50008 - Probability and Statistics - Lecture 6

Oliver Killane

07/03/22

Efficient Consistent Estimator

Lecture Recording

Lecture recording is available here

We can quantify how *good* estimators are. For example with the **Estimator Bias** (difference between the expected using the estimator and the parameter $bias(T) = E[T|\theta] - \theta$). We also wanto to quantify the **Efficiency of Estimators**.

Definition: Estimator Efficiency

Given two unbiased estimators $\hat{\Theta}(\underline{X})$ and $\tilde{\Theta}(\underline{X})$ where $\underline{X} = (X_1, \dots, X_n)$ (a sample containing n observations $X \dots$).

We can compare the mean, variances etc to determine which estimator is more efficient (typically lower variance)

$\hat{\Theta}$ is more efficient than $\tilde{\Theta}$ if:

$$\forall\theta Var_{\hat{\Theta}}(\hat{\Theta}|\theta) \leq Var_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta) \quad \text{or} \quad \exists\theta Var_{\hat{\Theta}}(\hat{\Theta}|\theta) < Var_{\tilde{\Theta}|\theta}(\tilde{\Theta}|\theta)$$

More efficient means less variance in estimates.

IF an estimator is more efficient than any other possible estimator, it is called **efficient**.

Example: Bias and Efficiency

Given a population with mean μ and variance σ^2. We have a sample:

$$\underline{X} = (X_1, \ldots, X_n)$$

We consider two extimators:

1. $\hat{M} = \overline{X}$ (the sample mean)
2. $\tilde{M} = X_1$ (the first observation in the sample)

We can compute the bias as for both:

1. The expected value of the sample mean is the population mean μ, hence $\hat{M}$ is unbiased.
2. The expected value of any observation is μ, so the first observation in the sample is also ubiased.

Next we can consider the variance.

For a single sample we know the variance will be σ^2, hence:

$$Var_{\tilde{M}}(\tilde{M}|\mu \text{ and } \sigma^2) = Var(X_1) = \sigma^2$$

However for the sample mean, we know can use the **Central Limit Theorem** to determine that the variance of the mean of a sample will be divided by the sample size.

$$Var_{\hat{M}}(\hat{M}|\mu \text{ and } \sigma^2) = Var(\overline{X}) = \frac{\sigma^2}{n}$$

Hence for all values of n, the variance of $\hat{M} \leq \tilde{M}$ (at $n = 1$ they are equal), so $\hat{M}$ is the more efficient estimator.

Definition: Estimator Consistency

A consistent estimator improves as the sample size grows. Formally:

$$\forall \epsilon > 0 \; P(|\hat{\Theta} - \theta|) \to 0 \quad \text{as} \quad n \to \infty$$

If $\hat{\Theta}$ is unbiased, then:

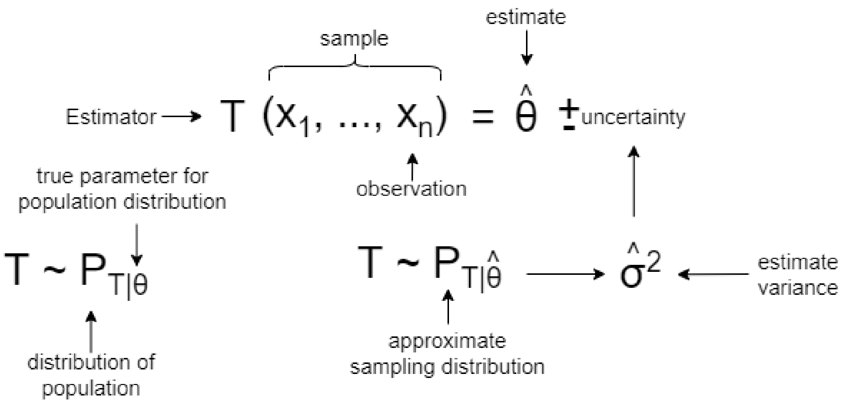
$$\lim_{n \to \infty} Var(\hat{\Theta}) = 0 \Rightarrow \hat{\Theta} \quad \text{is consistent}$$

Note: $\overline{X}$ (sample mean) is a consistent estimator for any population.

Confidence Intervals

Lecture Recording

Lecture recording is available here



In order to quantify our degree of uncertainty in an estimate $\hat{\theta}$, when the true value θ is unknown, we use use our estimate as the true value, to compute the distribution $P_{T|\hat{\theta}}$ (the approximate sampling distribution).

Known Variance

Confidence Interval

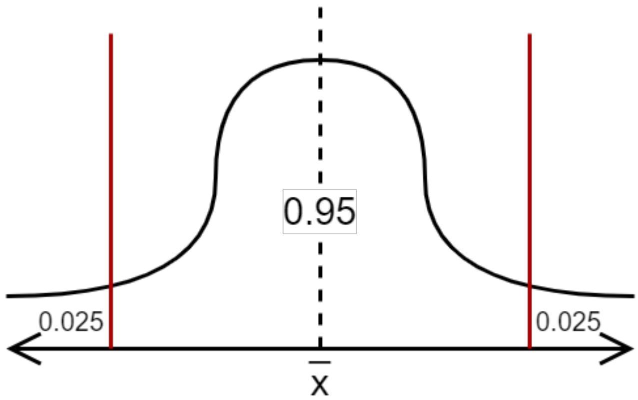
If we know the true variance of the population, then the sample mean would be distributed as:

$$\overline{X} \sim N\left(\overline{x}, \frac{\sigma^2}{n}\right)$$

If μ (population mean) $= \overline{x}$, then we can say that (using thestandard normal distribution) there is a 95% probability the observed statistic $\overline{X}$ is in the range:

$$\left[\overline{x} - 1.96\frac{\sigma}{n}, \overline{x} + 1.96\frac{\sigma}{n}\right]$$

(Double ended, 95% confidence interval for μ)



With the Standard Normal Distribution

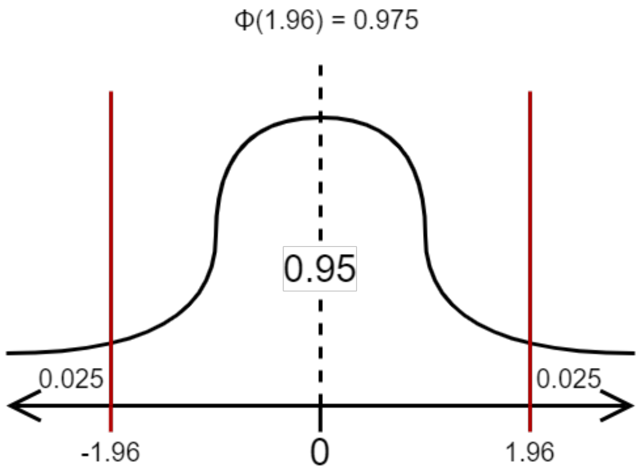
We can define any normal distribution in terms of the standard normal distribution.

$$X \sim N(\mu, \sigma^2) \Leftrightarrow Y = \frac{X - \mu}{\sigma} \Leftrightarrow Y \sim N(0, 1)$$

We can then use tables for the standard normal distribution, using $\Phi(z) = P(X \leq z)$ given $Z \in N(0, 1)$:

Note if you have sample size as part of the variance, $Y = \dfrac{X - \mu}{\left(\dfrac{\sigma}{\sqrt{n}}\right)}$.

For example in the previous confidence interval, we used the normal distribution to calculate the values.



Given the critical value z for the normal distribution e.g 1.96 for double-ended 95% confidence interval, we have:

Standard Normal	$X \sim N(0, 1)$	$[-z, z]$
Normal Distribution	$X \sim N(\mu, \sigma^2)$	$\mu - z\sigma, \mu + z\sigma$
Sample Mean	$\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$	$\left[\mu - z\dfrac{\sigma}{\sqrt{n}}, \mu + z\dfrac{\sigma}{\sqrt{n}}\right]$
Population mean	$\mu \sim N\left(\overline{X}, \dfrac{\sigma^2}{n}\right)$	$\left[\overline{x} - z\dfrac{\sigma}{\sqrt{n}}, \overline{x} + z\dfrac{\sigma}{\sqrt{n}}\right]$

Example: Employees Opinions on the Board

A corporation surveys employees on wether they think the board is doing a good job.

1000 employees are randomly selected, and 732 say the board is doing a good job. Find the 99% confidence interval for the proportion of the employees that think the board is doing a good job. Assume the variance is $\sigma^2 = 0.25$.

First we get the sample mean:

$$\overline{x} = \frac{732}{1000} = 0.732$$

Next we determine the standard deviation:

$$\sigma = \sqrt{0.25} = 0.5$$

We want to get the double-ended 99% interval, so each tail will have size 0.005. By using the standard normal distribution we have $\Phi(2.576) = 0.995$, so $z = 2.576$.

Hence we can calculate the interval as:

$$\begin{aligned}
\mu &= \left[\overline{x} - z\frac{\sigma}{\sqrt{n}}, \overline{x} + z\frac{\sigma}{\sqrt{n}}\right] \\
&= \left[0.732 - 2.576\frac{0.5}{\sqrt{1000}}, 0.732 + 2.576\frac{0.5}{\sqrt{1000}}\right] \\
&= \left[0.732 - 2.576\frac{0.5}{\sqrt{1000}}, 0.732 + 2.576\frac{0.5}{\sqrt{1000}}\right] \\
&\approx 0.732 \pm 0.0407
\end{aligned}$$

Unknown Variance

In a problem where we are trying to fit a normal distribution, but both the mean and variance are unknown.

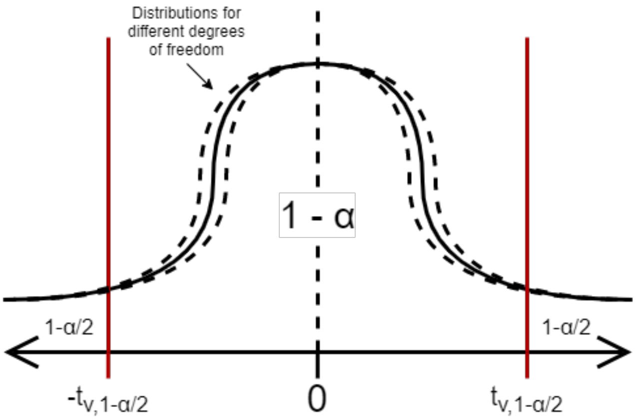
$$\text{Bias Corrected Variance } S_{n-1} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

We use the bias corrected variance of our sample, and as a result must use a different distribution to the normal distribution.

Normal Distribution (σ known)	**Studen't t distribution (σ unknown)**
$\dfrac{\overline{X} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)} \sim N(0,1)$	$\dfrac{\overline{X} - \mu}{\left(\frac{s_{n-1}}{\sqrt{n}}\right)} \sim t_{n-1}$

In the student's distribution we set degrees of freedom $\nu = n - 1$.

For a double ended confidence $(100-\alpha)\%$, we compute $t_{\nu=n-1,\ 1-\alpha/2}$ to find the critical values (the places where the tails start/ the α-quantile of t_ν).



$$\left[\overline{x} - t_{\nu=n-1,\ 1-\alpha/2} \times \frac{s_{n-1}}{\sqrt{n}},\ \overline{x} + t_{\nu=n-1,\ 1-\alpha/2} \times \frac{s_{n-1}}{\sqrt{n}}\right]$$

When using the tables for t values, we use the size we want (e.g 0.975 for 95% double-ended confidence interval), and then use the degrees of freedom $(n-1)$.