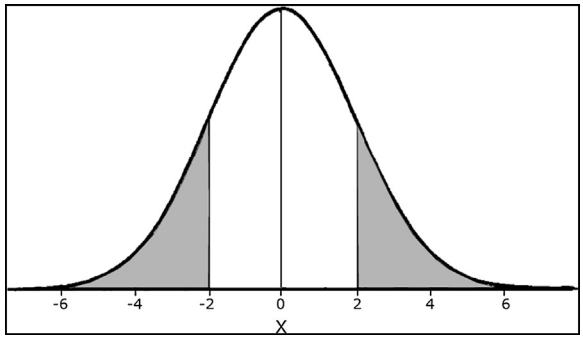

Figure 22.1 Normal curve with shaded areas indicating proportion under the curve beyond the value of *t*. That is, assuming normality, if the mean difference in the long run was truly 0, at least 43% of similar samples would have a mean difference >2. On this basis, we would not want to reject the null hypothesis.

The 0.05 value of *t* for 9 degrees of freedom (df) is 2.262, so that the 95 confidence limits of the mean are $2 \pm 2.262 \times 2.44 = -3.52$ to 7.52. Because 0 is included within these limits, the null hypothesis cannot be rejected. The *t* table can be seen online at http://www.sjsu.edu/faculty/gerstman/StatPrimer/t-table.pdf, the probabilities for the *t* values can be obtained from http://vassarstats.net/tabs.html, http://in-silico.net/statistics/ttest. The online programs http://www.graphpad.com/quickcalcs/ttest1.cfm, http://www.usablestats.com/calcs/2samplet, and http://easycalculation.com/statistics/ttest-calculator.php allow you to enter the data, and then perform the test.

Problem 22.1

The table below shows the peak flow rates (l/min) in asthmatic patients before and after exertion.

Subject	Before	After
1	320	297
2	235	200
3	322	220
4	376	334
5	286	210
6	254	255
7	381	338
8	397	341
9	299	227

Did exertion cause a decrease in peak flow rate? Would you reject the null hypothesis?

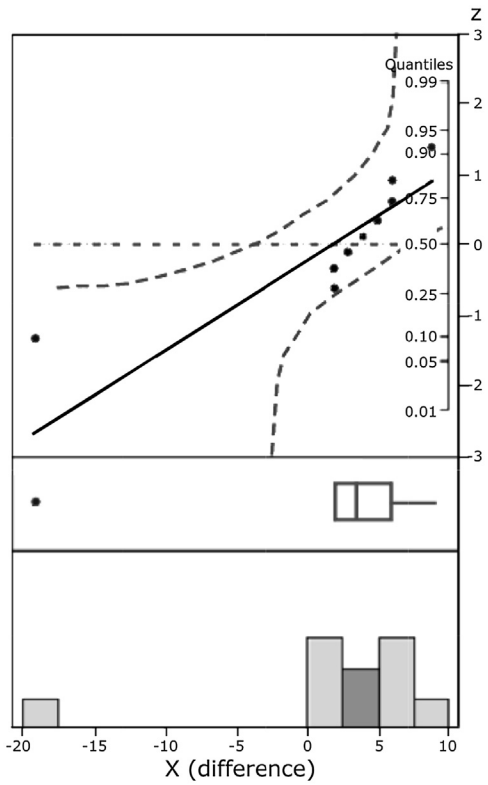

Figure 22.2 Upper panel: normal quantile plot with dashed lines showing 95% confidence limits. Middle panel: box plot. Lower panel: histogram of differences.

There are several points to notice about the paired t-test:

1. The numbers should be ratio or interval numbers, and the distribution of the set of differences should be approximately normal. This is not true here, as shown by observation and Figure 22.2.

There is one extreme outlier shown in all the panels, and the data do not fit a normal quantile plot.

It does not matter if each distribution is not normal; what matters is the distribution of the differences between paired measurements.

2. There must be justification for pairing based on intuitive reasoning or experience. For example, a blood sample divided into two portions should have the same constituents in each portion. Therefore, adding something to one of the pair constitutes the difference to be examined. If theory does not provide the answer, we can turn to experience. Prior work has shown that rat littermates reared on the same diet grow at very similar rates that are more closely matched than are rats from two different litters, or two different species. Patients who are being tested for airway reactivity tend to have the same reactivity when tested on different days, so that testing a patient before and then again after administering a potential airway irritant allows pairing. If, however, previous studies have not shown consistency of response to the same stimulus (diet, airway inhalant, etc.) then pairing should not be used. Any investigator using a paired design (shown below to be more efficient than an unpaired design) must be prepared to justify the use of pairing.
3. The calculated value for t is 0.82, and the probability of t is 0.43 (Figure 22.1).
4. t is a ratio of a difference between an observed mean and a population mean (numerator) to a measure of variability (denominator). The numerator is the signal and the denominator is the noise, so that t represents the signal-to-noise ratio. If the difference is small or the variability is big, t will be small, leading to inability to reject the null hypothesis. On the other hand, if the numerator is big relative to the denominator, t will be big, leading to rejection of the null hypothesis.
5. The numerator indicates the absolute difference between two measurements (the effect size), and its importance has to be judged on the physiological or clinical importance of its magnitude. A difference of 2 mg/dl of serum potassium is huge and potentially serious, whereas a difference of 2 mm Hg of systolic blood pressure is trivial and not clinically important. Whether a difference is big or small is not a statistical question but a matter of judgment by the investigator. The numerator reveals the importance of the measured difference.
6. The denominator, here the standard deviation of the mean, has variability determined by the variability within the population. For a given numerator to yield a high value of t and thus lead to rejection of the null hypothesis the denominator should be as small as possible for that set of measurements. This can sometimes be achieved by making the sampled population as homogeneous as possible. For example, there should be less variability of weight gain in the peanut experiment with rats from the same inbred species rather than from different species. At other

times minimize variability by avoiding outliers because the standard deviation is not a resistant measurement. Sometimes it is appropriate to transform the data by logarithmic or other transformation to avoid having long tails to the distribution and thus inflating the standard deviation. Another way of minimizing the denominator is to increase the sample size. Because variability is a function of $\sqrt{N}$, an increase in sample size decreases the standard deviation of the mean.

7. The probability of *t* therefore is based on a ratio of an observed difference to its variability. If *t* is large, so that it is unlikely that the mean difference is zero, then we may wish to reject the null hypothesis. We cannot be certain that the null hypothesis is false; the best we can do is estimate its probability of being false. If we reject the null hypothesis but are wrong (as shown by future work) we commit a Type I error. It is our choice as to what probability to use to minimize a Type I error. Conventionally, the 95% confidence limits are used, giving a 5% chance of making a Type I error, and this is often called "statistical significance."
   a. In normal conversation, significance implies importance, but that is not true in statistics. Its meaning is confined to the chances of making a Type I error, whether or not the observed difference is physiologically or clinically important (see Chapter 10).
   b. The 5% (or 0.05) figure for significance is arbitrary. One percent gives a smaller chance of making a Type I error and rejecting the null hypothesis falsely, 10% has a greater chance of making a Type I error. The 5% level means that a Type I error occurs about 1/20 times, and most people find that psychologically pleasing. On the other hand, if an investigator is doing a number of screening tests on different types of peanut preparation, he or she might well use the 10% cut off value to decide which types to study further.
8. To summarize: If *t* is significant, that is, we think it reasonable to reject the null hypothesis, we still need to evaluate the absolute magnitude of the difference. If small, we may elect to ignore it. For example, 1 million hypertensive people are tested with a new antihypertensive drug. The mean difference before and after the drug is 2 mm Hg, but because the standard deviation of the mean is incredibly small, *t* is significant. But that does not mean that the difference is important, and the company manufacturing the drug might elect not to market it.
9. If *t* is significant, it argues for a difference between the pairs, but does not prove that the difference observed was due to the experiment. There might have been factors outside our control or knowledge that were the causes. Perhaps the 9/10 rats fed raw peanuts and gained more weight than their littermates were kept warm and slept a lot, whereas their littermates were kept cold and made to be active, so that they burned up more calories. We would see a difference that was *associated* with the type of peanuts, but not due to it.
10. If *t* is small, first examine the numerator—the difference. If it is small and unimportant, then we really do not care that we have not reached statistical significance. Effectively, the difference does not matter to us. If the difference is large enough

to be important, but *t* is too low for significance, then examine the denominator. Is the large variability due to inhomogeneity that can be reduced? Is it due to a non-normal distribution that can be normalized? Is it practical (cost, time, manpower) to increase sample size? If none of these remedies is possible, it may be possible to do a nonparametric analysis (see below). Failure to reject the null hypothesis does not mean that the difference was zero, but merely that you have not proved it is not.

11. The one outlier should have been picked up before the analysis was done. Why did 9/10 rats gain more weight on raw peanuts whereas 1 rat not only lost weight, but lost a great deal of weight? Was there an error in weighing or entering the data? Perhaps whoever weighed that rat misread the scale. Did that rat differ in any way? Rats can get pneumonia or tuberculosis, and if that rat was ill it was not validly a member of the group. The effect of removing the outlying pair from Table 22.1 is shown in Figure 22.3.

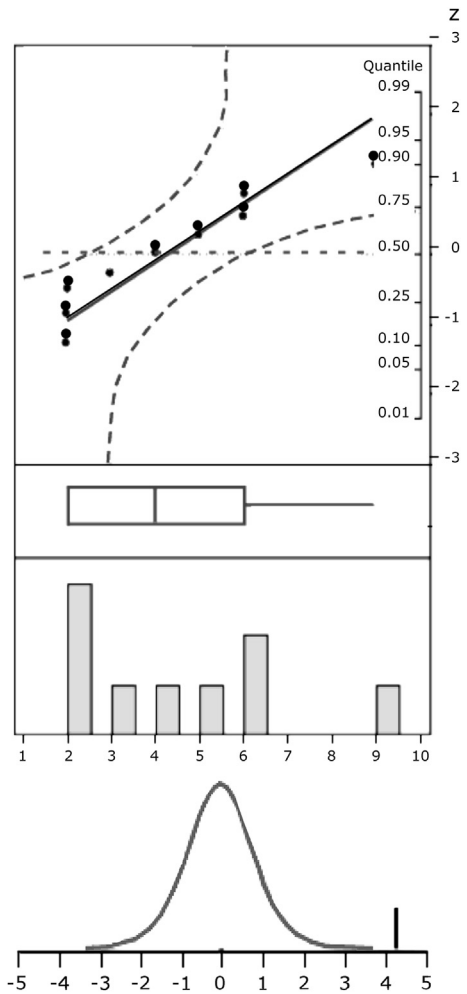

Figure 22.3 Upper panel: quantile plot. Second panel: box plot. Third panel: histogram. Fourth panel: t-test for the nine paired observations.

Removing the outlier has changed the mean difference from 2 to 4.33, standard deviation from 7.72 to 2.40, standard deviation of the mean from 2.44 to 0.80, and t from 0.82 to 5.42. That is, the mean difference is now 5.42 standard deviations of the mean from a hypothesized mean of 0, and this would happen with a probability of only 0.00006. The 95% confidence limits become $4.33 \pm 0.80 \times 2.306 = 2.49$ to 6.17. These limits do not include 0, confirming the statistical significance of the test.

12. If t is significant, it does not matter if the distribution was not normal. Uninformed reviewers often make this error. In fact, if t is significant under these circumstances, it would be even more statistically significant if the distribution were made more normal.

13. Because the normal curve is symmetrical, the tails of the curve that suggest that the null hypothesis can be rejected contain equal areas. If the null hypothesis is true, there is only a 0.025 (2.5%) probability that the sample mean will lie more than $t_{0.05}\, s_{\overline{X}}$ above the mean and another 0.025 probability that it will lie less than $t_{0.05}\, s_{\overline{X}}$ below the mean. The two probabilities together add up to 0.05. Whether to use two areas or only one area depends upon the alternative hypothesis H_A. Remember that if the null hypothesis H_O is rejected, an alternative hypothesis has to be accepted. There are three alternative hypotheses:

$$H_A : \overline{X} \neq \mu;$$

$$H_A : \overline{X} > \mu;$$

$$H_A : \overline{X} < \mu.$$

The first hypothesis states that an excessive deviation from $\mu = 0$ in either direction will lead to rejection of the null hypothesis, the second states that a mean significantly above $\mu = 0$ will lead to rejection of the null hypothesis, and the third states that a mean significantly below $\mu = 0$ will lead to rejection of the null hypothesis. The first is known as the two-tailed test, and the other two each as a one-tailed test. Because the area in a one-tailed test is half that of a two-tailed test, a given value of t will give a probability for a one-tailed test that is half that of a two-tailed test, for example, 0.025 instead of 0.05. It is thus easier to achieve statistical significance for a one-tailed than for a two-tailed test.

When is a one-tailed test permissible? The answer depends in part on what we are looking for. In the raw versus roasted peanut experiment, we might have had no prior guesses as to how the results would turn out, so that a two-tailed test would be appropriate. However, even if we had expected raw peanuts to be better, they could have turned out to be worse. Some proposed treatments are actually harmful, not helpful. Therefore, a two-tailed test is appropriate. An example of how predictions can be wrong can be found by examining the CAST trial (CAST, 1989; Ruskin, 1989). The drug flecainide had been shown to be useful in treating and preventing ventricular arrhythmias

in experimental animals and in some humans with a normal myocardium, and was being used extensively in clinical practice to treat ventricular arrhythmias in patients after myocardial infarction. To investigate further and to legitimize an accepted practice, the CAST trial randomized patients to a control group or one of several newer antiarrhythmic agents, including flecainide. An interim analysis after about one-third of the patients had been admitted to the study showed to everyone's dismay that four times as many patients had died in the flecainide arm than the control arm of the study. The study was abruptly halted.

One-tailed tests are occasionally used. If a peanut producer is testing the growth potential of a new species of peanut, there is interest in producing it commercially only if it is better than the standard type. Therefore, the manufacturer is interested only in a mean growth potential greater than the standard, that is, only in the upper tail. Remember that, it is easier to show significance in a one- than a two-tailed test, because to exceed 5% of the area under the curve in one direction takes a smaller deviation from the zero population mean. For example, in a two-tailed test with $N = 10$, the value of t that corresponds to 0.025 of the area under the curve for a total of 0.05 for both tails is 2.262, but for 0.05 in one tail is 1.833 (See Figure 22.4).

There is little use for one-tailed tests in biology or medicine. We can never be sure that an adverse response might not occur, and so should consider a deviation in either direction. There is no justification for using a one-tailed rather than a two-tailed test simply because it is easier to show significance. It may, however, be used when only one alternative hypothesis is of interest, for example, the new treatment will be used only if it does not cause harm. If it is, the decision to use a one-tailed test must be made before the experiment is done to avoid unconscious bias.

14. Although most two-sided tests are symmetrical, with 2.5% of the area under the curve more than 1.96 σ above and below the mean, this is not an absolute

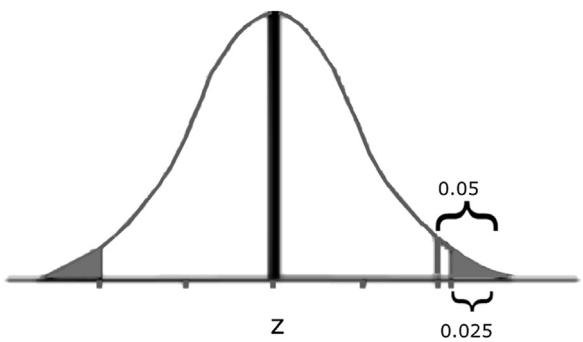

Figure 22.4 The shaded areas in the two tails make up 0.05 of the total area under the curve, as does the area in the upper tail demarcated by the left of the two lines. The second line is closer to the mean, so that for a given value of α, a smaller value of t is needed for a one-tailed test than a two-tailed test.

requirement. A statistical test provides support for a decision, and decisions have consequences. Blind adherence to a standard method may not be effective. As discussed in detail by Moyé (2000), there are times when there is more concern for H_A: $\overline{X} < \mu$ than for H_A: $\overline{X} > \mu$. As an example, he discusses testing a new treatment for diabetes mellitus. There is a current standard treatment that is effective in reducing the risk of cardiovascular disease, and its harmful side effects are uncommon and readily recognized and dealt with. A new treatment is proposed and tested. The investigator might be more concerned with an increase in harmful side effects than in an improvement in treatment effect. If in the planning stage the investigator has decided to make the critical value of the Type I error α 0.05, the decision could be made to apportion 0.03 to the tail that indicates harm and 0.02 to the tail that indicates benefit. The decision then is to reject the null hypothesis in favor of harm if $t \leq -1.88$, and in favor of benefit if $t \geq 2.05$. This concept applies to unpaired t-tests, and to significance tests in general.

There is no reason not to do the test as usual, and then interpret the observed t value differently for the two alternative hypotheses.

15. The P value of a test such as the t-test has a composite origin, depending as it does on the observed difference (effect size), sample size, sample variability, and by implication the shape of the distribution and the presence or absence of outliers.

Sample Size for Paired Test

To determine what sample size is needed to show significance in any paired experiment, we have to specify the standard deviation of the data, usually taken from similar experiments or a pilot study, the effect size (difference from zero) that is required, and the designated value of α, usually 0.05 or 0.01. Sample size may be calculated online at http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html, http://www.dssresearch.com, http://www.maths.surrey.ac.uk/cgi-bin/stats/sample/twomean.cgi, and http://www.sample-size.net.

Unpaired t-Test

To compare two unmatched groups, use an unpaired t-test. The numbers in the two groups may be different, but even if they are the same a paired test cannot be done unless pairing is justified.

Return to the data shown in Table 22.1, and assume that the experiment compared two separate groups of young rats, one group fed with added raw peanuts, the other with added roasted peanuts. No pairing is done. To do the required calculations, first calculate the mean and standard deviation for each group. These values for the means are: Raw 57.9 g, Roasted 55.9 g, and for the standard deviations are: Raw 5.59 g and Roasted 4.75 g.

The t ratio then becomes the difference between the means divided by the measure of variability. The variability is, however, derived from two separate variabilities, one for each group. The variability of the difference between two groups is greater than either one alone, because we have to allow for one mean being unusually low and the other unusually high. To calculate the combined variability, calculate a common or pooled variance s_p^2 that is the weighted average of the two variances (Chapter 3). This can be calculated as $s_p^2 = \frac{\sum (X_{1i}-\overline{X}_1)^2+\sum (X_{2i}-\overline{X}_2)^2}{N_1+N_2}$.

Then the standard deviation of the difference between two means is $S_{\overline{X}_1-\overline{X}_2} = \sqrt{\frac{s_p^2}{N_1}+\frac{s_p^2}{N_2}}$.

(This formula implies no correlation between the members of the two groups. If there is correlation, the formula needs to be modified.)

Therefore, $t = \frac{\overline{X_1}-\overline{X_2}}{\sqrt{\left(\frac{s_p^2}{N_1}+\frac{s_p^2}{N_2}\right)}}$.

(Strictly speaking, the numerator should be $\left|\overline{X_1}-\overline{X_2}\right|-\mu$, because the hypothesis is that the difference between the two means is not significantly different from the population mean difference of zero. It is simpler to omit μ from the formula.)

The value obtained for t is tested against $N-2$ df, because two groups are involved.

For the data in Table 22.1, the difference between the means is 2. The argument is that if in the long run there is no difference between the means of the two groups, then the mean difference will be zero. If this null hypothesis is true, how often will a difference of 2 arise? The way to determine this is to relate the difference between the means to a measure of variability. If variability is such that a difference of 2 can arise frequently, then we will not reject the null hypothesis. If, however, a difference is unlikely to occur with that population, then we can reject the null hypothesis. For the data in Table 22.1 tested as an unpaired experiment, $t = \frac{2}{2.32} = 0.8621$, df = 18, and P = 0.40. About 40% of the time if we drew two samples with $N = 10$ from this population, the means would differ by 2 or more. We therefore do not reject the null hypothesis.

The unpaired t-test can be done online at http://www.graphpad.com/quickcalcs/ttest1.cfm, http://www.usablestats.com/calcs/2samplet, and http://easycalculation.com/statistics/ttest-calculator.php.

Problem 22.2

Reanalyze the peak flow rate data as if they were two different groups of subjects. How do the results differ from a paired test on the same data? Explain why the results are so different.

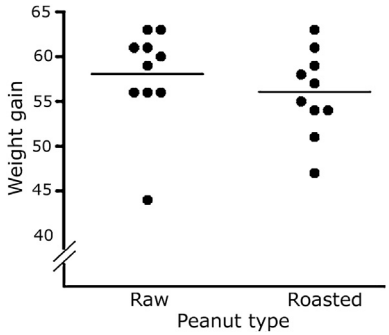

Figure 22.5 Distribution of data from raw peanuts (left) and roasted peanuts (right).

Requirements for unpaired t-test:

1. First consider if the data meet the requirements for an unpaired t-test. Two of these are similar to the requirements for a paired test, namely that the numbers are ratio numbers and each distribution is normal. As seen in Figure 22.5, the second requirement has not been met, but for the moment ignore that.
   Another requirement not needed for the paired test is that the variances should be similar in the two data sets. The way of determining this will be described later.
2. Once again, the 44 g weight gain for the rat fed raw peanuts in the 10th pair is unduly low. Exclude this measurement and repeat the unpaired t-test with 9 in the raw group and 10 in the roasted group. The means become 59.44 g (raw) and 55.90 g (roasted), with respective standard deviations of 2.88 and 4.74, respectively. Now $t = 1.939$ with 17 df, and $P = 0.0693$. The P value for rejecting the null hypothesis is much bigger than when the paired test was done—0.0693 versus 0.0006—attesting to the greater sensitivity of the paired test *provided it is legitimate to do it.*
3. The remaining considerations about importance versus significance, how to interpret the numerator and the denominator, and what to do about a nonsignificant P value are the same as for the paired test.

We always need to ask if the observed difference, whether significant or not, is important. This is a value judgment made by workers in that field who ought to know if a difference of 3.54 g weight gain is important. It is possible to quantitate this difference by assessing what contribution the diet makes to total variability. If we take the total variability of the 19 weight gains (excluding the aberrant value), we can ask how much this is reduced if we allow for differences due to the diet. This subject will be taken up more fully in Chapter 25 on ANOVA, but we can estimate the relative reduction, termed ω^2, as (Hays, 1973)

$$\omega^2 = \frac{t^2 - 1}{t^2 + N_1 + N_2 - 1}.$$

Thus for the unpaired peanut data,

$$\omega^2 = \frac{1.939^2 - 1}{1.939^2 + 10 + 9 - 1} = 0.1268$$

which can be interpreted as showing that about 13% of the variability of weight gains can be accounted for by differences in diet.

Unequal Variances

The logic of the t-test requires a pooled variance from the two groups with comparable variances as a way of obtaining a more representative variance from a larger total sample. If the variances are very different, then their weighted average represents neither sample, with the potential for distorting the value for the standard deviation of the mean that is the denominator for the t-test. If the sample sizes and the variances are very different, the smaller sample has a disproportionate effect on the pooled value because we are dividing by the square root of N. To test the hypothesis of equal variances, divide the larger variance by the smaller variance to obtain the variance ratio termed F (Chapter 8). This will be discussed fully in the Chapter 25 on Analysis of Variance. Suffice it to state that it is possible to determine whether the observed F ratio is far enough removed from the population ratio of 1 that the hypothesis of equal variances can be rejected.

If the variances are significantly different, use a modified t-test or a non-parametric test. If a nonparametric test, it should not be the Mann—Whitney U test (see below) because if the group variances are very different, Type I error rates may be too high or too low; if too low, the risk of inflating the Type II error is increased (Kasuya, 2001; Neuhäuser, 2002; Ruxton, 2006). Modified t-tests appear in statistics programs as the Welch or the Satterthwaite test, the latter being preferable. In both tests, first determine the ratio d (analogous to t) by

$$d = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}.$$

In the Welch test, the value d is assessed by a distribution that depends on the values for the F ratio $\frac{s_1^2}{s_2^2}$, N_1, and N_2, as determined by tables (see Pearson and Hartley, 1966) or provided by the computer program. It may be performed online at http://www.graphpad.com/quickcalcs/ttest1.cfm.

The Satterthwaite method uses the same value of d, but uses the t distribution with the df being not $N_1 + N_2 - 2$ but rather

$$\nu = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}\right)^2}{\frac{\left(\frac{s_1^2}{N_1}\right)^2}{N_1 - 1} - \frac{\left(\frac{s_2^2}{N_2}\right)^2}{N_2 - 1}}.$$

Because this will usually not be an integer, the next smallest integer is used. There is no need to compare the two variances before doing the test, and many authorities prefer to perform the test routinely.

As an example, return to the peanut data tested in unpaired groups. The two means are 57.9 and 55.9, and the variances are 31.21 and 22.48, with $N = 10$ in each group. Then d becomes

$$d = \frac{57.9 - 55.9}{\sqrt{\left(\frac{31.21}{9} + \frac{22.48}{9}\right)}} = \frac{2}{2.44} = 0.82 \text{ (as before).}$$

$$v = \frac{\left(\frac{31.21}{10} + \frac{22.48}{10}\right)^2}{\frac{\left(\frac{31.21}{10}\right)^2}{9} - \frac{\left(\frac{22.48}{10}\right)^2}{9}} = \frac{28.83}{1.08 + 0.56} = 17.57 \text{ and so P} = 0.40.$$

The probability of rejecting the null hypothesis of 0.40 is not very different from the value of 0.43 obtained when the differences between the variances were ignored. That is because these two variances are not significantly different; the one aberrant measurement had more effect on the mean difference than on the variances. Some online tests provide the option for using unequal variances: http://studentsttest.com/, http://in-silico.net/statistics/ttest, http://graphpad.com/quickcalcs/ttest1.cfm, http://vassarstats.net/, and http://www.quantitativeskills.com/sisa/statistics/t-test.htm.

In general, the t-test is robust and tolerates moderate departures from the basic requirements. Nonnormality of the distribution is more serious than differences in variances, and it is worse to have the larger variance associated with the smaller group than with the larger group. Finally, lack of normality or inequality of variances is much worse for small than large samples, and, if practical, samples with over 15 in each group are needed to minimize Type I errors (Ramsey, 1980). In addition, nonnormality and unequal variances greatly diminish the power of the t-test (Rosner, 1995).

Sample Size for Unpaired Test

Arguments similar to those used in Chapter 11 apply to the two-sample t-test. Online calculations may be done at http://www.stat.ubc.ca/~rollin/stats/ssize/n2.html, http://www.graphpad.com/quickcalcs/ttest1.cfm, http://www.danielsoper.com/statcalc3/calc.aspx?id=47, and http://www.sample-size.net.

Conclusion

Statistical testing is valuable in emphasizing the variability of the measurements. We can cautiously draw conclusions from the tests as long as we use them in support of reasonable

biological hypotheses, allow for the possibility of Type II errors, and make sure that any experiment has sufficient power to allow sensible conclusions to be drawn. This does not mean that we discard unexpected results, but merely that these need stronger confirmation.

Nonparametric or Distribution Free Tests

Parametric tests such as the t-test lose efficiency, sometimes drastically, when the distributions are severely nonnormal because of skewing, outliers, kurtosis, or grossly unequal variances. They can be replaced by several robust tests that are referred to as distribution free or nonparametric tests. The two tests to be described below, when applied to normal distributions, are very efficient. The relative efficiency of two tests is determined by the ratio of the sample sizes needed to achieve the same power for a given significance level and a given difference from the null hypothesis (Healy, 1994). When the two distributions are normal, the distribution free tests are about 95% as efficient as the t-test (Wilcox, 1996; Siegal and Castellan, 1988). When the distributions are grossly abnormal, then the distribution free tests have greater efficiency. The main nonparametric test to replace the paired t-test is the Wilcoxon-signed rank test, and the major replacement for the unpaired t-test is the Mann—Whitney U test.

The Wilcoxon-Signed Rank Test

As for the paired t-test, the paired values for each group are set out, and the difference between each pair is calculated. Then these differences are ranked from the smallest (1) to the biggest (N), ignoring the sign of the difference; **differences of zero are not ranked**. Any tied ranks are averaged. Once the ranking has been done, the negative signs are put back, and the sums of the negative and the positive ranks are calculated.

The basic theory is that if the paired sets are drawn from the same population, then there will be some small, some medium, and some large positive differences, and approximately the same number of small, medium, and large negative differences. Therefore, the sums of the negative and positive ranks should be about the same. If we can calculate the sampling distribution of T, the smaller of these two sums (positive vs negative) for any value of N, then we can determine if one of those sums is so much smaller than the other that the null hypothesis should be rejected. Although the test is part of statistical computer packages, an example to illustrate the principle is shown in Table 22.2.

The three smallest differences are each 2. Because these account for the first 3 ranks, in the fourth column average them $\frac{1+2+3}{3}$ and assign each a rank of 2. The next value, 3, occupies the next rank, the fourth rank. Similarly the two 6 differences occupy ranks 7 and 8, but being equal are each assigned a rank of 7.5. The fifth column shows the same ranks, but now the positive and negative ranks are identified. The sums of the positive and negative ranks are different. If the null hypothesis were true, the two sums should be similar. Calculations or tables show that the probability of such a

Table 22.2 Wilcoxon-signed rank test used for paired peanut data

Raw peanuts R	Roasted peanuts P	Difference D	Rank	Signed rank
61	55	6	7.5	+7.5
60	54	6	7.5	+7.5
56	47	9	9	+9
63	59	4	5	+5
56	51	5	6	+6
63	61	2	2	+2
59	57	2	2	+2
56	54	2	2	+2
44	63	−19	10	−10
61	58	3	4	+4
		$\Sigma X_i = 20$ $\overline{X_D} = 2$		$\Sigma+ = 45$ $\Sigma- = 10$

difference in signed rank sums based on listing all the possible combinations of the signed ranks for the sample size studied is 0.082. This is not at the conventional level of statistical significance, but is much closer to it than the probability of 0.43 obtained by the paired test. If the Wilcoxon test is done for the nine pairs after excluding the one aberrant pair, the probability from the Wilcoxon test is 0.0039, not as striking as the 0.0006 from the paired t-test but still a good reason to reject the null hypothesis.

The Wilcoxon-signed rank test does not give any result if N is ≤ 5.

If the number of pairs is >10, T approximates a normal distribution and we do not need special tables to test the null hypothesis. Because the sum of the first N numbers is $\frac{N(N+1)}{2}$, if the sums of the negative and positive ranks were equal, each would be $\frac{N(N+1)}{4}$. Therefore, test the difference between the observed and expected value of T $\left(T - \frac{N(N+1)}{4}\right)$ by dividing by the standard deviation of T

$$\sigma_T = \sqrt{\frac{N(N-1)(2N+1)}{24}}$$

Therefore, use the z table to test.

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N-1)(2N+1)}{24}}}$$

Pratt (1959) pointed out that ignoring the zeros may produce paradoxical probabilities, and proposed ranking the differences including the zeros, then dropping the zeros when summing the negative and positive ranks, and using the tables of probabilities for the total number of observations, including the zeros.

Online calculations can be done at http://www.socscistatistics.com/tests/signedranks/Default2.aspx, http://www.sdmproject.com/utilities/?show=Wilcoxon, and http://vassarstats.net/wilcoxon.html.

Problem 22.3
Perform a Wilcoxon test on the data from Problem 22.1.

The Sign Test

This is a simpler and less powerful version of the Wilcoxon test. It is used when the data are ordinal or nominal (or categorical). For example, 11 observers rate two different bacteriological stains A and B for clarity. Each observer records a preference: if A is better than B, the result is +, and if B is better than A, the result is −. On the null hypothesis that there is no difference between the two stains, there should be as many negative results as positive results. If there are more of one sign than the other, then the departure from the null hypothesis can be tested using the binomial distribution for $p = 0.5$. Hypothetical data are presented in Table 22.3.

The results show a preference for A in 9 out of 11 trials. On the null hypothesis of no difference between the stains we expect 5 or 6. The question then is to determine if 9 is an unusual event if the null hypothesis is true. To see how this decision is reached, examine Figure 22.6.

Adding 9, 10, and 11 together gives a probability of 0.0328, and we would probably reject the null hypothesis. This is the probability of one tail of the distribution. However, a finding of 0, 1, or 2 would also lead to a rejection of the null hypothesis. Because the designation of + or − is arbitrary, finding either 0, 1, 2, 9, 10, or 11 + would occur with a probability of $2 \times 0.0328 = 0.0656$. This is still evidence against the null hypothesis, although not quite as strong.

Table 22.3 Sign test

Observer	Result
1	+
2	+
3	−
4	+
5	+
6	+
7	−
8	+
9	+
10	+
11	+

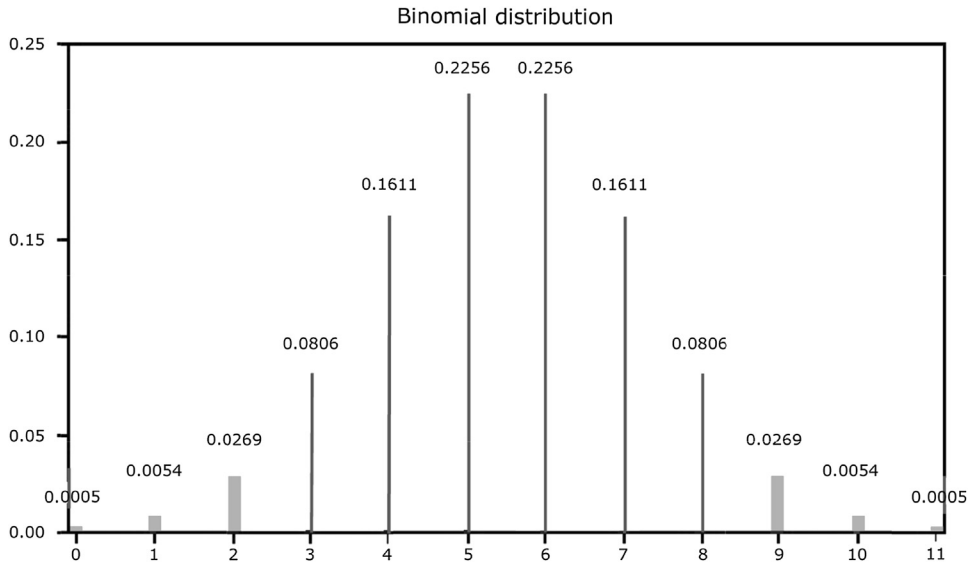

Figure 22.6 Binomial distribution for $N = 11$, P = 0.50.

Using this test in place of the Wilcoxon-signed rank test loses the information provided by the size of the differences and so produces a less powerful test. The test can be done easily with online programs http://www.graphpad.com/quickcalcs/binomial1.cfm, and http://www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html.

The Mann–Whitney U Test

In 1945, Wilcoxon developed a ranking test for comparing the positions of two distributions; he called the statistic T that was the sum of the ranks in the smaller group. Two years later, Mann and Whitney extended the theory and they called their statistic U. The two statistics are interconvertible.

$$U = N_1 N_2 - \frac{1}{2} N_1 (N_1 + 1) - T$$

Requirements

1. Each sample is drawn at random from its own population.
2. The values are independent of each other within each sample, and the two samples are independent of each other.
3. The measurement scale is at least ordinal.

The test can be done easily. Consider two groups, A with n_1 members and B with n_2 members, each drawn at random from the same distribution. Because the two sets of measurements come from the same distribution, pool them into a single set and then

rank them from the smallest, with a rank of 1, to the largest with a rank of $n_1 + n_2$. Then add up the ranks in each group separately. Intuitively, each group should have similar proportions of low ranks, medium-sized ranks, and high ranks, so that if n_1 and n_2 are equal the sums of the ranks in the two groups should be equal or nearly so. If n_2 is twice as big as n_1, for example, then the sum of the ranks of n_2 should be about twice as large as the sum of ranks from n_1. The more the sums of ranks in the two groups differ from the expected proportion, the less likely is it that the null hypothesis that they come from the same distribution is true. The possible combinations can be enumerated and the probability of any discrepancy between the sums in the two data sets can be ascertained.

The critical values of rank sums for possible combinations of n_1 and n_2 have been calculated, are given in standard tables, and are available in standard computer programs.

If N is >20 in either group use the normal approximation

$$z = \frac{U - \frac{N_1 N_2}{2}}{\sqrt{\frac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}.$$

This test is available in computer programs, but an example will clarify the method (Table 22.4).

The probability of such a discrepancy is 0.2240, and suggests that we cannot reject the null hypothesis. For the unpaired t-test, the probability of rejecting the null hypothesis was 0.4336. The Mann–Whitney test is closer to rejecting the null hypothesis.

When two or more values are tied, the sum of ranks is modified by averaging the tied ranks. For example, the fourth and fifth measurements are each 54, so allocate them each a rank of 4.5. (If the tied measurements are in the same group, it does not matter if we average their ranks or not, because the sum of ranks 4 and 5 is the same as the sum of ranks 4.5 and 4.5. If the tied ranks are in different groups, then the ranks must be averaged.) Average all the sets of tied measurements. The value of T is usually corrected for ties

Table 22.4 Mann–Whitney test using peanut data

Raw peanuts R	Rank R	Roasted peanuts P	Rank P
61	16	55	6
60	14	54	4.5
56	8	47	2
63	19	59	12.5
56	8	51	3
63	19	61	16
59	12.5	57	10
56	8	54	4.5
44	1	63	19
61	16	58	11
	ΣR = 121.5		ΣP = 88.5

but the correction factor is usually unimportant; there are several types of correction possible (Conover, 1980; Krauth, 1988; Rosner, 1995). The whole test can be done online at http://www.socscistatistics.com/tests/mannwhitney/Default2.aspx, http://www.wessa.net/rwasp_Reddy-Moores%20Wilcoxon%20Mann-Witney%20Test.wasp, and http://vassarstats.net/utest.html. A clear description of how these tests are derived and used appears in several publications (LaVange and Koch, 2006; Noether, 1976; Bland, 1995).

Problem 22.4
Perform a Mann–Whitney test on the data from Problem 22.1.

ADVANCED CONCEPTS

Comparing Two Coefficients of Variation

Sometimes we are interested in comparing the coefficients of variation of two groups. This can be done in two ways. If the logarithms of the data are normally distributed, then the ratio

$$F = \frac{s^2_{\log X_1}}{s^2_{\log X_2}}$$

can be evaluated from standard F tables. If the data are normally distributed, however, their logarithms will not be normally distributed, so use

$$Z = \frac{\mathrm{CV}_1 - \mathrm{CV}_2}{\sqrt{\left(\frac{\mathrm{CV}_p^2}{N_1-1} + \frac{\mathrm{CV}_p^2}{N_2-1}\right)\left(0.5 + \mathrm{CV}_p^2\right)}},$$

where $\mathrm{CV}_p = \frac{\mathrm{CV}_1(N_1-1)+\mathrm{CV}_2(N_2-1)}{N_1+N_2-1}$ is the weighted mean of the two coefficients of variation CV_i (Zar, 2010).

The Paired t-Test Implies an Additive Model

The paired t-test implies the model

$$X_{i2} = X_{i1} + \alpha + \varepsilon_i \text{ or } X_{i2} - X_{i1} = \alpha + \varepsilon$$

where X_{i1} and X_{i2} are the two members of each pair, α is the mean difference between them (the effect of the treatment), and ε_i is the error associated with each difference. On the other hand, in any given study the relationship might be multiplicative:

$$X_{i2} = \alpha X_{i1} + \varepsilon_i.$$

In this model, the effect of the treatment is to increase each value for X by a factor α. The difference between these two models is unimportant if all the X_i values are close together, but assumes importance if X_i varies widely. For example, Table 22.5 shows data based on hypothetical norepinephrine concentrations (pg/ml) before and after dialysis.

The initial data are in the first column, and the final values for the additive model are in the second column. The differences between the two are similar for each pair (column 3), with a mean difference of 31.11 and a narrow standard deviation; it is reasonable to reject the null hypothesis that this difference is not significantly different from a mean difference of zero. If we postulate a multiplicative model with about a 10% decrease, as shown in column 5, then the actual decreases vary widely, with a mean of 36.22 but a standard deviation of 32.03, which suggests a skewed distribution as well as a difficulty in rejecting the null hypothesis. On the other hand, taking the ratio of the two gave a mean of 0.896 with a very small standard deviation and standard error, making it easier to reject the null hypothesis that the ratio was 1.

Motulsky (2009) recommended that instead of setting out the data to display proportional differences, as shown in Table 22.5, the two members of the pair should be set out as a ratio of $\frac{treated}{control}$. The disadvantage to working with ratios is that they are asymmetric; below 1 the range can be only from 1 to 0, where above 1 the ratio can in theory be

Table 22.5 Effects of additive and multiplicative models

	Fixed additive model			Fixed multiplicative model		
Initial	Final	Fixed difference	Final proportional difference	Proportional % difference	Actual difference	Final/ initial ratio
1	2	3	4	5	6	7
847	817	30	762	10	85	0.90
794	766	28	699	12	95	0.88
439	400	39	399	9	40	0.91
254	220	34	231	9	23	0.91
245	218	27	218	11	27	0.89
174	143	31	151	13	23	0.87
140	112	28	129	8	11	0.92
119	87	32	105	12	14	0.88
81	50	31	73	10	8	0.90
		Mean 31.11 sd 3.69 se 1.23		Mean 10.44	Mean 36.22 sd 32.03 se 10.68	Mean 0.896 sd 0.017 se 0.0056

any value >1. To overcome this, he recommended using the logarithms of the ratios. A zero value means no change, a negative value means a decrease, and a positive value means an increase. If this is done for the initial and final data in Table 22.6, the results in the final column are reproduced, and are interpreted as showing that the ratio is significantly below 1 so that there has been a consistent decrease from initial to final measurements.

Confidence Limits for Medians

We usually perform nonparametric tests either for ordinal data or for nonnormally distributed ratio numbers. For the latter, the descriptive summaries include the median. Sometimes we may want to determine the confidence limits for the median, or the difference between two medians. Gardner and Altman (1995) describe a conservative method for these calculations.

The $100(1-\alpha)\%$ confidence interval for the population interval requires calculating the lower (R_L) and upper (R_U) ranks, assuming the data are arranged in order from smallest to largest:

$$R_L = \frac{N}{2} - \left(z_{1-\alpha/2}\sqrt{\frac{N}{2}} \right)$$

$$R_U = 1 + \frac{N}{2} + \left(z_{1-\alpha/2}\sqrt{\frac{N}{2}} \right).$$

For the data presented in Tables 3.3, 3.4, and 3.16, there were $N = 53$ measurements. Then for 95% confidence limits,

$$R_L = \frac{53}{2} - \left(1.96\sqrt{\frac{53}{2}} \right) = 16.4$$

$$R_U = 1 + \frac{53}{2} + \left(1.96\sqrt{\frac{53}{2}} \right) = 37.6.$$

Return to the array of measurements and locate the 16th and the 38th measurements as the nearest integers. These are 1.18 and 1.32, which are the required limits of the median.

To set limits for a percentile, modify the above equation to

$$R_L = N_p - \left(1.96\sqrt{N_p(1-p)} \right)$$

$$R_U = 1 + N_p + \left(1.96\sqrt{N_p(1-p)} \right),$$

where p is the percentile. Thus for the 75th percentile, $p = 0.75$, and the 95% confidence limits are

$$R_L = 53 \times 0.75 - \left(1.96\sqrt{53 \times 0.75(1 - 0.75)}\right) = 33.6$$

$$R_U = 1 + 53 \times 0.75 + \left(1.96\sqrt{53 \times 0.75(1 - 0.75)}\right) = 46.93.$$

The measurements corresponding to ranks 34 and 47 are 1.20 and 1.53, respectively.

Online calculations can be performed at http://www.mountaingoatsoftware.com/tools/velocity-range-calculator.

Calculating the confidence limits of the difference between two medians is not often wanted, but can be done by the bootstrap technique (Chapter 37) or by the online program at http://www.wessa.net/rwasp_bootstrapplot1.wasp.

Ranking Transforms

If the measurements in the two groups are ranked but then a classical parametric t-test is done on the ranks, a robust test results that can be used for abnormal distributions (Iman, 1974). This approach is supported by Healy (1994).

The Meaning of the Mann–Whitney Test

An issue that causes confusion is what the Mann–Whitney test indicates. Because the measurements have been turned into ranks, the test cannot allow us to compare the means of the two distributions. Does it compare medians? Consider two groups A and B of equal size with the following measurements (Table 22.6):

The sum of ranks for group A is $1 + 2 + 3 + 10 + 11 + 12 = 39$, and the sum of ranks for B is $4 + 5 + 6 + 7 + 8 + 9 = 39$. The null hypothesis of equality of the sums of ranks of the two groups is obvious, but what is equal? The median of group A is 8.55 and the median of group B is 7.85. These are fairly close to each other. If, however, we make the three largest measurements in group A 127, 158, and 192, the sums of ranks are unaltered, but the median of group A becomes 65.7, much greater than the median of group B. Therefore, the test does not compare medians, although with less dispersed distributions it does serve this purpose. What the test actually does is to compare the equality of mean ranks, and thus, by inference, of the distributions. However, like all tests it must not be used without thought, as the example above shows. Many texts point

Table 22.6 Two distributions compared

A: 1.3, 2.7, 4.4,		12.7, 15.8, 19.2
B:	5.9, 7.0,7.2, 8.5, 9.0, 11.4	

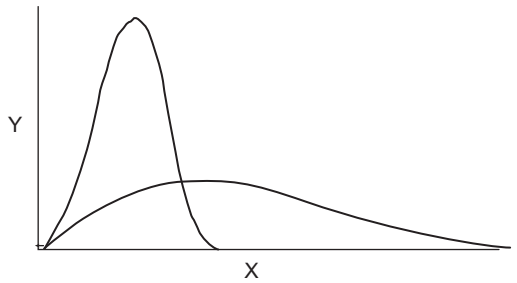

Figure 22.7 Comparison between two distributions.

out that the Mann–Whitney test is most useful when the only difference between the two groups is a measure of location. More formally, that the unspecified distributions of two groups X and Y differ only in location, such that $X = Y + \mathrm{d}$, where d is a constant. This may be true for some distributions, but not for others. For example, Healy (1994) has pointed out that many biochemical and endocrine distributions share a common start, for example, a low or zero value, but then the control and experimental groups differ in shape (Figure 22.7).

Under these circumstances, the curves differ by more than location.

APPENDIX

1. Often a publication provides the data as the mean $(\overline{X})$, standard deviation (s), and number of observations (N), but does not give the individual measurements. You may want to perform an unpaired t-test that compares your own data set with the published data, but how can you do this without having all the data? Let us assume that the published data for cerebral arterial pulsatility in normal neonates has mean 34.3, standard deviation 4.1, and 37 observations. Your own data in neonates with heart disease have values of 51.6, 9.2, and 22, respectively.

To calculate t we need the differences between the two means, which we have, and the standard deviation of that difference which we need to obtain. To derive the required values, you can make use of known relationships:

a. $s^2 = \frac{\sum (X_i - \overline{X})^2}{N-1}$ and so $\sum (X_i - \overline{X})^2 = (N-1)s^2$ (Columns 3 and 6 below)

b. Then the pooled variance can be calculated from Eqn (19.13), using total in column 7: $s_p^2 = \frac{2382.57}{57} = 41.80$

1	2	3	4	5	6	7
Group	N	$N-1$	$\overline{X}$	s	s^2	$(N-1)s^2$
1	37	36	34.3	4.1	16.81	605.16
2	22	21	51.6	9.2	84.64	1777.44
Total	59	57	Difference = 17.3			2382.6

Based on these calculations,

$$s_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{41.8}{37}+\frac{41.8}{22}} = 1.74, \text{ and}$$

$$t = \frac{51.6-34.3}{1.74} = 9.94. \text{ Degrees of freedom } 57, \ P < 0.0001.$$

As you will realize, this simple arithmetic can be automated, and it can be obtained by entering the means, standard deviations, and number of measurements in a computer program. A simple online program can be obtained at http://graphpad.com/quickcalcs/ttest1.cfm?Format=SD. Should you wish to calculate values for $\sum X_i$, multiply the mean by N, and to obtain $\sum X_i^2$ just add $\frac{\left(\sum X_i\right)^2}{N}$ to $(N-1)s^2$.

Problem 22.5

To make sure that you understand the unpaired t-test, try the following problem, and then check the results with the online application.

Group	Number	Mean	Standard Deviation
1	197	14.7	2.9
2	39	19.3	3.3

You should get $t = 6.92$.

REFERENCES

Bland, M., 1995. An Introduction to Medical Statistics. Oxford University Pess, Oxford.

CAST, 1989. Preliminary report: effect of encainide and flecainide on mortality in a randomized trial of arrhythmia suppression after myocardial infarction. The Cardiac Arrhythmia Suppression Trial (CAST) investigators. N. Engl. J. Med. 321, 406–412.

Conover, W.J., 1980. Practical Nonparametric Statistics. John Wiley & Sons, New York.

Gardner, M.J., Altman, D.G., 1995. Statistics with Confidence—Confidence Intervals and Statistical Guidelines. British Medical Journal, London.

Hays, W.H., 1973. Statistics for the Social Sciences. Holt, Rinehart and Winston, Inc., New York.

Healy, M.J.R., 1994. Statistics from the Inside. 12. Non-normal data. Arch. Dis. Child. 70, 158–163.

Iman, R.L., 1974. Use of a t-statistic as an approximation to the exact distribution of the Wilcoxon signed ranks test statistic. Commun. Stat. 3, 795–806.

Kasuya, E., 2001. Mann–Whitney U test when variances are unequal. Anim. Behav. 61, 1247–1249.

Krauth, J., 1988. Distribution-free Statistics. An Application-oriented Approach. Elsevier, Amsterdam.

Lavange, L.M., Koch, G.G., 2006. Rank score tests. Circulation 114, 2528–2533.

Mitchell, H.H., Burroughs, W., Beadles, H.P., 1936. The significance and accuracy of biological values of proteins computed from nitrogen metabolism data. J. Nutr. 11, 257–274.

Motulsky, H.J., 2009. Statistical Principles: The Use and Abuse of Logarithmic Axes [Online]. Available: http://www.graphpad.com/faq/file/1487logaxes.pdf.

Moyé, L.A., 2000. Statistical Reasoning in Medicine. The Intuitive P-Value Primer. Springer-Verlag, New York.

Neuhäuser, M., 2002. Two-sample tests when variances are unequal. Anim. Behav. 63, 823–825.

Noether, G.E., 1976. Introduction to Statistics. A Nonparametric Approach. Houghton Miflin Co., Boston.
Pearson, E.S., Hartley, H.O., 1966. Biometrika Tables for Statisticians. Cambridge University Press, Cambridge.
Pratt, J.W., 1959. Remarks on zeros and ties on the Wilcoxon signed rank procedures. J. Stat. Assoc. 54, 655–667.
Ramsey, P.H., 1980. Exact type I error rates for robustness of student's t test with unequal variances. J. Educ. Stat. 5, 337–349.
Rosner, B., 1995. Fundamentals of Biostatistics. Duxbury Press, New York.
Ruskin, J.N., 1989. The cardiac arrhythmia suppression trial (CAST). N. Engl. J. Med. 321, 386–388.
Ruxton, G.D., 2006. The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney U test. Behav. Ecol. 17, 688–690.
Siegal, S., Castellan Jr., N.J., 1988. Nonparametric Statistics for the Behavioral Sciences. McGraw-Hill, New York.
Wilcox, R.R., 1996. Statistics for the Social Sciences. Academic Press, San Diego.
Zar, J.H., 2010. Biostatistical Analysis. Prentice Hall, Upper Saddle River, NJ.