Mathematical Statistics

Class 7. Population means difference. Population proportions difference. Student's t-distribution.

MDI, September 2022.

Difference of population means, known population variance

Assume we have two independent samples: $X = X_1, \dots, X_n \sim f(\mu_1, \sigma_1^2)$, and $Y = Y_1, \dots, Y_m \sim f(\mu_2, \sigma_2^2)$, and we explicitly know variances. We are interested in estimation of their means difference, parameter $\theta = \mu_1 - \mu_2$. Let's introduce $\hat{\theta} = \bar{X} - \bar{Y}$, the point estimator of θ. It has following properties:

- $\mathbb{E}(\hat{\theta}) = \mathbb{E}(\bar{X}) - \mathbb{E}(\bar{Y}) = \mu_1 - \mu_2 = \theta$, so $\hat{\theta}$ is an unbiased estimator of θ.
- As X and Y are independent samples we can write down simplified formula for the variance of $\hat{\theta}$:

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

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 $\bar{Y} \sim \mathcal{N}(\mu_2, \frac{\sigma_2^2}{m})$. Because sum of two normal random variables is a normal random variable, we obtain distribution of $\hat{\theta}$:

$$\hat{\theta} \sim \mathcal{N}\left(\theta, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$$

If we want to obtain a confidence interval for $\theta = \mu_1 - \mu_2$, we perform a classic procedure:

$$\begin{aligned} 1 - \alpha = P(L < \mu_1 - \mu_2 < U) = P(-U < -\theta < -L) = \\ = P\left(\frac{\hat{\theta} - U}{\operatorname{Var}(\hat{\theta})} < \frac{\hat{\theta} - \theta}{\operatorname{Var}(\hat{\theta})} < \frac{\hat{\theta} - L}{\operatorname{Var}(\hat{\theta})}\right) \end{aligned}$$

We can notice that the fraction in the middle of the last inequality is standard normal variable $Z \sim \mathcal{N}(0, 1)$. Because the density function of that distribution is symmetric, statisticians prefer to make symmetric bounds for obtained random variable:

$$1 - \alpha = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right)$$

We can estimate constant value $z_{\alpha/2}$ from the statistical table or any calculator. As we are done with that, we can find out the bounds L and U for the confidence interval for population mean μ itself:

$$L = \bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \tag{1}$$

$$U = \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \tag{2}$$

Finally, the $(1-\alpha)100\%$ confidence interval for the unknown population mean μ:

$$\boxed{\mu \in \left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}\right)} \tag{3}$$

Difference of population proportions

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, $n > 30$. 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, $m > 30$. We are interested in estimation of the parameter $\theta = p_1 - p_2$.

We introduce the point estimator $\hat{\theta} = \frac{k}{n} - \frac{r}{m} = \hat{p}_1 - \hat{p}_2$, which is the difference between two sample proportions. The properties of the estimator are:

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

If $n,\ m > 30$ then as a consequence of the Central Limit Theorem we have $\hat{p}_1 \sim \mathcal{N}\left(p_1, \frac{p_1(1-p_1)}{n}\right)$ and $\hat{p}_2 \sim \mathcal{N}\left(p_2, \frac{p_2(1-p_2)}{m}\right)$. Because sum of two normal random variables is a normal random variable, we obtain distribution of $\hat{\theta}$:

$$\hat{\theta} \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right) \tag{4}$$

Then again, classic procedure:

$$\begin{aligned} 1 - \alpha = P(L < p_1 - p_2 < U) = P(-U < -\theta < -L) = \\ = P\left(\frac{\hat{\theta} - U}{\text{Var}(\hat{\theta})} < \frac{\hat{\theta} - \theta}{\text{Var}(\hat{\theta})} < \frac{\hat{\theta} - L}{\text{Var}(\hat{\theta})}\right) \end{aligned}$$

If all necessary conditions are fulfilled, and (4) is true, then the fraction $\frac{\hat{\theta}-\theta}{\text{Var}(\hat{\theta})}$ behaves as Standard Normal random variable $Z \sim \mathcal{N}(0,1)$. And the last equation can be seen again as:

$$1 - \alpha = P\left(-z_{\alpha/2} < Z < z_{\alpha/2}\right).$$

We estimate constant $z_{\alpha/2}$ from the table, according to our choice of confidence level. After that is done, we can write down bounds for required confidence interval:

$$L = \hat{p}_1 - \hat{p}_2 - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$$
$$U = \hat{p}_1 - \hat{p}_2 + z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}},$$

where we change p_1 and p_2 to their point estimates, because we do not know the true parameters, and point estimates are the only thing we have in disposal.

The $(1-\alpha)100\%$ Confidence Interval for the difference of population proportions:

$$\boxed{p_1 - p_2 \in (L, U)} \tag{5}$$

Problems

1. Sample of Small Business Centre clients considering starting a business were questioned. Of a random sample of 94 males, 50 received assistance in business planning. Of an independent random sample of 68 females, 40 received assistance in business planning. Find a 99% confidence interval for the difference between the population proportion of male and female clients who received assistance in business planning.

Sampling distribution of the sample mean with unknown population variance

Assume we have a sample $X = X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. Our object of interest is unknown population mean μ, we want to construct a confidence interval, or to perform some hypothesis testing. The main problem is that we can not use the previous asymptotic transition scheme, because of unknown variance, which plays its role in the formula:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1).$$

Motivation: to use the Student's t-distribution and cancel out the unknown variance.

$$1-\alpha = P(L < \mu < U) = P(-U < -\mu < -L) = P\left(\frac{\bar{X} - U}{\frac{S}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{S}{\sqrt{n}}}\right) \tag{6}$$

The fraction in the middle of inequality is exactly a Student's t-variable. Because of the fact that this distribution is also symmetric, as standard normal, statisticians prefer to make symmetric bounds as well:

$$1 - \alpha = P(-t_{\alpha/2} < t(n-1\ df) < t_{\alpha/2})$$

We can estimate $t_{\alpha/2}$ from the table or any calculator. Once we are done with that, we can find out the bounds L and U for the confidence interval for population mean μ itself.

$$L = \bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}} \tag{7}$$

$$U = \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}} \tag{8}$$

And finally the $(1-\alpha)100\%$ confidence interval for the unknown population mean μ:

$$\boxed{\mu \in \left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)} \tag{9}$$

Problems

1. A random sample of 5 states gave the following areas (in 1000 square miles): 147, 84, 24, 85, 159. Stating any assumptions you need to make, find the 95% confidence interval for the mean area for all 50 states in the United States.

2. In a study of maximal aerobic capacity, 12 women were used as subjects, and one measurement that was blood plasma volume. The following data give their blood plasma volumes in liters:

   3.15 2.99 2.77 3.12 2.45 3.85 2.99 3.87 4.06 2.94

   Assume that these are observations of a normally distributed random variable X that has mean μ and standard deviation σ. Find a 90% confidence interval for μ.

3. A random sample of 5 observations from a normal distribution with mean μ and variance σ^2 gives a sample mean 100. An Independent random sample of size 10 from the same population has sample variance 9. Find a 90% confidence interval for the population mean.