The Sampling Distribution of The Difference of Means From Two Independent Samples

We are using the **hypothesis testing** technique to help us make decisions. To this end, we always need to compute the ***p*-value** based on the sample data. Remember that the p-value is defined as:

$$\mathbb{P}\left(\text{observing the data we have or more extreme} \,|\, H_0 \text{ is true}\right)$$

In the case of comparing means from two independent samples, our **null and alternative hypotheses** are:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \Rightarrow \delta = \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 \neq \mu_2 \Rightarrow \delta = \mu_1 - \mu_2 \neq 0 \end{cases}$$

We can see that the estimator for the difference δ is basically $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$. Therefore, in order to calculate the p-value, we need to figure out the distribution of $\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2$.

If the assumptions are met, we have:

$$\boldsymbol{X}_1 \sim \mathcal{N}(\mu_1, \sigma_1^2) \text{ and } \boldsymbol{X}_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$$

Then we know that:

$$\bar{\boldsymbol{X}}_1 \sim \mathcal{N}\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \text{ and } \bar{\boldsymbol{X}}_2 \sim \mathcal{N}\left(\mu_2, \frac{\sigma_2^2}{n_2}\right)$$

Then if you read enough **Extra Reading Material**, we can find that:

$$\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2 \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

If we know the population variances, we should have:

$$\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathcal{N}(0, 1) \tag{1}$$

and we are done. That is all we need. However, we generally do not know σ_1^2 and/or σ_2^2. The intuitive thing to do is to replace them with their estimators (sample variances):

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

When we do that, we introduce some errors, and the formula (2) no longer follows the standard normal distribution. We need to look into this in more details.

1 Equal Variances: $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$

If we have a common variance: $\sigma_1^2 = \sigma_2^2 = \sigma^2$, then formula (1) becomes:

$$\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim \mathcal{N}(0, 1) \tag{3}$$

However, we do not know the value of σ^2. Intuitively, we need to replace σ^2 with the sample variance. When we do that, intuitively, the formula might follow a $\boldsymbol{t}$-distribution. There are still two things we need to decide or guess.

First, ***what would be a good estimator for*** σ^2***?*** We definitely should use the sample variances for that purpose. Let's denote this estimator $\boldsymbol{S}^2$. Then what should $\boldsymbol{S}^2$ look like? Is it $\boldsymbol{S}_1^2$? Or is it $\boldsymbol{S}_2^2$? Or is it some sort of combination of $\boldsymbol{S}_1^2$ and $\boldsymbol{S}_2^2$? Since we have a common variance here, both $\boldsymbol{S}_1^2$ and $\boldsymbol{S}_2^2$ are the unbiased estimator for the common variance σ^2. Using either one should be fine. However, since we have the information of two samples already, our intuition is that we could get a much more accurate estimation for σ^2 by combining them.

Second, ***what would be the*** **degree of freedom** (df) ***of the distribution?*** Since we have two samples, one with the degree of freedom $n_1 - 1$ and the other $n_2 - 1$. Intuitively, our guess for the degree of freedom would be $(n_1 - 1) + (n_2 - 1) = n_1 + n_2 - 2$.

I hope you feel that we are almost there. Once we replace σ^2 with $\boldsymbol{S}^2$, formula (3) becomes like this:

$$\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\boldsymbol{S}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{4}$$

Our guess is that formula (4) should follow a $\boldsymbol{t}$-distribution with a degree of freedom $n_1 + n_2 - 2$. Well ... maybe we should say that we want formula (4) to have a distribution of $\boldsymbol{\mathcal{T}(n_1 + n_2 - 2)}$. Towards this goal, can do some algebraic manipulations of formula (4):

$$\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\boldsymbol{S}^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\boldsymbol{S}^2} \cdot \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = \frac{\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\sqrt{\boldsymbol{S}^2}}$$

$$= \frac{\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\sqrt{\frac{\boldsymbol{S}^2}{\sigma^2}}}$$

$$= \frac{\frac{(\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2 \cdot \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}}{\sqrt{\frac{(n_1 + n_2 - 2)\boldsymbol{S}^2}{\sigma^2} \cdot \frac{1}{n_1 + n_2 - 2}}} \tag{5}$$

Now recall the definition of the random variable:

$$\boldsymbol{T} = \frac{\boldsymbol{Z}}{\sqrt{\boldsymbol{U}/\nu}} \sim \boldsymbol{\chi^2}(\nu)$$

where $\boldsymbol{Z}$ is the standard normal random variable and $\boldsymbol{U}$ is a χ^2 random variable with a degree of freedom ν. Now let's look at formula (5). The

numerator is the same as formula (3), which is the standard normal. If the term $\frac{(n_1+n_2-2)\boldsymbol{S}^2}{\sigma^2}$ follows $\boldsymbol{\chi^2}(n_1+n_2-2)$, then formula (5) would, by definition, follow $\boldsymbol{\mathcal{T}}(n_1+n_2-2)$. That would be very convenient for us. Note that:

$$\frac{(n_1-1)\boldsymbol{S}_1^2}{\sigma_1^2} = \frac{(n_1-1)\boldsymbol{S}_1^2}{\sigma^2} \sim \boldsymbol{\chi^2}(n_1-1)$$

and

$$\frac{(n_2-1)\boldsymbol{S}_1^2}{\sigma_2^2} = \frac{(n_2-1)\boldsymbol{S}_2^2}{\sigma^2} \sim \boldsymbol{\chi^2}(n_2-1)$$

Therefore, we can see that:

$$\frac{(n_1-1)\boldsymbol{S}_1^2}{\sigma^2} + \frac{(n_2-1)\boldsymbol{S}_2^2}{\sigma^2} = \frac{(n_1-1)\boldsymbol{S}_1^2+(n_2-1)\boldsymbol{S}_2^2}{\sigma^2} \sim \boldsymbol{\chi^2}(n_1+n_2-2)$$

Therefore, if we let $(n_1+n_2-2)\boldsymbol{S}^2 = (n_1-1)\boldsymbol{S}_1^2+(n_2-1)\boldsymbol{S}_2^2$, then formula (5) would follow $\boldsymbol{\chi^2}(n_1+n_2-2)$. That is:

$$\boldsymbol{S}^2 = \frac{(n_1-1)\boldsymbol{S}_1^2+(n_2-1)\boldsymbol{S}_2^2}{n_1+n_2-2}$$

which we called **the pooled estimator** for the common variance, which is basically the weighted average of the two sample variances with their degree of freedoms as the weights. We often denote the poole estimate as s_p^2.

Under the null hypothesis is true where $\mu_1-\mu_2=0$, our test statistic in this case becomes:

$$t = \frac{\bar{x_1}-\bar{x_2}}{\sqrt{s_p^2\cdot\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim \boldsymbol{\mathcal{T}}(n_1+n_2-2)$$

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

Once we compute the test statistics based on our sample data, the p-value can be calculated using the PDF of the $\boldsymbol{t}$-distribution with the specified degree of freedom.

2 Unequal Variances: $\sigma_1^2 \neq \sigma_2^2$

When there is not a common variance, that is, $\sigma_1^2 \neq \sigma_2^2$, we have to figure out the distribution of formula (2):

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

However, this is a lot more difficult task to do. The numerator is not the problem, because we already see that it follows a normal distribution. The problem is the term $\frac{\boldsymbol{S}_1^2}{n_1} + \frac{\boldsymbol{S}_2^2}{n_2}$ inside the square root of the denominator. We can see that it is in the form of:

$$\sum_{i=1}^{n} k_i \boldsymbol{S}_i^2 \tag{6}$$

where $k_i = \frac{1}{\nu_i + 1}$ and ν_i is the degree of freedom of $\boldsymbol{S}_i$. Formula (6) is often seen in statistics, and people need to deal with it frequently. However, the PDF cannot be expressed analytically. Therefore, we often use approximation to deal with it.

Intuitively, if we could use a $\boldsymbol{\chi^2}$-distribution as the approximation, then formula (2) would still follow a $\boldsymbol{t}$-distribution. Fortunately, we can. Therefore, under the null hypothesis is true where $\mu_1 - \mu_2 = 0$, our test statistic in this case simply becomes:

$$t = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim \boldsymbol{\mathcal{T}}(\nu), \text{ where } \nu \approx \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

This is the **Welch–Satterthwaite** approximation. We may come back to see how exactly this approximation is derived in the future. Whenever we are using this formula, we are performing a **Welch's t-test**.