

## Two-sample hypothesis tests

As we noted when discussing regression models, in many situations we have two natural subgroups in the data, and we're interested in how they compare to each other (there's nothing limiting ourselves to only two subgroups, and the techniques we're going to talk about generalize to more than two groups, but that's beyond the scope of this course). That is, we're interested in *two-sample* (or *independent samples*) hypotheses and tests.

- (1) gender discrimination on salaries
- (2) “gender gap” in political opinions
- (3) effectiveness of medical treatments (light therapy and jaundice)
- (4) comparative volatility of stock exchanges

Let  $\mu_1$  and  $\mu_2$  be the mean values of the random variable of interest for groups 1 and 2, respectively. Are they different from each other?

We have, in fact, already talked about a way to answer this question. Let  $y_i$  be the value of the random variable of interest for the  $i^{\text{th}}$  observation, and define an indicator variable  $x_i$  for the two groups equaling 0 for group 1 and 1 for group 2. Then, fit the regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

The estimated constant term is the estimate of  $E(Y)$  when  $x_i = 0$ ; that is, it is the sample mean for group 1,  $\bar{Y}_1$ , estimating  $\mu_1$ . The estimated slope term is the estimate of the change in  $E(Y)$  when  $x$  changes from 0 to 1; that is, it is the difference between the sample mean for group 2 and the sample mean for group 1,  $\bar{Y}_2 - \bar{Y}_1$ , estimating  $\mu_2 - \mu_1$ . The  $t$ -statistic testing  $\beta_1 = 0$  is thus a test of whether  $\mu_2 - \mu_1 = 0$ , or  $\mu_1 = \mu_2$ .

We also can structure this test by analogy with the one-sample situation. As always, the null hypothesis gets the benefit of the doubt; since we have no reason to believe that the two groups are different *a priori*, the natural hypotheses are

$$H_0 : \mu_1 = \mu_2$$

versus

$$H_a : \mu_1 \neq \mu_2.$$

The natural estimate of the difference between the population means is the difference in the sample means,  $\bar{Y}_1 - \bar{Y}_2$ , and that will be the kernel of the test statistics we'll now derive. We're assuming that each population is (roughly) normally distributed, so that we can appeal to  $t$ -statistic constructions.

First, let's assume that the variances are the same in the two groups; i.e.,  $\sigma_1^2 = \sigma_2^2 = \sigma^2$ , or homoscedasticity. In this case we should use all of the data values to estimate the common  $\sigma^2$ ; this is done using the *pooled estimate of variance*,

$$s_{\text{pooled}}^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

(note that this is just a weighted average of the group variances, weighted by (one less than) the sample sizes). Now, by analogy with the one-sample problem, we test the hypotheses above using a  $t$ -statistic,

$$t = \frac{|\bar{Y}_1 - \bar{Y}_2|}{s_{\text{pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

which follows a  $t$ -distribution on  $n_1 + n_2 - 2$  degrees of freedom. This statistic is identical to the  $t$ -statistic for the slope coefficient in the regression model given above.

What if the assumption of constant variance is not reasonable? That is, what if we have heteroscedasticity? This is called the Behrens-Fisher problem, and remarkably enough (given how simple the problem is) there is no unambiguously supported solution to it. The most commonly used solution is to construct a test statistic that doesn't assume constant variance, and then approximate it with a  $t$ -distribution by estimating appropriate degrees of freedom. The test statistic has the form

$$t = \frac{|\bar{Y}_1 - \bar{Y}_2|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}},$$

and it is compared to a  $t$ -distribution with degrees of freedom equaling

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1}}.$$

This is usually not an integer, but computer packages can calculate tail probabilities for nonintegral  $t$ -distributions. Presumably we would choose which test to believe based on the degree of our belief in whether heteroscedasticity exists. **Minitab** gives two tests of heterogeneity of variance, which test the hypotheses

$$H_0 : \sigma_1^2 = \sigma_2^2$$

versus

$$H_a : \sigma_1^2 \neq \sigma_2^2.$$

If the tests indicate a significant difference in variances between the groups, that suggests that the Behrens–Fisher test is more trustworthy than the Gosset test. If the sample sizes in the two groups are (close to) equal, it turns out that the Gosset test is relatively insensitive to nonconstant variance, so it usually won’t make much of a difference.

What if the normality assumption is not appropriate? If we have large samples in both groups, this is not really a problem, since we can appeal to the Central Limit Theorem to justify our  $t$ -tests. What if at least one of our samples is small, however, and normality isn’t reasonable? Just as was true in the one-sample case, there are nonparametric tests that can be used. These tests are valid when the usual  $t$ -test assumptions do not hold, but they are less likely to identify a genuine effect than the  $t$ -tests when the normality assumption does hold.

The first of these tests is (Mood’s) median test. This test calculates the median of the joint sample of both groups. If there was no difference in location between the two groups, we would expect that about half of the values in each sample would be above the joint median, and about half would be below it. A pattern where one sample has most of its values above the joint median, while the other has most below the joint median, indicates a difference in location. A chi-squared ( $\chi^2$ ) test is used to assess the significance of the pattern. The hypotheses being tested here are

$H_0$  : The distribution of values in group 1 is the same as that in group 2

versus

$H_a$  : The distribution of values in group 1 is different from that in group 2.

Note that the distributions in the two groups aren’t specified; the only question is whether they are the same or different.

A second nonparametric test is the Mann–Whitney rank sum test. Roughly speaking, this is similar to a  $t$ -test, except that instead of using the true data values, the data used are the ranks of the values in the joint sample. Generally speaking, this test is more powerful than the median test. It also tests the hypotheses

$H_0$  : The distribution of values in group 1 is the same as that in group 2

versus

$H_a$  : The distribution of values in group 1 is different from that in group 2.

If the tests indicate a difference in distributions between the two groups, you then have to try to figure out what characterizes those differences: different locations, different variability, different shapes, and so on. This would involve looking at boxplots and/or histograms of the values from the two groups, and comparing statistical summaries. From a practical point of view, the tests are best able to detect shifts (differences) in the medians of the two groups, so comparing medians will often suggest how the distributions in the two groups differ.

It's important to see the distinction between this two-sample situation and the paired-samples situation we discussed earlier. The two-sample problem is characterized by the presence of two **independent** samples. There should be nothing that ties together the first observation in one sample with the first observation in the second sample; reordering the observations for one group while not doing so for the other would result in no loss of information. On the other hand, the paired-samples situation is characterized by there having been two values obtained from each individual observation in the sample (the observations here are generally objects that are the same for each variable — a specific company measured twice, a specific country measured twice, etc.). The first observations for each variable are tied together, since they both came from the same observation, and reordering one variable would make it impossible to see that connection (resulting in loss of information). Thus, while the paired-samples situation occurs when two variables are measured on one group of observations, the two-sample situation occurs when one variable is measured on two different groups of observations. While paired-sample tests compare variables, two-sample tests compare groups.

It's important to remember that in this two-sample situation variability is coming from both samples. Thus, you can end up with seemingly strange situations. Say our two groups had  $\bar{Y}_1 = 10$ ,  $s_1 = 40$ ,  $n_1 = 100$ ,  $\bar{Y}_2 = -0.5$ ,  $s_2 = 40$ ,  $n_2 = 100$ . Is the mean of the first sample significantly different from 0? The answer is yes:

$$t = \frac{10 - 0}{40/\sqrt{100}} = 2.50.$$

Since the sample mean of the second group is less than zero, you might think that this means that the two means are significantly different from each other, but that's not the case:

$$t = \frac{10 - (-0.5)}{40\sqrt{\frac{2}{100}}} = 1.86.$$

How can it be that  $\bar{Y}_1$  is significantly greater than 0, but not significantly greater than a mean that is **less** than zero? One way to think about it is that it's because we don't

know for sure that  $\mu_2$  really is less than or equal to zero; a confidence interval for  $\mu_2$  is  $(-8.44, 7.44)$ . This just reinforces once again that sample estimates are not the same as population parameters, since they have random variability associated with them.