Interval Estimators

Rosario Barone

Tor Vergata University of Rome

Statistical tools for decision making

Undergraduate Degree in Global Governance

A.Y. 2023/2024

Remark: Standardization of the Normal Distribution

- To facilitate comparisons and calculations, we often standardize the normal distribution.
- Standardization involves transforming individual data points to a standard scale called Z-scores.
- The Z-score represents the number of standard deviations a data point is from the mean.
- Formula for standardization: $Z = \frac{X - \mu}{\sigma}$, where X is the data point, μ is the mean, and σ is the standard deviation.
- $Z \sim N(0, 1)$.

Point estimates → Interval estimates

- Our focus so far has been on point estimates of a parameter and their variances
- Useful when the estimator is approximately normal
- Their relevance is much less obvious when the distribution of the estimator is non normal or in case of small sample size
- Starting point for Interval estimates: Point estimates.

Pivotal quantity

Pivot

Let $X = (X_1, X_2, \dots, X_n)$ be a random sample from a distribution dependent on a parameter (or vector of parameters) θ. Consider a random variable $g(X, \theta)$ whose distribution remains constant for all θ. This random variable g is referred to as a *pivotal quantity* or a *pivot*.

- It is a function of observations and unobservable parameters such that the function's probability distribution does not depend on the unknown parameters.
- The function and its value can depend on the parameters of the model, but its distribution must not.

Interval estimator

Interval estimator

Consider a model $f(x;\theta)$ for data X. Then a pivot $G = g(X;\theta)$ is a function of X and θ that has known distribution independent of θ, invertible as a function of θ for each value of X. That is, given a region $\mathcal{A}$ such that $P(g(X;\theta) \in \mathcal{A}) = 1 - \alpha$, we can find a region $\mathcal{R}_\alpha(X, \mathcal{A})$ of the parameter space such that

$$1 - \alpha = P(z(X;\theta) \in \mathcal{A}) = P(\theta \in \mathcal{R}_\alpha(X, \mathcal{A}))$$

- Under repeated sampling, $\mathcal{R}_\alpha(x, \mathcal{A})$ is a realization of the random variable $\mathcal{R}_\alpha(X, \mathcal{A})$ that contains the true θ with probability $1 - \alpha$.

Confidence Intervals Construction: choice of the Pivot

- Exact pivot are rare, but there are numerous approximate pivots

- Let $X = (X_1, X_2, \ldots, X_n)$ be a random sample of size n. Let $f(X, \theta)$ be a statistical model and let $T(\theta)$ be an estimator for θ with variance $V(\theta)$

- We define the pivot $G(\theta) = \frac{T(\theta) - \theta}{V(\theta)^{1/2}}$.

- For large n, whatever the value of θ, G is approximately standard Normal and so G is a pivot.

Confidence Intervals Construction

$$P(G(\theta) \leq g) = P\left(\frac{T(\theta) - \theta}{V(\theta)^{1/2}} \leq g\right)$$

Then

$$P\left(g_{\alpha/2} \leq \frac{T(\theta) - \theta}{V(\theta)^{1/2}} \leq g_{1-\alpha/2}\right) = 1 - \alpha$$

where $g_{\alpha/2}$ is the quantile of the standard normal distribution, that is $\Phi(z_{\alpha/2}) = \alpha/2$. Then:

$$P(V(\theta)^{1/2} g_{\alpha/2} \leq T(\theta) - \theta \leq V(\theta)^{1/2} g_{1-\alpha/2}) = 1 - \alpha$$

and

$$P(T(\theta) - V(\theta)^{1/2} g_{1-\alpha/2} \leq \theta \leq T(\theta) - V(\theta)^{1/2} g_{\alpha/2}) = 1 - \alpha$$

Confidence Intervals Construction

- The random interval whose endpoints are $T(\theta) - V(\theta)^{1/2} g_{1-\alpha/2}$ and $T(\theta) - V(\theta)^{1/2} g_{\alpha/2}$ contains θ with probability $(1-\alpha)$. This interval is called *an approximate* $(1-\alpha)\%$ *confidence interval for* θ or *a confidence interval for* θ *with coverage probability* $(1-\alpha)\%$.

- If g is symmetric, $-g_{1-\alpha/2} = g_{\alpha/2}$, we may rewrite:

$$T(\theta) \pm V(\theta)^{1/2} g_{\alpha/2}$$

- **Caution:** It is incorrect to state that the confidence interval contains the true value of θ with a probability of $(1-\alpha)$. The interval may or may not "contain" the parameter value. It is not a matter of probability. The $(1-\alpha)\%$ confidence is related to the reliability of the estimation method, but not to the specific calculated interval.

Confidence Interval for the Mean (σ Known)

- Suppose to estimate the population mean (μ) with a known population standard deviation (σ). We define sample mean estimator $\bar{X}$ and

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

- the confidence interval is given by:

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- Where $\bar{X}$ is the sample mean, $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level (value of the quantile function at $\alpha/2$), σ is the known population standard deviation, and n is the sample size.

Example

- Let's consider an example to illustrate the computation and interpretation of a confidence interval.
- Suppose we want to estimate the average height (μ) of a population of adults based on a sample of 10 individuals.
- We measure the heights of the individuals in the sample and calculate a sample mean of 170 cm. We know that the standard deviation is 5 cm.
- Assuming a 95% confidence level, we can use the formula for a confidence interval to compute the interval estimate.

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Example

- Using the given values, the confidence interval is:

$$170 \pm 1.96\left(\frac{5}{\sqrt{10}}\right)$$

- Calculating the values, we get a confidence interval of (166.78, 173.22) cm.
- This means that we can be 95% confident that the true average height of the population lies within this interval: if we were to take repeated samples and construct confidence intervals using the same method, approximately 95% of those intervals would contain the true population mean.

Confidence Interval for the Mean (σ Unknown)

Generally, the population standard deviation σ, like the population mean μ, is unknown. Therefore, to obtain a confidence interval for the population mean, we rely on the sample statistics $\bar{X}$ and S. Then the statistic

$$t = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$$

follows a Student's t-distribution with $(n-1)$ degrees of freedom.

- If the random variable X does not have a normal distribution, the statistic t still approximately follows a Student's t-distribution due to the Central Limit Theorem.

Student's t Distribution

Probability Density Function (pdf)

The probability density function of a random variabe X distributed as a Student's t-distribution with d degrees of freedom is given by:

$$f(x) = \frac{\Gamma\left(\frac{d+1}{2}\right)}{\sqrt{d\pi}\Gamma\left(\frac{d}{2}\right)}\left(1+\frac{x^2}{d}\right)^{-\frac{d+1}{2}}$$

where Γ is the gamma function.

Moments

- $E(X) = Me(X) = Mo(X) = 0$
- $Var(X) = \frac{d}{d-2}$ if $d > 2$

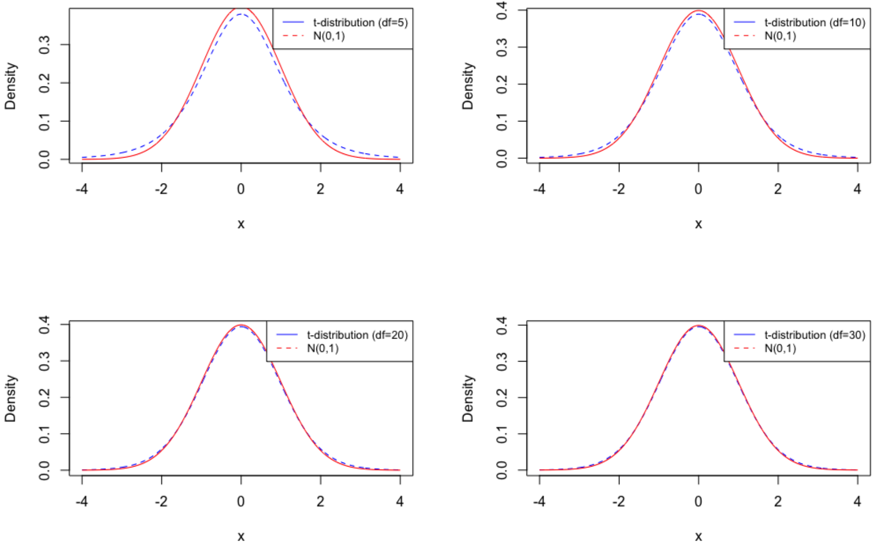
Student's t Distribution

- The Student's t-distribution with parameter n (degrees of freedom) governs the random variable

  $$T_d = \frac{Z}{\sqrt{K/d}},$$

  where Z and K are two independent random variables following the standard normal distribution $\mathcal{N}(0,1)$ and the chi-square distribution $\chi^2(n)$ with d degrees of freedom, respectively.

- The Student's t-distribution has a shape very similar to that of the standard normal distribution. However, the density plot appears flatter, and the area under the tails is greater than that of the standard normal distribution due to the fact that σ is unknown and is estimated by S. The uncertainty about σ results in greater variability.


t-distribution (df=5)
N(0,1)
Density
0.0
0.1
0.2
0.3
-4
-2
0
2
4
x
t-distribution (df=10)
N(0,1)
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
t-distribution (df=20)
N(0,1)
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x
t-distribution (df=30)
N(0,1)
Density
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
x

Confidence Interval for the Mean (σ Unknown):

- In the case of estimating the population mean (μ) with Unknown population standard deviation (σ), the confidence interval is given by:

$$\bar{X} - t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + t_{n-1,\alpha/2} \cdot \frac{S}{\sqrt{n}}$$

- Where $\bar{X}$ is the sample mean, $t_{n-1,\alpha/2}$ is the critical value corresponding to a cumulative area of $(1-\alpha)$ in the Student's t-distribution with $(n-1)$ degrees of freedom, S is the unknown population standard deviation, and n is the sample size.

Example

- Suppose we want to estimate the average height (μ) of a population of adults based on a sample of 10 individuals.
- We measure the heights of the individuals in the sample and calculate a sample mean of 170 cm and a sample standard deviation of 5.2 cm.
- Assuming a 95% confidence level, we can use the formula for a confidence interval to compute the interval estimate.

$$\bar{X} \pm t_{n-1,\alpha} \frac{S}{\sqrt{n}}$$

Example

- Using the given values, the confidence interval is:

$$170 \pm 2.26 \left(\frac{5.2}{\sqrt{10}} \right)$$

- Calculating the values, we get a confidence interval of (166.28, 173.72) cm.
- This means that we can be 95% confident that the true average height of the population lies within this interval: if we were to take repeated samples and construct confidence intervals using the same method, approximately 95% of those intervals would contain the true population mean.

Confidence Interval for a proportion

- Given a population whose elements possess a certain characteristic according to a given proportion indicated by the unknown parameter π, it is possible to construct a confidence interval for π based on the corresponding point estimator, given by the sample proportion $p = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$G(\pi) = \frac{p - \pi}{\sqrt{\frac{p(1-p)}{n}}} \approx \mathcal{N}(0, 1)$$

- the confidence interval is given by:

$$p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}$$

- Where p is the estimated proportion, $z_{\alpha/2}$ is the critical value from the standard normal distribution corresponding to the desired confidence level (value of the quantile function at $\alpha/2$) and n is the sample size.

Confidence Interval for the sample variance S

- If $x_1, x_2, \ldots, x_n$ is a random sample taken from a normal population with mean μ and variance σ^2, then if the sample variance is denoted by S^2, the random variable

  $$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

  has a chi-squared (χ^2) distribution with $n-1$ degrees of freedom.

Chi-Squared Distribution

Definition

If X^2 is a random variable following a chi-squared distribution with k degrees of freedom, we write $X^2 \sim \chi^2(k)$, with probability Density Function (PDF):

$$f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{(k/2)-1} e^{-x/2}$$

- $x > 0$
- k: Degrees of freedom.
- $\Gamma(\cdot)$: Gamma function.
- $E(X) = k$
- $Var(X) = 2k$

Confidence Interval for the sample variance S

We know that

$$X^2 = \frac{(n-1)S^2}{\sigma^2}$$

has a chi-squared (χ^2) distribution with $n-1$ degrees of freedom. The confidence interval is developed as :

$$\chi^2_{\alpha,n-1} \leq X^2 \leq \chi^2_{1-\alpha,n-1}$$

$$\chi^2_{\alpha,n-1} \leq \frac{(n-1)S}{\sigma^2} \leq \chi^2_{1-\alpha,n-1}$$

$$\frac{1}{\chi^2_{1-\alpha,n-1}} \leq \frac{\sigma^2}{(n-1)S} \leq \frac{1}{\chi^2_{\alpha,n-1}}$$

$$\frac{(n-1)S}{\chi^2_{1-\alpha,n-1}} \leq \sigma^2 \leq \frac{(n-1)S}{\chi^2_{\alpha,n-1}}$$