Math Revision Session

Statistics (6): Estimation

Jukina HATAKEYAMA

The University of Osaka, Department of Economics

February 17, 2025

1. Point Estimation

2. Interval Estimation
   - Confidence Interval
   - Confidence Interval for Mean (Known Variance)
   - Confidence Interval for Mean (Unknown Variance)
   - When Population Distribution is Unknown
   - Constructing a Confidence Interval

1. Point Estimation

2. Interval Estimation
   - Confidence Interval
   - Confidence Interval for Mean (Known Variance)
   - Confidence Interval for Mean (Unknown Variance)
   - When Population Distribution is Unknown
   - Constructing a Confidence Interval

Point Estimation

In statistics, **point estimation** refers to the process of estimating an unknown parameter using a single value, known as an **estimate**. The function used to generate this estimate is called an **estimator**.

Definition: Let θ be an unknown parameter. An estimator $\hat{\theta}$ is a function of the sample data used to approximate θ. The realised value of $\hat{\theta}$ for a given sample is called an estimate.

Unbiasedness

An estimator $\hat{\theta}$ is said to be **unbiased** for a parameter θ if:

$$\mathbb{E}[\hat{\theta}] = \theta.$$

This means that the expected value of the estimator equals the true parameter value, on average, over repeated sampling.

Consistency

An estimator $\hat{\theta}_n$ is **consistent** for a parameter θ if:

$$\hat{\theta}_n \xrightarrow{p} \theta \quad \text{as } n \to \infty.$$

This means that as the sample size increases, the estimator converges in probability to the true parameter value.

For any $c > 0$,

$$Pr(|\hat{\theta} - \theta| > c) \to 0 \quad \text{as } n \to \infty.$$

Efficiency

An estimator $\hat{\theta}$ is **efficient** if it has the smallest variance among all unbiased estimators of θ.

Mathematically, for two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, $\hat{\theta}_1$ is more efficient than $\hat{\theta}_2$ if:

$$\mathsf{Var}(\hat{\theta}_1) < \mathsf{Var}(\hat{\theta}_2).$$

The Cramér-Rao lower bound provides a theoretical lower limit for the variance of an unbiased estimator.

Cramér-Rao Lower Bound (CRLB)

The **Cramér-Rao lower bound** provides a theoretical lower bound on the variance of any unbiased estimator. It states that for an unbiased estimator $\hat{\theta}$ of a parameter θ, the variance satisfies:

$$\mathsf{Var}(\hat{\theta}) \geq \frac{1}{I(\theta)},$$

where $I(\theta)$ is the **Fisher information**, given by:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log L(\theta; X)\right].$$

This bound indicates the best possible precision an unbiased estimator can achieve.

Properties of Estimators: Bias

The **bias** of an estimator $\hat{\theta}$ is defined as the difference between its expected value and the true parameter value:

$$\text{Bias}(\hat{\theta}) = \mathbb{E}[\hat{\theta}] - \theta.$$

Interpretation:

- If $\text{Bias}(\hat{\theta}) = 0$, the estimator is **unbiased**.
- If $\text{Bias}(\hat{\theta}) \neq 0$, the estimator is **biased**.

A small bias may be acceptable if it significantly reduces variance, as seen in the bias-variance tradeoff.

Mean Squared Error (MSE)

The **mean squared error (MSE)** of an estimator $\hat{\theta}$ is given by:

$$\mathsf{MSE}(\hat{\theta}) = \mathbb{E}[(\hat{\theta} - \theta)^2].$$

The MSE can be decomposed as:

$$\mathsf{MSE}(\hat{\theta}) = \mathsf{Var}(\hat{\theta}) + (\mathbb{E}[\hat{\theta}] - \theta)^2.$$

This decomposition shows the trade-off between variance and bias in an estimator.

1 Point Estimation

2 Interval Estimation
   - Confidence Interval
   - Confidence Interval for Mean (Known Variance)
   - Confidence Interval for Mean (Unknown Variance)
   - When Population Distribution is Unknown
   - Constructing a Confidence Interval

Confidence Interval (CI)

A **confidence interval** is a range of values, derived from a sample, that is used to estimate the true value of an unknown population parameter. The interval provides an estimate of the uncertainty or variability of the population parameter based on the sample data.

Confidence Coefficient

The **confidence coefficient** refers to the probability that the confidence interval will contain the true population parameter. For example, a 95% confidence interval suggests that 95% of similarly constructed intervals would contain the true parameter value.
The confidence coefficient is typically expressed as a percentage, such as 90%, 95%, or 99%.

Confidence Interval Formula

The formula for calculating a confidence interval is:

$$CI = \hat{\mu} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

where:

- $\hat{\mu}$ is the sample mean,
- $z_{\alpha/2}$ is the critical value from the standard normal distribution,
- σ is the population standard deviation (or sample standard deviation if unknown),
- n is the sample size.

Confidence Interval for Mean (Known Variance)

In interval estimation, we estimate a parameter using a range of values, called a **confidence interval**.

For a normal population with known variance σ^2, the $(1-\alpha)100\%$ confidence interval for the mean μ 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,
- σ is the known population standard deviation,
- n is the sample size.

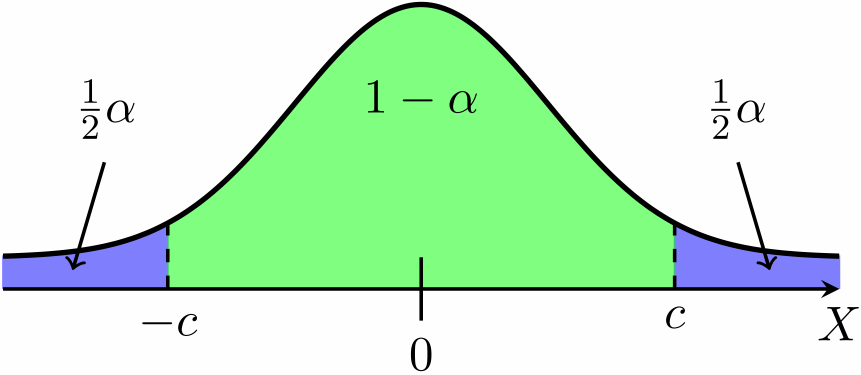
This interval provides a range where the true mean μ is likely to lie with probability $(1-\alpha)$.

- Now, we consider the interval estimation of an unknown population mean μ. We have 10 observations and know the population variance. We set the confidence coefficient to $1 - \alpha = 0.95$.
- Let $\bar{X}$ denote the sample mean, which follows a normal distribution with mean μ and variance 2.5:

$$\bar{X} \sim N_{\mathbb{R}}(\mu, 2.5)$$

Standardising the sample mean, we have:

$$Z = \frac{\bar{X} - \mu}{\sqrt{2.5}} \sim N_{\mathbb{R}}(0, 1)$$



$$Pr(-c \leq Z \leq c) = 1 - \alpha$$

$$Pr(Z \leq c) = \Phi(c) = 1 - \frac{\alpha}{2}$$

If we set $\alpha = 0.05$, then $Pr(Z \leq c) = \Phi(c) = 1 - 0.05/2 = 0.975$. This means that the probability of Z being less than or equal to c is 0.975

To perform interval estimation, that is, to create a confidence interval with a given confidence coefficient, we need to determine the point c that satisfies the conditions above. When Z follows a standard normal distribution, c can be found from the standard normal distribution table.
$\Rightarrow c \simeq 1.96$

$$\begin{aligned}
0.95 &= \Pr(-1.96 \leq Z \leq 1.96) \\
&= \Pr\left(-1.96 \leq \frac{\bar{X} - \mu}{\sqrt{2.5}} \leq 1.96\right) \\
&= \Pr\left(\bar{X} - 1.96 \times \sqrt{2.5} \leq \mu \leq \bar{X} + 1.96 \times \sqrt{2.5}\right)
\end{aligned}$$

If we write the realised value of the random variable, $\bar{X}$, as $\bar{x}$, then we have the confidence interval as:

$$\bar{x} - 1.96 \times \sqrt{2.5} \leq \mu \leq \bar{x} + 1.96 \times \sqrt{2.5}$$

If $\bar{x}$ from the 10 observations is 100, then we have:

$$96.9 \leq \mu \leq 103.1$$

Confidence Interval for Mean (Unknown Variance)

- Now, we consider the interval estimation of an unknown population mean μ. We have 10 observations and do not know the population variance. Instead, we use the sample variance S^2. We set the confidence coefficient to $1 - \alpha = 0.95$.
- Let $\bar{X}$ denote the sample mean, which follows a normal distribution with mean μ and sample variance S^2:

  $$\bar{X} \sim N_{\mathbb{R}}(\mu, \frac{S^2}{n})$$

  Standardising the sample mean, we have:

  $$t = \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$$

  where t_{n-1} is the t-distribution with $n - 1$ degrees of freedom.

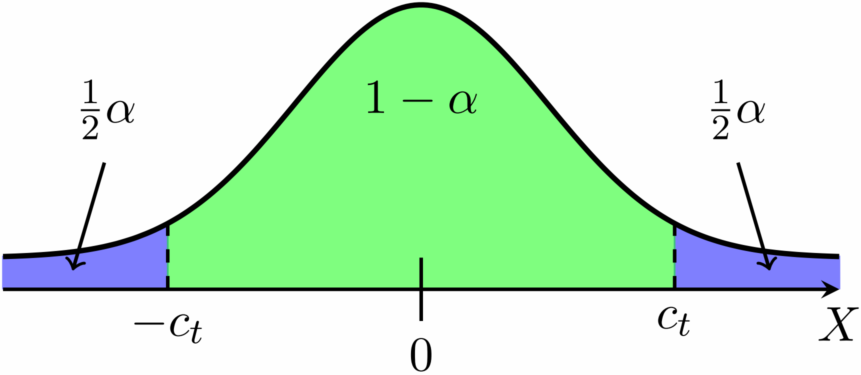
Relationship Between Sample Mean and Sample Variance

- The sample mean $\bar{X}$ and sample variance S^2 are not independent, as they are both calculated from the same sample.
- However, when the population follows a normal distribution, the ratio of the sample mean and sample variance follows at-distribution.
- Specifically, if the population is normally distributed as $N(\mu, \sigma^2)$, the sample mean $\bar{X}$ and sample variance S^2 are related as:

  $$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_\nu$$

  where $\nu = n - 1$ is the degrees of freedom.
- In this case, both the sample mean and sample variance follow the t-distribution simultaneously, which implies they are not independent.

NOTE: t distribution



$$Pr(-c_t \leq t \leq c_t) = 1 - \alpha$$

$$Pr(t \leq c_t) = F_{t_{n-1}}(c_t) = 1 - \frac{\alpha}{2}$$

If we set $\alpha = 0.05$, then $Pr(t \leq c_t) = F_{t_{n-1}}(c_t) = 1 - 0.05/2 = 0.975$. This means that the probability of t being less than or equal to c_t is 0.975.

To perform interval estimation, that is, to create a confidence interval with a given confidence coefficient, we need to determine the point c_t that satisfies the conditions above. When t follows a t-distribution with $n-1$ degrees of freedom, c_t can be found from the t-distribution table.

$\Rightarrow c_t \simeq 2.262$ (for $n = 10$, with 9 degrees of freedom)

$$
\begin{aligned}
0.95 &= \Pr(-2.262 \leq t \leq 2.262) \\
&= \Pr\left(-2.262 \leq \frac{\bar{X} - \mu}{\sqrt{S^2/n}} \leq 2.262\right) \\
&= \Pr\left(\bar{X} - 2.262 \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.262 \times \frac{S}{\sqrt{n}}\right)
\end{aligned}
$$

If we write the realised value of the random variable, $\bar{X}$, as $\bar{x}$, then we have the confidence interval as:

$$
\bar{x} - 2.262 \times \frac{S}{\sqrt{n}} \leq \mu \leq \bar{x} + 2.262 \times \frac{S}{\sqrt{n}}
$$

If $\bar{x}$ from the 10 observations is 100 and the sample standard deviation $S = 5$, then we have:

$$
96.13 \leq \mu \leq 103.87
$$

When Population Distribution is Unknown

- When the population distribution is unknown, the sample mean $\bar{X}$ can still be approximated by a normal distribution due to the Central Limit Theorem (CLT), provided that the sample size is large enough.
- The CLT states that, regardless of the population distribution, the distribution of the sample mean $\bar{X}$ will tend to a normal distribution as the sample size n increases:

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \xrightarrow[n\to\infty]{\mathrm{d}} N(0,1)$$

- This approximation holds even if the population variance σ^2 is unknown. In this case, we can use the sample variance S^2 to estimate σ^2.
- For large n, the sampling distribution of the sample mean is approximately normal, and we can use a normal distribution for constructing confidence intervals or performing hypothesis tests:

$$\bar{X} \pm Z_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

How to Construct a Confidence Interval (1)

1. **Find the variance of the estimator:** First, calculate the variance of the estimator $\hat{\theta}$ based on the sample data.
2. **Standardise the estimator:** Next, standardise the estimator to obtain a standardised random variable. This can be done using:

$$Z = \frac{\hat{\theta} - \theta}{\text{Standard Error}}$$

3. **Substitute unknown parameters:** If the variance of the estimator contains unknown parameters, replace these parameters with their consistent estimators. For instance, replace the population variance with the sample variance if the population variance is unknown.
4. **Determine the quantile:** Using the distribution that the standardised random variable follows, determine the quantile based on $1 - \alpha$. For example, if the standardised variable follows a normal distribution, the critical value can be obtained from the standard normal distribution.
5. **Construct the confidence interval:** Finally, the confidence interval for the parameter θ is given by:

$$\hat{\theta} \pm \text{Critical Value} \times \text{Standard Error}$$