Lecture 9

Parameter Estimation, Least Squares Estimation, Maximum Likelihood Estimation, Interval Estimation

Parameter Estimation

Estimation theory is a branch of statistics that deals with estimating the values of parameters based on measured empirical data that has a random component.

The parameters describe an underlying physical setting in such a way that their value affects the distribution of the measured data. An estimator attempts to approximate the unknown parameters using the measurements.

Parameter Estimation

We will consider two different approaches to the Parameter Estimation:

- *point estimation*, which involves the use of sample data to calculate a single value (known as a point estimate) to serve as a "best guess" or "best estimate" of an unknown population parameter.
- *interval estimation*, which uses the sample data to calculate an interval of possible values of an unknown population parameter.

Parameter Estimation

For the Point Estimation, we will consider two methods generally applied in research:

- *Least Squares Estimation (LSE)*.
- *Maximum Likelihood Estimation (MLE)*.

Least Squares Estimation

The most important application of LSE is in data fitting.

The best fit in the least-squares sense minimizes the *sum of squared residuals* (a residual being: *the difference between an observed value, and the fitted value provided by a model*).

Least Squares Estimation

Least-squares problems fall into two categories:

- *linear* or *ordinary least squares*;
- *nonlinear* least squares,

depending on whether or not the residuals are linear in all unknowns.

Least Squares Estimation

The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution.

The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.

Least Squares Estimation

The objective of the LSE consists of adjusting the parameters of a model function to best fit a data set.

A simple data set consists of n points $(x_i, y_i),\ i = 1, \ldots, n,$ where x_i is an independent variable and y_i is a dependent variable whose value is found by observation.

Least Squares Estimation

The model function has the form $f(x, \boldsymbol{\theta})$, where $m < n$ adjustable parameters are held in the vector $\boldsymbol{\theta}$.

The goal of the LSE is to find the parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ for the model that "best" fits the data.

Least Squares Estimation

The fit of a model to a data point is measured by its residual, defined as the difference between the actual value of the dependent variable and the value predicted by the model:

$$r_i = y_i - f(x_i, \boldsymbol{\theta}).$$

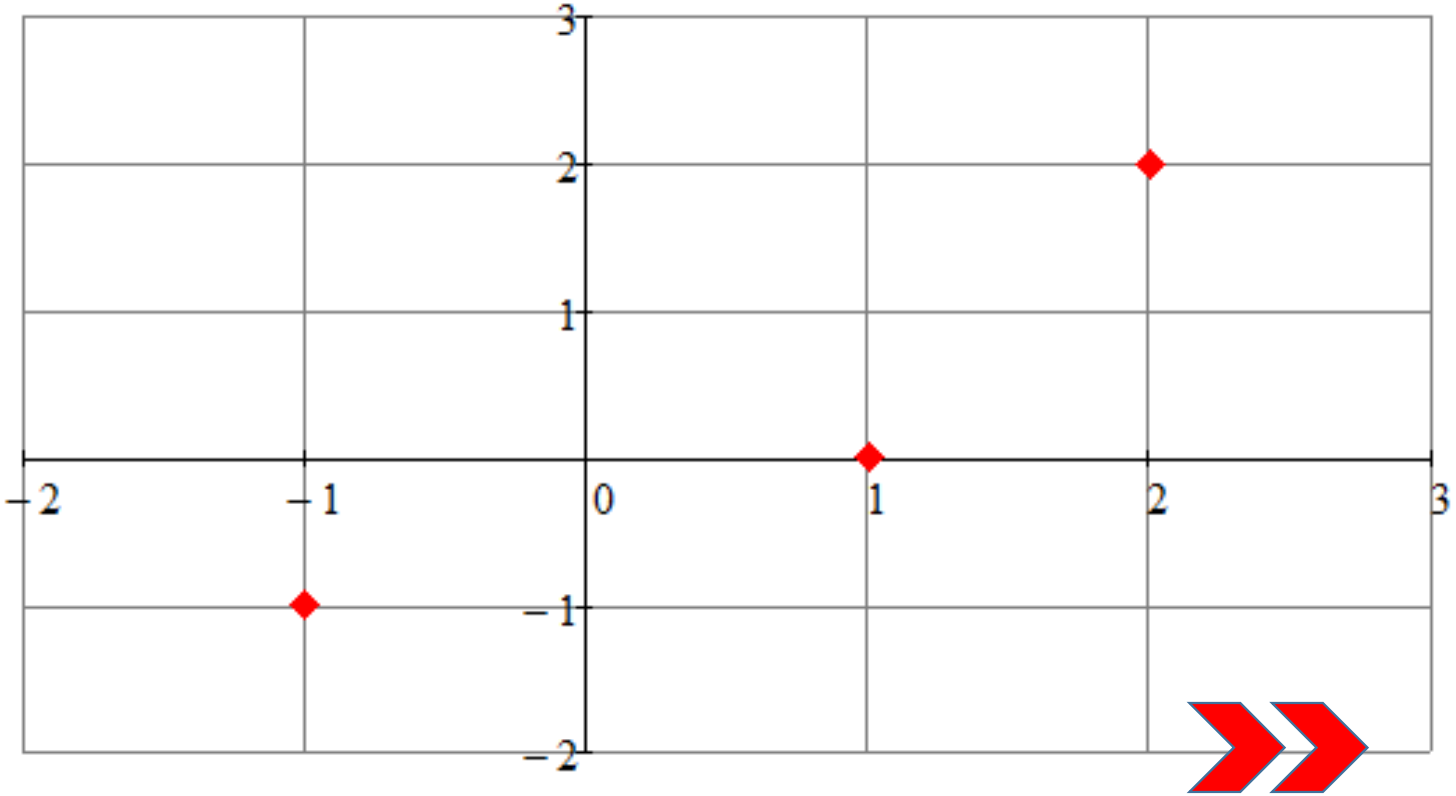
The LSE method finds the optimal parameter values by minimizing the function of the sum of squared residuals:

$$SSR(\boldsymbol{\theta}) = \sum_{i=1}^{n} r_i \,.$$

Least Squares Estimation

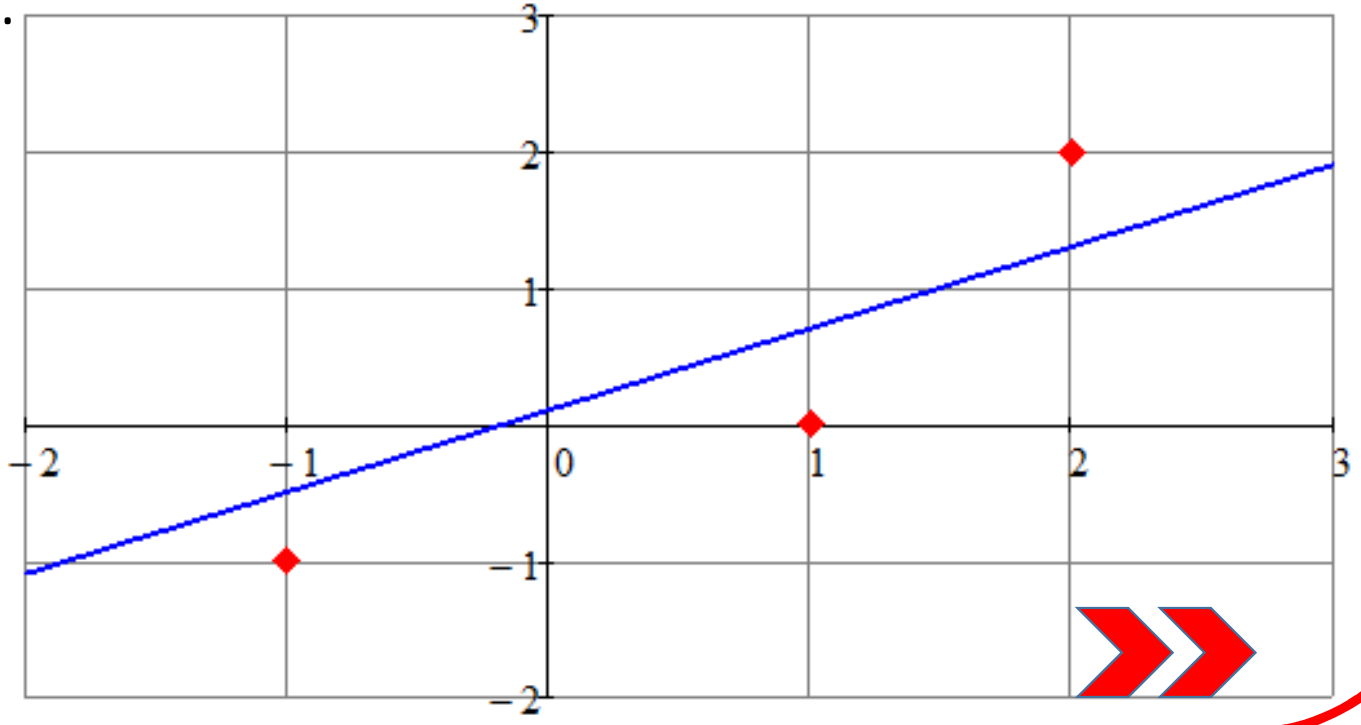
Consider an experiment resulting in three datapoints

$$X := \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix} \qquad Y := \begin{pmatrix} -1 \\ 0 \\ 2 \end{pmatrix}$$



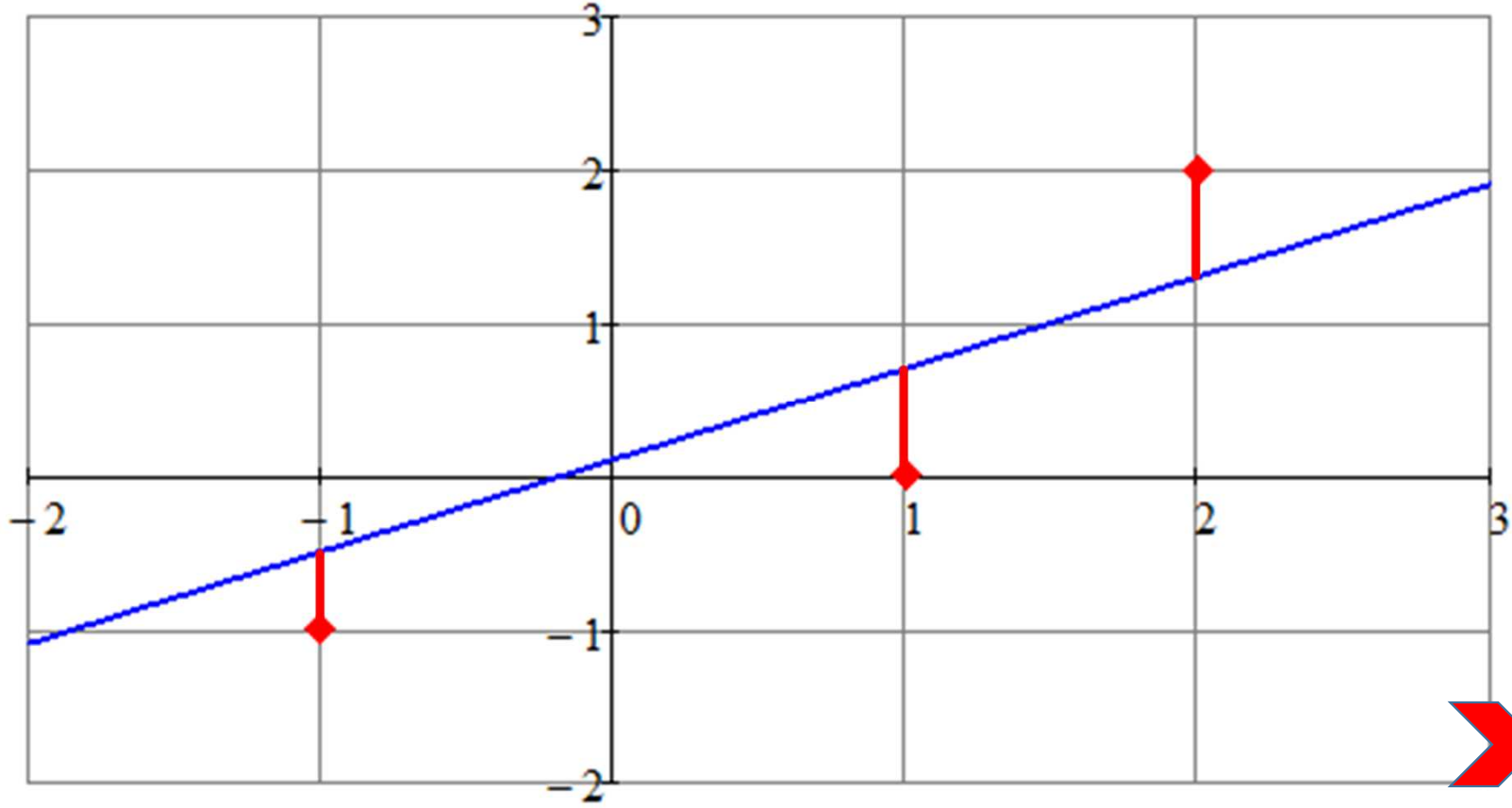
Least Squares Estimation

It is proposed, or assumed, that the model explaining the data is a line $f(x) = kx + b$, where k is a slope and b is an intercept of the line.



Least Squares Estimation

The residuals r_i depend on the values of parameters k and b:



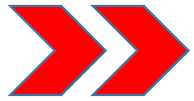
Least Squares Estimation

The sum of squared residuals function we need to minimize is given by:

$$SSR(k,b) = \sum_{i=1}^{3} \left(Y_i - (kX_i + b)\right)^2$$

Substituting X_i and Y_i with the observed values, we get

$$SSR(k,b) = 3b^2 - 2b + 4bk + 6k^2 - 10k + 5$$



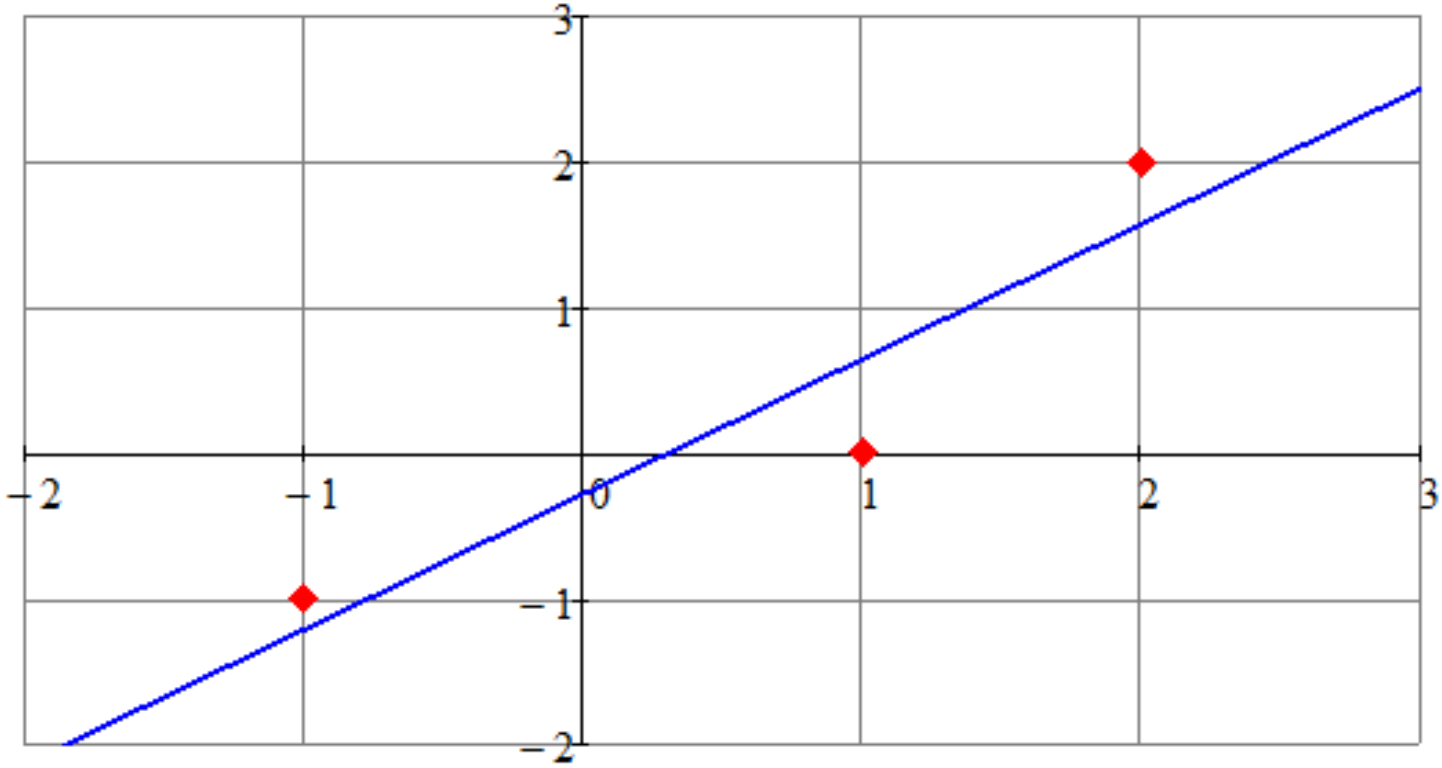
Least Squares Estimation

In order to find the minimum of the function $SSR(k, b)$, we need to set its partial derivatives with respect to k and b to zero, and solve the system of equations:

$$\begin{cases} \dfrac{\partial SSR}{\partial k} = 12k + 4b - 10 = 0; \\ \dfrac{\partial SSR}{\partial b} = 4k + 6b - 2 = 0. \end{cases}$$

Least Squares Estimation

Solving the system of equations for k and b, we obtain the estimates for the parameter values: $\hat{k} = 0.928$, $\hat{b} = -0.2857$.



Least Squares Estimation

The sum of squared residuals for the obtained values is

$$SSR(\hat{k}, \hat{b}) = 0.6429,$$

and this is the least value possible; all other sets of parameters will give us greater values of the SSR.

Least Squares Estimation

When estimating distribution parameters based on a random sample, the elements of the sample serve as X-values.

The Y-values might be obtained by the EDF, or, which is preferable, by calculating so called *median ranks*.

Least Squares Estimation

The Median Ranks method is used to obtain an estimate of the *cdf* for each element of the ordered sample.

The median rank $Y_i = MR$ for the element X_i of the ordered sample is obtained by solving the following equation for MR:

$$0.5 = \sum_{k=i}^{n} \binom{n}{k} MR^k (1 - MR)^{n-k}$$

where n is the sample size and i is the order number.

Least Squares Estimation

A more straightforward and easier (though, less accurate) method of estimating median ranks is so called Benard's Approximation:

$$MR_i = \frac{i - 0.3}{n + 0.4}, \qquad i = 1, \dots, n.$$

Maximum Likelihood Estimation

From a statistical point of view, the method of maximum likelihood estimation (MLE) is considered to be the most robust of the parameter estimation techniques.

The basic idea behind MLE is to obtain the most likely values of the parameters, for a given distribution, that will best describe the data.

Maximum Likelihood Estimation

If X is a continuous random variable with *pdf* $f(x, \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T$ is a vector of m unknown parameters which need to be estimated, and $x_1, x_2, \ldots, x_n$ are elements of the random sample, then

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i, \boldsymbol{\theta})$$

is called the *likelihood function*.

Maximum Likelihood Estimation

The goal of MLE is to find the values of the model parameters that maximize the likelihood function over the parameter space $\boldsymbol{\Theta}$, that is

$$\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \epsilon \boldsymbol{\Theta}} \hat{L}(\boldsymbol{\theta})$$

Maximum Likelihood Estimation

In practice, it is often convenient to work with the natural logarithm of the likelihood function, called the *log-likelihood*:

$$\Lambda(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f(x_i, \boldsymbol{\theta})$$

Since the logarithm is a monotonic function, the maximum of $\Lambda(\boldsymbol{\theta})$ occurs at the same value of $\widehat{\boldsymbol{\theta}}$ as does the maximum of $L(\boldsymbol{\theta})$.

Maximum Likelihood Estimation

By maximizing $\Lambda(\boldsymbol{\theta})$ which is much easier to work with than $L(\boldsymbol{\theta})$, the maximum likelihood estimators (MLE) $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ are the simultaneous solutions of m equations such that:

$$\frac{\partial \Lambda}{\partial \theta_j} = 0, \qquad j = 1,2, \ldots, m.$$

Maximum Likelihood Estimation

For some probability distributions, these equations can be explicitly solved for $\widehat{\boldsymbol{\theta}}$, but in general *no closed-form solution* to the maximization problem is known or available, and an MLE can only be found via *numerical optimization*.

Interval Estimation

The interval estimation profoundly differs from the point estimation. Instead of providing a single number that has virtually zero chances of being equal to the true value of the parameter, it gives you the range which covers true value with given probability.

Interval Estimation

One of the most widespread form of the interval estimation is confidence interval (CI), which proposes a range of plausible values for an unknown parameter (for example, the mean).

The interval has an associated confidence level that the true parameter is in the proposed range.

Interval Estimation

Given observations $x_1, x_2, \ldots, x_n$ and a confidence level γ, a valid confidence interval has a γ probability of containing the true underlying parameter.

The level of confidence can be chosen by the investigator.

Interval Estimation

Given observations $x_1, x_2, \ldots, x_n$ and a confidence level γ, (sometimes $\gamma \cdot 100\%$), a valid confidence interval has a γ probability of containing the true underlying parameter.

The level of confidence can be chosen by the investigator. Most commonly, the 95% confidence level is used. However, confidence levels of 90% and 99% are also often used in analysis.

Interval Estimation

Factors affecting the width of the confidence interval include the size of the sample, the confidence level, and the variability in the sample.

A larger sample will tend to produce a better estimate of the population parameter, when all other factors are equal. A higher confidence level will tend to produce a broader confidence interval.

Interval Estimation

Let's assume that an unbiased estimate $\hat{\theta}$ is obtained for a certain parameter θ.

We predefine some large value of γ, such that the random event with the probability γ could be regarded as almost certain.

Interval Estimation

We want to obtain such ε that

$$\Pr\left\{\left|\theta - \hat{\theta}\right| \leq \varepsilon\right\} = \gamma.$$

In other form

$$\Pr\left\{\hat{\theta} - \varepsilon \leq \theta \leq \hat{\theta} + \varepsilon\right\} = \gamma.$$

The values $\hat{\theta} - \varepsilon$ and $\hat{\theta} + \varepsilon$ are called lower and upper boundaries for the CI.

Interval Estimation

We will consider the interval estimation algorithms for the expected value (mean) and the variance of normally distributed random variable, based on a random sample.

Also, for the estimation of the mean value, we will consider two separate cases:

- true value of the variance is known *a priori*;
- true value of the variance is unknown.

Interval Estimation

If $x_1, x_2, \ldots, x_n$ is a sample from $N(\mu, \sigma^2)$ distribution, then the sample mean $\bar{x}$ has $N(\mu, {}^{\sigma^2}/_{n})$ distribution, and from the properties the normal distribution we know that $Z = \frac{\bar{x}-\mu}{\sigma}\sqrt{n}$ has an $N(0,1)$ distribution.

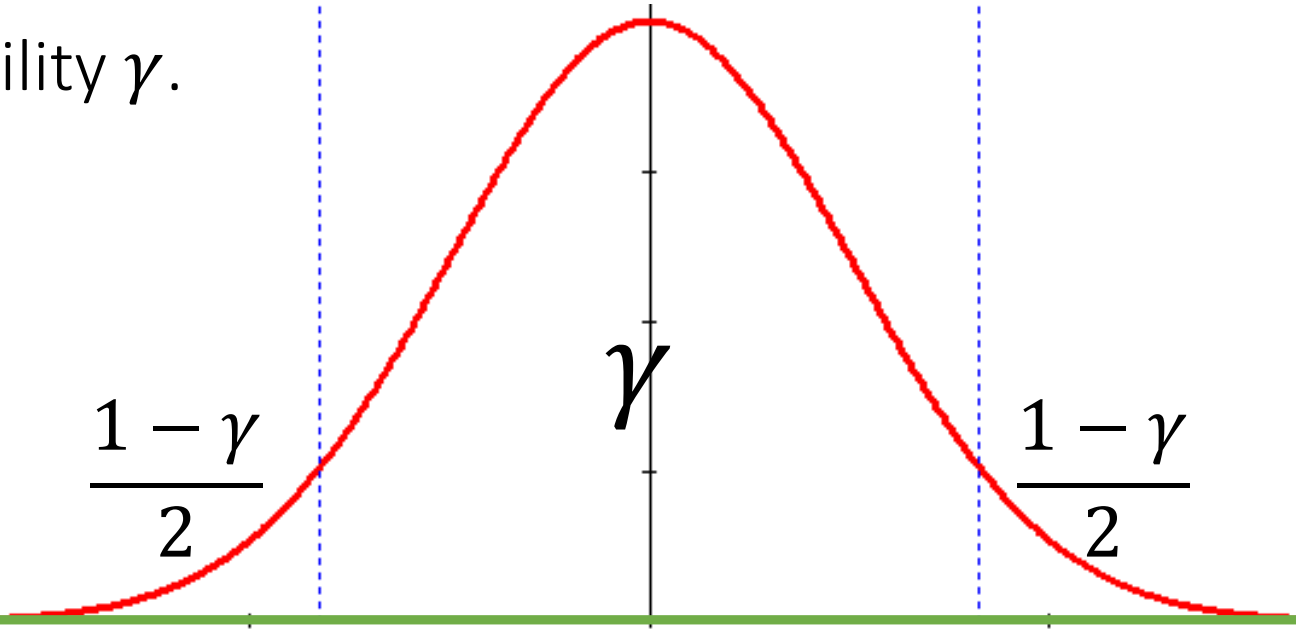
If z_L and z_U are chosen such that $\Pr\{z_L \leq Z \leq z_U\} = \gamma$ for a random variable Z with standard normal distribution, then

$$\gamma = \Pr\left\{z_L \leq \frac{\bar{x} - \mu}{\sigma}\sqrt{n} \leq z_U\right\} =$$

$$= \Pr\left\{z_L \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z_U \frac{\sigma}{\sqrt{n}}\right\} =$$

$$= \Pr\left\{\bar{x} - z_U \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} - z_L \frac{\sigma}{\sqrt{n}}\right\}$$

We have found that $L = \bar{x} - z_U \frac{\sigma}{\sqrt{n}}$ and $U = \bar{x} - z_L \frac{\sigma}{\sqrt{n}}$ satisfy the confidence interval definition: the interval (L, U) covers μ with probability γ.



The values z_U and z_L are the quantiles of the standard normal distribution. If $\alpha = 1 - \gamma$, then

$$z_L = z_{\alpha/2} \text{ and } z_U = z_{1-\alpha/2}.$$

Moreover, since normal distribution is symmetrical,

$$z_{\alpha/2} = -z_{1-\alpha/2}.$$

Summarizing, the $\gamma \cdot 100\%$ confidence interval for μ is:

$$I_\gamma = \left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

For example, if $\alpha = 0.05$, we use $z_{0.975} = 1.96$ and the 95% confidence interval is

$$I_{0.95} = \left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}; \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

If true value of the variance is not known, we can use its unbiased estimator instead:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 .$$

However, it is known that random variable $T = \frac{\bar{x}-\mu}{s}\sqrt{n}$ follows *Student's t-distribution* with $n-1$ degrees of freedom.

Following the similar reasoning, we obtain the $\gamma \cdot 100\%$ confidence interval for μ as

$$I_\gamma = \left(\bar{x} - t_{1-\alpha/2} \frac{s}{\sqrt{n}}; \bar{x} + t_{1-\alpha/2} \frac{s}{\sqrt{n}} \right)$$

where $t_{1-\alpha/2}$ is $1 - \alpha/2$ quantile of *Student's t-distribution* with $n - 1$ degrees of freedom.

The confidence interval for the variance is obtained by the similar reasoning:

$$\Pr\{\sigma_L^2 \leq \sigma^2 \leq \sigma_U^2\} = \gamma = 1 - \alpha$$

$$I_\gamma = (\sigma_L^2; \sigma_U^2)$$

However, there are some differences.

It was found that if a random variable X follows normal distribution $N(\mu, \sigma^2)$, and sample mean $\bar{x}$ follows $N(\mu, {}^{\sigma^2}/_{n})$ distribution, then the following relation is true:

$$(n-1)s^2 = \sigma^2 \chi^2_{n-1},$$

where χ^2_{n-1} denotes *chi-squared distribution* with $n-1$ degrees of freedom.

To calculate lower and upper boundaries of the CI for the variance, we must find the quantiles of chi-squared distribution:

$$\Pr\{\chi^2 \leq \chi_L^2\} = \Pr\{\chi^2 \geq \chi_U^2\} = \frac{1-\gamma}{2} = \frac{\alpha}{2}.$$

Since chi-squared distribution is not symmetrical, the quantiles for upper and lower boundaries should be calculated separately.

After some transformations we obtain

$$\sigma_L^2 = \frac{(n-1)s^2}{\chi^2_{n-1;1-\alpha/2}}; \quad \sigma_U^2 = \frac{(n-1)s^2}{\chi^2_{n-1;\alpha/2}},$$

where $\chi^2_{n-1;\alpha/2}$ and $\chi^2_{n-1;1-\alpha/2}$ are $\frac{\alpha}{2}$-quantile and $\left(1-\frac{\alpha}{2}\right)$-quantile of the chi-squared distribution with $n-1$ degrees of freedom.

The values of the quantiles for the Student's t-distribution and chi-squared distribution are often tabulated, similar to the quantiles of standard normal distribution.

Also, they can be obtained in Mathcad:

$$z_p = qnorm(p, 0{,}1)$$
$$t_p = qt(p, n-1)$$
$$\chi^2_p = qchisq(p, n-1)$$

Textbook Assignment

F.M. Dekking et al. *A Modern Introduction to…*

- Chapters 21-24, pp. 313-373