9

COMPARING TWO GROUPS

Chapters 6 and 7 described how to make inferences about the population mean, and other measures of location, associated with a single population of individuals or things. This chapter extends these methods to situations where the goal is to compare two groups. For example, Table 2.1 reports data from a study on changes in cholesterol levels when participants take an experimental drug. But of fundamental interest is how the changes compare to individuals who receive a placebo instead. Example 4 in section 6.2 described an experiment on the effect of ozone on weight gain among rats. The two groups in this study consisted of rats living in an ozone environment and ones that lived in an ozone-free environment. Do weight gains differ for these groups, and if they do, how might this difference be described? Two training programs are available for learning how to invest in stocks. To what extent, if any, do these training programs differ? How does the reading ability of children who watch thirty hours or more of television per week compare to children who watch ten hours or less? How does the birth weight of newborns among mothers who smoke compare to the birth weight among mothers who do not smoke? In general terms, if we have two independent variables, how might we compare them?

9.1 Comparing the means of two independent groups

When trying to detect and describe differences between groups, by far the most common strategy is to use means. We begin with a classic method designed for two independent groups. By independent groups is meant that the observations in the first group are independent of the observations in the second. In particular, the sample means for the two groups, say $\bar{X}_1$ and $\bar{X}_2$, are independent. So, in the example dealing with weight gain among rats, it is assumed that one group of rats is exposed to an ozone environment, and a separate group of rats, not associated with the first group, is exposed to an ozone-free environment. This is in contrast to using, for example, the same rats under both conditions, or using rats from the same litter, in which case the sample means might be dependent.

The two-sample Student's *t*-test

The classic and best-known method for comparing the means of two independent groups is called the *two-sample Student's t-test*. Here we let μ_1 and μ_2 represent the two population means, and the corresponding standard deviations are denoted by σ_1 and σ_2. The goal is to test

$$H_0 : \mu_1 = \mu_2, \tag{9.1}$$

the hypothesis that the population means are equal. It turns out that we can get exact control over the probability of a Type I error if the following three assumptions are true:

- Random sampling
- Normality
- Equal variances. That is, $\sigma_1 = \sigma_2$, which is called the *homogeneity of variance* assumption.

Before describing how to test the hypothesis of equal means, first consider how we might estimate the assumed common variance. For convenience, let σ_p^2 represent the common variance and let s_1^2 and s_2^2 be the sample variances corresponding to the two groups. Also let n_1 and n_2 represent the corresponding sample sizes. The typical estimate of σ_p^2 is

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}. \tag{9.2}$$

For the special case where the sample sizes are equal, meaning that $n_1 = n_2$, s_p^2 is just the average of the two sample variances. That is,

$$s_p^2 = \frac{s_1^2 + s_2^2}{2}.$$

Now consider the problem of testing the null hypothesis of equal means. Under the assumptions already stated, the probability of a Type I error will be exactly α if we reject the null hypothesis when

$$|T| \geq t, \tag{9.3}$$

where

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}, \tag{9.4}$$

and t is the $1 - \alpha/2$ quantile of Student's t-distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom, which is read from table 4 in appendix B. An exact $1 - \alpha$ confidence interval for the difference between the population means, under the same assumptions, is

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}. \tag{9.5}$$

Example 1

Salk (1973) conducted a study where the general goal was to examine the soothing effects of a mother's heartbeat on her newborn infant. Infants were

Table 9.1 Weight gain, in grams, for large babies

Group 1 (heartbeat)			
Subject	Gain	Subject	Gain
1	190	11	10
2	80	12	10
3	80	13	0
4	75	14	0
5	50	15	−10
6	40	16	−25
7	30	17	−30
8	20	18	−45
9	20	19	−60
10	10	20	−85

$n_1 = 20, \bar{X}_1 = 18.0, s_1 = 60.1, s_1/\sqrt{n_1} = 13$

Group 2 (heartbeat)							
Subject	Gain	Subject	Gain	Subject	Gain	Subject	Gain
1	140	11	25	−21	−50	31	−130
2	100	12	25	−22	−50	32	−155
3	100	13	25	−23	−60	33	−155
4	70	14	30	−24	−75	34	−180
5	25	15	30	−25	−75	35	−240
6	20	16	30	−26	−85	36	−290
7	10	17	45	−27	−85		
8	0	18	45	−28	−100		
9	−10	19	−45	29	−110		
10	−10	20	−50	30	−130		

$n_2 = 36, \bar{X}_2 = -52.1, s_2 = 88.4, s_2/\sqrt{n_2} = 15$

placed in a nursery immediately after birth and they remained there for four days except when being fed by their mothers. The infants were divided into two groups. The first was continuously exposed to the sound of an adult's heartbeat; the other group was not. Salk measured, among other things, the weight change of the babies from birth to the fourth day. Table 9.1 reports the weight change for the babies weighing at least 3,510 grams at birth. As indicated, the sample standard deviations are $s_1 = 60.1$ and $s_2 = 88.4$. The estimate of the assumed common variance is

$$s_p^2 = \frac{(20-1)(60.1^2)+(36-1)(88.4^2)}{20+36-2} = 6,335.9.$$

So

$$T = \frac{18-(-52.1)}{\sqrt{6,335.9\left(\frac{1}{20}+\frac{1}{36}\right)}} = \frac{70.1}{22.2} = 3.2.$$

The sample sizes are $n_1 = 20$ and $n_2 = 36$, so the degrees of freedom are $\nu = 20+36-2 = 54$. If we want the Type I error probability to be $\alpha = .05$, then $1-\alpha/2 = .975$, and from table 4 in appendix B, $t = 2.01$. Because $|T| = 3.2$, which is greater than 2.01, reject H_0 and conclude that the means differ.

That is, we conclude that among all newborns we might measure, the average weight gain would be higher among babies exposed to the sound of a heartbeat compared to those that are not exposed. By design, the probability that our conclusion is in error is .05, assuming normality and homoscedasticity. The .95 confidence interval for $\mu_1 - \mu_2$, the difference between the population means, is

$$[18 - (-52.1)] \pm 2.01\sqrt{6,335.9\left(\frac{1}{20} + \frac{1}{36}\right)} = (25.5, 114.7).$$

This interval does not contain zero, and it indicates that the difference between the means is likely to be at least 25.5, so again you would reject the hypothesis of equal means.

Violating assumptions: When does Student's *t* perform well?

There are two conditions where the assumption of normality, or equal variances, can be violated and yet Student's t appears to continue to perform well in terms of Type I errors and accurate confidence intervals. The homoscedasticity assumption can violated if both distributions are normal and the sample sizes are equal, provided the sample sizes are not overly small, say less than 8 (Ramsey, 1980). As for non-normality, Student's t appears to perform well in terms of Type I error probabilities provided the two distributions are identical. That is, not only do they have the same means, they have the same variances, the same amount of skewness, the tails of the distribution are identical, and so on. So if we were to plot the distributions, the plots would be exactly the same. If, for example, you want the probability of a Type I error to be .05, generally, the actual Type I error probability will be less than or equal to .05.[1]

Conditions where Student's *t* performs poorly

If the goal is to test the hypothesis of equal means, without being sensitive to other ways the groups might differ, Student's t can be unsatisfactory in terms of Type I errors and accurate confidence intervals when sampling from normal distributions with unequal sample sizes and unequal variances. Problems due to unequal variances are exacerbated when sampling from non-normal distributions instead, and now concerns arise even with equal sample sizes (e.g., Algina, et al., 1994; Wilcox, 1990). When dealing with groups that differ in skewness, again problems with controlling the probability of a Type I error occur, and the combination of unequal variances and different amounts of skewness makes matters worse. Some degree of unequal variances, as well as mild differences in skewness, can be tolerated. But the extent to which this is true, based on the data under study, is difficult to determine in an accurate manner.

1. A key reason is that if we sample an observation from each group, and if the groups have the same skewness, the distribution of the difference between these two observations is symmetric. We saw in chapter 7 that for a symmetric distribution, Type I error probabilities larger than the specified α level can be avoided.

Example 2

Recall from chapter 6 that one of way of determining the distribution of T under normality is to use simulations. That is, to generate data from a normal distribution, compute T, and repeat this process many times. With the aid of a computer, we can extend this method when sampling from two distributions that have equal means but which differ in terms of skewness and have unequal variances. In particular, imagine that we sample 40 observations from a standard normal distribution and 60 observations from the distribution shown in figure 9.1, and then we compute T. Repeating this process 1000 times provides a fairly accurate indication of the distribution of T when the null hypothesis of equal means is true. Figure 9.2 shows a plot of the results plus the distribution of T assuming normality. Under normality, and with a Type I error probability

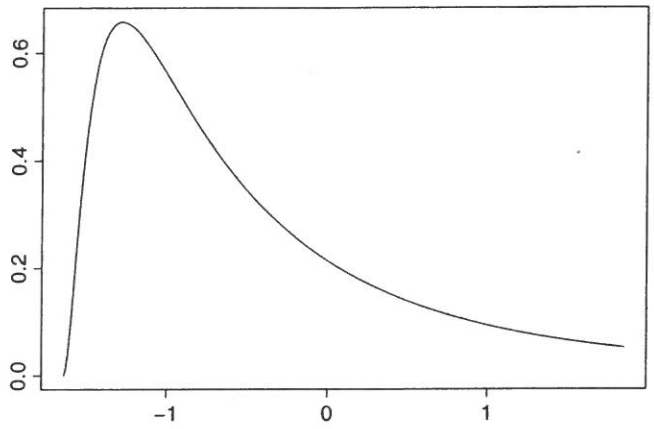

Figure 9.1 A skewed distribution with a mean of zero.

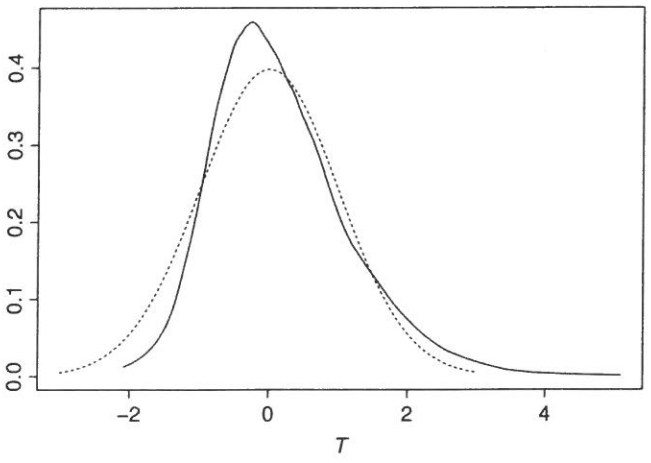

Figure 9.2 The distribution of T when sampling 40 values form a standard normal and 60 values form the distribution in figure 9.1. Also shown is the distribution of T when both groups have normal distribution. This illustrates that differences in skewness can have an important impact on T.

of $\alpha = .05$, Student's T rejects the hypothesis of equal means if $T \leq -2.002$ or if $T \geq 2.002$. But figure 9.2 indicates that we should reject if $T \leq -1.390$ or if $T \geq 2.412$, values that differ substantially from what would be used under normality.

Some authorities might criticize this last example on the grounds that if groups differ in terms of the variances and the amount of skewness they have, surely the means differ as well. That is, they would argue that Type I errors are not an issue in this case. But even if we accept this point of view, this last illustration can be seen to create concerns about power, and it indicates that confidence intervals based on Student's t can be relatively inaccurate.

Usually, a basic requirement of any method is that with sufficiently large sample sizes, good control over the Type I error probability and accurate confidence intervals will be obtained. There are theoretical results indicating that under general conditions, when the goal is to compare the means without being sensitive to other features of the distribution (such as unequal variances), Student's t can be unsatisfactory regardless of how large the sample sizes might be (Cressie and Whitford, 1986). Exceptions are when the sample sizes are equal and when both groups have identical distributions. This means that Student's t provides a valid test of the hypothesis that the distributions are identical, but it can be unsatisfactory when computing confidence intervals for the difference between the means or when testing the hypothesis that groups have equal means.

Finally, in terms of Type II errors and power, Student's t can perform very poorly, compared to alternative techniques, when outliers tend to occur. The presence of outliers does not necessarily mean low power, but the reality is that power might be increased substantially when comparing groups with something other than the means, as will be illustrated. Some additional concerns about Student's t are summarized in Wilcox (2003, 2005). One of these concerns is that unequal variances and differences in skewness can create power problems as well.

Why testing assumptions can be unsatisfactory

Some commercial software now contains a test of the assumption that two groups have equal variances. The idea is that if the hypothesis of equal variances is not rejected, one would then use Student's t. But a *basic principle* is that failing to reject a null hypothesis is not, by itself, compelling evidence that the null hypothesis should be accepted or that the null hypothesis is approximately true. Accepting the null hypothesis is only reasonable if the probability of rejecting (power) is sufficiently high to ensure that differences that have practical importance will be detected. If there is a low probability of detecting a difference that is deemed important, concluding that no difference exists is difficult to defend. In the case of Student's t, would a test of the assumption of equal variances have enough power to detect a situation where unequal variances causes a problem? All indications are that the answer is no (e.g., Markowski and Markowski, 1990; Moser, et al., 1989; Wilcox, et al., 1986; Zimmerman, 2004; Hayes and Cai, 2007). Presumably exceptions occur if the sample sizes are sufficiently large, but it is unclear how we can be reasonably certain when this is the case. Part of the problem is that the extent to which the variances can differ, without having a major impact on the Type I error probability, is a complicated function of the sample sizes, and the extent to which groups differ in

terms of skewness, and the likelihood of observing outliers. Testing the hypothesis that data have a normal distribution is another strategy that might be followed. But when do such tests have enough power to detect departures from normality that are a concern? The answer is not remotely clear and so this approach cannot be recommended at this time. A better strategy is to use more modern methods that perform reasonably well under normality, but which continue to perform well under non-normality or when groups have unequal variances.

There are many alternatives to Student's t when comparing groups. Because testing assumptions seems dubious, how can we tell whether some alternative technique might give a substantially different sense about whether and how the groups differ? Currently, the only known strategy that answers this question in an adequate manner is to simply try alternative methods, some of which are outlined later in this chapter. However, a criticism of applying many methods is that control over the probability of at least one Type I error can become an issue. This issue, and methods for dealing with it, are described and illustrated in chapter 11.

Interpreting Student's t when we reject

Despite its many practical problems, Student's t does have a positive feature. If we reject, this is a good indication that the distributions differ in some manner. This is because when the distributions do not differ, it controls the probability of a Type I error fairly well. But even though the method is designed to compare means, in reality it is also sensitive to differences in variances and skewness. As previously noted, some would argue that if the distributions differ, surely the means differ. However, when we reject, it is unclear whether the main reason is due to differences between the means. The main reason could be differences between the variances or skewness. Moreover, rejecting with Student's t raises concerns about whether the confidence interval, given by equation (9.5), is reasonably accurate. In summary, when rejecting with Student's t, it is reasonable to conclude that the groups differ in some manner. But when Student's t indicates that groups differ, there are concerns that the nature of the difference is not being revealed in a reasonably accurate manner. And when Student's t fails to reject, this alone is not compelling evidence that the groups do not differ in any important way.

Dealing with unequal variances: Welch's test

Many methods have been proposed for comparing means when the population variances (σ_1^2 and σ_2^2) differ. None are completely satisfactory. Here we describe one such method that seems to perform reasonably well compared to other techniques that have been derived when attention is restricted to comparing means. Popular commercial software now contains this method, which was derived by Welch (1938).

Recall from chapter 5 that the sampling distribution of the sample mean has variance σ^2/n, which is called the squared standard error of the sample mean. For the situation at hand, the difference between the sample means, $\bar{X}_1 - \bar{X}_2$, also has a sampling distribution, and the corresponding mean of this difference is $\mu_1 - \mu_2$, the difference between the population means. Roughly, this means that if we were repeat a study millions of times, and if we averaged the differences between the sample means resulting from each study, we would get $\mu_1 - \mu_2$, the di.erence between the population means.

Put another way, on average, over many studies, $\bar{X}_1 - \bar{X}_2$ estimates $\mu_1 - \mu_2$. Moreover the variance (or squared standard error) of the difference between the sample means can be shown to be

$$\text{VAR}(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Also recall from chapter 6 that under normality, if we standardize a variable by subtracting its mean, and then dividing by its standard error, we get a standard normal distribution. That is, if a variable has a normal distribution, then in general,

$$\frac{\text{variable} - \text{population mean of the variable}}{\text{standard error of the variable}}, \tag{9.6}$$

will have a standard normal distribution. Here the variable of interest is $\bar{X}_1 - \bar{X}_2$, the difference between the sample means, which has a population mean of $\mu_1 - \mu_2$. Consequently, based on the equation for the squared standard error, $\text{VAR}(\bar{X}_1 - \bar{X}_2)$, it follows that

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

has a standard normal distribution. *If* the hypothesis of equal means is true, then $\mu_1 - \mu_2 = 0$, in which case this last equation becomes

$$\frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

which again has a standard normal distribution. As usual, the population variances are rarely known, but they can be estimated with the sample variances, in which case this last equation becomes

$$W = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{9.7}$$

where, as before, s_1^2 and s_2^2 are the sample variances corresponding to the two groups being compared; this is the test statistic used by Welch's test.

When the hypothesis of equal means is true, W will have, approximately, a standard normal distribution if the sample sizes are sufficiently large, thanks to the central limit theorem. That is, we can determine how large W must be to reject the hypothesis of equal means using values in table 1 in appendix B. But in general, W will not have a normal distribution, so some other approximation of an appropriate critical value is required. Welch's approach to this problem is implemented in the following manner. For convenience, let

$$q_1 = \frac{s_1^2}{n_1} \text{ and } q_2 = \frac{s_2^2}{n_2}. \tag{9.8}$$

As was done with Student's t, table 4 in appendix B is used to determine a critical value, t, but now the degrees of freedom are

$$\nu = \frac{(q_1 + q_2)^2}{\frac{q_1^2}{n_1 - 1} + \frac{q_2^2}{n_2 - 1}}. \tag{9.9}$$

Under normality, W has, approximately, a Student's t-distribution with degrees of freedom given by equation (9.9). That is, reject the hypothesis of equal means if $|W| \geq t$. The $1-\alpha$ confidence interval for the difference between the means, $\mu_1 - \mu_2$, is

$$(\bar{X}_1 - \bar{X}_2) \pm t\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}. \tag{9.10}$$

Example 3

Tables 2.1 and 2.2 report data on the effectiveness of a drug to lower cholesterol levels. For the data in table 2.1, corresponding to the group that received the experimental drug, the sample size is $n_1 = 171$, the sample variance is $s_1^2 = 133.51$, and the sample mean is $\bar{X}_1 = -9.854$. For the group that received the placebo, $n_2 = 177$, $s_2^2 = 213.97$, and $\bar{X}_2 = 0.124$. To apply Welch's test, compute $q_1 = 133.51/171 = 0.78076$ and $q_2 = 213.97/177 = 1.20887$, in which case the degrees of freedom are

$$\nu = \frac{(0.78076 + 1.20887)^2}{\frac{0.78076^2}{171-1} + \frac{1.2088^2}{177-1}} = 332.99.$$

The test statistic is $W = 7.07$, the $\alpha = .05$ critical value is 1.967, and because $|7.07| \geq 1.967$, reject the null hypothesis.

Student's t versus Welch's test

Some brief comments about the relative merits of Student's t versus Welch's Test should be made. When comparing groups that do not differ in any manner, there is little reason to prefer Student's t over Welch's test. But if the distributions differ in some way, such as having unequal variances. The choice of method can make a practical difference. Welch's test reduces problems with unequal variances, given the goal of comparing means, but it does not eliminate them. Differences in skewness remain a concern, and, as is the case with all methods based on means, outliers can destroy power. So, when rejecting with Welch's test, like Student's t-test, it is reasonable to conclude that the distributions differ in some manner, but there is uncertainty about whether the main reason has to do with differences between the population means; the primary reason could be unequal variances or differences in skewness. And when we fail to reject, this could be because the groups differ by very little, but another possibility is that power is low due to sample sizes that are too small, differences in skewness, or outliers.

In fairness, there are situations where Student's t correctly concludes that groups differ in some manner when Welch's test does not. This can happen because Student's t can be more sensitive to certain types of differences, such as unequal variances.

A positive feature of Welch's method is that with sufficiently large sample sizes, it will control the probability of a Type I error given the goal of comparing means, and it provides accurate confidence intervals as well, assuming random sampling only. This is in contrast to Student's t, which does not satisfy this goal when the sample sizes are unequal and the groups differ in skewness. A rough explanation is that under random sampling, regardless of whether the groups differ, Welch's test uses a correct estimate of the standard error associated with the difference between the means, $\bar{X}_1 - \bar{X}_2$, but there are conditions where this is not the case when using Student's t (Cressie and

Whitford, 1986). As previously noted, an exception is when groups have identical distributions. So again, an argument for considering Student's t is that if it rejects, a good argument can be made that the groups differ in some manner. A very rough rule is that a method that uses the correct standard error is likely to have more power than a method does not. So here, the expectation is that Welch's test will tend to have more power than Student's t, but exceptions are encountered where Student's t rejects and Welch's method does not.

Comments about outliers when comparing means

Any method for comparing groups based on means runs the risk of relatively low power. As noted in previous chapters, outliers can inflate the sample variances which in turn can result in low power, and there is some possibility that the mean will poorly reflect what is typical. Outliers also have other consequences relevant to power that might not be immediately obvious but which are illustrated by the next example.

Example 4

Imagine that an experimental drug is under investigation and that there is concern that it might damage the stomach. For illustrative purposes, suppose the drug is given to a sample of rats, a placebo is given to a control group, and the results are as follows:

Experimental drug: 4, 5, 6, 7, 8, 9, 10, 11, 12, 13

Placebo: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.

The goal is to determine whether the average amount of stomach damage differs for these two groups. The corresponding sample means are $\bar{X}_1 = 8.5$ and $\bar{X}_2 = 5.5$ and $T = 2.22$. With $\alpha = .05$, the critical value is $t = 2.1$, so Student's t would reject the hypothesis of equal means and conclude that the first group has a larger population mean than the second (because the first group has the larger sample mean). Now, if we increase the largest observation in the first group from 13 to 23, the sample mean increases to $\bar{X}_1 = 9.5$. So the difference between $\bar{X}_1$ and $\bar{X}_2$ has increased from 3 to 4 and this would seem to suggest that we have stronger evidence that the population means differ and in fact the first group has the larger population mean. However, increasing the largest observation in the first group also inflates the corresponding sample variance, s_1^2. In particular, s_1^2 increases from 9.17 to 29.17. The result is that T *decreases* to $T = 2.04$ and we no longer reject. That is, increasing the largest observation has more of an effect on the sample variance than the sample mean in the sense that now we are no longer able to conclude that the population means differ. Increasing the largest observation in the first group to 33, the sample mean increases to 10.5, the difference between the two sample means increases to 5 and now $T = 1.79$. So again we do not reject and in fact our test statistic is getting smaller. It is left as an exercise to show that a similar result is obtained when using Welch's test. This illustration provides another perspective on how outliers can mask differences between population means.