Hypotheses Testing 'Cookbook'.

Gleb Karpov

October, 2022

1. Hypotheses about population mean, known variance

Prerequisites:

- Random Sample $(x_1, \ldots, x_n)$
- Population variance, σ^2, — is known (!)
- Either $n > 30$ - then CLT works fine, if not - assumption that population is normally distributed, *i.e.* $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

We want to test hypothesis $H_0 : \mu = \mu_0$ versus alternative $H_1 : \mu > \mu_0$.

Let us assume, that $\bar{x}$ is a mean value of the sample we have. Then, p-value is the probability for random variable sample mean $\bar{X}$ be even more extreme than $\bar{x}$ assuming that H_0 is true, i.e.:

$$p - \text{value} = P_{H_0}(\bar{X} > \bar{x}). \tag{1}$$

If conditions above are fulfilled, we need just to work out how to calculate probability in Eq. (1). That can be done sweet and simple:

$$P_{H_0}(\bar{X} > \bar{x}) = P_{H_0}\left(\frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} > \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = P(Z > z_{\text{score}}).$$

So, basically, one need to compute z_{score} of such test: $z_{\text{score}} = \frac{\bar{x}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$, and then calculate probability for standard normal variable Z being greater than test statistic.

After obtaining $p -$ value decision is made by comparing it with significance level α in a usual way.

2. Hypotheses about population proportion, large sample

Let's assume we have random sample: $x_1, \ldots, x_n$, with k positive answers, where each X_i is Bernoulli random variable with probability of success p_1, $n > 30$. We are interested in testing hypotheses about parameter p — population proportion.

If $n > 30$ then, as a consequence of the *Central Limit Theorem*:

$$\hat{p} \sim \mathcal{N}\left(p, \frac{p(1-p)}{n}\right) \tag{2}$$

We want to test hypothesis $H_0 : p = p_0$ versus alternative $H_1 : p > p_0$.

Let us assume, that $\tilde{p} = \frac{k}{n}$ is an observable proportion in the only sample we have. Then, p-value is the probability for random variable sample mean $\hat{p}$ be even more extreme than $\tilde{p}$ assuming that H_0 is true, i.e.:

$$p - \text{value} = P_{H_0}(\hat{p} > \tilde{p}). \tag{3}$$

If conditions above are fulfilled, we need just to work out how to calculate probability in Eq. (3). That can be done sweet and simple:

$$P_{H_0}(\hat{p} > \tilde{p}) = P_{H_0}\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} > \frac{\tilde{p} - p}{\sqrt{\frac{p(1-p)}{n}}}\right) = P(Z > z_{\text{score}}).$$

Then, one need to compute probability that standard normal variable Z is greater then z_{score} of the test, $z_{\text{score}} = \frac{\tilde{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$. Assuming that H_0 is true, we substitute value $p = p_0$, i.e. the value of population proportion we believe in.

Test decision is being carried out in a usual manner: by comparing p-value and α, or by comparison of coordinates: z_α and z_{score}.

3. Hypotheses about population mean, unknown variance

Prerequisites: Random Sample $(x_1, \dots, x_n)$ taken from the normal population, *i.e.* $X_i \sim \mathcal{N}(\mu, \sigma^2)$.

We want to test hypothesis $H_0 : \mu = \mu_0$ versus alternative $H_1 : \mu > \mu_0$.

Let us assume, that $\bar{x}$ is a mean value of the sample we have. Then, p-value is the probability for random variable sample mean $\bar{X}$ be even more extreme than $\bar{x}$ assuming that H_0 is true, i.e.:

$$p - \text{value} = P_{H_0}(\bar{X} > \bar{x}). \tag{4}$$

If conditions above are fulfilled, we need just to work out how to calculate probability in Eq. (4). That can be done sweet and simple:

$$P_{H_0}(\bar{X} > \bar{x}) = P_{H_0}\left(\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} > \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}\right) = P(t_{(n-1)df} > t_{\text{score}}). \tag{5}$$

So, basically, one need to compute t_{score} of such test: $t_{\text{score}} = \frac{\bar{x}-\mu_0}{\frac{S}{\sqrt{n}}}$, and then calculate probability for the Student's variable t be greater than test statistic.

After obtaining $p-$ value decision is made by comparing it with significance level α in a usual way.

3.1 Unknown variance, large sample

Please, note, that when number of degrees of freedom is large *enough* (say, more than 100, however some sources claim that even more than 30 is already large enough), then t-distribution behaves as Standard Normal distribution. In this case we can use another transformation:

$$P_{H_0}(\bar{X} > \bar{x}) = P_{H_0}\left(\frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}} > \frac{\bar{x} - \mu_0}{\frac{S}{\sqrt{n}}}\right) = P(Z > z_{\text{score}}). \tag{6}$$

And, by thus, to work with z_{score} and z_α instead of their analogs from t-distribution.

4. Hypotheses about difference of population means

4.1 Common prerequisites

Let us introduce $\bar{X}, \bar{Y}$ – sample means, random variables (as we used to). So it means that each time we have new sample X or Y, the value of their sample means very likely could be different.

But we have just two samples in our disposal! So let's introduce *observable* sample means $\bar{x}$ and $\bar{y}$, which are just constants, so-called *realizations* of corresponding random variables $\bar{X}$ and $\bar{Y}$.

4.1.1 Distributions of difference of sample means

Assume we have two independent samples: $X = X_1, \dots, X_n \sim f(\mu_1, \sigma_1^2)$, $Y = Y_1, \dots, Y_m \sim f(\mu_2, \sigma_2^2)$. We know that if $n,\ m > 30$ then it follows from the Central Limit Theorem that $\bar{X} \sim \mathcal{N}(\mu_1, \frac{\sigma_1^2}{n})$ and $\bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{m})$. Let's look at the properties of the random variable $\bar{X} - \bar{Y}$:

- $\mathbb{E}(\bar{X} - \bar{Y}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_1 - \mu_2$.
- As X and Y are independent samples we can write down simplified formula for the variance of $\bar{X} - \bar{Y}$:

$$\text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X} - \bar{Y}) = \text{Var}(\bar{X}) + \text{Var}(\bar{Y}) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}.$$

Because sum of two normal random variables is a normal random variable, we obtain distribution of $\bar{X} - \bar{Y}$:

$$\bar{X} - \bar{Y} \sim \mathcal{N}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right). \tag{7}$$

4.1.2 Two-tailed test

Let us consider two-tailed test. We want to test null hypothesis $H_0 : \mu_1 = \mu_2$ versus alternative $H_1 : \mu_1 \neq \mu_2$. When calculating p-value we need to cover two extreme cases: be greater than the difference which we observe now and less than the same difference but negative.

$$p\text{-value} = P_{H_0}\left(\bar{X} - \bar{Y} > |\bar{x} - \bar{y}|\right) + P_{H_0}\left(\bar{X} - \bar{Y} < -|\bar{x} - \bar{y}|\right). \tag{8}$$

4.1.3 One-tailed test

There are two possible variants for one-tailed tests.

1. If want to test null hypothesis $H_0 : \mu_1 = \mu_2$ versus alternative $H_1 : \mu_1 - \mu_2 > 0$, then such test is called *right-tailed* test, i.e. 'bad' values to our point of view are in the right (positive) part of distribution density. We can calculate p-value in this case as follows:
$$p\text{-value} = P_{H_0}\left(\bar{X} - \bar{Y} > \bar{x} - \bar{y}\right). \tag{9}$$
2. If want to test null hypothesis $H_0 : \mu_1 = \mu_2$ versus alternative $H_1 : \mu_1 - \mu_2 < 0$, then such test is called *left-tailed* test, i.e. 'bad' values to our point of view are in the left (negative) part of distribution density. We can calculate p-value in this case as follows:
$$p\text{-value} = P_{H_0}\left(\bar{X} - \bar{Y} < \bar{x} - \bar{y}\right). \tag{10}$$

4.2 Known variances

Assume we have two independent samples: $X = X_1, \ldots, X_n \sim f(\mu_1, \sigma_1^2)$, $Y = Y_1, \ldots, Y_m \sim f(\mu_2, \sigma_2^2)$, and we explicitly know variances.

4.2.1 Two-tailed test

Let us consider one of the components of the p-value in Eq. (8) and assume that $\bar{x} - \bar{y} > 0$. We will use transformation to the standard normal variable:

$$P_{H_0}\left(\bar{X} - \bar{Y} > \bar{x} - \bar{y}\right) = P_{H_0}\left(\frac{\bar{X} - \bar{Y} - \cancelto{0}{(\mu_1 - \mu_2)}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} > \frac{\bar{x} - \bar{y} - \cancelto{0}{(\mu_1 - \mu_2)}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}\right) = P\left(Z > \underbrace{\frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}}_{z_{\text{score}}}\right) \tag{11}$$

Because of the symmetry of the Standard Normal distribution density, we can combine Eq. (8) and Eq. (11) as:

$$\begin{aligned} p\text{-value} =& 2P\left(Z > z_{\text{score}}\right), \text{ if } z_{\text{score}} > 0 \\ p\text{-value} =& 2P\left(Z < z_{\text{score}}\right), \text{ if } z_{\text{score}} < 0 \end{aligned} \tag{12}$$

From there one can perform comparison of p-value with test significance level α, or make comparison of scores: $|z_{\text{score}}|$ and $z_{\alpha/2}$ ($z_{\alpha/2}$ is such point that $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$).

4.2.2 One-tailed test

Let us consider *right-tailed* test: $H_0 : \mu_1 = \mu_2$ versus alternative $H_1 : \mu_1 > \mu_2$. In this case p-value coincides with the result of Eq. (11):

$$p\text{-value} = P_{H_0}\left(\bar{X} - \bar{Y} > \bar{x} - \bar{y}\right) = P\left(Z > \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}\right) = P\left(Z > z_{\text{score}}\right)$$

From there one can perform comparison of p-value with test significance level α, or make comparison of scores: z_{score} and z_α (z_α is such point that $P(Z > z_\alpha) = \alpha$).

4.3 Unknown but equal variances

Assume we have two independent samples: $X = X_1, \ldots, X_n \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $Y = Y_1, \ldots, Y_m \sim \mathcal{N}(\mu_2, \sigma_2^2)$. we do not know variances explicitly, but assume that they are equal: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

We need to introduce new entity to help us in construction of t-variable. This is a **pooled variance**:

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{(m+n-2)}, \tag{13}$$

where S_x^2 and S_y^2 are sample variances of sample X and Y respectively.

Then the following random variable behaves as Student's t-variable with $(m+n-2)$ d.f.:

$$\boxed{\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{m+n}{mn}}} \sim t_{(m+n-2)}} \tag{14}$$

4.3.1 Two-tailed test

Let us consider one of the components of the p-value in Eq. (8) and assume that $\bar{x}-\bar{y} > 0$. We will use transformation to the Student's t-variable:

$$P_{H_0}\left(\bar{X}-\bar{Y} > \bar{x}-\bar{y}\right) = P_{H_0}\left(\frac{\bar{X}-\bar{Y}-\cancelto{0}{(\mu_1-\mu_2)}}{S_p\sqrt{\frac{m+n}{mn}}} > \frac{\bar{x}-\bar{y}-\cancelto{0}{(\mu_1-\mu_2)}}{S_p\sqrt{\frac{m+n}{mn}}}\right) = P\left(t > \underbrace{\frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{m+n}{mn}}}}_{t_{\text{score}}}\right) \tag{15}$$

Because of the symmetry of the Student's t-distribution density, we can combine Eq. (8) and Eq. (15) as:

$$\begin{aligned} p\text{-value} =& 2P\left(t_{(n+m-2)} > t_{\text{score}}\right), \text{ if } t_{\text{score}} > 0 \\ p\text{-value} =& 2P\left(t_{(n+m-2)} < t_{\text{score}}\right), \text{ if } t_{\text{score}} < 0 \end{aligned} \tag{16}$$

From there one can perform comparison of p-value with test significance level α, or make comparison of scores: $|t_{\text{score}}|$ and $t_{\alpha/2}$ ($t_{\alpha/2}$ is such point that $P(t_{(n+m-2)} > t_{\alpha/2}) = \frac{\alpha}{2}$).

4.3.2 One-tailed test

Let us consider *right-tailed* test: $H_0 : \mu_1 = \mu_2$ versus alternative $H_1 : \mu_1 > \mu_2$. In this case p-value coincides with the result of Eq. (15):

$$p\text{-value} = P_{H_0}\left(\bar{X}-\bar{Y} > \bar{x}-\bar{y}\right) = P\left(t > \frac{\bar{x}-\bar{y}}{S_p\sqrt{\frac{m+n}{mn}}}\right) = P\left(t_{(n+m-2)} > t_{\text{score}}\right) \tag{17}$$

From there one can perform comparison of p-value with test significance level α, or make comparison of scores: t_{score} and t_α (t_α is such point that $P(t_{(n+m-2)} > t_\alpha) = \alpha$).

5. Hypotheses about difference of population proportions, large samples.

5.1 Common prerequisites

Let's assume we have two independent samples: $X_1, \ldots, X_n$, with k positive answers, where each X_i is Bernoulli random variable with probability of success p_1 (population proportion). Also sample $Y_1, \ldots, Y_m$, with r positive answers, where each Y_j is Bernoulli random variable with probability of success p_2.

We already know statistics $\hat{p}_1, \hat{p}_2$ – sample proportions, being random variables in their nature. But when we work with specific case we have just two samples. So let's introduce *observable* sample proportions $\tilde{p}_1$ and $\tilde{p}_2$, which are constants, *realizations* of corresponding random variables $\hat{p}_1$ and $\hat{p}_2$.

We need to introduce new variable to help us in Hypotheses Testing. This is a **pooled proportion**:

$$p_{\text{pool}} = \frac{k+r}{n+m} = \frac{\#\text{ of positive answers in both samples}}{\text{Total number of responses in both samples}}. \tag{18}$$

5.2 Distribution of difference of sample proportions

If $n,\ m > 30$ then as a consequence of the Central Limit Theorem we have

$$\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right), \qquad \hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right).$$

Let us look at the properties of the random variable $\hat{p}_1 - \hat{p}_2 = \frac{k}{n} - \frac{r}{m}$, which is the difference between two sample proportions:

- $\mathbb{E}(\hat{p}_1 - \hat{p}_2) = \mathbb{E}(\hat{p}_1) - \mathbb{E}(\hat{p}_2) = p_1 - p_2$.
- $\text{Var }(\hat{p}_1 - \hat{p}_2) = \text{Var}(\hat{p}_1) + \text{Var}(\hat{p}_2) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}$.

Because sum of two normal random variables is a normal random variable, we obtain distribution of $\hat{p}_1 - \hat{p}_2$:

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right) \tag{19}$$

5.3 Two-tailed test

Let us consider two-tailed test. We want to test null hypothesis $H_0 : p_1 = p_2$ versus alternative $H_1 : p_1 \neq p_2$. When calculating p-value we need to cover two extreme cases: for the difference of sample proportions be greater than the difference which we observe right now and less than the same difference but negative.

$$p\text{-value} = P_{H_0}\left(\hat{p}_1 - \hat{p}_2 > |\tilde{p}_1 - \tilde{p}_2|\right) + P_{H_0}\left(\hat{p}_1 - \hat{p}_2 < -|\tilde{p}_1 - \tilde{p}_2|\right). \tag{20}$$

Let us consider one of the components of the p-value in Eq. (20) and assume that $\tilde{p}_1 - \tilde{p}_2 > 0$. We will use transformation to the Standard Normal variable:

$$P_{H_0}\left(\hat{p}_1 - \hat{p}_2 > \tilde{p}_1 - \tilde{p}_2\right) = P_{H_0}\left(\underbrace{\frac{\hat{p}_1 - \hat{p}_2 - \cancelto{0}{(p_1 - p_2)}}{\sqrt{p_{\text{pool}}(1-p_{\text{pool}})\frac{m+n}{mn}}}}_{Z\sim\mathcal{N}(0,1)} > \underbrace{\frac{\bar{x} - \bar{y} - \cancelto{0}{(p_1 - p_2)}}{\sqrt{p_{\text{pool}}(1-p_{\text{pool}})\frac{m+n}{mn}}}}_{z_{\text{score}}}\right) = P\left(Z > z_{\text{score}}\right) \tag{21}$$

Because of the symmetry of the Standard Normal distribution density, we can combine Eq. (20) and Eq. (21) as:

$$\begin{aligned} p\text{-value} =& 2P\left(Z > z_{\text{score}}\right), \text{ if } z_{\text{score}} > 0 \\ p\text{-value} =& 2P\left(Z < z_{\text{score}}\right), \text{ if } z_{\text{score}} < 0 \end{aligned} \tag{22}$$

Then one can perform comparison of p-value with test significance level α, or make comparison of scores: $|z_{\text{score}}|$ and $z_{\alpha/2}$ ($z_{\alpha/2}$ is such point that $P(Z > z_{\alpha/2}) = \frac{\alpha}{2}$).

5.4 One-tailed test

There are two possible variants for one-tailed tests.

1. If want to test null hypothesis $H_0 : p_1 = p_2$ versus alternative $H_1 : p_1 - p_2 > 0$, then such test is called *right-tailed* test. We can calculate p-value in this case as follows:

$$p\text{-value} = P_{H_0}\left(\hat{p}_1 - \hat{p}_2 > \tilde{p}_1 - \tilde{p}_2\right). \tag{23}$$

2. If want to test null hypothesis $H_0 : p_1 = p_2$ versus alternative $H_1 : p_1 - p2 < 0$, then such test is called *left-tailed* test. We can calculate p-value in this case as follows:

$$p\text{-value} = P_{H_0}\left(\hat{p}_1 - \hat{p}_2 < \tilde{p}_1 - \tilde{p}_2\right). \tag{24}$$

Let us consider *right-tailed* test. In this case p-value coincides with the result of Eq. (21):

$$p\text{-value} = P_{H_0}\left(\hat{p}_1 - \hat{p}_2 > \tilde{p}_1 - \tilde{p}_2\right) = P\left(Z > \frac{\bar{x} - \bar{y}}{\sqrt{p_{\text{pool}}(1-p_{\text{pool}})\frac{m+n}{mn}}}\right) = P\left(Z > z_{\text{score}}\right)$$

From there one can perform comparison of p-value with test significance level α, or make comparison of scores: z_{score} and z_α (z_α is such point that $P(Z > z_\alpha) = \alpha$).