8. Statistical Inference

8.1. Estimation (Estimointi)

A statistic is called an estimator (estimaattori), if it is used in order to assess the value of a population parameter (such as μ, σ^2, Π). The process of doing that is called estimation and the resulting value estimate (estimaatti). Parameters to be estimated are often denoted by θ and their estimators by t.

A point estimate (piste-estimaatti) is a single value used as a parameter estimate. An interval estimate (väliestimaatti) defines a confidence interval (luottamusväli), which is expected to contain the true population parameter with high probability (e.g. 95%, 99%).

Example. The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is an estimator of the population mean μ. Any concrete value of $\bar{X}$, such as $\bar{x} = 5$, is called an estimate.

8.1.1. Point Estimation (Piste-estimointi)

The idea of point estimation is to find the best possible estimate of the unknown population parameter θ. Good estimators have the following properties:

1. Unbiasedness (harhattomuus):
   An estimator t is unbiased, if $E(t) = \theta$. (t is asymptotically unbiased (asymptoottisesti harhaton), if $E(t) \to \theta$ for $n \to \infty$).

2. Consistency (tarkentuvuus):
   An estimator t is consistent if its probability of being close to θ increases as the sample size increases, that is, $V(t) \overset{n\to\infty}{\longrightarrow} 0$.

3. Efficiency (tehokkuus):
   An estimator t is efficient, if it is unbiased and has the smallest variance of all unbiased estimators.

4. Sufficiency (tyhjentävyys):
   An estimator t is sufficient, if it uses all information from the sample observations.

Example.

The most common estimators of the expected value μ, the population variance σ^2, the population proportion Π, and the population correlation ρ are

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{for } \mu,$$

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \qquad \text{for } \sigma^2,$$

$$\hat{P} = 100\frac{\#X_i \text{ of type A}}{n} \qquad \text{for } \Pi,$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{for } \rho,$$

because they are unbiased, consistent, and sufficient, and for normally distributed variables also efficient.

Similarly, the population median of unsymmetric populations is estimated by the sample median. For symmetric distributions the sample mean is a more efficient, but less robust estimator of the population median than the sample median.

8.1.2. Interval Estimation (Väliestimointi)

The random interval (t_n, T_n) is called confidence interval for the parameter θ at confidence level $1-\alpha$ $(0 < \alpha < 1)$, (luottamusväli luottamustasolla $1 - \alpha$), if

$$P(t_n \leq \theta \leq T_n) = 1 - \alpha,$$

that is, the interval $[t_n, T_n]$ covers the true population parameter θ with probability $1-\alpha$. The idea of interval estimation is to find statistics t_n and T_n such that the interval $[t_n, T_n]$ is as narrow as possible, with the true population parameter θ in its center.

Confidence Interval for μ

a) σ^2 known

Let us first assume that the population variance σ^2 is known and that $X_1, \ldots, X_n$ are randomly sampled from a normal distribution $N(\mu, \sigma^2)$, that is,

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \ \sim \ N\left(\mu, \frac{\sigma^2}{n}\right).$$

So in about 68% of all cases the sample mean $\bar{X}$ remains within a distance of $\sigma_{\bar{X}} := \sigma/\sqrt{n}$ from the population mean or expected value μ. σ_X is therefore called the standard error of the mean (keskiarvon keskivirhe).

Standardizing yields

$$Z := \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1).$$

Let us introduce $z_{\alpha/2}$ as the value such that $\Phi\left(z_{\alpha/2}\right) = 1 - \frac{\alpha}{2}$, that is, the probability that a standard normally distributed random variable exceeds $z_{\alpha/2}$ is $100\frac{\alpha}{2}\%$.

Now, by symmetry of the normal distribution:

$$
\begin{aligned}
&P\left(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}\right) = 1 - \alpha \quad \Leftrightarrow \\
&P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha \quad \Leftrightarrow \\
&P\left(-z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha \quad \Leftrightarrow \\
&P\left(-\left(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \leq -\mu \leq -\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)\right) = 1 - \alpha \\
&\Leftrightarrow \; P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.
\end{aligned}
$$

A $100(1-\alpha)\%$ confidence interval for μ, when σ^2 is known, is therefore:

$$CI = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right).$$

That is, the probability of finding the true population mean μ outside this interval is only $100\alpha\%$.

Example.

Let $X \sim N(\mu, 9)$. A random sample of size $n = 36$ yields $\bar{X} = 10$. We wish to find a 95% confidence interval for μ.

Now $\alpha = 0.05$ implying $1-\alpha/2 = 0.975$, such that $z_{0.05/2}$=NORMSINV(0.975)=1.96.

A 95% confidence interval for μ, using

$$CI = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right),$$

is therefore:

$$\left(10 - 1.96 \cdot \frac{\sqrt{9}}{\sqrt{36}},\ 10 + 1.96 \cdot \frac{\sqrt{9}}{\sqrt{36}}\right) \approx (9, 11).$$

The probability that the true population mean μ will be outside this range is only 5%.

Sample-Size Determination
(Otoskoon määräämisestä)

Consider a random sample $X_1, \ldots, X_n$ with $X_i \sim N(\mu, \sigma^2)$, such that $\bar{X} \sim N(\mu, \sigma^2/n)$. We know by virtue of the law of large numbers that by increasing the sample size n we can eventually get the sample mean $\bar{X}$ arbitrarily close to the population mean μ. We also know from our discussion of the $100(1-\alpha)\%$ confidence interval for μ that

$$P\left(|\bar{X}-\mu| \leq z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \;\Leftrightarrow\; P\left(|\bar{X}-\mu| > z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = \alpha,$$

so the margin of error (virhemarginaali) for $d = |\bar{X}-\mu|$ in $100(1-\alpha)\%$ of all samples is $d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

The minimum required sample size, such that the sample mean $\bar{X}$ deviates from its corresponding population parameter μ by more than a prespecified value d only with probability α, is therefore:

$$n = \frac{\sigma^2}{d^2} \cdot z^2_{\alpha/2}.$$

Example.
Let $X \sim N(\mu, 9)$ and assume we wish to find the minimum required sample size such that in 95% of all cases the sample mean deviates from the population mean by no more than 1 unit. Now $\alpha = 0.05$, such that $1 - \frac{\alpha}{2} = 0.975$ and $z_{0.05/2}$=NORMSINV(0.975)=1.96.

$P(|\bar{X} - \mu| > 1) \leq 0.05$ requires then:

$$n \geq \frac{9}{1^2} \cdot 1.96^2 = 34.57 \approx 35.$$

b) σ^2 unknown

If the population variance σ^2 is unknown, we need to estimate it with the sample variance

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

The standard error of the mean $\sigma_{\bar{X}} = \sigma/\sqrt{n}$ is now estimated by $s_{\bar{X}} = S/\sqrt{n}$.

Recall from our discussion of the Student t-distruibution, that

$$T_n = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

A $100(1-\alpha)\%$ confidence interval for μ may therefore be obtained by replacing σ^2 with S^2 and $z_{\alpha/2}$ with $t_{\alpha/2}(n-1)$=TINV$(\alpha; n-1)$,* that is,

$$\left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right).$$

*$t_{\alpha/2}(n-1)$ denotes the value of T for which $P(T > t_{\alpha/2}(n-1)) = \frac{\alpha}{2} \quad \Leftrightarrow \quad P(|T| > t_{\alpha/2}(n-1)) = \alpha$.

Example.

A shopkeeper claims that the eggs for sale have an average weight of 50g. A client selects 30 eggs randomly and gets $\bar{X} = 45$g and $S = 6$g. Do we believe what the shopkeepers claims?

Consider a 95% confidence interval for μ: $\alpha = 0.05$, $t_{0.05/2}(29)$=TINV(0.05; 29)=2.045, and a 95% confidence interval for μ is:

$$\left(45 - 2.045 \cdot \frac{6}{\sqrt{30}},\ 45 + 2.045 \cdot \frac{6}{\sqrt{30}}\right) \approx (42.76,\ 47.24).$$

Now, $50 \notin (42.76, 47.24)$, which leads us to doubt the shopkeepers claim.

Note. Recall that the Student-t distribution approaches the standard normal for $n \to \infty$, such that *for large sample sizes* we may replace $t_{\alpha/2}(n-1)$ with $z_{\alpha/2}$ in order to obtain as a $100(1-\alpha)\%$ confidence interval for μ:

$$\left(\bar{X} - z_{\alpha/2}\frac{S}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right).$$

Example: $45 \pm 1.96 \cdot \frac{6}{\sqrt{30}} = (42.85,\ 47.15)$.

Confidence Interval for Π

Recall from our discussion of the sampling distribution of the sample proportion Π, that

$$\hat{P} \overset{as.}{\sim} N\left(\Pi, \frac{\Pi(100-\Pi)}{n}\right), \text{ where}$$

$$\hat{P} = \frac{100}{n}\sum_{i=1}^{n} I_i \quad \text{with } I_i = \begin{cases} 1 & \text{for } X_i \text{ of type A} \\ 0 & \text{otherwise.} \end{cases}$$

Replacing Π with its estimator $\hat{P}$ yields for the standard error of the sample proportion for large sample sizes:

$$s_{\hat{P}} = \sqrt{\frac{\hat{P}(100-\hat{P})}{n}}.$$

A $100(1-\alpha)\%$ confidence interval for Π is therefore:

$$\left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(100-\hat{P})}{n}},\ \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(100-\hat{P})}{n}}\right).$$

The minimum required sample size, such that the sample proportion $\hat{P}$ differs from the population proportion Π beyond a given percentage d only with probability α is

$$n = \frac{\hat{P}(100-\hat{P})}{d^2} \cdot z^2_{\alpha/2}.$$

Example.

Consider a sample of 64 products from a production line, 26 of which are defect, that is,

$$\hat{P} = 100 \cdot \frac{26}{64}\% = 40.625\%.$$

Let's find a 95% confidence interval for Π. Recalling $z_{0.05/2}$=NORMSINV(0.975)=1.96, and using

$$CI = \left(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{P}(100-\hat{P})}{n}},\ \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{P}(100-\hat{P})}{n}} \right)$$

yields as a 95% confidence interval for Π:

$$\left(40.625 - 1.96\sqrt{\frac{40.625 \cdot 59,375}{64}},\ 40.625 + 1.96\sqrt{\frac{40.625 \cdot 59,375}{64}} \right)$$

$$\approx\ (28.6,\ 52.7).$$

That is, with 95% confidence we can state that the fraction of defect products is somewhere betweeen 28.6% and 52.7%. The risk that the real percentage is outside this range is only 5%.