22.4UC2

Numerical Computation and Statistics in Engineering

Unit 4: Estimation

Estimators

Two different type of inferences (here for example about the mean) can be made from a sample:

- one can estimate the true mean of the population; or
- one can try to decide whether the true mean exceeds same value or lies within some interval.

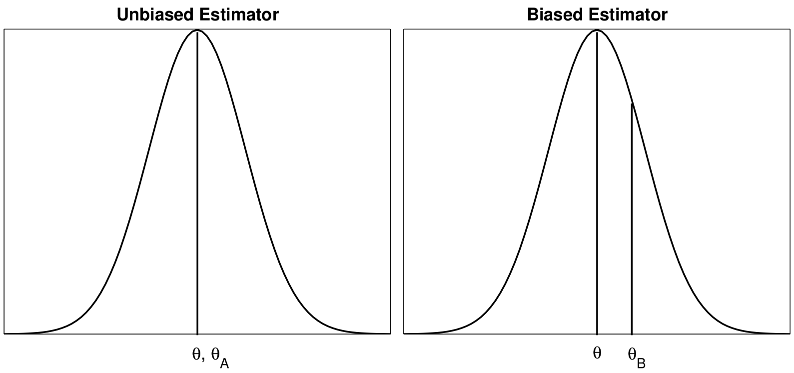
Suppose we want to estimate a population parameter θ (say mean, standard deviation, or $P(\text{‘}x \le d\text{’})$). A *point estimator* for theta is a rule that tells us how to compute from the sample data a single value $\hat{\theta}$ (also called a point estimator) that will serve as an estimator for θ.

An *interval estimator* is a rule computing an interval to estimate θ.

Example. If $x_1, \ldots, x_n$ is a random sample from a population then $\bar{x}$ is a point estimator for the true population mean, whereas $[\bar{x} - s, \bar{x} + s]$ is an interval estimator for the population mean. □

Bias

An estimator $\hat{\theta}$ is called *unbiased* if $E(\hat{\theta}) = \theta$. The *bias* of an estimator is $B = E(\hat{\theta}) - \theta$.



MVUE

In addition to unbiasedness we hope for a small standard deviation (or variance) of the probability distribution of $\hat{\theta}$. An *unbiased* estimator which has minimum variance among all unbiased estimators is called a *minimum variance unbiased estimator* (MVUE).

If such a MVUE does *not* exist one prefers the estimator which minimises the *mean squared error*

$$E[(\theta - \hat{\theta})^2].$$

Note that

$$\begin{aligned} E[(\theta - \hat{\theta})^2] &= E(\theta^2) - 2\theta E(\hat{\theta}) + E(\hat{\theta}^2) \\ &= \theta^2 - 2\theta E(\hat{\theta}) + \mathrm{var}_{\hat{\theta}} + E(\hat{\theta})^2 \\ &= B^2 + \mathrm{var}_{\hat{\theta}}. \end{aligned}$$

MVUE

In particular, if $B = 0$ then

- the mean squared error is equal to the variance of $\hat{\theta}$, and
- the estimator $\hat{\theta}$ that yields the smallest mean squared error is also a MVUE for θ.

Example. If x has binomial distribution with parameters n and p, then $\frac{x}{n}$ is an unbiased estimator for p.

Indeed, since $E(x) = np$ it follows that $E(\frac{x}{n}) = \frac{1}{n}E(x) = \frac{1}{n}np = p$. □

MVUE

Example. If s^2 is the standard deviation of a random sample from an *infinite* population then $E(s^2) = \sigma^2$, the true population standard deviation, hence s^2 is an unbiased estimator for σ (regardless of the nature of the sampled population). Here we use that $s^2 = \frac{1}{n-1}[\sum_i x_i^2 - \frac{1}{n}(\sum_i x_i)^2]$, and the fact that for any random variable, $E(y^2) = \sigma_y^2 + E(y)^2$.

$$\begin{aligned} E(s^2) &= \frac{1}{n-1}\left[\sum_i E(x_i^2) - \frac{1}{n}E[\sum_i x_i)^2]\right] \\ &= \frac{1}{n-1}\left[\sum_i (\sigma^2 + \mu^2) - \frac{1}{n}(\sigma^2_{\sum_i x_i} + E(\sum_i x_i)^2)\right] \\ &= \frac{1}{n-1}\left[n\sigma^2 + n\mu^2 - \frac{1}{n}\cdot n\sigma^2 - \frac{1}{n}(n\mu)^2\right] \\ &= \frac{1}{n-1}[n\sigma^2 - \sigma^2] = \sigma^2. \end{aligned}$$

□

An Example

Consider the following three density functions:

$$\begin{aligned}
\rho(x) &= \frac{1}{2\pi\sigma^2} e^{-(x-\mu)^2/(2\sigma^2)} && \text{for } -\infty < x < \infty, \\
\rho(x) &= \frac{1}{\pi(1+(x-\mu)^2)} && \text{for } -\infty < x < \infty, \\
\rho(x) &= \frac{1}{2c} && \text{for } -c \le x-\mu \le c, \text{ and } 0 \text{ else.}
\end{aligned}$$

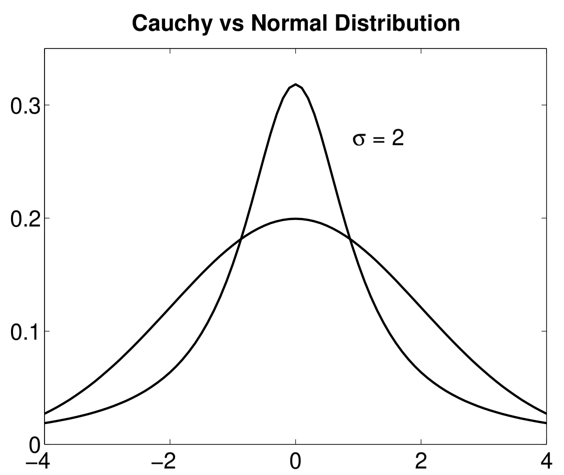
The first is the normal distribution, the second the Cauchy distribution, and the third the uniform distribution. All three have mean μ.

In theory we have at least three estimators for μ from a given sample, namely $\bar{x}$ (mean), $\tilde{x}$ (median), $\bar{x}_e$ (average between the two extreme observations).

An Example

- If the sample comes from a normal distribution, $\bar{x}$ is the best of the estimators as it is the MVUE.
- If the sample comes from a Cauchy distribution then $\bar{x}$ and $\bar{x}_e$ are bad estimators, whereas $\tilde{x}$ is quite good (the MVUE is not known). $\bar{x}$ is bad because it is sensitive to outliers, and the heavy tails of the Cauchy distribution will make such outliers very probable.
- If the distribution is uniform then $\bar{x}_e$ is the best estimator. x_e is sensitive to outliers, but the lack of tails makes such observations impossible.

An Example



Maximum Likelihood Estimators

Let $x_1, \dots, x_n$ be a random sample. The *likelihood* of the sample is defined as

- $L = P(x_1, \dots, x_n) = \prod_i P(x_i)$

  if the x_i are discrete random variables;
- $L = \rho(x_1, \dots, x_n) = \prod_i \rho(x_i)$

  if the x_i are continuous random variables. (Note that $\rho(x_1, \dots, x_n)$ is the density function of

$$F(x_1, \dots, x_n) = P(t_1 \le x_1, \dots, t_n \le x_n)\,.)$$

The *maximum likelihood estimator* for θ (or a list of parameters $\theta_1, \dots, \theta_k$) is the estimator $\hat\theta$ (or $\hat\theta_1, \dots, \hat\theta_n$) that *maximises* L.

In practice one often maximises the logarithm of $\rho(x_1, \dots, x_n)$, which is easier to calculate and gives the same estimator since the logarithm function is strictly increasing.

Maximum Likelihood Estimators

Example. Let $x_1, \dots, x_n$ be a random sample of n observations of a random variable x with exponential density function

$$\rho(x) = \begin{cases} \frac{e^{-\frac{x}{\beta}}}{\beta} & \text{if } 0 \le x < \infty, \\ 0 & \text{else.} \end{cases}$$

What is the maximum likelihood estimator $\hat\beta$ for β?

The joint density function is $L(\beta) = \frac{1}{\beta^n} e^{\sum_i -x_i/\beta}$, and

$$\ln L = -n \ln \beta + \sum_i -x_i/\beta$$

Setting $\frac{d \ln L}{d\beta}$ equal to 0 gives

$$\frac{\sum_i x_i}{\beta^2} - \frac{n}{\beta} = 0$$

or $\beta = \frac{1}{n}\sum_i x_i$. Thus $\hat\beta = \bar x$ is the maximum likelihood estimator for β. □

Maximum Likelihood Estimators

Example. What is the maximum likelihood estimator of the success probability θ of a random sample from a population with binomial probability distribution?

Here $L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x}$, and we maximise

$$\ln L(\theta) = \ln \binom{n}{x} + x \ln\theta + (n-x)\ln(1-\theta)\,.$$

Then

$$\frac{d \ln L}{d\theta} = 0 + \frac{x}{\theta} - \frac{n-x}{1-\theta}\,,$$

and thus $x - \theta x = \theta n - \theta x$, or $\theta = \frac{x}{n}$. □

Example. On 20 cold days a student gets his car started on the third, first, fifth, first, second, third, first, seventh, second, fourth, eighth, fourth, third, first, fifth, sixth, second, first, second, and sixth try. If the distribution of this random variable is modelled by a geometric probability distribution, what is the maximum likelihood estimator for θ?

Maximum Likelihood Estimators

The probability for success in the xth try is

$$\theta(1-\theta)^{x-1}$$

for $x = 1, 2, 3, \ldots$. Then $L(\theta) = \prod_i \theta(1-\theta)^{x_i - 1} = \theta^n (1-\theta)^{(\sum_i x_i) - n}$, and $\ln L(\theta) = n \ln\theta + (\sum_i x_i - n)\ln(1-\theta)$, thus

$$\frac{d\,L}{d\,\theta} = \frac{n}{\theta} - \frac{\sum_i x_i - n}{1-\theta}\,.$$

The necessary condition for a maximum is thus

$$n - n\theta = \theta \sum_i x_i - n\theta\,,$$

or $\theta = \frac{n}{\sum_i x_i} = \bar{x}^{-1}$.

For our data, $n = 10$ and $\bar{x} = 3.35$, so that an estimator is given by 0.299.

□

The Confidence Coefficient

We continue with interval estimators. The two numbers computed by an interval estimator are the endpoints of the *confidence interval*. The *confidence coefficient* for a confidence interval is the probability that the interval will contain the true (to be estimated) parameter.

As an example we consider the case when $\hat{\theta}$ is approximately normally distributed with mean $E(\hat{\theta}) = \theta$ and error (standard deviation) $\sigma_{\hat{\theta}}$. Then

$$z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}}$$

is approximately a standard random variable. We are looking for values z' such that

$$P(-z' \le z \le z') = 1 - \alpha\,,$$

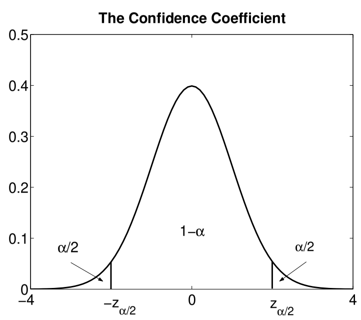
for $1-\alpha$ the confidence coefficient of the interval $[-z', z']$. From the graph

The Confidence Coefficient

we see that $z' = z_{\alpha/2}$, which is the unique z' such that $P(z \le z') = \alpha/2$. Substituting back the definition of z we find that for given confidence coefficient $1-\alpha$ the confidence interval for θ is

$$[\hat{\theta} - z_{\alpha/2}\sigma_\theta, \hat{\theta} + z_{\alpha/2}\sigma_\theta]\,,$$

where $z_{\alpha/2}$ is the unique z' such that $P(z \le z') = \alpha/2$ for the normally distributed random variable z with mean 0 and standard deviation 1.



Estimating the Mean

If the sampling size n is large ($n \geq 30$) then $\bar{x}$, the sampling mean, is approximately normally distributed with mean $E(\bar{x}) = \mu$, the true population mean, and $\sigma_{\bar{x}} = \sigma/\sqrt{n}$. Thus $\bar{x}$ is an unbiased estimator for μ, and $\bar{x}$ is also the MVUE for μ. Since the distribution of $\bar{x}$ is approximately normal we can use the previous analysis to get the $(1-\alpha)100\%$ confidence interval

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

for μ, where $z_{\alpha/2}$ is the z-value that locates from $-\infty$ to $z_{\alpha/2}$ an area $\alpha/2$ under the standard normal density function.

Estimating the Mean

If the population is smaller, or if the value of σ has to be approximated by the sample deviation (sample error) s, then the t-distribution with $n-1$ degrees of freedom replaces the normal distribution so that the $(1-\alpha)100\%$ confidence interval becomes

$$\bar{x} \pm t_{\alpha/2}\sigma_{\bar{x}} = \bar{x} \pm t_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

for μ, where $t_{\alpha/2}$ is the t-value that locates from $-\infty$ to $t_{\alpha/2}$ an area $\alpha/2$ under the density function of the t-distribution with $n-1$ degrees of freedom.

Estimating the Mean

Example. Time between server failures is recorded and for a sample of 20 failures the values $\bar{x} = 1500$ hours and $s = 210$ hours are computed. What is the 95% confidence interval for the mean based on this sample?

To apply the theory we have to assume a normal distribution for the time between server failures. Then $z = \frac{x-\bar{x}}{s/\sqrt{n}}$ has a t-distribution and for $\alpha = 0.05$ we find the interval

$$\bar{x} \pm t_{\alpha/2}\frac{s}{\sqrt{n}} = 1500 \pm 2.093\frac{210}{\sqrt{20}} = 1500 \pm 98.282\,.$$

$\square$

Estimating the Mean

Example. If a random sample of size 20 of a normal population with standard deviation 12.3 has mean 83.2, then we construct the following confidence intervals:

- 90%: $83.2 \pm z_{0.05}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.645\frac{12.3}{\sqrt{20}} = 83.2 \pm 4.524$
- 95%: $83.2 \pm z_{0.025}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.960\frac{12.3}{\sqrt{20}} = 83.2 \pm 5.390$
- 99%: $83.2 \pm z_{0.005}\frac{12.3}{\sqrt{20}} = 83.2 \pm 2.576\frac{12.3}{\sqrt{20}} = 83.2 \pm 7.084$

If the standard deviation of the population has to be estimated as well, and if 12.3 is an estimate based on the sample then the confidence intervals change as follows:

- 90%: $83.2 \pm t_{19,0.05}\frac{12.3}{\sqrt{20}} = 83.2 \pm 1.729\frac{12.3}{\sqrt{20}} = 83.2 \pm 4.755$
- 95%: $83.2 \pm t_{19,0.025}\frac{12.3}{\sqrt{20}} = 83.2 \pm 12.093\frac{12.3}{\sqrt{20}} = 83.2 \pm 5.756$
- 99%: $83.2 \pm t_{19,0.005}\frac{12.3}{\sqrt{20}} = 83.2 \pm 2.861\frac{12.3}{\sqrt{20}} = 83.2 \pm 7.868$

□

Estimating the Mean

Example. Readings from a machine show the following values:

11.3968	4.1666	0.8273	18.4765	7.8963
6.3828	6.4634	2.0181	5.2051	6.4615
16.7264	1.1679	3.2379	0.2825	3.0543
3.4829	0.6679	2.5763	6.3852	1.5892

What is the 95% confidence interval for the mean?

Here $\bar{x} = 5.4232$, and $s = 5.0388$, thus the 95% confidence interval is

$$\bar{x} \pm t_{19,0.025}\frac{s}{\sqrt{20}} = 5.4232 \pm 2.358\,.$$

(The numbers are random numbers for an exponential distribution with mean 4.) In theory our results do not apply since the sampling size is too small and the distribution is not bell-shaped! □

Estimating the Mean

Example. A company selling easy-to-assemble furniture wants to determine how long it takes to assemble a chest of drawers Bialitt. Using the following data (in minutes) gathered from 25 volunteers we construct a 95% confidence interval.

84.3487	59.6883	95.5066	98.7535	70.0706
116.8183	116.7833	92.2473	99.5458	96.4928
89.2658	107.5158	81.2337	136.6637	90.2721

Here $\bar{x} = 95.6804$ and $s = 19.1003$. Thus the interval is

$$\bar{x} \pm t_{14,0.025}\frac{s}{\sqrt{15}} = 95.6804 \pm 10.5784\,.$$

(The data are random numbers from a normal distribution with mean 93 and standard deviation 20.) □

Estimating the Difference Between Means

If $\bar{x}_1$ and $\bar{x}_2$ are the values of the means and standard deviation of independent random samples of size n_1 and n_2 from normal populations with known standard deviations σ_1 and σ_2 respectively, then

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2 + \sigma_2^2}{n_1 + n_2}}$$

is a $(1-\alpha)100\%$ confidence interval for the difference between the means $\mu_1 - \mu_2$. (By the central limit theorem this confidence interval can also be used for independent random samples from non-normal populations if $n_1, n_2 \geq 30$, or for even smaller samples when the density functions of the populations are known to be bell-shaped.)

Estimating the Difference Between Means

If σ_1 and σ_2 are to be estimated by the samples standard deviations then the $(1-\alpha)100\%$ confidence interval for the difference between the means is

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2, n_1+n_2-2} s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

where

$$s_P = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}.$$

(s_P is called the pooled estimator for σ^2, and is unbiased.) Here the only assumption is that the two populations are normal.

Estimating the Difference Between Means

Example. Two machines make wires. 10 measurements taken in 1 minute intervals from both machines show the following diameters:

I:	1.0429	1.0627	0.9203	0.9280	1.0286
	0.9800	1.0345	1.0408	1.0356	1.0645
II:	1.0001	1.0157	0.9439	0.9794	0.9753
	0.9319	0.9877	0.9483	1.0225	0.9558

What is a 95% confidence interval for $\mu_1 - \mu_2$?

Estimating the Difference Between Means

Here $n_1 = n_2 = 10$, $\bar{x}_1 = 1.0138$, $s_1 = 0.0526$, $\bar{x}_2 = 0.9761$, and $s_2 = 0.0309$. Thus $x_1 - x_2 = 0.0377$ and $s_P^2 = \frac{9}{18}(s_1^2 + s_2^2) = 0.00186$.

The interval is thus

$$0.0377 \pm t_{18,0.025} s_P \sqrt{\frac{2}{10}} = 0.0377 \pm 2.101 \cdot 0.0136 = 0.0377 \pm 0.0287\,.$$

(The data are random numbers for normal distributions with mean and standard deviations 1.0, 0.05, and 0.98, 0.03 respectively.)

The analysis shows slightly more: Since we are 95% confident that the difference between the means is within the interval

$$[0.0090, 0.0664]$$

(which does *not* contain 0), we are also 95% confident that the means of the diameters of wires produced by the two machines differ. □

Estimating Proportions

We often try to estimate proportions, probabilities, or percentages such as faulty transistors, faulty lights, etc. If the sample size is large the corresponding random variable has approximately a binomial distribution (even though we often sample without replacement), and can be approximated by a normal distribution.

Thus, if θ denotes the true probability and $\hat{\theta} = \frac{x}{n}$ its estimate derived from a sample of size n then we can assert with $(1-\alpha)100\%$ confidence that the error we make is less than

$$z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\,.$$

Indeed, $z = \frac{x-\mu}{\sigma} = \frac{x-n\theta}{\sqrt{n\theta(1-\theta)}}$ is approximately a standard normal random variable and thus

$$P(-z_{\alpha/2} \leq \frac{x-n\theta}{\sqrt{n\theta(1-\theta)}} \leq z_{\alpha/2}) \leq 1-\alpha\,.$$

Estimating Proportions

Approximating θ by $\hat{\theta}$ under the radical and solving for θ gives

$$\begin{aligned} & P\left(-z_{\alpha/2}\sqrt{n\hat{\theta}(1-\hat{\theta})} \leq n\theta - x \leq z_{\alpha/2}\sqrt{n\hat{\theta}(1-\hat{\theta})}\right) \\ = \; & P\left(\frac{x}{n} - z_{\alpha/2}\frac{\sqrt{n\hat{\theta}(1-\hat{\theta})}}{n} \leq \theta \leq \frac{x}{n} + z_{\alpha/2}\frac{\sqrt{n\hat{\theta}(1-\hat{\theta})}}{n}\right) \\ = \; & P\left(\hat{\theta} - z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{n}}\right)\,. \end{aligned}$$

Estimating Proportions

Example. A study is made to determine the proportion of people aged between 16 and 25 that use use the internet. If 316 out of 400 young people use the internet, what is a 95% confidence interval for $p = \frac{316}{400} = 0.79$?

Here $n = 400$, $\hat{\theta} = 0.79$, and $z_{0.025} = 1.960$. With 95% confidence the maximum error we make is

$$1.960\sqrt{\frac{0.79 \cdot 0.21}{400}} \approx 0.0399\,,$$

i.e., the interval is 0.79 ± 0.0399. □

Estimating Differences in Proportions

We often estimate differences between proportions (differences between males and females in favour of a certain candidate, difference between the percentage of faulty transistors manufactured by two machines, etc.).

If we have two samples x_1 and x_2 of size n_1 and n_2 respectively, then $\hat{\theta}_1 - \hat{\theta}_2 = \frac{x_1}{n_1} - \frac{x_2}{n_2}$ is an estimator for the difference between the two proportions. If both n_1 and n_2 are large then $\hat{\theta}_1 - \hat{\theta}_2$ is approximately normally distributed with mean $\theta_1 - \theta_2$ the difference between the true proportions, and variance

$$\frac{\theta_1(1-\theta_1)}{n_1} + \frac{\theta_2(1-\theta_2)}{n_2}\,.$$

Putting everything together and estimating θ_1 and θ_2 by $\hat{\theta}_1$ and $\hat{\theta}_2$ we find the $(1-\alpha)100\%$ confidence interval

$$(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}\,.$$

Estimating Differences in Proportions

Example. Voters are questionned after they went to the ballots. Out of 212 male voters 76 voted for candidate A, and out of 179 female voters 57 voted for the same candidate. What is a 99% confidence interval for the difference between the percentages of voters voting for candidate A?

Here $n_1 = 212$, $\hat{\theta}_1 = \frac{76}{212} = 0.3585$ and $n_2 = 179$, $\hat{\theta}_2 = \frac{57}{179} = 0.3184$. The endpoints of the invterval are thus

$$\begin{aligned}
&(\hat{\theta}_1 - \hat{\theta}_2) \pm z_{\alpha/2}\sqrt{\tfrac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \tfrac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} \\
&= (0.3585 - 0.3184) \pm 2.575\sqrt{\frac{0.3585 \cdot 0.6415}{212} + \frac{0.3184 \cdot 0.6816}{179}} \\
&= 0.0401 \pm \sqrt{0.0011 + 0.0012} \\
&= 0.0401 \pm 0.0479\,.
\end{aligned}$$

The interval is thus $[-0.0078, 0.0880]$, which includes 0. This means that the case that there is *no* difference in voting habits among males and females is included. □