Confidence intervals. Part I

Statistics

Anton Afanasev

Higher School of Economics

DSBA 211
December 10, 2022

Seminar Overview

1. Quiz
2. Interval estimation
   - Motivation
   - Definition
   - How to derive
3. Confidence intervals for population mean
   - Population variance is known
   - Population variance is unknown
   - Independence of sample mean and sample variance
   - Accuracy of estimation
4. Confidence intervals for population variance
   - Population mean is unknown
   - Population mean is known

Quiz

Please find maximum likelihood estimation of θ using sample $X_1, \ldots, X_n$, generated from a normal distribution with parameters:

1. $\mu = 0, \sigma^2 = \theta^2$,
2. $\mu = \theta, \sigma^2 = 2\theta$.

What good of a point estimate?

Example

- Let $X \sim \mathcal{N}(0, 9)$.
- As a result of experiment, we have a sample with 3 observations:

$$x_1 = 4, \quad x_2 = 5, \quad x_3 = 6.$$

- Our best point estimate of $\mu = 0$ is:

$$\overline{x} = \frac{4+5+6}{3} = 5,$$

which is pretty far away from real μ.

- In order to give a quantitative perspective on how confident we are that the resulting value of a point estimate is close to real parameter, interval estimation is introduced.

Confidence intervals

- Let $X_1, \dots, X_n$ be a random sample from a population with parameter θ.

Definition

Confidence interval for θ with confidence level $1 - \alpha$ is a pair of random variables $L(X_1, \dots, X_n)$ and $U(X_1, \dots, X_n)$, such that:

$$\mathsf{P}\left(L(X_1, \dots, X_n) \le \theta \le U(X_1, \dots, X_n)\right) = 1 - \alpha.$$

- α – significance level.
- Sometimes confidence interval for θ with confidence level $1 - \alpha$ is defined as

$$\mathsf{P}\left(L(X_1, \dots, X_n) \le \theta \le U(X_1, \dots, X_n)\right) \ge 1 - \alpha,$$

which is used when probabilities are partially identified, e.g. discrete cases.

How to find confidence intervals?

Algorithm

1. Determine, which point estimator you will be using:
   - $\mu \to \overline{X}$ (mean),
   - $\sigma^2 \to S^2$ (variance),
   - $p \to \widehat{P}$ (proportion),
   - $\theta \to \widehat{\theta}$ (arbitrary parameter).

   Make sure you know the distribution of that point estimator.
2. Find a pivot function h, which depends only on the sample and estimated parameter. It must have a table distribution:
$$h(X_1, \ldots, X_n; \theta) \sim Z, \chi^2_k, t_k, F_{p,k}, \text{etc.}$$
3. Constrain pivot function with critical values:
$$\mathsf{P}\left(x_{1-\alpha/2} \leq h(X_1, \ldots, X_n; \theta) \leq x_{\alpha/2}\right) = 1 - \alpha.$$
4. Express θ in the inequality above.

Confidence interval for population mean μ

Population variance σ^2 is known

- Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance σ^2. The value of σ^2 is known.
- Pivot function in this case is the standardized $\overline{X}$:

$$h(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \overset{\text{CLT}}{\sim} \mathcal{N}(0, 1).$$

- Confidence interval of μ with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

$$\mathsf{P}\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Confidence interval for population mean μ

Population variance σ^2 is known

- $(1-\alpha)\cdot 100\%$ confidence interval for μ can be written as:

$$(\mu)_{1-\alpha} \in \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

- $\dfrac{\sigma}{\sqrt{n}}$ is called standard error of $\overline{X}$:

$$\text{S.E.}\left(\overline{X}\right) = \frac{\sigma}{\sqrt{n}}.$$

- Overall, for symmetric distributions of pivot functions, confidence intervals have the following view:

$$(\text{Param})_{1-\alpha} \in \text{Point Est.} \pm \text{Crit. Value}\left(\frac{\alpha}{2}\right) \times \text{S.E.}(\text{Point Est.})\,.$$

Confidence interval for population mean μ

Population variance σ^2 is known

- Let's simulate experiments to calculate sample mean and see, how confidence intervals are constructed.
- Refer to the 1st block in the link:

  Confidence intervals for μ

- Common misconception:
  - This is NOT a correct interpretation of confidence intervals:

    "$1-\alpha$ is a probability that μ belongs to CI".
  - This is:

    "$1-\alpha$ is a probability that CI contains μ",

  since μ is a constant, while CI itself is a random variable.

Problem 1

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

Problem 2

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

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

Confidence interval for population mean μ

Population variance σ^2 is unknown

- Let $X_1, \ldots, X_n$ be a random sample from a population, distributed as $\mathcal{N}\left(\mu, \sigma^2\right)$. The value of σ^2 is unknown.
- Old pivot function has unknown parameter:

$$h(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

- Let's substitute σ with its estimate $S = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2}$:

$$h(X_1, \ldots, X_n; \mu) = \frac{\overline{X} - \mu}{S / \sqrt{n}}.$$

- Both $\overline{X}$ and S are random variables $\Rightarrow h(X_1, \ldots, X_n; \mu)$ is not normally distributed.

Confidence interval for population mean μ

Population variance σ^2 is unknown

- Fisher's lemma:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- Applying Fisher's lemma to $h(X_1, \ldots, X_n; \mu)$:

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

- Confidence interval of μ with confidence level $1-\alpha$ then:

$$\mathsf{P}\left(-t_{n-1;\, \alpha/2} \leq \frac{\overline{X} - \mu}{S/\sqrt{n}} \leq t_{n-1;\, \alpha/2}\right) = 1 - \alpha.$$

$$\mathsf{P}\left(\overline{X} - t_{n-1;\, \alpha/2} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{n-1;\, \alpha/2} \cdot \frac{S}{\sqrt{n}}\right) = 1 - \alpha.$$

Confidence interval for population mean μ

Population variance σ^2 is unknown

- $(1-\alpha)\cdot 100\%$ confidence interval for μ can be written as:

$$(\mu)_{1-\alpha} \in \overline{X} \pm t_{n-1;\,\alpha/2} \cdot \frac{S}{\sqrt{n}}.$$

- $\dfrac{S}{\sqrt{n}}$ is called estimated standard error of $\overline{X}$:

$$\text{E.S.E.}\left(\overline{X}\right) = \frac{S}{\sqrt{n}}.$$

- If n is large ($n > 30$), we can use z-values, instead of t-values:

$$t_{n-1;\,\alpha/2} \xrightarrow[n\to\infty]{} z_{\alpha/2} \quad \text{and} \quad S \xrightarrow[n\to\infty]{} \sigma.$$

- For simulations refer to the 2nd block in the link:

Confidence intervals for μ

Problem 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.

$\overline{X}$ and S^2 are independent

- Assuming that sample $X_1, \ldots, X_n$ is derived from a normal population, statistics $\overline{X}$ and S^2 are independent.
- This allows us to state that

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1},$$

since numerator Z and denominator $\sqrt{\chi^2}$ should be independent to create a t-distribution.

$\overline{X}$ and S^2 are independent

Proof

- Let's consider vector of residuals:

$$\mathbf{X} = \begin{pmatrix} X_1 - \overline{X} & \dots & X_n - \overline{X} \end{pmatrix}^\top .$$

- Each component of $\mathbf{X}$ is uncorrelated with $\overline{X}$:

$$\mathsf{Cov}(X_j - \overline{X}, \overline{X}) = \mathsf{Cov}(X_j, \overline{X}) - \mathsf{Cov}(\overline{X}, \overline{X}) = \frac{\sigma^2}{n} - \frac{\sigma^2}{n} = 0,$$

  thus making $\overline{X}$ and $\mathbf{X}$ uncorrelated.
- If components of multivariate normal distribution are uncorrelated $\Longrightarrow$ they are independent.
- $\overline{X}$ and $\mathbf{X}$ are independent as components of vector $\begin{pmatrix} \overline{X} & \mathbf{X} \end{pmatrix}^\top$.
- Inherently, $\overline{X}$ is independent with $\mathbf{X}^\top \mathbf{X} = (n-1)S^2$.

Uncorrelatedness in multivariate normal distribution

- On the example of bivariate case.
  Let vector $\begin{pmatrix} X & Y \end{pmatrix}^\top$ be bivariate normal with joint p.d.f.:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right),$$

  where $\rho = \mathsf{Corr}(X, Y)$.

- Let $\rho = 0$:

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{1}{2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right) =$$

$$= \frac{1}{\sqrt{2\pi}\sigma_X} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right] \cdot \frac{1}{\sqrt{2\pi}\sigma_Y} \exp\left[-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right] =$$

$$= f_X(x) \cdot f_Y(y).$$

- X and Y are independent by definition.

Accuracy of estimation

- Accuracy of estimation e is a half-width of the corresponding confidence interval.
- For μ with known σ^2:

$$(\mu)_{1-\alpha} \in \overline{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

$$e = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}.$$

- How many observations required to get the accuracy e:

$$z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \leq e,$$

$$n \geq \frac{z_{\alpha/2}^2 \cdot \sigma^2}{e^2}.$$

Problem 4

The reaction time of a patient to a certain stimulus is known to have a standard deviation of 0.05 seconds. How large a sample of measurements must a psychologist take in order to be 95% confident and 99% confident, respectively, that the error in the estimate of the mean reaction time will not exceed 0.01 seconds?

Problem 5

1. A student constructed two 95% confidence intervals for unknown parameter $\theta : (-\infty, 4.2)$ and $(3.5, \infty)$. What could be the confidence of the interval (3.5, 4.2)?
2. A student constructed two 95% confidence intervals for unknown parameter $\theta : (-5, 4.2)$ and $(3.5, 7)$. What could be the confidence of the interval (3.5, 4.2)?

Confidence interval for population variance σ^2

Population mean μ is unknown

- Let $X_1, \ldots, X_n$ be a random sample from a population, distributed as $\mathcal{N}\left(\mu, \sigma^2\right)$. The value of μ is unknown.
- Unbiased estimator of σ^2 is:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 .$$

- Pivot function is given by Fisher's lemma:

$$h(X_1, \ldots, X_n; \sigma^2) = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

- Confidence interval of σ^2 with confidence level $1-\alpha$ then:

$$\mathsf{P}\left(\chi^2_{n-1;\,1-\alpha/2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{n-1;\,\alpha/2}\right) = 1 - \alpha.$$

Confidence interval for population variance σ^2

Population mean μ is unknown

$$\mathsf{P}\left(\frac{1}{\chi^2_{n-1;\,1-\alpha/2}} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{\chi^2_{n-1;\,\alpha/2}}\right) = 1-\alpha.$$

$$\mathsf{P}\left(\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\,1-\alpha/2}}\right) = 1-\alpha.$$

- $(1-\alpha)\cdot 100\%$ confidence interval for σ^2 can be written as:

$$\left(\sigma^2\right)_{1-\alpha} \in \left(\frac{(n-1)S^2}{\chi^2_{n-1;\,\alpha/2}};\frac{(n-1)S^2}{\chi^2_{n-1;\,1-\alpha/2}}\right).$$

- For simulations refer to the 1st block in the link:

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

Problem 6

A manufacturer bonds a plastic coating to a metal surface. A random sample of nine observations on the thickness of this coating is taken from a week's output. The sample thickness (in millimeters) were as follows:

$$19.8 \quad 21.2 \quad 18.6 \quad 20.4 \quad 21.6 \quad 19.8 \quad 19.9 \quad 20.3 \quad 20.8$$

Assuming that the population distribution is normal, find a 90% confidence interval for the population variance.

Confidence interval for population variance σ^2

Population mean μ is known

- Let $X_1, \ldots, X_n$ be a random sample from a population, distributed as $\mathcal{N}\left(\mu, \sigma^2\right)$. The value of μ is known.
- Unbiased estimator of σ^2 is:

$$\varsigma^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \mu)^2 .$$

- Pivot function:

$$h(X_1, \ldots, X_n; \sigma^2) = \frac{n\varsigma^2}{\sigma^2} \sim \chi^2_n.$$

- Confidence interval of σ^2 with confidence level $1 - \alpha$ then:

$$\mathsf{P}\left(\chi^2_{n;\ 1-\alpha/2} \leq \frac{n\varsigma^2}{\sigma^2} \leq \chi^2_{n;\ \alpha/2}\right) = 1 - \alpha.$$

Confidence interval for population variance σ^2

Population mean μ is known

$$\mathsf{P}\left(\frac{n\varsigma^2}{\chi^2_{n;\ \alpha/2}} \leq \sigma^2 \leq \frac{n\varsigma^2}{\chi^2_{n;\ 1-\alpha/2}}\right) = 1 - \alpha.$$

- $(1-\alpha)\cdot 100\%$ confidence interval for σ^2 can be written as:

$$\left(\sigma^2\right)_{1-\alpha} \in \left(\frac{n\varsigma^2}{\chi^2_{n;\ \alpha/2}}; \frac{n\varsigma^2}{\chi^2_{n;\ 1-\alpha/2}}\right).$$

- For simulations refer to the 2[nd] block in the link:

Confidence intervals for σ^2, p, $p_x - p_y$, $\mu_x - \mu_y$

Look at the time!