CHAPTER SIX

ANALYSIS OF CONTINUOUS VARIABLES: COMPARING MEANS

In the last chapter, we addressed the analysis of discrete variables. Much of the statistical analysis in medical research, however, involves the analysis of **continuous variables** (such as cardiac output, blood pressure, and heart rate) which can assume an infinite range of values. As with discrete variables, the statistical analysis of continuous variables requires the application of specialized tests. In general, these tests compare the means of two (or more) data sets to determine whether the data sets differ significantly from one another.

There are four situations in biostatistics where we might wish to compare the means of two or more data sets. Each situation requires a different statistical test depending on whether the data is normally or non-normally distributed about the mean (Figure 6-1).

1. When we wish to compare the observed mean of a data set with a standard or normal value, we use the **test of hypothesis** or the **sign test**.
2. When we wish to determine whether the mean in a single patient group has changed as a result of a treatment or intervention, the **single-sample, paired t-test** or **Wilcoxon signed-ranks test** is appropriate.
3. When we are evaluating the means of two different groups of patients, we use the **two-sample, unpaired t-test** or **Wilcoxon rank-sum test**.
4. When multiple comparisons are required to determine how one therapy differs from several others, we employ **analysis of variance** (**ANOVA)**. The use of multiple comparisons is discussed in the next chapter.

If we wish to compare: ↓	**Normally Distributed Data**	**Non-Normally Distributed Data**
A mean with a normal value	**Test of Hypothesis**	**Sign Test**
Paired observations within a single patient group	**Single-sample, paired t-test**	**Wilcoxon signed-ranks test**
Means from two different patient groups	**Two-sample, unpaired t-test**	**Wilcoxon rank-sum test**
Multiple patient groups	**ANOVA**	**Nonparametric ANOVA**

Figure 6-1: Analysis of Continuous Variables

COMPARING MEANS

There are three factors which determine whether an observed sample mean is different from another mean or normal value. First, the larger the difference between the means, the more likely the difference has not occurred by chance. Second, the smaller the variability in the data about the mean, the more likely the observed sample mean represents the true mean of the population-at-large. The **standard deviation** represents the variability of the data about the mean. The smaller the standard deviation, the smaller the variability of the data about the mean. Third, the larger the sample size, the more accurately the sample mean will represent the true population mean. The **standard error of the mean** estimates how closely the sample mean approximates the true population mean. As the sample size increases (and approaches the size of the population), the standard error of the mean approaches zero.

THE t DISTRIBUTION: ANALYSIS OF NORMALLY DISTRIBUTED DATA

The **t distribution** is a probability distribution which is frequently used to evaluate hypotheses regarding the means of continuous variables. It is commonly referred to as "**Student's t-test**" after William Gosset, a mathematician with the Guinness Brewery, who in 1908 noted that if one samples a normally distributed bell-shaped population, the sample observations will also be normally distributed (assuming the sample size is greater than 30). Unfortunately, company policy forbade employee publishing and he was forced to use the pseudonym "Student." He named the distribution, "t", and defined a measure of the difference between two means known as the "**critical ratio**" or "**t statistic**" which followed the t distribution:

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

> where $\bar{x}$ = the mean of the sample observations, μ = the mean of the population μ, and se = the standard error of the sample mean (which is equal to the sample standard deviation divided by the square root of the sample size, n).

The t distribution is similar to the standard normal (z) distribution (discussed in Chapter Two) in that it is symmetrically distributed about a mean of zero. Unlike the normal distribution, however, which has a standard deviation of 1, the standard deviation of the t distribution varies with an entity known as the **degrees of freedom**. Since the t-test plays a prominent role in many statistical calculations, degrees of freedom is an important statistical concept. Degrees of freedom is related to sample size and indicates the number of observations in a data set that are free to vary. For example, if we make n observations and calculate their mean, we are free to change only n - 1 of the observations if the mean is to remain the same as once we have done so, we will automatically know the value of the nth observation. The degrees of freedom for a data set are therefore equal to the sample size (n) - 1. Whereas there is only one standard normal (z) distribution, there is a separate t distribution for each possible degree of freedom from 1 to ∞. As with the normal distribution, critical values of the t-statistic can be obtained from t distribution tables (found in any statistics textbook) based on the desired significance level (p-value) and the degrees of freedom.

Appropriate use of the t distribution requires that three assumptions are met. The first assumption is that the observations follow a normal (gaussian or "bell-shaped") distribution; that is, they are evenly distributed about the true population mean. If the observations are not normally distributed, the t-statistic is not accurate and should not be used. As a general rule, if the median differs markedly from the mean, the t-test should not be used. The second assumption is that the **variances** (the standard deviations squared) of the two groups being compared, although unknown, are equal. The third assumption is that the observations occur independently (i.e., an observation in one group does not influence the occurrence of an observation in the other group).

THE t DISTRIBUTION AND CONFIDENCE INTERVALS

In Chapter Three, we saw that **confidence intervals** could be calculated for any mean in order to evaluate how confident we were that our sample mean represented the true population mean. In order to calculate the 95% confidence interval for a mean we used the following equation:

$$95\% \text{ confidence interval} = \text{mean} \pm \text{approximately } 2 \text{ se}$$

Remember that this calculated the approximate confidence interval. In reality, the exact multiplying factor for the standard error of the mean depends on the sample size and degrees of freedom. Using the t distribution, we can calculate the exact 95% confidence interval for a particular mean ($\bar{x}$) and standard deviation (sd) as follows (the critical value of t is obtained from a t distribution table based on the desired significance level and degrees of freedom present):

$$95\% \text{ confidence interval} = \bar{x} \pm t \cdot se = \bar{x} \pm t \cdot \frac{sd}{\sqrt{n}}$$

For example, suppose we studied 87 intensive care unit (ICU) patients and found that the mean ICU LOS (length of stay) was 6.7 days with a standard deviation of 19.1 days. If we wish to determine how closely our observed mean ICU LOS approximates the true mean LOS for all ICU patients with 95% confidence, we would determine the critical value of t for a significance level of 0.05 (5% chance of a Type I error) and 86 (87-1) degrees of freedom. The critical value of t for these parameters, as obtained from a t distribution table, is 1.99. Calculating the confidence interval as above we obtain:

$$95\%\ \text{confidence interval} = 6.7 \pm 1.99 \cdot \frac{19.1}{\sqrt{87}} = 6.7 \pm 4.07\ \text{days}$$

Therefore, we can be 95% confident that, based on our study data, the interval from 2.63 to 10.77 days contains the true mean LOS for all ICU patients.

ANALYSIS OF NORMALLY DISTRIBUTED CONTINUOUS VARIABLES

The t-test is commonly used in statistical analysis. It is an appropriate method for comparing two groups of continuous data which are both normally distributed. The most commonly used forms of the t-test are the **test of hypothesis**, the **single-sample, paired t-test**, and the **two-sample, unpaired t-test**.

TEST OF HYPOTHESIS

Suppose we know from previous experience that the normal mean LOS for all ICU patients in our hospital is 3.2 days and we wish to compare this to our study mean of 6.7 days to determine whether these two means differ significantly. To do so, we would use a form of the t-test known as the **test of hypothesis** in which we compare a single observed mean with a standard or normal value. For this comparison, the null and alternate hypotheses would be:

Null hypothesis: the normal and study ICU LOS are not different
Alternate hypothesis: the normal and study ICU LOS are different

Using a significance level of 0.05 and n-1 (86) degrees of freedom we noted above that the critical value of t is 1.99. The t statistic would then be calculated for the test of hypothesis as follows:

$$t = \frac{\bar{x} - \mu}{sd / \sqrt{n}} = \frac{6.7 - 3.2}{19.1/\sqrt{87}} = \frac{3.5}{2.05} = 1.7$$

where $\bar{x}$ = the study ICU LOS of 6.7 days, μ = the normal ICU LOS of 3.2 days, sd = the standard deviation of 19.1 days*, and n = 87

* Remember that one of the inherent assumptions in using the t-test is that the sample and population have the same variance (and therefore the same standard deviation)

Since 1.7 does not exceed the critical value of 1.99, we cannot reject the null hypothesis and must conclude that the means, although different, do not differ significantly. This is consistent with what we found using the 95% confidence interval in which we found that there was a 95% probability that the true population mean (3.2 days in this example) was between 2.63 and 10.77 days. If the normal ICU LOS in our hospital was actually 2.4 days (instead of 3.2 days), the critical value of t obtained from the above equation would be (6.7-2.4)/2.05 or 2.10. Since a t-statistic of 2.10 exceeds the critical value of 1.99, we would accept our alternate hypothesis and state that our study ICU LOS was significantly greater than the normal ICU LOS with at least 95% confidence. We would also know that this is true by realizing that an ICU LOS of 2.4 days lies outside of our 95% confidence interval of 2.63 to 10.77 days. Thus, t-tests and confidence intervals are just two different methods of determining the same statistical information.

The broadness of the above 95% confidence interval demonstrates that there is considerable variability in the data. If the sample variability was less (i.e., the standard deviation was smaller), the 95% confidence interval would be narrower and we would be more likely to find that our two means were significantly different. For example, had out standard deviation been only 5 days, instead of 19.1 days, t would have been 6.53 rather than 1.7. This is clearly greater than the critical value of 1.99 necessary to accept the alternate hypothesis and state that a significant difference exists with at least 95% confidence. Thus, the greater the variability in the sample data, the greater the difficulty in proving that the sample mean is different from a normal value. This also holds true when one is comparing two groups of patients. The more variable the observations in each group, the harder it is to prove that they are different.

SINGLE-SAMPLE, PAIRED T-TEST

If we wish to evaluate an intervention or treatment and have paired observations (such as pre- and post-intervention) on a single group of patients, we can use a **single-sample, paired t-test** to determine whether the paired observations are significantly different from one another. In the test of hypothesis, we compared a single mean with a standard value. In calculating the paired t-test, because we are interested in the differences between pairs of data, the mean of the differences between the pairs replaces the mean of the observations. Otherwise, the calculation of the t-statistic remains the same.

Consider the following data on arterial oxygen tension (PaO_2) measurements in 10 patients before and after the addition of positive end-expiratory pressure (PEEP). Our null and alternate hypotheses would be as follows:

Null hypothesis: there is no change in PaO_2 with the addition of PEEP.
Alternate hypothesis: there is a change in PaO_2 with the addition of PEEP.

Patient	Pre-PEEP PaO_2 (torr)	Post-PEEP PaO_2 (torr)	Difference
1	55	70	15
2	48	62	14
3	50	68	18
4	67	80	13
5	72	85	13
6	68	78	10
7	42	65	23
8	55	69	14
9	61	86	25
10	73	91	18
mean	59.1	75.4	16.3
sd	10.7	9.9	4.7

Using a significance level of 0.05 and n-1 (9) degrees of freedom, the critical value of t, obtained from a t distribution table, is 2.26. The t statistic for the single-sample, paired t-test is then calculated as follows:

$$t = \frac{\text{mean of the differences}}{\text{standard error of the mean of the differences}} = \frac{16.3}{4.7/\sqrt{10}} = 10.9$$

Since 10.9 is greater than the critical value of 2.26 required to identify a difference with a significance level of 0.05 (95% confidence of not committing a Type I error), we would reject the null hypothesis of no difference and state that the addition of PEEP significantly increases PaO_2. Since our calculated critical value of 10.9 markedly exceeds 2.26, the actual significance level (or p-value) is really much smaller than 0.05. In fact, the actual significance level associated with a critical value of 10.9 and 9 degrees of freedom is < 0.0001.

TWO SAMPLE, UNPAIRED T-TEST

One of the most common uses of the t-test is to compare observations from two separate groups of patients to determine whether the two groups are significantly different with respect to a particular variable of interest. To make such a comparison, we use a **two-sample, unpaired t-test**. As with the other forms of the t-test, it is assumed that the data are normally distributed.

Consider the following data on ICU charges for 10 age and procedure matched patients from the Preoperative Evaluation study. We wish to determine whether performing a preoperative evaluation on the patient (placement of a pulmonary artery catheter the night before surgery) results in higher ICU charges than no preoperative evaluation (pulmonary artery catheter placement in the operating room prior to surgery). Note that the mean and median of each group is similar, and that the variances are also similar. The first two assumptions of the t-test are therefore met and the use of a t-test to analyze this data is appropriate.

	Study Patients	Control Patients
	$5,199	$21,929
	$25,451	$8,065
	$9,774	$5,832
	$3,995	$28,213
	$14,226	$16,039
	$30,415	$1,995
	$3,919	$20,877
	$23,936	$25,374
	$20,029	$2,544
	$15,533	$15,257
mean	$14,948	$14,613
median	$13,380	$15,648
sd	$9,580	$9,550
variance	$91,776,400	$91,202,500

Null hypothesis: preoperative evaluation does not affect ICU charges
Alternate hypothesis: preoperative evaluation either increases or decreases ICU charges.

The degrees of freedom in a two-sample t-test are calculated as (n_1 + n_2 -2) since we are free to vary only n - 1 of the observations in each of the two groups. Assuming a significance level of 0.05 and 18 degrees of freedom (10+10-2), the critical value of t is 2.1. The t statistic for a two-sample, unpaired t-test is given by the following modification of the t-test equation:

$$t = \frac{\overline{x}_1 - \overline{x}_2}{s_{r_p}\sqrt{1/n_1 + 1/n_2}}$$

where s_{r_p} = the **pooled standard deviation** for both groups which is calculated as follows:

$$s_{r_p} = \sqrt{\frac{(n_1-1)(sd_1)^2 + (n_2-1)(sd_2)^2}{n_1 + n_2 - 2}}$$

Thus, we calculate the t statistic for the above data as:

$$s_{r_p} = \sqrt{\frac{(9)(9580)^2 + (9)(9550)^2}{18}} = 9565$$

$$t = \frac{(14948 - 14613)}{9565} = \frac{335}{9565} = 0.035$$

Since t does not exceed the critical value of 2.1 necessary to detect a significant difference with 95% confidence, we must accept the null hypothesis and state that performing a preoperative evaluation prior to the patient's operation does not result in a significant difference in ICU charges. Note that in the example above, our alternate hypothesis determined that we needed to perform a **two-tailed** test. We do not know *a priori* whether preoperative evaluation will either increase or decrease ICU charges and thus need to investigate both possibilities. If we were only interested in identifying whether preoperative evaluation increased ICU charges, our alternate hypothesis would reflect that and be stated as:

Alternate hypothesis: preoperative evaluation increases ICU charges

We would therefore only need to perform a **one-tailed** test as we would only be interested in identifying differences in one direction. Remember that a two-tailed test requires more statistical evidence to detect a significant difference than does a one-tailed test because the two-tailed test must look in both directions (i.e., are charges increased? and are charges decreased?). The critical value of t reflects this difference in that the critical value necessary to detect a difference for a one-tailed test is smaller than that for a two-tailed test. Also remember that if we only perform a one-tailed test (i.e., preoperative evaluation increases ICU charges), but preoperative evaluation actually decreases ICU charges, we will commit a Type II error because our alternate hypothesis will not account for this possibility. We will thus be forced to accept the null hypothesis and conclude that there is no difference when, in fact, there really is. In common practice, therefore, most researchers perform the statistically more strenous two-tailed test in order to be sure of detecting any difference that may be present and avoiding Type II errors.

TRANSFORMATION OF NON-NORMALLY DISTRIBUTED DATA

As we have noted, one of the important assumptions in the use of t-tests is that the data is normally distributed. This is not always the case, however. Data may inherently not be normally distributed about the mean, or significant outliers may be present which skew the data in one direction or another. If this is true, the use of t-tests is inappropriate. In the situation in which the observations in a data set do not follow a normal distribution there are two options: we can either "**transform**" the data to a scale which is normally distributed, or we can use **non-parametric methods** of analysis which do not assume a normal distribution.

There are several methods for transforming non-normally distributed data to a normal distribution. **Logarithmic transformations** involve calculating the logarithm or using the square root of each data point to create a “transformed” data set. This will frequently minimize the effect of outlying data points and convert the data to a more normal distribution. Statistical tests utilizing the t distribution can then be used on the transformed, and now normally distributed, data to evaluate the observations. **Rank transformations** involve ordering the observations from smallest to largest and assigning a rank to each observation from 1 to n beginning with the smallest observation. Rank-ordering of data is the basis of the non-parametric statistical methods.

One potential problem with analyzing data through transformation is that the results of the statistical analysis actually refer to the transformed data and not the original data. In most situations, this point is not important. It may, however, make interpretation of the results difficult as the units of transformed data will be expressed in terms of the logarithm or square root that was used in the transformation. Further, the transformed data may lead to conclusions which are not supported by the original, non-transformed data and vice versa. Because of these potential problems with transformation of data, the more commonly used alternative in analyzing non-normally distributed data is to use **non-parametric methods** of analysis.

NON-PARAMETRIC METHODS OF ANALYSIS

In the presence of non-normally distributed data, the alternative to transformation is to use **non-parametric methods**. These tests, unlike those which use the t distribution, do not assume that the data are normally distributed. They are statistically stronger than t-tests in the presence of skewed data, but are weaker than t-tests in the presence of normally distributed data. The most commonly used non-parametric methods are the **sign test**, the **Wilcoxon signed-ranks test**, and the **Wilcoxon rank-sum test** (**Mann-Whitney U test**).

THE SIGN TEST

If we wish to compare a single non-normally distributed mean with a known standard or normal value, the **sign test** is a useful method (note that if the data were normally distributed, we would use the **test of hypothesis** in this situation).

The sign test is based on the fact that if the observations were normally distributed and were not different from the standard or normal value, their median would closely approximate the standard or normal value (the median of the population-at-large). Given this, there would be a probability of 0.50 that any observation was either greater than or less than the standard value. If the observed values differ from the standard value significantly (and therefore do not represent the population-at-large), the probability will be skewed to one side or the other of 0.50. To perform the sign test, we must first compare each of our n observations with the standard or normal value (the population median) and sum the number of comparisons (N) in which the observation exceeds the normal value. The probability of such an occurrence is given by:

$$\text{probability} = \frac{(0.5)^{N}(0.5)^{n-N}}{(n-N)\ !}$$

For example, what is the probability of 10 ICU patients all having an ICU LOS greater than the normal LOS of 2.7 days (i.e., N=10)? The probability of this event is obtained by:

$$\text{probability} = \frac{(0.5)^{10}(0.5)^{10-10}}{(10-10)\ !} = \frac{(0.00098)(1)}{(1)} = 0.00098$$

The actual probability of such an occurrence is therefore 1 in 1000. As with **Fisher's exact test**, all of the more extreme cases must also be calculated, when appropriate, and each of the resultant probabilities summed to obtain the total probability of occurrence. Note that this is an example of a one-tailed test; we are asking whether a mean is greater than the known value. If we wanted to test whether the ICU LOS for the same 10 patients was either less than or greater than the normal value, we would want to perform a two-tailed test. For the sign test, the two-tailed probability is given by doubling the one tailed probability.

The sign test can also be used to evaluate paired sets of data. Given any two values, there is a probability of 0.50 that the difference between the values is positive. Similarly, there is a probability of 0.50 that the difference is negative. Consider the following data from the Preoperative Evaluation study comparing ICU LOS for preoperative evaluation (study) patients and for age and procedure matched patients (control patients) who did not undergo preoperative evaluation. The data means and medians are not similar and are therefore not normally distributed being skewed by patient pair 6. To appropriately analyze this data, a non-parametric method such as the sign test must be used. Our null and alternate hypotheses would be as follows:

Null hypothesis: the ICU LOS for study and control patients is the same.
Alternate hypothesis: the ICU LOS for study patients and control patients are different.

Patient Pairs	Study Patient ICU LOS	Control Patient ICU LOS	Difference
1	4	2	2
2	13	4	9
3	3	1	2
4	9	1	8
5	2	2	0
6	**178**	**1**	**177**
7	2	3	-1
8	8	4	4
9	4	1	3
10	3	4	-1

Given 10 pairs of age and procedure-matched patients, 8 sets of paired data had a positive difference while 2 sets had a negative difference. Using the sign test, and calculating each of the more extreme possible outcomes, the total probability of this occurring is:

$$P_{total} = \frac{(0.5)^8(0.5)^2}{2!} + \frac{(0.5)^9(0.5)^1}{1!} + \frac{(0.5)^{10}(0.5)^0}{0!}$$

$$P_{total} = 0.00049 + 0.00098 + 0.00098 = 0.0025$$

Since we did not know *a priori* whether the two groups differed and in which direction, we need to perform a two-tailed test. Thus, the probability that the two groups have the same mean ICU LOS is (0.0025)(2) = 0.005 or 1 in 200. We therefore reject the null hypothesis and accept our alternate hypothesis stating that the groups differ significantly with a probability of 0.005.

WILCOXON SIGNED RANKS TEST

When we are evaluating a single group of patients and wish to evaluate the difference between two paired samples which are non-normally distributed, the **Wilcoxon signed-ranks test** is most commonly used. This is the non-parametric equivalent of the **single sample, paired t-test**. The Wilcoxon-signed ranks test evaluates the hypothesis that the medians of the two sets of observations are the same. It is almost as powerful as the t-test when the observations are normally distributed and is more powerful than the t-test when the observations are not.

Consider the following data on right ventricular end-diastolic volume index (RVEDVI) before and after patients received a 500 mL normal saline infusion. If the data were normally distributed, we would use the single-sample, paired t-test. Upon inspection of the raw data, however, it is obvious that the data are not normally distributed and contain several outliers (particularly patients 3 and 7) which skew the data. To confirm this, we can compare the mean and median of each sample. Since the mean and median are not similar, the data are non-normally distributed and a t-test is inappropriate. We would therefore use the Wilcoxon signed-ranks test to compare the means of the pre- and post-infusion data.

Patient	RVEDVI Pre-bolus	RVEDVI Post-bolus	Difference	Rank	Signed Rank
1	62	83	21	7	7
2	54	60	6	1	1
3	182	172	-10	2	-2
4	73	101	28	9	9
5	47	72	25	8	8
6	32	47	15	5	5
7	120	160	40	10	10
8	90	79	-11	3	-3
9	85	105	20	6	6
10	55	67	12	4	4
mean	80.0	94.6			4.5
median	67.5	81.5			4.5
sd	43.7	41.5			4.5

To perform the Wilcoxon signed-ranks test, we first calculate the difference between each pair of observations. These differences are then ordered from smallest to largest (ignoring the sign) and ranked from 1 to n beginning with the smallest observation. The sign of the difference (either positive or negative) is then assigned to that difference's rank. We then use the "signed ranks" of the observations to calculate the t statistic. The mean and median of the ranks are both 4.5. The ranked data are now normally distributed and use of the t-test is appropriate whereas it would not have been on the original, non-normally distributed data. Assuming we wish to detect a difference with 95% confidence (significance level of 0.05) and have 9 degrees of freedom (10-1 pairs of data), the critical value of t is 2.26. The t statistic is then calculated on the ranks as follows:

$$t = \frac{\text{mean of ranks}}{\text{standard error of the mean of ranks}} = \frac{4.5}{4.5/\sqrt{10}} = 3.2$$

Since our calculated t exceeds the critical value of t (2.26) for a significance level of 0.05, we would reject the null hypothesis and accept the alternate hypothesis, concluding that a 500 mL normal saline infusion does result in a significant increase in RVEDVI with a significance level of < 0.05.

WILCOXON RANK-SUM TEST (MANN-WHITNEY U TEST)

If we are comparing the means of two independent groups, and the data are non-normally distributed, the **Wilcoxon rank-sum test** (also known as the **Mann-Whitney U test**) is appropriate. This is the non-parametric equivalent of the two-sample, unpaired t-test. The Wilcoxon rank-sum test analyzes the equality of the sample medians rather than the means (as is done in the two-sample, unpaired t-test). It is similar to the Wilcoxon signed-ranks test for paired data in that the raw observations are ranked from smallest to largest and the mean ranks are then compared using the t statistic for two independent samples (two-sample, unpaired t-test). The difference in the ranking for this test is that the observations of both groups are considered as one group for the purpose of assigning the ranks. The theory of the test is that if the two samples are similar, their medians will also be similar and the mean ranks will be equal. If one mean rank is larger, then that sample must have larger observations (and therefore a larger median) than the other. The Wilcoxon rank-sum test then quantitates how different the two mean ranks are using the t-statistic.

Consider the following data on total hospital charges and ICU LOS (length of stay) for age and procedure matched patients from the Preoperative Evaluation study. The data are non-normally distributed containing several outliers (patient pairs 4 and 6) which skew the data (note the differences between the mean and median for each group).

To use the Wilcoxon rank-sum test, the data from both groups are combined and ranked from smallest to largest. Thus, all 20 of the patient charges have been considered as a group and ranked from 1 to 20 (see below). Similarly, all of the LOS measurements have been considered as a group and ranked from 1 to 20. Note that for the LOS measurements, some of the observations have the same value. In this situation, the ranking is performed by using the mean rank for each of the ties. For example, four patients had a LOS of 1 day. Normally, these patients would receive the ranks 1, 2, 3, and 4. Since we must treat each observation equally, each of the four patients with LOS of 1 day is assigned the average rank of 2.5 [(1+2+3+4)/4=2.5]. Similarly, there are four patients with a LOS of 2 days. These patients would normally receive the ranks 5, 6, 7, and 8. We assign each of these patients the average rank of 6.5 [(5+6+7+8)/4=6.5].

Patient Pairs	Study Pt Charges	Rank	Control Pt Charges	Rank	Study Pt LOS	Rank	Control Pt LOS	Rank
1	$28,004	**8**	$19,750	**3**	4	**14**	2	**6.5**
2	$44,768	**13**	$39,684	**12**	13	**19**	4	**14**
3	$23,727	**5**	$22,954	**4**	3	**10**	1	**2.5**
4	$330,989	**19**	$26,818	**7**	9	**18**	1	**2.5**
5	$11,172	**1**	$18,998	**2**	2	**6.5**	2	**6.5**
6	$838,282	**20**	$36,673	**11**	178	**20**	1	**2.8**
7	$28,334	**9**	$101,016	**18**	2	**6.5**	3	**10**
8	$53,260	**15**	$70,198	**17**	8	**17**	4	**14**
9	$34,446	**10**	$26,248	**6**	4	**14**	1	**2.5**
10	$57,966	**16**	$49,287	**14**	3	**10**	4	**14**
mean	**$145,095**	**11.6**	**$41,163**	**9.4**	**22.2**	**13.0**	**2.7**	**7.5**
median	**$39,607**		**$31,746**		**4**		**2**	
sd	**$261,068**	**6.1**	**$31,746**	**5.8**	**54.9**	**5.0**	**2.21**	**5.1**

We will first compare the patient charges using the Wilcoxon rank-sum test. To calculate the t statistic, we must first calculate the **pooled standard deviation of the ranks** which is calculated as in the two-sample unpaired t-test, but using the standard deviations of the ranks:

$$s_{r_p} = \sqrt{\frac{(9)(6.1)^2 + (9)(5.8)^2}{10+10-2}} = 5.95$$

The degrees of freedom in this case are calculated by (n_1 + n_2 -2). For a significance level of 0.05 and 18 degrees of freedom (10+10-2), the critical value of t is 2.1. The t statistic for the Wilcoxon rank-sum test is calculated as follows:

$$t = \frac{\text{rank}_1 - \text{rank}_2}{s_{r_p}\sqrt{1/n_1 + 1/n_2}} = \frac{11.6 - 9.4}{(5.95)(0.447)} = 0.83$$

Based on this, we cannot reject our null hypothesis of no difference as the calculated value of t does not exceed the critical value of 2.1. The mean hospital charges are therefore not significantly different between the groups despite the difference of over $100,000 between the means. The reason that this is true is related to the large variability that exists in the charges as noted by the large standard deviations. The large variability in the data makes it difficult to prove that a difference exists between the groups.

We now can compare the LOS measurements, again using the Wilcoxon rank-sum test. Using the same significance level (0.05) and degrees of freedom (18), the critical value of t for this comparison is still 2.1. The t statistic is calculated as follows:

$$s_{r_p} = \sqrt{\frac{(9)(5.0)^2 + (9)(5.1)^2}{10+10-2}} = 5.05$$

$$t = \frac{13.5 - 7.5}{(5.05)(0.447)} = 2.26$$

As t exceeds the critical value of 2.1, we reject the null hypothesis of no difference and state that the mean ICU LOS of preoperative evaluation patients is significantly greater than that of control patients with a significance level of < 0.05.

Interestingly, if we use the two sample, unpaired t-test to analyze this data, the null hypothesis is not rejected and the significant difference is missed (a Type II error). This is because of the non-normal distribution of the data and the inequality of the sample variances, both of which violate the basic assumptions of the t-test, making it inaccurate and leading to a false conclusion. This emphasizes the importance of performing the appropriate statistical test if valid conclusions are to be made.

The Wilcoxon rank-sum test is especially useful in evaluating data from ordered categories (such as tumor staging classifications) which have non-numerical scores. For example, how would we compare two treatments whose outcome is documented as either "worse", "unchanged", or "improved?” One option is to compress the data into two categories (such as "improved" and "not improved") so that we can use a chi-square test or Fisher’s exact test to compare the groups. By doing so, however, we sacrifice data (reducing our observations to two categories instead of three categories) and gain less information from the study. If we apply ranks to the outcome categories, however, and evaluate these ranks with the Wilcoxon rank-sum test, no data is sacrificed and we use the full potential of the data collected.

SUGGESTED READING

1. Dawson-Saunders B, Trapp RG. Basic and clinical biostatistics (2nd Ed). Norwalk, CT: Appleton and Lange, 1994.
2. Wassertheil-Smoller S. Biostatistics and epidemiology: a primer for health professionals. New York: Springer-Verlag, 1990.
3. Campbell MJ, Machin D. Medical statistics: a commonsense approach. Chichester: John Wiley and Sons Ltd, 1990.
4. Moses LE, Emerson JD, Hosseini H. Analyzing data from ordered categories. NEJM 1984; 311:442-8.
5. O'Brien PC, Shampo MA. Statistical considerations for performing multiple tests in a single experiment. 4. performing multiple statistical tests on the same data. Mayo Clin Proc 1988; 63:1043-45.
6. O'Brien PC, Shampo MA. 5. One sample of paired observations (paired t-test). Mayo Clin Proc 1981; 56:324-6.
7. O'Brien PC, Shampo MA. 6. Comparing two samples (the two-sample t-test). Mayo Clin Proc 1981; 56:393-4.
8. Godfrey K. Comparing the means of several groups. NEJM 1985; 313:1450-6.
9. Conover WJ, Iman RL. Rank transformations as a bridge between parametric and nonparametric statistics. Am Stat 1981;35:124-129.