STATISTICS

Module 2 *of*
MATHEMATICS AND STATISTICS

Marina Santacroce

marina.santacroce@unibocconi.it
Department of Decision Sciences, Bocconi University

L16: Interval estimation.

References: S.Ross *Introduction to* Probability and Statistics for Engineers and Scientists: 7.3, 7.4, 7.5

Outline

Estimation of a population parameters

One of the main goals of inferential statistics is the estimation of the population parameters by observing the data of a random sample.

- $X_1, \dots X_n$, are sampled from a population with distribution F_θ
- F_θ is specified up to a vector of *unknown parameters* θ

$\Rightarrow$ use the sample to make inference about the unknown parameters.

- Point estimation method (last lecture)

$\Rightarrow$ Interval estimation method (today)

Confidence interval for the mean of a normal population (with known variance)

$X_1, \ldots X_n$ is a random sample from a normal population with unknown mean μ and known variance σ^2.

- The goal is an interval estimation for the unknown parameter μ based on the sample $X_1, \ldots X_n$.
- Chosen a level of confidence $1-\alpha$, $\alpha \in (0,1)$, we determine the related confidence interval by using the distribution of the statistics $\overline{X}$ as follows:
- $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \qquad \Longrightarrow \qquad \mathbb{P}(|\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}| \leq z_{\frac{\alpha}{2}}) = 1-\alpha$

  with $z_{\frac{\alpha}{2}}$ such that $\mathbb{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$, for a standard normal Z.

$$\Rightarrow \mathbb{P}(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \overline{X} - \mu \leq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$$\Rightarrow \mathbb{P}(\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

$100(1-\alpha)$ per cent of the intervals $[\overline{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$ contain the true value μ of the parameter

if $\overline{x}$ is the observed value $\Rightarrow$ the interval estimate $[\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$.

Example: computation of an interval estimate

Example: Consider a signal whose value is an unknown constant μ which is transmitted 9 times from a location A to a location B. We assume the signal received at B is $X = \mu + Y$ where the noise Y is normal with mean 0 and variance 4.
We observe the 9 values of the normal random sample $X_1, \dots X_9$ received at B:

$$5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5$$

and we want to find the *95 percent confidence (two sided) interval estimate for μ*.

- Since $\overline{x} = 9$ and $z_{0.025} = 1.96$, i.e. $\mathbb{P}(Z > 1.96) = 0.025$,

$$[9 - 1.96\frac{2}{3}, 9 + 1.96\frac{2}{3}] = [7.69, 10.31]$$

  is the *95 percent confidence (two sided) interval estimate for μ* computed on the observations.

- $\overline{X} \sim \mathrm{N}(\mu, \frac{4}{9}) \quad \Longrightarrow \quad \mathbb{P}(\frac{|\overline{X}-\mu|}{\frac{2}{3}} \leq 1.96) = 0.95$

  Recall

$$\mathbb{P}(\overline{X} - 1.96\frac{2}{3} \leq \mu \leq \overline{X} + 1.96\frac{2}{3}) = 0.95$$

$\Rightarrow$ 95 per cent of the intervals $[\overline{X} - 1.96\frac{2}{3}, \overline{X} + 1.96\frac{2}{3}]$ contain the true value μ of the parameter.

One-sided confidence intervals for the mean of a normal population (with known variance)

- We find a one-sided (upper) interval estimate for the unknown parameter μ.

  Chosen $1-\alpha$, the related confidence one-sided *upper* interval by using the distribution of the statistics $\overline{X}$ is determined as follows:

- $\overline{X} \sim \mathrm{N}(\mu, \frac{\sigma^2}{n}) \quad \Longrightarrow \quad \mathbb{P}(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_\alpha) = 1-\alpha$

  $\mathbb{P}(\overline{X}-\mu \leq z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-\alpha \quad \Rightarrow \quad \mathbb{P}(\mu \geq \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-\alpha$

$\Rightarrow$ $100(1-\alpha)$ per cent of the intervals $[\overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$ contain the value μ

- if $\overline{x}$ is the observed value $\Rightarrow$ the one-sided upper interval estimate

$$[\overline{x} - z_\alpha \frac{\sigma}{\sqrt{n}}, \infty)$$

- Analogously, using $\mathbb{P}(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \geq -z_\alpha) = 1-\alpha$, one can find the one-sided *lower* interval estimate for the unknown parameter μ:

$$(-\infty, \overline{x} + z_\alpha \frac{\sigma}{\sqrt{n}}]$$

Confidence interval for the mean of a normal population (with unknown variance)

Let $X_1, \ldots X_n$ be a random sample from a normal population with *unknown* mean μ and *unknown* variance σ^2.

- The goal is an interval estimation for μ based on the sample $X_1, \ldots X_n$.
- We determine the $100(1-\alpha)$ percent confidence interval by using that $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $n-1$ degrees of freedom,

$$\mathbb{P}\left(\frac{|\overline{X}-\mu|}{\frac{S}{\sqrt{n}}} \leq t_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

  with $t_{\frac{\alpha}{2},n-1}$ such that $\mathbb{P}(T_{n-1} > t_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$.

$$\Rightarrow \quad \mathbb{P}(\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}) = 1-\alpha$$

- $100(1-\alpha)\%$ of the intervals $[\overline{X} - t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}, \overline{X} + t_{\frac{\alpha}{2},n-1}\frac{S}{\sqrt{n}}]$ contain μ
- if $\overline{x} = \frac{1}{n}\sum_{i:1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i:1}^{n}(x_i - \overline{x})^2$ are the observed sample mean and sample variance $\quad\Rightarrow\quad$ the interval estimate

$$[\overline{x} - t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2},n-1}\frac{s}{\sqrt{n}}]$$

Example: computation of an interval estimate

Example: Consider the signal with unknown value μ transmitted 9 times from A to B. Recall the signal received at B is $X = \mu + Y$ where the noise Y is normal with mean 0 and unknown variance.
We received at B the 9 values:

$$5, 8.5, 12, 15, 7, 9, 7.5, 6.5, 10.5$$

and we want to find the *95 percent confidence (two sided) interval estimate for μ*.

- Since $\overline{x} = 9$, $s^2 = \frac{\sum_{i:1}^{9} x_i^2 - 9\overline{x}^2}{8} = 9.5$, $s = 3.082$ and $t_{0.025,8} = 2.306$,

  $$[9 - 2.306\frac{3.082}{3}, 9 + 2.306\frac{3.082}{3}] = [6.63, 11.37]$$

  is the *95 percent confidence (two sided) interval estimate for μ* computed on the observations.

- By comparing it to the one we computed when the variance was known, we observe that

  $$[6.63, 11.37] \quad \text{is sensibly larger than} \quad [7.69, 10.31]$$

  - the observed sample variance is greater than the value assumed known for the variance ($s = 3.082 > \sigma = 2$)
  - the t-distribution has a bigger variation compared to the standard normal.

One-sided confidence intervals for the mean of a normal population (with unknown variance)

- The one-sided upper interval estimate for μ with confidence level $1-\alpha$, can be found from

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \leq t_{\alpha,n-1}\right) = 1-\alpha \quad \Rightarrow \quad \mathbb{P}\left(\mu \geq \overline{X} - t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

- If $\overline{x} = \frac{1}{n}\sum_{i:1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i:1}^{n}(x_i - \overline{x})^2$ are the observed sample mean and sample variance $\Rightarrow$ the one-sided upper interval estimate

$$\left[\overline{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}, \infty\right)$$

- Similarly, using

$$\mathbb{P}\left(\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \geq -t_{\alpha,n-1}\right) = 1-\alpha \quad \Rightarrow \quad \mathbb{P}\left(\mu \leq \overline{X} + t_{\alpha,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha,$$

it can be found that the one-sided lower interval estimate for μ with confidence level $1-\alpha$ is:

$$\left(-\infty, \overline{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}\right]$$

Interval estimation for the mean of a population with known and unknown variance

If the population is not normal and if the sample is reasonably large

by the CLT

$$\sqrt{n}(\overline{X} - \mu)/\sigma \qquad \text{and} \qquad \sqrt{n}(\overline{X} - \mu)/S$$

will be approximately distributed as a standard normal and t with $n - 1$ degrees of freedom.

Therefore, the interval estimates obtained in the case of known variance

- two-sided $[\overline{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \overline{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$
- one-sided upper $[\overline{x} - z_{\alpha} \frac{\sigma}{\sqrt{n}}, \infty)$ and lower $(-\infty, \overline{x} + z_{\alpha} \frac{\sigma}{\sqrt{n}}]$

and the interval estimates obtained in the case of unknown variance

- two-sided $[\overline{x} - t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}, \overline{x} + t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}]$
- one-sided upper$[\overline{x} - t_{\alpha,n-1} \frac{s}{\sqrt{n}}, \infty)$ and lower $(-\infty, \overline{x} + t_{\alpha,n-1} \frac{s}{\sqrt{n}}]$

will still be approximate $100(1 - \alpha)$ percent confidence interval estimates for the unknown mean μ of the population.

Confidence intervals for the variance of a normal population

Let $X_1, \dots X_n$ be a random sample from a normal population with *unknown* mean μ and *unknown* variance σ^2.

- The goal is an interval estimation for σ^2 based on the sample $X_1, \dots X_n$.
- We determine the $100(1-\alpha)$ percent confidence interval by using that $(n-1)\frac{S^2}{\sigma^2}$ has a χ^2-distribution with $n-1$ degrees of freedom,

$$\mathbb{P}\left(\chi^2_{1-\frac{\alpha}{2},n-1} \le (n-1)\frac{S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right) = 1-\alpha$$

with $\chi^2_{\frac{\alpha}{2},n-1}$ such that $\mathbb{P}(\chi^2_{n-1} > \chi^2_{\frac{\alpha}{2},n-1}) = \frac{\alpha}{2}$.

$$\Rightarrow \quad \mathbb{P}\left((n-1)\frac{S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le (n-1)\frac{S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right) = 1-\alpha$$

- $100(1-\alpha)\%$ of the intervals $[(n-1)\frac{S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, (n-1)\frac{S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}]$ contain σ^2
- if $\overline{x} = \frac{1}{n}\sum_{i:1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i:1}^{n}(x_i - \overline{x})^2$ are the observed sample mean and sample variance $\Rightarrow$ the interval estimate for the variance σ^2 is $[(n-1)\frac{s^2}{\chi^2_{\frac{\alpha}{2},n-1}}, (n-1)\frac{s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}]$

Confidence intervals for the variance of a normal population

Let $X_1, \dots X_n$ be a random sample from a normal population with *known* mean μ and *unknown* variance σ^2.

- The goal is an interval estimation for σ^2 based on the sample $X_1, \dots X_n$.
- We determine the $100(1-\alpha)$ percent confidence interval by using that $\sum_{i:1}^{n} \frac{(X_i-\mu)^2}{\sigma^2}$ has a χ^2-distribution with n degrees of freedom,

$$\mathbb{P}\left(\chi^2_{1-\frac{\alpha}{2},n} \leq \sum_{i:1}^{n} \frac{(X_i-\mu)^2}{\sigma^2} \leq \chi^2_{\frac{\alpha}{2},n}\right) = 1-\alpha$$

Similarly to the previous case we will find

$$\left[\sum_{i:1}^{n}(X_i-\mu)^2/\chi^2_{\frac{\alpha}{2},n}, \sum_{i:1}^{n}(X_i-\mu)^2/\chi^2_{1-\frac{\alpha}{2},n}\right]$$

From the observed $\sum_{i:1}^{n}(x_i-\mu)^2$, the interval estimates for σ^2

$$\left[\sum_{i:1}^{n}(x_i-\mu)^2/\chi^2_{\frac{\alpha}{2},n}, \sum_{i:1}^{n}(x_i-\mu)^2/\chi^2_{1-\frac{\alpha}{2},n}\right]$$

Confidence intervals for the difference in means of two normal populations

Let $X_1, \dots X_n$ and $Y_1, \dots Y_m$ be two random samples from two normal populations with *unknown* means μ_1 and μ_2 and *known* variances σ_1^2 and σ_2^2.

- The goal is an interval estimation for $\mu_1 - \mu_2$ based on the samples $X_1, \dots X_n$ and $Y_1, \dots Y_m$.
- We determine the $100(1-\alpha)$ percent confidence interval by using that $\overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \sigma_1^2/n + \sigma_2^2/m)$.

$$\mathbb{P}\left(-z_{\frac{\alpha}{2}} \leq \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}} \leq z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

  and thus the two-sided interval for $\mu_1 - \mu_2$ will be

$$\left[\overline{X} - \overline{Y} - z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n + \sigma_2^2/m},\ \overline{X} - \overline{Y} + z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n + \sigma_2^2/m}\right]$$

- if $\overline{x} = \frac{1}{n}\sum_{i:1}^{n} x_i$ and $\overline{y} = \frac{1}{m}\sum_{i:1}^{m} y_i$ are the observed sample means $\Rightarrow$ the interval estimate

$$\left[\overline{x} - \overline{y} - z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n + \sigma_2^2/m}, \overline{x} - \overline{y} + z_{\frac{\alpha}{2}}\sqrt{\sigma_1^2/n + \sigma_2^2/m}\right]$$

Confidence intervals for the difference in means of two normal populations

Let $X_1, \ldots X_n$ and $Y_1, \ldots Y_m$ be two random samples from two normal populations with *unknown* means μ_1 and μ_2 and *unknown* (common) variance σ^2.
We find an interval estimation for $\mu_1 - \mu_2$ based on the samples $X_1, \ldots X_n$ and $Y_1, \ldots Y_m$.

Let us denote $S_p^2 = \frac{(n-1)S_1^2+(m-1)S_2^2}{n+m-2}$, and recall

$$\overline{X} - \overline{Y} \sim \mathrm{N}(\mu_1 - \mu_2, \sigma^2(1/n + 1/m)) \quad \text{and} \quad (n+m-2)\frac{S_p^2}{\sigma^2} \sim \chi^2_{n+m-2}$$

$$\Rightarrow \qquad \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2(\frac{1}{n}+\frac{1}{m})}} \sim t_{n+m-2}.$$

$$\mathbb{P}\Big(-t_{\frac{\alpha}{2},n+m-2} \le \frac{\overline{X}-\overline{Y}-(\mu_1-\mu_2)}{\sqrt{S_p^2(1/n+1/m)}} \le t_{\frac{\alpha}{2},n+m-2}\Big) = 1-\alpha$$

and the two-sided interval for $\mu_1 - \mu_2$ is

$$\Big[\overline{X} - \overline{Y} - t_{\frac{\alpha}{2},n+m-2} S_p \sqrt{1/n + 1/m},\ \overline{X} - \overline{Y} + t_{\frac{\alpha}{2},n+m-2} S_p \sqrt{1/n + 1/m}\Big].$$

- $\overline{x}$, $\overline{y}$ and $s_p^2 = \frac{(n-1)s_1^2+(m-1)s_2^2}{n+m-1}$ are the observed statistics, the estimate is

$$\Big[\overline{x} - \overline{y} - t_{\frac{\alpha}{2},n+m-2} s_p \sqrt{1/n + 1/m}, \overline{x} - \overline{y} + t_{\frac{\alpha}{2},n+m-2} s_p \sqrt{1/n + 1/m}\Big]$$

Approximate interval estimation for the mean in a Bernoulli population

Suppose in a population *each individual independently has a certain characteristic*, with some unknown probability p. We observe a random sample $X_1, \ldots, X_n$ where $X_i = 1$ if the individual i in the sample has the characteristic and 0 otherwise.

$X = \sum_{i:1}^{n} X_i \Rightarrow$ *number of individuals in the sample with the characteristic*;

$\hat{p} = \overline{X} \Rightarrow$ *proportion of individuals in the sample with the characteristic.*

We find an *approximate* confidence interval for the unknown parameter p based on the observations $X_1, \ldots, X_n$.

- X is a binomial with parameters n and p;
- if n is large, by the CLT X is approximately normal $\frac{X-np}{\sqrt{np(1-p)}} \approx\sim \mathrm{N}(0,1)$

$$\mathbb{P}\left(\frac{|X - np|}{\sqrt{np(1-p)}} \le z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

- to obtain a confidence interval we approximate $\sqrt{np(1-p)}$ with $\sqrt{n\hat{p}(1-\hat{p})}$

$$\mathbb{P}\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n} \le p \le \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p}/n)}\right) \approx 1 - \alpha$$

which yields an approximate interval for p

$$[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}\,,\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})/n}]$$

Approximate interval estimation for the mean in a Bernoulli population

An interval estimate of

$$[\hat{p} - z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n}\,,\ \hat{p} + z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})/n}]$$

will be obtained by substituting to $\hat{p}$ the value observed in the sample.

Ex.: If $n = 100$ transistors are sampled and 80 meets the standards, the approximate 95% confidence interval estimate for p is

$$[0.8 - 1.96\sqrt{0.8(0.2)/100}\,,\ 0.8 + 1.96\sqrt{0.8(0.2)/100}] = [0.7216\,,\ 0.8784]$$

Similarly to the preceding cases, we can also find one-sided *approximate* intervals:

- Approximate *upper* confidence interval: $[\hat{p} - z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n}\,,\ 1]$
- Approximate *lower* confidence interval: $[0,\ \hat{p} + z_{\alpha}\sqrt{\hat{p}(1-\hat{p})/n}]$

Interval estimation for the mean in an exponential population

Suppose $X_1, \ldots, X_n$ is a random sample from an exponential population with unknown mean θ.

We find a $100(1-\alpha)$ percent confidence interval for the unknown parameter θ based on the observations $X_1, \ldots, X_n$.

$$\sum_{i:1}^{n} X_i \sim \text{gamma}(n, \tfrac{1}{\theta}) \quad \Rightarrow \quad \tfrac{2}{\theta}\sum_{i:1}^{n} X_i \sim \chi^2_{2n} \quad \text{(show)}$$

$$\mathbb{P}\left(\chi^2_{1-\frac{\alpha}{2},2n} \leq \frac{2}{\theta}\sum_{i:1}^{n} X_i \leq \chi^2_{\frac{\alpha}{2},2n}\right) = 1-\alpha$$

$$\mathbb{P}\left(2\sum_{i:1}^{n} X_i/\chi^2_{\frac{\alpha}{2},2n} \leq \theta \leq 2\sum_{i:1}^{n} X_i/\chi^2_{1-\frac{\alpha}{2},2n}\right) = 1-\alpha$$

which yields the $100(1-\alpha)$ percent confidence interval for θ

$$[2\sum_{i:1}^{n} X_i/\chi^2_{\frac{\alpha}{2},2n}\,,\; 2\sum_{i:1}^{n} X_i/\chi^2_{1-\frac{\alpha}{2},2n}]$$

One-sided $100(1-\alpha)$ percent confidence intervals will be:

- *Upper* confidence interval: $[2\sum_{i:1}^{n} X_i/\chi^2_{\alpha,2n}\,,\; \infty]$
- *Lower* confidence interval: $[0\,,\; 2\sum_{i:1}^{n} X_i/\chi^2_{1-\alpha,2n}]$