Confidence Intervals: possible scenarios.

Gleb Karpov

October, 2022

1. Estimating population mean. Population variance is known.

Prerequisites:

- Random Sample $(X_1, \ldots, X_n)$ of I.I.D. variables
- Population variance, σ^2, — is known (!)
- Either $n > 30$ - then CLT works fine, if not - assumption that population is normally distributed, *i.e.* $X_i \sim \mathcal{N}(\mu, \sigma^2$.

One of possible ways to obtain confidence interval:

$$1 - \alpha = P(L < \mu < U) = P(-U < -\mu < -L)$$

Also, if conditions are met, we can write transition to the standard normal variable:

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

And the previous equation takes form:

$$1 - \alpha = P\left(\frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} < Z < \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}}\right)$$

Let us consider right tail. Latter means that $P(Z > \frac{\bar{X}-L}{\frac{\sigma}{\sqrt{n}}}) = \alpha/2$. We call that point $z_{\alpha/2}$, *i.e.*, such point that to the right of it lies area $\alpha/2$. We can find it out through a table of normal distribution.

$$z_{\alpha/2} = \frac{\bar{X} - L}{\frac{\sigma}{\sqrt{n}}} \to L = \bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Then consider left tail. Latter means that $P(Z < \frac{\bar{X}-U}{\frac{\sigma}{\sqrt{n}}}) = \alpha/2$. This point would be $-z_{\alpha/2}$, *i.e.*, such point that to the left of it lies area $\alpha/2$.

$$-z_{\alpha/2} = \frac{\bar{X} - U}{\frac{\sigma}{\sqrt{n}}} \to U = \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

And we return to the initial statement of $(L < \mu < U)$:

$$\boxed{\mu \in \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)} \tag{1}$$

2. Estimating 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 estimation of the population parameter p — population proportion.

We introduce a point estimator $\hat{p} = \frac{k}{n}$, which we call a sample proportion.

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

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

Then, classic procedure:

$$1-\alpha = P(L < p < U) = P(-U < -p < -L) = P\left(\frac{\hat{p}-U}{\mathrm{Var}(\hat{p})} < \frac{\hat{p}-p}{\mathrm{Var}(\hat{p})} < \frac{\hat{p}-L}{\mathrm{Var}(\hat{p})}\right)$$

If all necessary conditions are fulfilled, and Eq. (2) is true, then the fraction $\frac{\hat{p}-p}{\mathrm{Var}(\hat{p})}$ behaves as Standard Normal random variable $Z \sim \mathcal{N}(0,1)$. So we can rewrite the last equation as:

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

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

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

where we change p to its point estimate $\hat{p}$, because we do not know the true parameter, and sample proportion is the only thing we have in disposal.

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

$$\boxed{p \in (L, U)} \tag{3}$$

3. Difference of population means, known population variance

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. We are interested in estimation of their means difference, parameter $\theta = \mu_1 - \mu_2$.

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 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 $\hat{\theta}$:

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

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

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

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

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

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:

$$L = \bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \tag{4}$$
$$U = \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \tag{5}$$

Finally, the $(1-\alpha)100\%$ confidence interval for the difference of population means $(\mu_1 - \mu_2)$:

$$\boxed{\mu \in \left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}, \bar{X} + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}} \right)} \tag{6}$$

4. 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 $\theta = p_1 - p_2$.

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

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

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

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

$$\hat{\theta} \sim \mathcal{N}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\right) \tag{7}$$

Then again, classic procedure:

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

If all necessary conditions are fulfilled, and (7) 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:

$$\boxed{p_1 - p_2 \in (L, U)} \tag{8}$$

5. Sampling distribution of the sample mean with unknown population variance

Let us assume we have a sample $X = X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$. 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.

$$1 - \alpha = P(L < \mu < U) = P(-U < -\mu < -L) = P\left(\frac{\bar{X} - U}{\frac{S}{\sqrt{n}}} < \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} < \frac{\bar{X} - L}{\frac{S}{\sqrt{n}}}\right) \tag{9}$$

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.

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

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

$$\boxed{\mu \in \left(\bar{X} - t_{\alpha/2}\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}\frac{S}{\sqrt{n}}\right)} \tag{10}$$