Mathematical Statistics

Class 5. Limiting distributions. Confidence intervals.

MDI, September 2022.

Convergence Theorems

Central Limit Theorem

Easy formulation: Let $X_1, \ldots, X_n$, be a collection of i. i. d. variables, taken from a distribution that has mean μ and finite variance σ^2. (Which gives us $E[X_i] = \mu$, and $\mathrm{Var}(X_i) = \sigma^2$). Let us define $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, and $Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. As n grows, the distribution of the random variable Z_n tends to the standard normal:

$$\boxed{Z_n \xrightarrow{n \to \infty} Z \sim \mathcal{N}(0, 1)}$$

Guru formulation: Let $X_1, \ldots, X_n$, be a collection of i. i. d. variables, taken from a distribution that has mean μ and finite variance σ^2. Let us define $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$. Let $G_n(x)$ denote the CDF of random variable $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$. Then, for any possible x, the following holds:

$$\lim_{n \to \infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy;$$

which basically says that random variable $Z_n = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$ has a limiting *standard normal distribution.*

Confidence intervals

- Using only a point estimate to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.
- We can throw a spear where we saw a fish but we are more likely to miss. If we toss a net in that area, we have a better chance of catching the fish.
- If we report a point estimate, we probably will not hit the exact population parameter. If we report a range of plausible values – a confidence interval – we have a good shot at capturing the parameter.
- Mathematically, indeed, if our point estimator $\hat{\theta}$ has continuous distribution, then $P_\theta\{\hat{\theta} = \theta\} = 0$.

Definition: (Confidence Interval). Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X with pdf $f(x;\theta)$. Let $0 < \alpha < 1$ be specified. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. We say that the interval (L, U) is a $(1-\alpha)100\%$ confidence interval for an unknown parameter θ if

$$1 - \alpha = P_\theta\{\theta \in (L, U)\}.$$

The probability that the interval includes θ is $1 - \alpha$, which is called the *confidence level* of the interval.

Estimating population mean. Population variance is known.

What do you need?

- Random Sample
- Population variance, σ^2, — is known (!)
- $n > 30$ - CLT works fine, if not - assumption that population is normally distributed,

One of possible ways to write:

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

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 $(L < \mu < U)$:

$$\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Problems:

1. Random sample of 40 students. The average resting heart-rate for the sample was 76.3 bpm. Assume the population std is 12.5 bpm. Construct a 99% CI for the average resting heart-rate of the population.

2. Manager of a restaurant wants to estimate the mean amount m that a visitor spends for a lunch. A sample contains 36 visitors. Sample mean is $\bar{x} = \$3.60$. Manager knows that the standard deviation for one visitor is \$0.72. Find the confidence level corresponding to the interval $(\$3.5; \$3.7)$.

3. A college admission officer for an MBA program has determined that historically candidates have undergraduate grade point averages that are normally distributed with std 0.45. A random sample of 25 applications from the current year is taken, yielding a sample mean grade average of 2.90.

   - Find a 95% CI for the population mean
   - Based on these sample results, a statistician computes for the population mean a CI running from 2.81 to 2.99. Find the probability content associated with this interval.

Estimating population proportion. Large sample.

Let's assume we have random sample: $X_1, \dots, X_n$, 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:

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

Then, classic procedure:

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

If all necessary conditions are fulfilled, and Eq. (1) 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:

$$\boxed{p \in (L, U)} \tag{2}$$

Problems

1. Soon after he took office in 1963, President Johnson was approved by 160 out of a sample of 200 Americans. With growing disillusionment over his Vietnam policy, by 1968 he was approved by only 70 out of a sample of 200 Americans. What is the 95% confidence interval for the percentage of all Americans who approved Johnson in 1968? In 1963?