MAT 2379, Introduction to Biostatistics

Chapter 12. Comparison of Two Independent Samples

In this chapter, we will compare the means corresponding to two independent populations. For this, we will use the following methods, called *methods of statistical inference*:
(a) interval estimation;
(b) hypothesis testing.
The case of dependent populations will be considered in Chapter 13. We skip Section 12.1.

12.2. Confidence Intervals and Tests for Means

We denote by μ_1, μ_2 the means of the two populations and by σ_1^2, σ_2^2 their variances. From the first population, we draw a sample of size n_1, whose mean is $\bar{X}_1$. From the second population, we draw a sample of size n_2, whose mean is $\bar{X}_2$. A point estimator for $\mu_1 - \mu_2$ is $\bar{X}_1 - \bar{X}_2$. A positive (respectively negative) observed value for this estimator is an indication that μ_1 might be larger (respectively smaller) than μ_2.

Example 1. (from p.219 of "Biostatistics. How it works" by Selvin) We want to examine if there is any difference in the final grade obtained in a statistics course between the male and female student populations. A sample of 37 male students has the mean $\bar{x}_1 = 85.738$. A sample of 30 female students has the mean $\bar{x}_2 = 89.4$.

A point estimate for $\mu_1 - \mu_2$ is $\bar{x}_1 - \bar{x}_2 = 85.738 - 89.4 = -3.662$. Since this a negative value, we might infer that μ_1 (the average grade for the male population) is smaller than μ_2 (the average grade for the student population). In what follows, we would like to refine this conclusion.

From the Central Limit Theorem, we know that the distribution of $\bar{X}_1$ is approximately normal with mean μ_1 and variance σ_1^2/n_1. Similarly, the distribution of $\bar{X}_2$ is approximately normal with mean μ_2 and variance σ_2^2/n_2. Since the populations are *independent*, we can deduce the following important fact:

The distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$

(the word "approximately" can be removed if the two populations are normal). By the standardization procedure, we obtain that:

$$\text{the distribution of } \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \text{ is approximately standard normal}$$

In practice, the variances σ_1^2 and σ_2^2 are unknown. Therefore, one has to replace them by suitable estimators. In order to do this, we have to know if the populations are normal or not. (If the populations are normal, we do not need large sample sizes; if they are not normal, we do.) Moreover, we need to know if the variances σ_1^2 and σ_2^2 are equal or not.

In summary, we have the following 4 cases:
Case (1). The two populations are normal with known variances σ_1^2 and σ_2^2.
Case (2). The two populations are normal with unknown and equal variances $\sigma_1^2 = \sigma_2^2$
Case (3). The two populations are normal with unknown and unequal variances $\sigma_1^2 \neq \sigma_2^2$

Case (4). The two populations are arbitrary and have unknown variances σ_1^2 and σ_2^2. (We need large sample sizes.)

We will only discuss Case (2).

Case (2). Inferences on $\mu_1 - \mu_2$: normal populations with equal variances

We consider the problem of comparing the means μ_1, μ_2 of two independent populations when the variances are equal (but unknown). We denote with σ^2 the common value for the two variances:

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

Note: To assess whether the populations are normal with equal variances, we draw the overlayed qq-plots together with the best fitted lines. If the plots appear to be approximately linear, and the slopes of the lines are approximately the same, we say that the two populations appear to be normal with equal variances. (See the R instructions for Chapter 12).

(a) Comparing the means: interval estimation

The random variable that we use for constructing a confidence interval for $\mu_1 - \mu_2$ is

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}} \tag{1}$$

which is known to have an approximate standard normal distribution. This variable contains the unknown variance σ^2 and cannot be used in this form. An estimator for σ^2 is

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

(called the *pooled variance estimator*). By replacing σ^2 with S_p^2 in formula (1) we are destroying the standard normal distribution. From the theory, it is known that:

$$T_0 = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \text{ has a } T \text{ distributionwith } n_1 + n_2 - 2 \text{ degrees of freedom}$$

The $(1 - \alpha)100\%$confidence interval for $\mu_1 - \mu_2$ is given by the formula:

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

where t is chosen from Table 17.4 such that $P(-t \leq T \leq t) = 1 - \alpha$ and T is a random variable with a T distribution with $n_1 + n_2 - 2$ degrees of freedom. If this interval contains only positive (respectively negative) values, then we can say that we are confident that μ_1 is larger (respectively smaller) than μ_2. If the interval contains 0, then we cannot conclude that there is a difference between the two means.

Example 2. (taken from p.366 of "Probability and Statistical Inference" by Hogg and Tanis) We want to compare the average scores μ_1 and μ_2 on a standardized test in mathematics taken by

students from large and small high schools. A sample of $n_1 = 9$ from large high schools yielded $\bar{x}_1 = 81.31$ and $s_1^2 = 60.76$. Another sample of size $n_2 = 15$ from small high schools yielded $\bar{x}_2 = 78.61$ and $s_2^2 = 48.24$. Compute a 95% confidence interval for the average difference between the scores of large and small high schools. Using this interval, can we say that there is enough evidence that the average score is higher in a large school than in a small school? Assume that the populations are normal with equal variances.

The pooled variance for the two samples is:

$$s_p^2 = \frac{8(60.76) + 14(48.24)}{9 + 15 - 2} = 52.79$$

From Table 17.4 we see that the value t such that $P(-t \le T \le t) = 0.95$ is $t = 2.074$, where T has a T distribution with 22 d.f. (Note that $P(T > t) = (1 - 0.95)/2 = 0.025$.) A 95% confidence interval for $\mu_1 - \mu_2$ is:

$$81.31 - 78.61 \pm 2.074\sqrt{(52.79)\left(\frac{1}{9} + \frac{1}{15}\right)}, \quad \text{or} \quad [-3.65;\ 9.05]$$

Since the interval contains 0 we cannot conclude that there is a difference between the two scores.

(b) Comparing the means: hypothesis testing

We consider the same problem of comparing two means when the variances are equal, but from the point of view of hypothesis testing. The observed value of the test statistic is

$$t_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(1/n_1 + 1/n_2)}}$$

which has a T distribution with $n_1 + n_2 - 2$ d.f. As in Chapter 11, we have the following three cases:

$$\textbf{Case I}: \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases} \qquad p - \text{value} = P\left(T \ge t_0\right)$$

$$\textbf{Case II}: \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 < \mu_2 \end{cases} \qquad p - \text{value} = P\left(T \le t_0\right)$$

$$\textbf{Case III}: \begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \ne \mu_2 \end{cases} \qquad p - \text{value} = 2P(T \ge |t_0|)$$

Example 2. (continued) Compare the means μ_1 and μ_2 from the point of view of hypothesis testing. Is there enough evidence that μ_1 is larger than μ_2?

We would like to test

$$H_0 : \mu_1 = \mu_2, \quad H_1 : \mu_1 > \mu_2$$

The observed value of the test statistic is

$$t_0 = \frac{81.31 - 78.61}{\sqrt{(52.79)\left(\frac{1}{9} + \frac{1}{15}\right)}} = 0.881$$

The $p - \text{value} = P(T \ge 0.881)$ is large (it lies between 0.1 and 0.25). We do not have enough evidence to say that μ_1 is larger than μ_2.