Estimation and Inference

Paolo Zacchia

Probability and Statistics

Lecture 5

Statistical estimation

- This Lecture introduces concepts of statistical **estimation**.
- Estimation is the process of making **evaluations** about the probability distributions that generate real world data.
- This is closely related to statistical **test of hypotheses**.
- This Lecture focuses on **parameter estimation**: the use of statistics to evaluate the parameters of a distribution.
- Both methods for parameter estimation introduced in this Lecture are grounded on axiomatic **statistical principles**.
- This Lecture covers **neither** so-called "Bayesian" methods for parameter estimation, **nor** methods for the evaluation of an unknown p.m.f. or p.d.f. (*nonparametric estimation*).

Parameter estimation

Definition 1

(Point) estimators, and their estimates. Any statistic, if used to make evaluations about certain features of a probability distribution, is called a *point estimator* (or more simply, an *estimator*). The sample realization of such a statistic is called an *estimate*.

Definition 2

Parameter space. The set of admissible values for the parameters θ is called *parameter space* and is usually denoted as $\boldsymbol{\Theta} \subseteq \mathbb{R}^K$.

Note about notation. Point estimators are usually denoted by a "hat" and possibly other subscripts. Hence:

$$\widehat{\theta} = \left(\widehat{\theta}_1, \ldots, \widehat{\theta}_K\right)$$

is a vector of point estimators for parameter vector $\theta = (\theta_1, \ldots, \theta_K)$. Analogously, $\widehat{\mu}$ and $\widehat{\sigma}$ are point estimators for the location and scale parameters μ and σ. The parameter space for $\widehat{\sigma}$ is $\mathbb{R}_+$.

The analogy principle

Statistical Principle 2. Analogy. The statistical *analogy principle* states that if the random variables that generate the sample and the parameters are related via some vector-valued function $\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})$ of dimension K, such that for $i = 1, \ldots, N$ a *zero moment condition* can be established:

$$\mathbb{E}\left[\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})\right] = \mathbf{0}$$

it follows that a point estimator for $\boldsymbol{\theta}$ can be obtained as the solution to the so-called *sample analogue* of the zero moment condition, that is the condition that equates the sample mean of $\boldsymbol{m}(\boldsymbol{x}_i; \boldsymbol{\theta})$ to zero.

$$\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{m}\left(\boldsymbol{x}_i; \widehat{\boldsymbol{\theta}}_{MM}\right) = \mathbf{0}$$

Here, the estimator $\widehat{\boldsymbol{\theta}}_{MM}$ is denoted by the subscript that identifies it as a **Method of Moments** (MM) estimator.

Intuition: if moments are a function of $\boldsymbol{\theta}$, match population moments with sample moments, and solve for $\boldsymbol{\theta}$ accordingly.

Example: estimates based on the mean

- For many distributions (Bernoulli, Poisson, normal, logistic, Laplace, etc.) the mean equals one parameter (p, λ, μ etc.).
- Let this be μ. The *zero moment condition* here is:

$$\mathbb{E}\left[X_i - \mu\right] = \mathbb{E}\left[m\left(X_i; \mu\right)\right] = 0$$

- . . . and therefore, the MM estimator is as follows.

$$\widehat{\mu}_{MM} = \frac{1}{N}\sum_{i=1}^{N} X_i = \bar{X}$$

- Things are analogous in a multivariate setting. Let the zero moment condition be, like in the multivariate normal case:

$$\mathbb{E}\left[\boldsymbol{x}_i - \boldsymbol{\mu}\right] = \mathbb{E}\left[\boldsymbol{m}\left(\boldsymbol{x}_i; \boldsymbol{\mu}\right)\right] = \mathbf{0}$$

- . . . and the MM estimator again follows easily.

$$\widehat{\boldsymbol{\mu}}_{MM} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i = \bar{\boldsymbol{x}}$$

Example: estimates based on the variance (1/2)

- Focus now on the **normal** distribution: suppose one wants to estimate σ^2. Two suitable *zero moment conditions* are:

$$\mathbb{E}\left[X_i - \mu\right] = \mathbb{E}\left[m_1\left(X_i; \mu, \sigma^2\right)\right] = 0$$
$$\mathbb{E}\left[\left(X_i - \mathbb{E}\left[X\right]\right)^2 - \sigma^2\right] = \mathbb{E}\left[m_2\left(X_i; \mu, \sigma^2\right)\right] = 0$$

  where the second condition also writes as follows.

$$\mathbb{E}\left[X_i^2\right] - \mu^2 - \sigma^2 = 0$$

- The MM estimators are $\widehat{\mu}_{MM} = \bar{X}$ (first equation) and:

$$\widehat{\sigma}^2_{MM} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2 = \frac{N-1}{N}S^2$$

  which is S^2 *rescaled* by the $(N-1)/N$ factor.

Example: estimates based on the variance (2/2)

- This can extend to other distribution. In the **logistic** case, for example, it is $\mathbb{V}\text{ar}[X_i] = \sigma^2\pi^2/3$ and so:

$$\widehat{\sigma}_{MM} = \sqrt{\frac{3}{\pi^2}\frac{N-1}{N}}S.$$

- This approach also extends to the **multivariate** case. The zero moment conditions of the multivariate normal are:

$$\mathbb{E}[\boldsymbol{x}_i - \boldsymbol{\mu}] = \mathbb{E}[\boldsymbol{m}_{\boldsymbol{\mu}}(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \mathbf{0}$$

$$\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - \mathbb{E}[\boldsymbol{x}_i]\,\mathbb{E}[\boldsymbol{x}_i]^{\mathrm{T}} - \boldsymbol{\Sigma}\right] = \mathbb{E}[\boldsymbol{m}_{\boldsymbol{\Sigma}}(\boldsymbol{x}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})] = \mathbf{0}$$

  or $\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\Sigma} = \mathbf{0}$ for the second set of equations.

- The MM estimators are $\widehat{\boldsymbol{\mu}}_{MM} = \bar{\boldsymbol{x}}$ (first set of equations) and a rescaled sample variance-covariance.

$$\widehat{\boldsymbol{\Sigma}}_{MM} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}} = \frac{N-1}{N}\boldsymbol{S}$$

Example: Gamma method of moments

- In sampling from the Gamma distribution with parameters α and β, the zero moment conditions are as follows.

$$\mathbb{E}\left[X_i - \frac{\alpha}{\beta}\right] = \mathbb{E}\left[m_1\left(X_i; \alpha, \beta\right)\right] = 0$$
$$\mathbb{E}\left[X_i^2 - \frac{\alpha\left(\alpha+1\right)}{\beta^2}\right] = \mathbb{E}\left[m_2\left(X_i; \alpha, \beta\right)\right] = 0$$

- Solving the sample analogs of the moment conditions, the MM estimators for α and β are returned as:

$$\widehat{\alpha}_{MM} = \frac{\bar{X}^2}{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2}$$
$$\widehat{\beta}_{MM} = \frac{\bar{X}}{\frac{1}{N}\sum_{i=1}^{N} X_i^2 - \bar{X}^2}$$

where $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ is the sample mean.

Example: method of moments and regression

- Recall the bivariate linear regression model developed in Lecture 3. The two parameters are as follows.

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \cdot \mathbb{E}[X_i]$$
$$\beta_1 = \frac{\mathbb{C}\text{ov}[X_i, Y_i]}{\mathbb{V}\text{ar}[X_i]}$$

- By substituting population with sample moments, the MM estimators for these parameters are obtained directly.

$$\widehat{\beta}_{0,MM} = \bar{Y} - \bar{X} \cdot \widehat{\beta}_{1,MM}$$
$$\widehat{\beta}_{1,MM} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}$$

- Note: only the assumption that the CEF of Y_i given X_i is linear is necessary to derive these estimators.

Likelihood function

A different approach to point estimation is **Maximum Likelihood Estimation** (MLE), which is based on the following concept.

Definition 3

The Likelihood Function. Suppose that a sample of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ is observed. For fixed values of those realizations, the *likelihood* function is defined as the joint mass or density function of the sample, $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta})$, as a function *of the parameters*; it is generally written as follows.

$$\mathcal{L}(\boldsymbol{\theta} | \mathbf{x}_1, \ldots, \mathbf{x}_N) = f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}) > 0$$

The likelihood function is by definition always positive because only values in the support of $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ can be observed.

The likelihood function resembles a probability function, but "is not quite" (it is a function of *parameters*, it does not integrate to 1).

The likelihood principle

The MLE is predicated upon the following axiomatic principle.

Statistical Principle 3. Likelihood. Suppose that two samples of observations, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, are observed, and they are obtained from distributions with the same unknown parameters $\boldsymbol{\theta}$. Suppose that the likelihood functions that are associated with the two realizations are proportional, in the sense that there exists a constant, expressed as a function of the observations $C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N)$, such that the two likelihood functions are always identical up to this constant:

$$\mathcal{L}(\boldsymbol{\theta}|\, \mathbf{x}_1, \ldots, \mathbf{x}_N) = C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N) \cdot \mathcal{L}(\boldsymbol{\theta}|\, \mathbf{y}_1, \ldots, \mathbf{y}_N)$$

where "always" means for every admissible value of the parameters $\boldsymbol{\theta}$. Then, any evaluation about the parameters should be identical in the two samples.

Interpretation of the likelihood principle

The likelihood principle has two main interpretations.

1. If, for all alternative pairs of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ it uniformly is:

$$C(\mathbf{x}_1, \ldots, \mathbf{x}_N; \mathbf{y}_1, \ldots, \mathbf{y}_N) = 1$$

   then two observations with the same value of the likelihood function imply identical "evaluations" about θ.

2. For any two values of the parameters θ' and θ'', the ratio

$$\frac{\mathcal{L}(\theta' | \mathbf{x}_1, \ldots, \mathbf{x}_N)}{\mathcal{L}(\theta'' | \mathbf{x}_1, \ldots, \mathbf{x}_N)} = \frac{\mathcal{L}(\theta' | \mathbf{y}_1, \ldots, \mathbf{y}_N)}{\mathcal{L}(\theta'' | \mathbf{y}_1, \ldots, \mathbf{y}_N)}$$

   must be constant across any different pairs of observations $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$.

Thus, the logical choice is to pick the value of θ that maximizes the likelihood function.

Maximum likelihood

- The **Maximum Likelihood Estimator** is the statistic:

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \mathcal{L}\left(\boldsymbol{\theta}|\, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)$$

  where the subscript MLE has an obvious meaning.

- Since the likelihood function is always positive, in practical settings it is often useful to maximize its logarithm instead, which is called the **log-likelihood function** instead.

$$\widehat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log \mathcal{L}\left(\boldsymbol{\theta}|\, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)$$

- In random samples, the following holds.

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_{MLE} &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \log \left[\prod_{i=1}^{N} f_{\boldsymbol{x}_i}\left(\mathbf{x}_i; \boldsymbol{\theta}\right)\right] \\ &= \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{i=1}^{N} \log f_{\boldsymbol{x}_i}\left(\boldsymbol{\theta}|\, \mathbf{x}_i\right)\end{aligned}$$

Example: the Bernoulli MLE (1/2)

- Consider a random sample drawn from $X_i \sim \text{Be}(p)$.
- The interest falls on p, with parameter space $\boldsymbol{\Theta} = [0, 1]$.
- The likelihood function here is:

$$\mathcal{L}(p|\, x_1, \ldots, x_N) = \prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}$$

- . . . while the log-likelihood function is as follows.

$$\log \mathcal{L}(p|\, x_1, \ldots, x_N) = \left(\sum_{i=1}^{N} x_i \right) \log(p) + \left(N - \sum_{i=1}^{N} x_i \right) \log(1-p)$$

Example: the Bernoulli MLE (2/2)

- The First Order Condition with respect to p is as follows.

$$\frac{d \log \mathcal{L}\left(\widehat{p}_{MLE} | \, x_1, \ldots, x_N\right)}{dp} = \frac{\sum_{i=1}^{N} x_i}{\widehat{p}_{MLE}} + \frac{N - \sum_{i=1}^{N} x_i}{1 - \widehat{p}_{MLE}} = 0$$

- This allows to verify that the MLE is the sample mean.

$$\widehat{p}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} X_i = \bar{X}$$

- Note that since $X_i \in \{0, 1\}$ it is $\bar{X} \in [0, 1]$, so the MLE is restricted to valid values in the parameter space.

- The Second Order Condition is as follows:

$$\frac{d^2 \log \mathcal{L}\left(p | \, x_1, \ldots, x_N\right)}{dp^2} = -\frac{\sum_{i=1}^{N} x_i}{p^2} - \frac{N - \sum_{i=1}^{N} x_i}{(1-p)^2} < 0$$

  $\widehat{p}_{MLE}$ is indeed the maximizer of the likelihood function.

Example: the univariate normal MLE (1/5)

- Consider a random sample drawn from $X_i \sim \mathcal{N}\left(\mu, \sigma^2\right)$.
- The interest falls on both parameters; as already argued the parameter space for σ^2 is restricted to $\boldsymbol{\Theta} = \mathbb{R}_+$.
- The likelihood function here is:

$$\begin{aligned}\mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right) \\ &= \frac{1}{\left(2\pi\sigma^2\right)^{\frac{N}{2}}} \exp\left(-\sum_{i=1}^{N} \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\right)\end{aligned}$$

- . . . while the log-likelihood function is as follows.

$$\begin{aligned}\log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) = &-\frac{N}{2}\log\left(2\pi\right) - \\ &-\frac{N}{2}\log\left(\sigma^2\right) - \sum_{i=1}^{N} \frac{\left(x_i - \mu\right)^2}{2\sigma^2}\end{aligned}$$

Example: the univariate normal MLE (2/5)

- The First Order Conditions, *evaluated at the solution*, are:

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \mu} = \sum_{i=1}^{N} \frac{x_i - \widehat{\mu}_{MLE}}{\widehat{\sigma}^2_{MLE}} = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \sigma^2} =$$

$$= -\frac{N}{2\widehat{\sigma}^2_{MLE}} + \sum_{i=1}^{N} \frac{\left(x_i - \widehat{\mu}_{MLE}\right)^2}{2\widehat{\sigma}^4_{MLE}} = 0$$

  which is a system of two equations in two unknowns.

- The solution is the pair of MLE estimators, which here are identical to their MM counterparts:

$$\widehat{\mu}_{MLE} = \frac{1}{N}\sum_{i=1}^{N} X_i \quad \& \quad \widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)^2.$$

Example: the univariate normal MLE (3/5)

- To check that this is indeed a maximum, one must study the determinant of the *Hessian matrix*.

$$\mathbf{H}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right) = \\ = \begin{bmatrix} \frac{\partial^2 \log \mathcal{L}\left(\mu,\sigma^2 \middle| x_1,\ldots,x_N\right)}{\partial \mu^2} & \frac{\partial^2 \log \mathcal{L}\left(\mu,\sigma^2 \middle| x_1,\ldots,x_N\right)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \log \mathcal{L}\left(\mu,\sigma^2 \middle| x_1,\ldots,x_N\right)}{\partial \sigma^2 \partial \mu} & \frac{\partial^2 \log \mathcal{L}\left(\mu,\sigma^2 \middle| x_1,\ldots,x_N\right)}{\partial (\sigma^2)^2} \end{bmatrix}$$

- These second-order partial derivatives are as follows.

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu^2} = -\frac{N}{\sigma^2}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \mu \partial \sigma^2} = -\sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^4}$$

$$\frac{\partial^2 \log \mathcal{L}\left(\mu, \sigma^2 \middle| x_1, \ldots, x_N\right)}{\partial \left(\sigma^2\right)^2} = \frac{N}{2\sigma^4} - \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{\sigma^6}$$

Example: the univariate normal MLE (4/5)

- *At the solution* the cross-derivatives are zero.

$$-\frac{1}{\widehat{\sigma}^2_{MLE}}\sum_{i=1}^{N}(x_i-\widehat{\mu}_{MLE})=0$$

- At the same time, *at the solution* the second derivative for σ^2 largely simplifies.

$$\frac{\partial^2 \log \mathcal{L}\left(\widehat{\mu}_{MLE},\widehat{\sigma}^2_{MLE}\middle|\, x_1,\ldots,x_N\right)}{\partial\left(\sigma^2\right)^2}=$$

$$=\frac{N}{2\widehat{\sigma}^4_{MLE}}-\sum_{i=1}^{N}\frac{\left(x_i-\widehat{\mu}_{MLE}\right)^2}{\widehat{\sigma}^6_{MLE}}$$

$$=\frac{N}{2\widehat{\sigma}^4_{MLE}}-\frac{N}{\widehat{\sigma}^4_{MLE}}$$

$$=-\frac{N}{\widehat{\sigma}^4_{MLE}}$$

Example: the univariate normal MLE (5/5)

- *At the solution* the Hessian matrix is thus as follows.

$$\mathbf{H}\left(\widehat{\mu}_{MLE}, \widehat{\sigma}^2_{MLE} \middle| x_1, \ldots, x_N\right) = \begin{bmatrix} -\dfrac{N}{\widehat{\sigma}^2_{MLE}} & 0 \\ 0 & -\dfrac{N}{2\widehat{\sigma}^4_{MLE}} \end{bmatrix}$$

  Clearly, its determinant is always positive.

- Also observe that there is at least *one* second-order partial derivative (that for μ) which is *always* negative.

- Thus, the conditions for a maximum in multivariate calculus are satisfied.

Example: the multivariate normal MLE (1/3)

- Consider a random sample drawn from $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.
- The likelihood function here is:

$$\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x}_1, \ldots, \mathbf{x}_N) =$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}|}} \exp\left(-\frac{(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2} \right)$$

$$= \frac{1}{\left[(2\pi)^K |\boldsymbol{\Sigma}|\right]^{\frac{N}{2}}} \exp\left(-\sum_{i=1}^{N} \frac{(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2} \right)$$

- . . . while the log-likelihood function is as follows.

$$\log \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \mathbf{x}_1, \ldots, \mathbf{x}_N) = -\frac{NK}{2} \log(2\pi) -$$

$$- \frac{N}{2} \log(|\boldsymbol{\Sigma}|) - \sum_{i=1}^{N} \frac{(\mathbf{x}_i - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})}{2}$$

Example: the multivariate normal MLE (2/3)

- It is useful to proceed in steps. The First Order Conditions with respect to $\boldsymbol{\mu}$, *evaluated at the solution*, are:

$$\begin{aligned}\frac{\partial \log \mathcal{L}\left(\widehat{\boldsymbol{\mu}}_{MLE}, \boldsymbol{\Sigma} | \, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)}{\partial \boldsymbol{\mu}} &= -\boldsymbol{\Sigma}^{-1} \sum_{i=1}^{N} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_{MLE}\right) \\ &= \mathbf{0}\end{aligned}$$

  notice that similarly as in the univariate case, this does not depend on $\boldsymbol{\Sigma}$ (here it helps that $\boldsymbol{\Sigma}$ is semi-definite positive).

- The solution for $\boldsymbol{\mu}$ is the MLE estimator, which again looks like its MM counterpart.

$$\widehat{\boldsymbol{\mu}}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i = \bar{\boldsymbol{x}}$$

Example: the multivariate normal MLE (3/3)

- With regard to $\boldsymbol{\Sigma}$, it is best to work with the First Order Conditions of its *inverse*:

$$\frac{\partial \log \mathcal{L}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma} | \, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)}{\partial \boldsymbol{\Sigma}^{-1}} = \frac{N}{2}\boldsymbol{\Sigma} - \sum_{i=1}^{N} \frac{\left(\mathbf{x}_i - \boldsymbol{\mu}\right)\left(\mathbf{x}_i - \boldsymbol{\mu}\right)^{\mathrm{T}}}{2}$$

- Setting it at zero, and evaluating the expression at $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ and *at the solution* for $\boldsymbol{\Sigma} = \widehat{\boldsymbol{\Sigma}}_{MLE}$, one obtains the MLE:

$$\widehat{\boldsymbol{\Sigma}}_{MLE} = \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)\left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}\right)^{\mathrm{T}}$$

  which is again like the MM version (the "rescaled" $\boldsymbol{S}$).

- Tedious algebraic work would reveal that $\left(\widehat{\boldsymbol{\mu}}_{MLE}, \widehat{\boldsymbol{\Sigma}}_{MLE}\right)$ indeed identify a maximum of the likelihood function.

Example: the Gamma MLE (1/2)

- Not always do the MM and MLE estimators coincide.
- Consider a random sample drawn from $X_i \sim \Gamma(\alpha, \beta)$.
- The likelihood function here is:

$$\mathcal{L}(\alpha, \beta | x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\Gamma(\alpha)} \beta^{\alpha} x_i^{\alpha-1} \exp(-\beta x_i)$$
$$= \left(\frac{\beta^{\alpha}}{\Gamma(\alpha)} \right)^{N} \left(\prod_{i=1}^{N} x_i^{\alpha-1} \right) \exp\left(-\beta \sum_{i=1}^{N} x_i \right)$$

- ... while the log-likelihood function is as follows.

$$\log \mathcal{L}(\alpha, \beta | x_1, \ldots, x_N) = N\alpha \log(\beta) -$$
$$- N \log[\Gamma(\alpha)] + (\alpha - 1) \sum_{i=1}^{N} \log(x_i) - \beta \sum_{i=1}^{N} x_i$$

Example: the Gamma MLE (2/2)

- The First Order Conditions are as follows.

$$\frac{\partial \log \mathcal{L}\left(\alpha, \beta | x_1, \ldots, x_N\right)}{\partial \alpha} = N \log\left(\beta\right) - \frac{N}{\Gamma\left(\alpha\right)} \frac{\partial \Gamma\left(\alpha\right)}{\partial \alpha} + \\ + \sum_{i=1}^{N} \log\left(x_i\right)$$

$$\frac{\partial \log \mathcal{L}\left(\alpha, \beta | x_1, \ldots, x_N\right)}{\partial \beta} = N \frac{\alpha}{\beta} - \sum_{i=1}^{N} x_i$$

- Like in the MM case, the solution must satisfy:

$$\frac{\widehat{\alpha}_{MLE}}{\widehat{\beta}_{MLE}} = \frac{1}{N} \sum_{i=1}^{N} X_i = \bar{X}$$

but otherwise the solution *lacks a closed form* and must be found *via numerical methods*, unlike in the MM case.

Example: no MLE for the "open" uniform! (1/2)

- Sometimes, the MLE does not even exist!
- Consider a random sample drawn from $X_i \sim \mathcal{U}(0, \theta)$ with *closed support*: $\mathbb{X} = [0, \theta]$.
- It is easy to see that $\mathbb{E}[X] = \theta/2$ and thus the MM is:

$$\widehat{\theta}_{MM} = \frac{2}{N} \sum_{i=1}^{N} X_i = 2\bar{X}.$$

- Instead, the MLE is the sample *maximum*: $\widehat{\theta}_{MLE} = X_{(N)}$.
- This can be verified by inspecting the likelihood function.

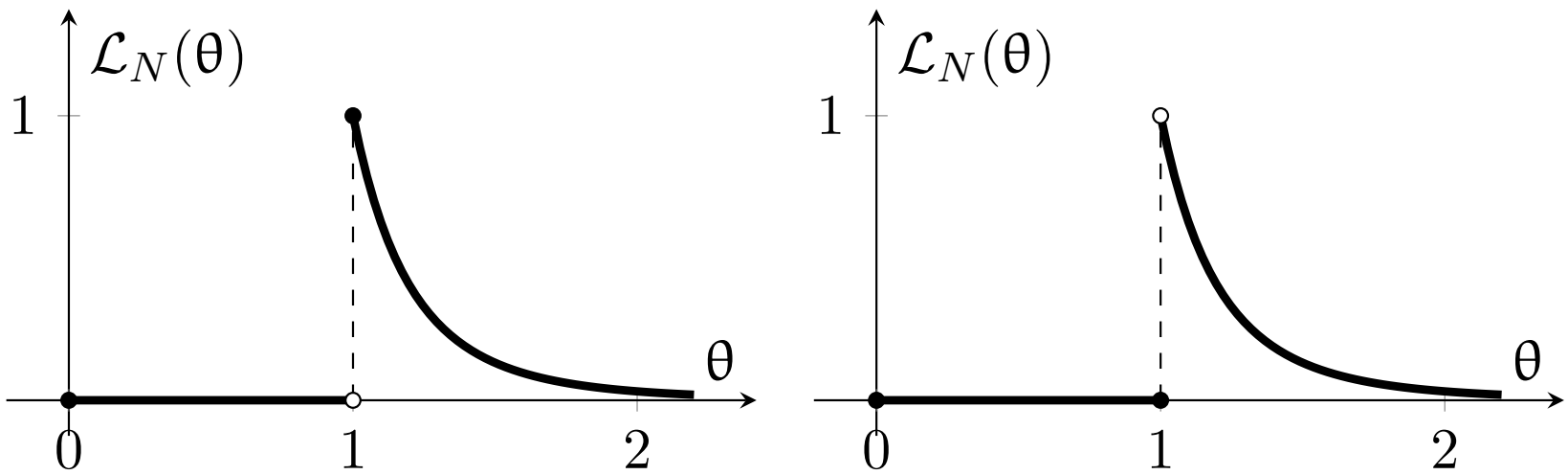
$$\mathcal{L}(\theta | x_1, \ldots, x_N) = \frac{1}{\theta^N} \cdot \mathbb{1}[0 \leq x_1, \ldots, x_N \leq \theta]$$

Example: no MLE for the "open" uniform! (2/6)

- And if the support is open, at least on the right: $\mathbb{X} = [0, \theta)$? One inequality in the identity function becomes strict now.

$$\mathcal{L}(\theta | x_1, \ldots, x_N) = \frac{1}{\theta^N} \cdot \mathbb{1}[0 \leq x_1, \ldots, x_N < \theta]$$

- As a consequence, there is no MLE any more!



- Note: $N = 5$ and $x_{(5)} = 1$ in both cases; however $\mathbb{X} = [0, \theta]$ in the left panel and $\mathbb{X} = [0, \theta)$ in the right panel. $\mathcal{L}_N(\theta)$ is shorthand notation for $\mathcal{L}(\theta | x_1, \ldots, x_N)$.

The invariance property of MLE (1/2)

All Maximum Likelihood Estimators have a convenient and important property. This property is not shared with MM estimators.

Theorem 1

Invariance of Maximum Likelihood Estimators. *Call* $\widehat{\boldsymbol{\theta}}_{MLE}$ *the Maximum Likelihood Estimator for a given parameter vector* $\boldsymbol{\theta}$*. Let* $\boldsymbol{\varphi} = \boldsymbol{g}(\boldsymbol{\theta})$ *be a* transformation *of parameter vector* $\boldsymbol{\theta}$*. The Maximum Likelihood estimator of* $\boldsymbol{\varphi}$ *is simply the corresponding transformation of the Maximum Likelihood Estimator of* $\boldsymbol{\theta}$*.*

$$\widehat{\boldsymbol{\varphi}}_{MLE} = \boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)$$

Proof.

The MLE of $\boldsymbol{\varphi}$ maximizes the so-called *induced likelihood function.*

$$\mathcal{L}^*\left(\boldsymbol{\varphi}\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_N\right) = \max_{\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{\varphi}\}} \mathcal{L}\left(\boldsymbol{\theta}\middle|\, \mathbf{x}_1, \ldots, \mathbf{x}_N\right)$$

(Continues...)

The invariance property of MLE (2/2)

Theorem 1

Proof.

(Continued.) Call such a maximum $\widehat{\boldsymbol{\varphi}}_{MLE}$, and observe that:

$$\begin{aligned}
\mathcal{L}^*\left(\widehat{\boldsymbol{\varphi}}_{MLE}\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right) &= \max_{\boldsymbol{\varphi}} \max_{\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{\varphi}\}} \mathcal{L}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right) \\
&= \max_{\boldsymbol{\theta}} \mathcal{L}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right) \\
&= \mathcal{L}\left(\widehat{\boldsymbol{\theta}}_{MLE}\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right) \\
&= \max_{\left\{\boldsymbol{\theta}:\boldsymbol{g}(\boldsymbol{\theta})=\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)\right\}} \mathcal{L}\left(\boldsymbol{\theta}\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right) \\
&= \mathcal{L}^*\left(\boldsymbol{g}\left(\widehat{\boldsymbol{\theta}}_{MLE}\right)\middle|\,\mathbf{x}_1,\ldots,\mathbf{x}_N\right)
\end{aligned}$$

where the first and last equalities follow from the definition of induced likelihood function, the second equality follows from the properties of iterated maximizations, while the remaining equalities follow from the definition of MLE. □

Example: the reparametrized normal MLE (1/2)

- Consider a random sample drawn from $X_i \sim \mathcal{N}\left(\mu, 1/\phi^2\right)$.
- Here the typical reparametrization of the scale parameter is adopted: $\phi^2 = 1/\sigma^2$ is the *precision* parameter.
- The likelihood function here is:

$$\begin{aligned}\mathcal{L}\left(\mu, \phi^2 \middle| x_1, \ldots, x_N\right) &= \prod_{i=1}^{N} \sqrt{\frac{\phi^2}{2\pi}} \exp\left(-\frac{\phi^2 \left(x_i - \mu\right)^2}{2}\right) \\ &= \left(\frac{\phi^2}{2\pi}\right)^{\frac{N}{2}} \exp\left(-\sum_{i=1}^{N} \frac{\phi^2 \left(x_i - \mu\right)^2}{2}\right)\end{aligned}$$

- . . . while the log-likelihood function is as follows.

$$\begin{aligned}\log \mathcal{L}\left(\mu, \phi^2 \middle| x_1, \ldots, x_N\right) &= -\frac{N}{2} \log\left(2\pi\right) + \\ &\quad + \frac{N}{2} \log\left(\phi^2\right) - \sum_{i=1}^{N} \frac{\phi^2 \left(x_i - \mu\right)^2}{2}\end{aligned}$$

Example: the reparametrized normal MLE (2/2)

- The First Order Conditions, *evaluated at the solution*, are:

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\phi}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \mu} =$$

$$= \widehat{\phi}^2_{MLE} \sum_{i=1}^{N} (x_i - \widehat{\mu}_{MLE}) = 0$$

$$\frac{\partial \log \mathcal{L}\left(\widehat{\mu}_{MLE}, \widehat{\phi}^2_{MLE} \middle| x_1, \ldots, x_N\right)}{\partial \phi^2} =$$

$$= \frac{N}{2\widehat{\phi}^2_{MLE}} + \sum_{i=1}^{N} \frac{(x_i - \widehat{\mu}_{MLE})^2}{2} = 0$$

- ...and it is easy to see that $\widehat{\phi}^2_{MLE} = 1/\widehat{\sigma}^2_{MLE}$ exactly.

$$\widehat{\phi}^2_{MLE} = N \left[\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2\right]^{-1}$$

Example: a reparametrized Gamma MLE (1/2)

- Consider a random sample drawn from $X_i \sim \Gamma(\alpha, 1/\theta)$. It is common to re-parametrize the Gamma distribution this way (β and θ are called *rate* and *scale* respectively).

- The likelihood function here is:

$$\mathcal{L}(\alpha, \theta | x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\Gamma(\alpha) \cdot \theta^{\alpha}} x_i^{\alpha-1} \exp\left(-\frac{1}{\theta} x_i\right)$$

$$= \left(\frac{\theta^{-\alpha}}{\Gamma(\alpha)}\right)^{N} \left(\prod_{i=1}^{N} x_i^{\alpha-1}\right) \exp\left(-\frac{1}{\theta} \sum_{i=1}^{N} x_i\right)$$

- ... while the log-likelihood function is as follows.

$$\log \mathcal{L}(\alpha, \theta | x_1, \ldots, x_N) = -N\alpha \log(\theta) -$$

$$- N \log[\Gamma(\alpha)] + (\alpha - 1) \sum_{i=1}^{N} \log(x_i) - \frac{1}{\theta} \sum_{i=1}^{N} x_i$$

Example: a reparametrized Gamma MLE (2/2)

- The First Order Conditions are as follows.

$$\frac{\partial \log \mathcal{L}\left(\alpha, \theta | x_1, \ldots, x_N\right)}{\partial \alpha} = -N \log\left(\theta\right) - \frac{N}{\Gamma\left(\alpha\right)} \frac{\partial \Gamma\left(\alpha\right)}{\partial \alpha} + \\ + \sum_{i=1}^{N} \log\left(x_i\right)$$

$$\frac{\partial \log \mathcal{L}\left(\alpha, \theta | x_1, \ldots, x_N\right)}{\partial \theta} = -\frac{N\alpha}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{N} x_i$$

- Again, the solution lacks a closed form, but the analysis of the conditions reveals that $\widehat{\theta}_{MLE} = 1/\widehat{\beta}_{MLE}$, and that the fixed relationship with the MLE for α is now as follows.

$$\widehat{\alpha}_{MLE} \widehat{\theta}_{MLE} = \frac{1}{N} \sum_{i=1}^{N} X_i = \bar{X}$$

Evaluation of estimators

- After an estimator is obtained, a natural question to ask is "how good is it" for evaluating the parameters of interest.

- To answer this, it is useful to remember that all estimators are statistics, and as such have sampling distributions.

- To *evaluate* different estimators one can study, for example, their key **moments**: mean and variance.

- The **mean** of an estimator can say how close an estimator gets, on average, to the parameters of interest.

- The **variance** of an estimator speaks about the "precision" and variability of the associated estimates.

- The **statistical properties** of estimators are multifaceted and multidimensional, and should be evaluated accordingly.

Mean Squared Error (MSE)

The following is a leading criterion for the evaluation of estimators.

Definition 4

Mean Squared Error (MSE). Consider an estimator $\widehat{\theta}$ for a given parameters of interest θ, where both $\widehat{\theta}$ and θ have dimension K. The *mean squared error* is defined as the following quantity:

$$\text{MSE} \equiv \mathbb{E}\left[\left(\widehat{\theta}-\theta\right)^{\text{T}}\left(\widehat{\theta}-\theta\right)\right] = \sum_{k=1}^{K}\mathbb{E}\left[\left(\widehat{\theta}_k-\theta_k\right)^2\right]$$

where $k = 1, \ldots, K$ indexes the parameters and associated estimators listed in θ and $\widehat{\theta}$, respectively.

Thus, for any vector of parameters and associated estimates the MSE is simply the sum of K elements of the form:

$$\text{MSE}_k = \mathbb{E}\left[\left(\widehat{\theta}_k-\theta_k\right)^2\right]$$

where both $\widehat{\theta}_k$ and θ_k are unidimensional.

Mean Squared Error: discussion

- Alternative criteria are possible, like say the following :

$$\mathrm{MAE}_k = \mathbb{E}\left[\left|\widehat{\theta}_k - \theta_k\right|\right]$$

  which is called **Mean Absolute Error** (MAE).

- The overwhelming majority of practical applications adopts the MSE largely thanks to the following decomposition:

$$\begin{aligned}
\mathrm{MSE}_k &= \mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right] + \mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2\right] \\
&= \mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right]\right)^2\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2\right] + \\
&\qquad + 2\,\mathbb{E}\left[\left(\widehat{\theta}_k - \mathbb{E}\left[\widehat{\theta}_k\right]\right)\left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)\right] \\
&= \mathbb{V}\mathrm{ar}\left[\widehat{\theta}_k\right] + \left(\mathbb{E}\left[\widehat{\theta}_k\right] - \theta_k\right)^2
\end{aligned}$$

  Above, the last element in the second line is easily shown to vanish to zero, as in a similar analysis in Lecture 1.

Bias and unbiasedness

The following is a desirable property of estimators.

Definition 5

Bias and unbiasedness. Consider a unidimensional estimator $\widehat{\theta}$ for some parameter of interest θ. Its *bias* is the quantity:

$$\text{Bias}_{\widehat{\theta}} \equiv \mathbb{E}\left[\widehat{\theta}\right] - \theta$$

and the estimator is *unbiased* if its bias is zero.

- Interpretation: an unbiased estimator returns the parameter of interest **on average** (across multiple samples).
- The MSE of an unbiased estimator equals its variance.
- Sometimes, unbiased estimators have larger variance than biased ones, that are said to be more **efficient**.
- This leads to a so-called **bias-variance trade-off**.

Example: MSE for the normal variance (1/2)

Consider a random sample drawn from a normally distributed random variable $X \sim \mathcal{N}(\mu, \sigma^2)$, and the following two pairs of alternative estimators for the two parameters:

$$\begin{pmatrix} \widehat{\mu}_1 \\ \widehat{\sigma}_1^2 \end{pmatrix} = \begin{pmatrix} \bar{X} \\ S^2 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \widehat{\mu}_2 \\ \widehat{\sigma}_2^2 \end{pmatrix} = \begin{pmatrix} \bar{X} \\ \frac{N-1}{N} S^2 \end{pmatrix}$$

where $\widehat{\sigma}_2^2 = \frac{N-1}{N} S^2$ is motivated by either the MM or MLE.

- The two estimators for μ are identical: one shall focus on those for σ^2.
- It was shown in Lecture 4 that $\mathbb{E}\left[S^2 - \sigma^2\right] = 0$, implying that $\widehat{\sigma}_1^2$ is an *unbiased* estimator of σ^2.
- This also implies that $\mathbb{E}\left[\frac{N-1}{N} S^2 - \sigma^2\right] = -\frac{1}{N}\sigma^2$: hence, $\widehat{\sigma}_2^2$ is a *biased* estimator of σ^2.

Example: MSE for the normal variance (2/2)

Yet the second setup has a lower MSE, as $\widehat{\sigma}_2^2$ is **more efficient** than $\widehat{\sigma}_1^2$! Compare their variances:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_1^2\right] - \mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_2^2\right] &= \mathbb{V}\mathrm{ar}\left[S^2\right] - \left(\frac{N-1}{N}\right)^2 \mathbb{V}\mathrm{ar}\left[S^2\right] \\
&= \frac{2N-1}{N^2}\, \mathbb{V}\mathrm{ar}\left[S^2\right] \\
&= \frac{(2N-1)\,\sigma^4}{N^2\,(N-1)^2}\, \mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] \\
&= \frac{2\,(2N-1)\,\sigma^4}{N^2\,(N-1)}
\end{aligned}$$

where the last line follows from the fact that if $W \sim \chi_\kappa^2$, then $\mathbb{V}\mathrm{ar}\left[W\right] = 2\kappa$ (here, $\kappa = N-1$). In addition:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_1^2\right] - \mathbb{V}\mathrm{ar}\left[\widehat{\sigma}_2^2\right] = \frac{2\,(2N-1)\,\sigma^4}{N^2\,(N-1)} > \frac{\sigma^4}{N^2} = \left\{\mathbb{E}\left[\widehat{\sigma}_2^2 - \sigma^2\right]\right\}^2$$

the higher precision compensates for unbiasedness in the MSE!

Best unbiased estimators

Definition 6

Best unbiased estimators. Consider the set of unbiased estimators $\widehat{\theta}$ of a certain parameter θ:

$$\mathbb{C}_\theta = \left\{ \widehat{\theta} : \mathbb{E}\left[\widehat{\theta}\right] = \theta \right\}$$

An estimator $\widehat{\theta}^*$ is called the *best unbiased estimator*, or the *uniform minimum variance unbiased estimator* of θ, if the following holds.

$$\mathbb{V}\text{ar}\left[\widehat{\theta}\right] - \mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] \geq 0 \quad \text{for all } \widehat{\theta} \in \mathbb{C}_\theta$$

In a multidimensional environment, if $\widehat{\boldsymbol{\theta}}$ is a vector of estimators that are all unbiased for a vector of parameters $\boldsymbol{\theta}$, this definition is recast in terms of a vector $\widehat{\boldsymbol{\theta}}^*$ of best unbiased estimators such that:

$$\mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}\right] - \mathbb{V}\text{ar}\left[\widehat{\boldsymbol{\theta}}^*\right] \geq \mathbf{0} \quad \text{for all } \widehat{\boldsymbol{\theta}} \in \mathbb{C}_{\theta_1} \times \ldots \times \mathbb{C}_{\theta_K}$$

meaning that the matrix on the left-hand side is positive semi-definite, and $\mathbb{C}_{\theta_k}$ is the set of unbiased estimators of θ_k for $k = 1, \ldots, K$.

Uniqueness of best unbiased estimators (1/2)

Theorem 2

Best unbiased estimators: uniqueness. *Let* $\widehat{\theta}^*$ *be a best unbiased estimator for some parameter* θ*. In this setting* $\widehat{\theta}^*$ *is unique.*

Proof.

Suppose that there is another best unbiased estimator $\widehat{\theta}^{**}$ that has the same expectation and variance as $\widehat{\theta}^*$. Define the estimator:

$$\widehat{\theta}' \equiv \frac{1}{2}\widehat{\theta}^* + \frac{1}{2}\widehat{\theta}^{**}$$

it is clear that $\mathbb{E}\left[\widehat{\theta}'\right] = \theta$. As per the variance, it must be that:

$$\begin{aligned}
\mathbb{V}\text{ar}\left[\widehat{\theta}'\right] &= \frac{1}{4}\,\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] + \frac{1}{4}\,\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right] + \frac{1}{2}\,\mathbb{C}\text{ov}\left[\widehat{\theta}^*, \widehat{\theta}^{**}\right] \\
&\leq \frac{1}{4}\,\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right] + \frac{1}{4}\,\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right] + \frac{1}{2}\left\{\mathbb{V}\text{ar}\left[\widehat{\theta}^*\right]\mathbb{V}\text{ar}\left[\widehat{\theta}^{**}\right]\right\}^{\frac{1}{2}} \\
&= \mathbb{V}\text{ar}\left[\widehat{\theta}^*\right]
\end{aligned}$$

(Continues...)

Uniqueness of best unbiased estimators (2/2)

Theorem 2

Proof.

(Continued.) In the previous display, the inequality in the second line follows from the properties of covariances and correlation, and the last line is due to the fact that $\widehat{\theta}^*$ and $\widehat{\theta}^{**}$ have the same variance by hypothesis.

Observe that the inequality must be replaced by an equality to avoid a contradiction! If the inequality were sharp, then $\widehat{\theta}^*$ would not be a best unbiased estimator, as $\widehat{\theta}'$ would improve it. To have an equality, it must be – again by the properties of covariances and correlations – that $\widehat{\theta}^{**}$ is a linear transformation of $\widehat{\theta}^*$, that is $\widehat{\theta}^{**} = a + b\widehat{\theta}^*$.

But then it must also be that $a = 0$, or else $\widehat{\theta}^{**}$ would be biased, and $b = 1$, since the following chain of equalities must also hold.

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}^*\right] = \mathbb{C}\mathrm{ov}\left[\widehat{\theta}^*, \widehat{\theta}^{**}\right] = \mathbb{C}\mathrm{ov}\left[\widehat{\theta}^*, b\widehat{\theta}^*\right] = b\,\mathbb{V}\mathrm{ar}\left[\widehat{\theta}^*\right]$$

Thus, $\widehat{\theta}^*$, $\widehat{\theta}^{**}$ and $\widehat{\theta}'$ are all identical estimators, that is, $\widehat{\theta}^*$ is the only best unbiased estimator. □

The Rao-Blackwell Theorem (1/2)

The following result links sufficient statistics with efficient unbiased estimators.

Theorem 3

The Rao-Blackwell Theorem. *Consider an environment where:*

- $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *is a sample drawn from some list of random vectors,*
- θ *is some parameter vector of interest,*
- $\widehat{\theta}$ *is any vector of unbiased estimators of* θ*, and*
- $\boldsymbol{t} = \boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *is a vector of statistics that are simultaneously sufficient for* θ*.*

Define the following statistic as a conditional expectation function.

$$\widehat{\theta}^* \equiv \mathbb{E}\left[\widehat{\theta} \middle| \boldsymbol{t}\right]$$

The statistic $\widehat{\theta}^*$ *is a uniformly better unbiased estimator of* θ*, that is, it is an unbiased estimator with lower variance than* $\widehat{\theta}$*.*

Proof.

(Continues...)

The Rao-Blackwell Theorem (2/2)

Theorem 3

Proof.

(Continued.) The Law of Iterated Expectations:

$$\theta = \mathbb{E}\left[\widehat{\theta}\right] = \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{E}\left[\widehat{\theta}\middle|\,\boldsymbol{t}\right]\right] = \mathbb{E}\left[\widehat{\theta}^{*}\right]$$

along with the Law of Total Variance:

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] &= \mathbb{V}\mathrm{ar}_{\boldsymbol{t}}\left[\mathbb{E}\left[\widehat{\theta}\middle|\,\boldsymbol{t}\right]\right] + \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\middle|\,\boldsymbol{t}\right]\right] \\ &= \mathbb{V}\mathrm{ar}\left[\widehat{\theta}^{*}\right] + \mathbb{E}_{\boldsymbol{t}}\left[\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\middle|\,\boldsymbol{t}\right]\right] \\ &\geq \mathbb{V}\mathrm{ar}\left[\widehat{\theta}^{*}\right]\end{aligned}$$

simultaneously show that *if* $\widehat{\theta}^{*}$ is an estimator of θ, it is unbiased and it also has a lower variance than $\widehat{\theta}$. The definition of sufficiency along with that of $\widehat{\theta}^{*}$, however, jointly imply that the latter is a legitimate estimator of θ, because its joint distribution by construction does not depend on θ. □

Use of the Rao-Blackwell Theorem, and example

- This result allows to "improve" existing estimators that are based on sufficient statistics or to "prove" that they cannot be improved!

- Example: consider sampling from the normal distribution, where the two statistics $(\bar{X}, S^2)$ are used as estimators of parameters (μ, σ^2).

- These statistics are sufficient, and thus can be expressed as per the Rao-Blackwell Theorem as conditional expectation functions of themselves (and quite trivial ones).

- Thus one cannot construct out of $(\bar{X}, S^2)$ a pair of unbiased estimators that is more efficient than $(\bar{X}, S^2)$ themselves.

- Yet as seen there may be estimators with lower MSE...

The Cramér-Rao Bound: introduction

- In order to evaluate estimators (both biased and unbiased) their variance must be compared against that of competing estimators.

- A fundamental result by H. Cramér, C. R. Rao and others allows to establish a **bound** for the variance of estimators.

- It implies that no estimator for the same parameter(s) and that a lower variance variance can be obtained.

- As lecture 6 shows, this result shines in asymptotic settings where MLE estimators usually hit the Cramér-Rao bound.

- The proof of this result is fairly simple yet best developed in steps, showing the univariate case first before moving to the multivariate one.

Cramér-Rao bound: general, univariate (1/4)

Theorem 4

Cramér-Rao Inequality (general) – univariate case. *Consider a sample drawn from a list of random variables* $(X_1, \ldots, X_N)$ *with joint p.m.f. or p.d.f. written as* $f(x_1, \ldots, x_N; \theta)$ *using shorthand notation. Also consider some parameter of interest* θ*, as well as some estimator* $\widehat{\theta} = \widehat{\theta}(X_1, \ldots, X_N)$ *for* θ*, such that its variance is finite and – in the continuous case – that the differentiation operation with respect to* θ *can pass through the expectation operator as shown below.*

$$\frac{\partial}{\partial \theta} \mathbb{E}\left[\widehat{\theta}\right] = \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \frac{\partial}{\partial \theta} \widehat{\theta}(x_1, \ldots, x_N) \cdot f(x_1, \ldots, x_N; \theta)\, dx_1 \ldots dx_N$$

In this setting the variance of $\widehat{\theta}$ *must satisfy the following inequality.*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] \geq \frac{\left(\frac{\partial}{\partial \theta} \mathbb{E}\left[\widehat{\theta}\right]\right)^2}{\mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f(X_1, \ldots, X_N; \theta)\right)^2\right]}$$

Proof.

(Continues...)

Cramér-Rao bound: general, univariate (2/4)

Theorem 4

Proof.

(Continued.) Defines the following transformed random variables.

$$U = \widehat{\theta}\left(X_1, \ldots, X_N\right)$$

$$V = \frac{\partial}{\partial \theta} \log f\left(X_1, \ldots, X_N; \theta\right)$$

The result follows as a natural property of covariances as proved in Lecture 3.

$$\mathbb{V}\mathrm{ar}\left[U\right] \geq \frac{\left[\mathbb{C}\mathrm{ov}\left[U, V\right]\right]^2}{\mathbb{V}\mathrm{ar}\left[V\right]}$$

Note that if $\mathbb{E}\left[V\right] = 0$ the above is recast as follows.

$$\mathbb{V}\mathrm{ar}\left[U\right] \geq \frac{\left[\mathbb{E}\left[UV\right]\right]^2}{\mathbb{E}\left[V^2\right]}$$

One must show that $\mathbb{E}\left[V\right] = 0$; this is done next for the continuous case only as the discrete case is analogous. **(Continues...)**

Cramér-Rao bound: general, univariate (3/4)

Theorem 4

Proof.
(Continued.)

$$\begin{aligned}
\mathbb{E}[V] &= \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_N;\theta)\right] \\
&= \mathbb{E}\left[\frac{1}{f(X_1,\ldots,X_N;\theta)}\frac{\partial}{\partial\theta}f(X_1,\ldots,X_N;\theta)\right] \\
&= \int_{\mathbb{X}_1}\cdots\int_{\mathbb{X}_N}\frac{\partial}{\partial\theta}f(x_1,\ldots,x_N;\theta)\,dx_1\ldots dx_N \\
&= \frac{\partial}{\partial\theta}\int_{\mathbb{X}_1}\cdots\int_{\mathbb{X}_N}f(x_1,\ldots,x_N;\theta)\,dx_1\ldots dx_N \\
&= \frac{\partial}{\partial\theta}\cdot 1 \\
&= 0
\end{aligned}$$

where the second line applies the chain rule while the fourth line follows from hypothesis. **(Continues...)**

Cramér-Rao bound: general, univariate (4/4)

Theorem 4

Proof.

(Continued.) In a similar vein, note that:

$$\begin{aligned}
\mathbb{E}\left[UV\right] &= \mathbb{E}\left[\widehat{\theta}\left(X_1,\ldots,X_N\right)\cdot\frac{\partial}{\partial\theta}\log f\left(X_1,\ldots,X_N;\theta\right)\right] \\
&= \mathbb{E}\left[\frac{\widehat{\theta}\left(X_1,\ldots,X_N\right)}{f\left(X_1,\ldots,X_N;\theta\right)}\frac{\partial}{\partial\theta}f\left(X_1,\ldots,X_N;\theta\right)\right] \\
&= \int_{\mathbb{X}_1}\ldots\int_{\mathbb{X}_N}\widehat{\theta}\left(x_1,\ldots,x_N\right)\cdot\frac{\partial}{\partial\theta}f\left(x_1,\ldots,x_N;\theta\right)dx_1\ldots dx_N \\
&= \frac{\partial}{\partial\theta}\int_{\mathbb{X}_1}\ldots\int_{\mathbb{X}_N}\widehat{\theta}\left(x_1,\ldots,x_N\right)\cdot f\left(x_1,\ldots,x_N;\theta\right)dx_1\ldots dx_N \\
&= \frac{\partial}{\partial\theta}\,\mathbb{E}\left[\widehat{\theta}\right]
\end{aligned}$$

and collecting terms, the postulated result is obtained. □

Cramér-Rao bound: general, multivariate (1/4)

Theorem 4

The Bound in the Multivariate case. *Consider a sample drawn from a list of random vectors* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *with joint p.m.f. or p.d.f. written as* $f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)$ *with shorthand notation. Also consider a vector of parameters of interest* θ *having length* K*, as well as some estimator* $\widehat{\theta} = \widehat{\theta}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *for* θ*, such that its variance is finite and – in the continuous case – that the differentiation operation with respect to* θ *can pass through the expectation operator as shown below.*

$$\begin{aligned}\frac{\partial}{\partial \theta^{\mathrm{T}}} \mathbb{E}\left[\widehat{\theta}\right] &= \\ &= \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \frac{\partial}{\partial \theta^{\mathrm{T}}} \widehat{\theta}(\mathbf{x}_1, \ldots, \mathbf{x}_N) \cdot f(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta) \, \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_N\end{aligned}$$

In this setting, the variance of $\widehat{\theta}$ *must satisfy the following inequality.*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] - \left[\frac{\partial}{\partial \theta^{\mathrm{T}}} \mathbb{E}\left[\widehat{\theta}\right]\right] \left[\boldsymbol{I}_N(\theta)\right]^{-1} \left[\frac{\partial}{\partial \theta^{\mathrm{T}}} \mathbb{E}\left[\widehat{\theta}\right]\right]^{\mathrm{T}} \geq \mathbf{0}$$

(Continues...)

Cramér-Rao bound: general, multivariate (2/4)

Theorem 4

(Continued.) *Here the inequality is interpreted in the sense that the $K \times K$ matrix on the left hand side of the inequality sign is positive semi-definite; in the expression, $\boldsymbol{I}_N(\theta)$ is defined as follows.*

$$\boldsymbol{I}_N(\theta) \equiv \mathbb{E}\left[\left(\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \theta)\right)\left(\frac{\partial}{\partial \theta} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \theta)\right)^{\mathrm{T}}\right]$$

Proof.

In analogy with the univariate case, define the random vectors:

$$\boldsymbol{u} = \widehat{\theta}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$$

$$\boldsymbol{v} = \frac{\partial}{\partial \theta} \log f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \theta)$$

that are related as follows by the properties of multivariate moments.

$$\mathbb{V}\mathrm{ar}[\boldsymbol{u}] - [\mathbb{C}\mathrm{ov}[\boldsymbol{u}, \boldsymbol{v}]][\mathbb{V}\mathrm{ar}[\boldsymbol{v}]]^{-1}[\mathbb{C}\mathrm{ov}[\boldsymbol{u}, \boldsymbol{v}]]^{\mathrm{T}} \geq \mathbf{0}$$

(Continues...)

Cramér-Rao bound: general, multivariate (3/4)

Theorem 4

Proof.

(Continued.) If $\mathbb{E}[\boldsymbol{v}] = \mathbf{0}$, the previous expression simplifies as:

$$\mathbb{V}\mathrm{ar}[\boldsymbol{u}] - \left[\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right]\right]\left[\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\right]\right]^{-1}\left[\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right]\right]^{\mathrm{T}} \geq \mathbf{0}$$

where $\mathbb{E}\left[\boldsymbol{v}\boldsymbol{v}^{\mathrm{T}}\right] = \boldsymbol{I}_N(\boldsymbol{\theta})$. Thus the main inequality of interest follows through if, in addition, the following can be shown.

$$\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right] = \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\,\mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]$$

Note that if the above relationship is proved, $\mathbb{E}[\boldsymbol{v}] = \mathbf{0}$ would follow easily by replacing $\boldsymbol{u}$ with the unit vector $\boldsymbol{\iota}_K = (1, \ldots, 1)^{\mathrm{T}}$ that has the same length K as $\boldsymbol{\theta}$.

To avoid repeating similar arguments as it was done (for illustrative purposes and for completeness) in the univariate case, only the more complex case involving $\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right]$ is shown here. **(Continues...)**

Cramér-Rao bound: general, multivariate (4/4)

Theorem 4

Proof.
(Continued.)

$$
\begin{aligned}
\mathbb{E}\left[\boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}\right] = & \\
&= \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \cdot \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \log f\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right)\right] \\
&= \mathbb{E}\left[\frac{\widehat{\boldsymbol{\theta}}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)}{f\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right)} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}\right)\right] \\
&= \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \widehat{\boldsymbol{\theta}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) \cdot \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_N \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \int_{\mathbb{X}_1} \ldots \int_{\mathbb{X}_N} \widehat{\boldsymbol{\theta}}\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right) \cdot f\left(\mathbf{x}_1, \ldots, \mathbf{x}_N; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x}_1 \ldots \mathrm{d}\mathbf{x}_N \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]
\end{aligned}
$$

Thus, $\mathbb{E}\left[\boldsymbol{v}\right] = \mathbf{0}$ as well as the main result both follow. □

Interpreting the Cramér-Rao bound

- The expressions of the Cramér-Rao inequalities surely look formidable, and it is worthwhile to analyze them carefully.
- In the univariate case, the main determinant of the bound is the denominator, called **Fisher information number**:

$$\mathcal{I}_N(\theta) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_N;\theta)\right)^2\right]$$

  this number is a different function of the parameter θ for each possible distribution that generates the data.
- The name "information" derives from the interpretation of $\mathcal{I}_N(\theta)$ as the overall "amount of knowledge" that a certain distribution $f(X_1,\ldots,X_N;\theta)$ can provide on θ: the higher the number, the lower the bound on the variance of θ.
- Its multivariate analogue is clearly matrix $\boldsymbol{I}_N(\boldsymbol{\theta})$, which is unsurprisingly called **Fisher information matrix**.

Simplifying the Cramér-Rao bound

While the expressions may still appear difficult to operationalize in practice, they can be simplified in a number of different ways.

1. If the estimators are *unbiased*, the two terms:

$$\frac{\partial}{\partial \theta}\,\mathbb{E}\left[\widehat{\theta}\right] \quad \text{and} \quad \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\,\mathbb{E}\left[\widehat{\boldsymbol{\theta}}\right]$$

   clearly reduce to 1 and to the identity matrix $\mathbf{I}$ respectively.

2. If the sample is *random*, both the information number and the information matrix can be simplified as it is expressed in the theorem discussed next.

3. Additional simplifications are possible under fairly general conditions that are detailed later.

Cramér-Rao bound in random samples (1/4)

Theorem 5

Cramér-Rao Inequality (random sample) – univariate case. *In the (univariate) setup of Theorem 4, if the sample is random the inequality can be expressed as follows:*

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] \geq \frac{\left(\frac{\partial}{\partial\theta}\,\mathbb{E}\left[\widehat{\theta}\right]\right)^2}{N\cdot\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X;\theta\right)\right)^2\right]}$$

where $f_X\left(x;\theta\right)$ is the p.m.f. or p.d.f. that generates the sample.

Multivariate case. *In the (multivariate) setup of Theorem 4, if the sample is random the inequality is based on the following version of the information matrix:*

$$\boldsymbol{I}_N\left(\boldsymbol{\theta}\right) = N\cdot\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right]$$

where $f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)$ is the p.m.f. or p.d.f. that generates the sample.

Cramér-Rao bound in random samples (2/4)

Theorem 5

Proof.

In the univariate case, observe that:

$$\begin{aligned}
\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f\left(X_1,\ldots,X_N;\theta\right)\right)^2\right] &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{N} f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)^2\right] \\
&= N\cdot\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X;\theta\right)\right)^2\right]
\end{aligned}$$

(Continues...)

Cramér-Rao bound in random samples (3/4)

Theorem 5

Proof.

(Continued.) In the previous display, the first line follows from i.i.d. sampling, the second line is a simple manipulation, both the third and fourth lines are based on the linear properties of expectations and independence, since terms of the following form for $i \neq j$:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X_i;\theta\right)\right)\left(\frac{\partial}{\partial\theta}\log f_X\left(X_j;\theta\right)\right)\right] = 0$$

must be equal to the product of the respective means and thus to zero, while the fifth line follows from identically distributed observations.

The derivation in the multivariate case is analogous: the central step is again the one that invokes independence across observations.

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_i;\theta\right)\right)\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}\left(\boldsymbol{x}_j;\theta\right)\right)^{\mathrm{T}}\right] = \mathbf{0}$$

The multivariate derivation follows. **(Continues...)**

Cramér-Rao bound in random samples (4/4)

Theorem 5

Proof.

(Continued.)

$$
\begin{aligned}
\boldsymbol{I}_N(\theta) &= \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{N} f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)\left(\frac{\partial}{\partial\theta}\log\prod_{i=1}^{N} f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)\left(\sum_{i=1}^{N}\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N}\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)^{\mathrm{T}}\right] \\
&= \sum_{i=1}^{N}\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x}_i;\theta)\right)^{\mathrm{T}}\right] \\
&= N\cdot\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x};\theta)\right)\left(\frac{\partial}{\partial\theta}\log f_{\boldsymbol{x}}(\boldsymbol{x};\theta)\right)^{\mathrm{T}}\right]
\end{aligned}
$$

This completes the proof. □

Cramér-Rao bound: more simplifications (1/3)

- Additional simplifications are possible under the hypothesis that the derivatives taken with respect to the parameters of interest can pass through the expectation operator *twice*.

- In the univariate case this implies that:

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f_X\left(X;\theta\right)\right)^2\right] = -\,\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X\left(X;\theta\right)\right]$$

- . . . and in the more general multivariate case the analogous result is known as the **information matrix equality**.

$$\mathbb{E}\left[\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right)^{\mathrm{T}}\right] =$$

$$= -\,\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]$$

Cramér-Rao bound: more simplifications (2/3)

The proof is given already in the general multivariate case.

$$
\begin{aligned}
\mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{\partial}{\partial \boldsymbol{\theta}} 1 \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathbb{X}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \int_{\mathbb{X}} \frac{\partial f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \mathrm{d}\mathbf{x} \\
&= \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \int_{\mathbb{X}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \frac{\partial}{\partial \boldsymbol{\theta}^{\mathrm{T}}} \left[\frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)\right] \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \left[\frac{\partial^2 \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) + \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \frac{\partial f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right] \mathrm{d}\mathbf{x} \\
&= \int_{\mathbb{X}} \frac{\partial^2 \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x} + \\
&\quad + \int_{\mathbb{X}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} \frac{\partial \log f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}^{\mathrm{T}}} f_{\boldsymbol{x}}\left(\mathbf{x}; \boldsymbol{\theta}\right) \mathrm{d}\mathbf{x}
\end{aligned}
$$

Cramér-Rao bound: more simplifications (3/3)

Therefore, if all the occasions for simplification apply:

1. unbiasedness of the estimator(s);
2. random (i.i.d.) samples;
3. second derivatives able to pass through expectations;

it follows that the Cramér-Rao Inequality can be written in the univariate case as:

$$\mathbb{V}\mathrm{ar}\left[\widehat{\theta}\right] \geq -\frac{1}{N}\left\{\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\log f_X\left(X;\theta\right)\right]\right\}^{-1}$$

and in the multivariate case as follows.

$$\mathbb{V}\mathrm{ar}\left[\widehat{\boldsymbol{\theta}}\right] + \frac{1}{N}\left\{\mathbb{E}\left[\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}}\log f_{\boldsymbol{x}}\left(\boldsymbol{x};\boldsymbol{\theta}\right)\right]\right\}^{-1} \geq \mathbf{0}$$

Example: the Poisson Cramér-Rao bound (1/3)

- Consider a random sample where all the observations are drawn from some Poisson distribution.
- Recall that if $X \sim \text{Pois}(\lambda)$, $\mathbb{E}[X] = \lambda$ and $\mathbb{V}\text{ar}[X] = \lambda$.
- Hence, both $\bar{X}$ and S^2 are *unbiased* estimators of λ!
- A natural question is: which of the two estimators is better according to the MSE criterion? It is necessary to calculate their variance.
- For the sample mean's case this is simple: $\mathbb{V}\text{ar}\left[\bar{X}\right] = \lambda/N$.
- However, calculating the *variance of the sample variance* is not as straightforward.

Example: the Poisson Cramér-Rao bound (2/3)

One can calculate the information number directly as follows.

$$
\begin{aligned}
\mathcal{I}_N(\lambda) &= N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\log\frac{\exp(-\lambda)\cdot\lambda^X}{X!}\right)^2\right] \\
&= N \cdot \mathbb{E}\left[\left(\frac{\partial}{\partial\lambda}\left(-\lambda + X\log(\lambda) - \log(X!)\right)\right)^2\right] \\
&= N \cdot \mathbb{E}\left[\left(-1 + \frac{X}{\lambda}\right)^2\right] \\
&= N\left(1 - \frac{2}{\lambda}\cdot\mathbb{E}[X] + \frac{1}{\lambda^2}\cdot\mathbb{E}\left[X^2\right]\right) \\
&= N\left(\frac{\lambda+\lambda^2}{\lambda^2} - 1\right) \\
&= \frac{N}{\lambda}
\end{aligned}
$$

Example: the Poisson Cramér-Rao bound (3/3)

Alternatively, the information number can be calculated via the second derivative.

$$\begin{aligned}
\mathcal{I}_N(\lambda) &= -N \cdot \mathbb{E}\left[\frac{\partial^2}{\partial \lambda^2} \log \frac{\exp(-\lambda) \cdot \lambda^X}{X!}\right] \\
&= -N \cdot \mathbb{E}\left[\frac{\partial}{\partial \lambda}\left(-1 + \frac{X}{\lambda}\right)\right] \\
&= -N \cdot \mathbb{E}\left[-\frac{X}{\lambda^2}\right] \\
&= \frac{N}{\lambda}
\end{aligned}$$

Thus, $\bar{X}$ as an estimator of λ hits the Cramér-Rao bound, and the question can be resolved in its favor even without knowing the variance of S^2!

Example: the normal Cramér-Rao bound (1/3)

The normal distribution has two parameters: so, the Cramér-Rao bound must be evaluated in the multivariate setup. In this case it is more convenient to work with the Hessian matrix.

$$\boldsymbol{I}_N\left(\mu,\sigma^2\right) =$$
$$= -N\cdot\mathbb{E}\begin{bmatrix} \frac{\partial^2}{\partial\mu^2}\log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] & \frac{\partial^2}{\partial\mu\partial\sigma^2}\log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] \\ \frac{\partial^2}{\partial\sigma^2\partial\mu}\log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] & \frac{\partial^2}{\partial(\sigma^2)^2}\log\left[\frac{1}{\sigma}\phi\left(\frac{X-\mu}{\sigma}\right)\right] \end{bmatrix}$$

Here $\phi(\cdot)$ denotes as usual the p.d.f. of the standard normal. The matrix can be evaluated taking steps not unlike those in the MLE analysis of the normal distribution.

$$\boldsymbol{I}_N\left(\mu,\sigma^2\right) = -N\,\mathbb{E}\begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{X-\mu}{\sigma^4} \\ -\frac{X-\mu}{\sigma^4} & \frac{1}{2\sigma^4}-\frac{(X-\mu)^2}{\sigma^6} \end{bmatrix} = \begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}$$

Example: the normal Cramér-Rao bound (2/3)

It is obvious that $\bar{X}$ is an unbiased estimator and its variance $\mathbb{V}\mathrm{ar}\left[\bar{X}\right] = \sigma^2/N$ attains the Cramér-Rao bound.

As far as the estimation of parameter σ^2 is concerned instead, the estimator S^2 is unbiased. While its variance is:

$$\mathbb{V}\mathrm{ar}\left[S^2\right] = \frac{\sigma^4}{(N-1)^2}\,\mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] = \frac{2\sigma^4}{N-1}$$

it does not attain the Cramér-Rao bound, which is calculated as $2\sigma^4/N$ for unbiased estimators of σ^2.

Consider the rescaled and biased estimator of the scale parameter $\widehat{\sigma}^2 = \frac{N-1}{N}S^2$. Its variance is as follows.

$$\mathbb{V}\mathrm{ar}\left[\frac{N-1}{N}S^2\right] = \frac{\sigma^4}{N^2}\,\mathbb{V}\mathrm{ar}\left[(N-1)\,\frac{S^2}{\sigma^2}\right] = \frac{2\,(N-1)\,\sigma^4}{N^2}$$

Example: the normal Cramér-Rao bound (3/3)

However, the estimator $\widehat{\sigma}^2 = \frac{N-1}{N} S^2$ is biased, and thus the bias needs to enter the calculation of the Cramér-Rao bound.

By calling $\mathcal{I}_N(\sigma^2)$ the bottom-right element of the information matrix $\boldsymbol{I}_N(\mu, \sigma^2)$, the bound is expressed as:

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\frac{N-1}{N} S^2\right] &\geq \frac{1}{\mathcal{I}_N(\sigma^2)} \left(\frac{\partial}{\partial \sigma^2}\, \mathbb{E}\left[\frac{N-1}{N} S^2\right]\right)^2 \\
&= \frac{2\sigma^4}{N} \left[\frac{\partial}{\partial \sigma^2} \left(\frac{N-1}{N} \sigma^2\right)\right]^2 \\
&= \frac{2\,(N-1)^2\,\sigma^4}{N^3}
\end{aligned}$$

and not even in this case it is attained.

For both estimators, the actual value of the Cramér-Rao bound is equal to $\frac{N-1}{N}$ times their effective variance.

Attainment of the Cramér-Rao Bound (1/2)

A quite agile theorem helps determine under what conditions can an unbiased estimator attain the Cramér-Rao Bound.

Theorem 6

Attainment of the Cramér-Rao Bound – univariate case. *In the (univariate) setup of Theorem 4, if* $\widehat{\theta}$ *is an unbiased estimator of* θ*, it attains the Cramér-Rao bound if and and only if:*

$$a_N(\theta)\left[\widehat{\theta}-\theta\right]=\frac{\partial}{\partial\theta}\log f_{X_1,\ldots,X_N}(x_1,\ldots,x_N;\theta)$$

for some function $a_N(\theta)$ *of the parameter.*

Proof.

Recall the proof of Theorem 4 as well as the properties of covariances and correlations: the equality is attained only if U (the estimator) is a linear function of V (the derivative of the "logarithmic" joint p.m.f. or p.d.f. of the sample, i.e. the log-likelihood function).

By the Cauchy-Schwarz Inequality this observation can be phrased as $a(U-\mathbb{E}[U])=V$. As a can be a function of θ, write it as $a_N(\theta)$. $\square$

Attainment of the Cramér-Rao Bound (2/2)

The result easily extends to the multivariate setup.

Theorem 6

Multivariate case. *In the (multivariate) setup of Theorem 4, if* $\widehat{\boldsymbol{\theta}}$ *is an unbiased estimator of* $\boldsymbol{\theta}$*, it attains the Cramér-Rao bound if and and only if:*

$$\mathbf{A}_N(\boldsymbol{\theta})\left[\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\right]=\frac{\partial}{\partial\boldsymbol{\theta}}\log f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta})$$

for some $K\times K$ *matrix* $\mathbf{A}_N(\boldsymbol{\theta})$ *which is a function of the parameters.*

Proof.

Similarly to the univariate case, the equality is only attained if $\boldsymbol{u}$ is a linear function of $\boldsymbol{v}$, i.e. $\mathbf{A}(\boldsymbol{u}-\mathbb{E}[\boldsymbol{u}])=\boldsymbol{v}$ where $\mathbf{A}=\mathbf{A}_N(\boldsymbol{\theta})$. □

Example: attainment of the normal bound

Consider again random sampling from the normal distribution. The derivative of the joint p.d.f. corresponds to the MLE First Order Conditions. Write them according to Theorem 6.

$$\begin{bmatrix} \sum_{i=1}^{N} \frac{(x_i - \mu)}{\sigma^2} \\ \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{2\sigma^4} - \frac{N}{2\sigma^2} \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{N}{\sigma^2} & 0 \\ 0 & \frac{N}{2\sigma^4} \end{bmatrix}}_{=\mathbf{A}_N(\mu,\sigma^2)} \begin{bmatrix} \frac{1}{N}\sum_{i=1}^{N} x_i - \mu \\ \sum_{i=1}^{N} \frac{(x_i - \mu)^2}{N} - \sigma^2 \end{bmatrix}$$

This decomposition not only shows again that $\bar{X}$ is an unbiased estimator of μ which attains the bound: it also reveals that the *only* unbiased estimator of σ^2 that attains the bound is:

$$\tilde{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N} (X_i - \mu)^2 .$$

This estimator is usually *unfeasible* since it requires the unlikely ex-ante perfect knowledge of the location parameter μ.

Tests of hypotheses

- In the practice of statistics, estimates are routinely used to assess some ex ante **hypotheses** about the parameters.
- The methods by which these evaluations are performed fall under the name of **tests of hypotheses**.
- To conduct such tests, researchers first formulate some **null hypothesis** H_0, generally written as:

$$\mathrm{H}_0 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0$$

  where $\boldsymbol{\theta}$ are the parameters of interest, and $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$ is the subset of the parameter space allowed by H_0.

- Conversely, the **alternative hypothesis** H_1 is a statement that negates H_0:

$$\mathrm{H}_1 : \ \boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$$

  where $\boldsymbol{\Theta}_0^c$ is the complement of $\boldsymbol{\Theta}_0$ in the parameter space.

Example: test on the normal mean

The simplest case of an hypothesis test is that about the value of a single parameter, say the location parameter of the normal distribution. In this case, H_0 and H_1 read respectively as:

$$\mathrm{H}_0 : \ \mu = C \qquad\qquad \mathrm{H}_1 : \ \mu \neq C$$

where $|C| < \infty$ is a finite value. This is a typical instance of a **two-sided test**.

In a slightly more nuanced case, H_0 and H_1 are represented by two complementary inequalities. For example, if $\boldsymbol{\Theta} = \mathbb{R}$:

$$\mathrm{H}_0 : \ \mu \geq C \qquad\qquad \mathrm{H}_1 : \ \mu < C$$

or vice versa. This is instead the case of a **one-sided test**.

A common scenario is that of $C = 0$; in this case, the one-sided test is a test about the sign of the parameter.

Example: test on the regression slope

A typical example of test is the one about the slope parameter of the linear regression model. In this case, the two hypotheses read, respectively:

$$\mathrm{H}_0 : \ \beta_1 = C \qquad\qquad \mathrm{H}_1 : \ \beta_1 \neq C$$

for the **two-sided test**, and:

$$\mathrm{H}_0 : \ \beta_1 \geq C \qquad\qquad \mathrm{H}_1 : \ \beta_1 < C$$

or vice versa for the **one-sided test**.

In this specific case, the test for $C = 0$ is about the relevance of the exogenous variable X_i as an **explanation** of the endogenous variable Y_i (or as a "predictor" as it is sometimes said).

Example: test on the exponential's parameter

Lastly, the parameter space for parameter λ of the exponential distribution is the set of positive values. Thus, the test with:

$$\mathrm{H}_0 : \ 0 < \lambda \leq C \qquad\qquad \mathrm{H}_1 : \ \lambda > C$$

is a proper formulation for testing whether the waiting time of some phenomenon of interest that can be modeled through the exponential distribution, is lower or higher than some positive number $C > 0$.

Example: test on the multivariate normal mean

Suppose that one is analyzing a phenomenon being modeled via the bivariate normal distribution, and is wondering whether the means of the two random variables involved (name them X and Y) are equal. In this case, the two hypotheses are formulated as:

$$\mathrm{H}_0 : \ \mu_X - \mu_Y = 0 \qquad\qquad \mathrm{H}_1 : \ \mu_X - \mu_Y \neq 0$$

which is a well-defined restriction in the parameter space.

Next, consider the multivariate normal case, where it might be interesting to verify whether all location parameters are equal to some specified value. Here, the hypotheses are:

$$\mathrm{H}_0 : \ \mu_k = C_k \qquad\qquad \mathrm{H}_1 : \ \mu_k \neq C_k$$

for $k = 1, \ldots, K$, where the restricted set $\boldsymbol{\Theta}_0$ is a point in $\mathbb{R}^K$.

Test on normal variances

The parameter σ^2 of the normal distribution can also be tested, knowing that the parameter space shall be restricted to positive numbers. Accordingly, a test can be formulated as:

$$\mathrm{H}_0 : \ 0 < \sigma^2 \leq C \qquad\qquad \mathrm{H}_1 : \ \sigma^2 > C$$

where, again, the constant $C > 0$ must be positive.

And if interest falls on the relationship between the variances of two independent normal random variables X and Y? A test for this environment can be formulated as:

$$\mathrm{H}_0 : \ \frac{\sigma_X^2}{\sigma_Y^2} \leq C \qquad\qquad \mathrm{H}_1 : \ \frac{\sigma_X^2}{\sigma_Y^2} > C$$

where σ_X^2 and σ_Y^2 are the variances of X and Y, respectively; it is typical to construct tests on the *equality of the two variances* for $C = 1$. Naturally, two-sided tests about specific values of the ratio are perfectly possible.

The procedure of hypothesis tests

A test of hypothesis is conventionally conducted as follows.

1. The researcher establishes the two alternative hypotheses, H_0 and H_1.

2. The researcher identifies *ex-ante* those values of the sample realizations that are associated with acceptance of the null hypothesis, and rejection of the alternative hypothesis: this is called the **acceptance region**.

   The research also identifies those values that are associated with acceptance of the alternative hypothesis, and rejection of the null hypothesis: the **rejection region**. The two sets must be complementary in the support of the sample.

3. The researcher examines the sample, thus taking a decision according to the criteria established at point 2. above.

Test statistics

The process of defining the acceptance and rejection regions of a test can be simplified through appropriate *statistics* set for this purpose.

Definition 7

Test statistic. In the context of some given test of hypothesis, a *test* statistic $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a statistic having support $\mathbb{T}$ and whose sample realization value is written as $t = T(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ which is such that, for a dichotomous partition of the support $\mathbb{T}_0 \cup \mathbb{T}_0^c = \mathbb{T}$, the test is resolved as follows.

$$t \in \begin{cases} \mathbb{T}_0 & \Rightarrow \mathrm{H}_0 \text{ is accepted, and } \mathrm{H}_1 \text{ is rejected} \\ \mathbb{T}_0^c & \Rightarrow \mathrm{H}_1 \text{ is accepted, and } \mathrm{H}_0 \text{ is rejected} \end{cases}$$

Here, $\mathbb{T}_0$ and $\mathbb{T}_0^c$ are respectively called the **acceptance region** and the **rejection region** associated with the test statistic.

The sampling distribution of a test statistic, and thus the probability that it falls in either $\mathbb{T}_0$ or $\mathbb{T}_0^c$, shall vary alongside $\boldsymbol{\theta}$. This way, a test statistic is informative about the parameters. The choice of $\mathbb{T}_0$ or $\mathbb{T}_0^c$ is ultimately arbitrary, but guided by conventions informed by theory.

Type I and Type II errors

Definition 8

Type I Error. In the framework of a test of hypotheses, the *type I* error is the circumstance whereby the null hypothesis is *rejected*, and the alternative hypothesis is *accepted*, while the null hypothesis is *true*.

Definition 9

Type II Error. In the framework of a test of hypotheses, the *type II* error is the circumstance whereby the null hypothesis is *accepted*, and the alternative hypothesis is *rejected*, while the null hypothesis is *false*.

The various outcomes of a test are commonly schematized as follows.

	$t \in \mathbb{T}_0$	$t \in \mathbb{T}_0^c$
$\theta \in \Theta_0$	Correct decision	Type I error
$\theta \in \Theta_0^c$	Type II error	Correct decision

Ideally, there are no errors: but this is impossible. Hence, a *trade-off* between Type I and Type II errors arises.

Power, level and size of a tests

Definition 10

Power Function. The probability that the test statistic falls in the rejection region, as a function of $\boldsymbol{\theta}$, is the *power* function of a test.

$$\mathbb{P}_T(\boldsymbol{\theta}) = \mathbb{P}(t \in \mathbb{T}_0^c; \boldsymbol{\theta}) = 1 - \mathbb{P}(t \in \mathbb{T}_0; \boldsymbol{\theta})$$

Clearly, a power function expresses the probability to commit a Type I error if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, and it equals one minus the probability to commit a Type II error if $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0^c$.

This notion, in turn, is instrumental in the following definitions.

Definition 11

Level of a test. Given a number $\alpha \in [0, 1]$, a test with power function $\mathbb{P}_T(\boldsymbol{\theta})$ has *confidence level* α if $\mathbb{P}_T(\boldsymbol{\theta}) \leq \alpha$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$.

Definition 12

Size of a test. Given a number $\alpha \in [0, 1]$, a test with power function $\mathbb{P}_T(\boldsymbol{\theta})$ has *size* α if $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} \mathbb{P}_T(\boldsymbol{\theta}) = \alpha$.

Size, level and testing conventions

- Given the Type I vs. Type II error trade-off, the convention in statistical practice is to "fix" the probability of a Type I error to be reasonably small.
- Specifically, the *confidence level* α of tests is conventionally fixed at values such as 0.1, 0.05, and 0.01.
- The smaller is α, the more credible is the outcome of a test when H_0 is rejected.
- However, trying to further shrink the probability of Type I errors may lead to a higher probability of Type II errors.
- Therefore, one should select those tests having a maximum probability of rejecting H_0 when it is true that is exactly α: the *size* of the test.
- In typical practical applications, the conceptual distinction between *level* and *size* is of little consequence.

Example: level and size in tests for the mean

- Consider a test about the mean of a certain distribution of interest, call it μ.
- In the two-sided case, $\boldsymbol{\Theta}_0 = \{C\}$ and $\boldsymbol{\Theta}_0^c = \mathbb{R} \setminus \{C\}$, hence there is no practical distinction between level and size.
- In the one-sided case, however, if H_0 states that the mean is smaller or equal than a constant C, $\boldsymbol{\Theta}_0 = (-\infty, C]$ and $\boldsymbol{\Theta}_0^c = (C, \infty)$, and vice versa.
- Therefore, for a fixed confidence level α there are different rejection probabilities for different values in $\boldsymbol{\Theta}_0$.
- In practice, typically, in such tests the maximum rejection probability is achieved at $\mu = C$.

The p-value

A statistic called **p-value** is typically reported among the outcomes of a test of hypothesis.

Definition 13

The p-value. In a test of hypothesis with given size α, a *p-value* is a statistic $P = P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, it is as follows.

$$\mathbb{P}(P(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \alpha) \leq \alpha$$

- Intuition: the smaller the p-value associated with a sample, the smaller is the probability to observe that sample when *the null hypothesis is true*.
- Hence, a smaller p-value is interpreted in terms of less favorable evidence in favor of the null hypothesis.
- This concept enables researchers to assess the outcomes of tests on a continuous (rather than dichotomous) scale.
- Usually, p-values are calculated as the probability to obtain test statistics that are *even less favorable to the null hypothesis* than the actual realization t.

One-sided test on the normal mean (1/5)

- Suppose that a researcher collects a random sample drawn from a normal distribution to conduct a test about μ.
- Suppose *for now* that the researcher knows parameter σ^2.
- Also let, *for now*, the test to be one-sided.

$$\mathrm{H}_0 : \mu \leq 0 \qquad\qquad \mathrm{H}_1 : \mu > 0$$

- In this setting, the *standardized* sample mean is a statistic which follows the standard normal distribution.
- Thus, a logical test protocol is to reject the null hypothesis if the *observed* standardized sample mean exceeds a certain **critical value**, call it z^*. The test would be resolved, for a given $\mu_0 \leq 0$, as follows.

$$\sqrt{N}\,\frac{\bar{X} - \mu_0}{\sigma} \begin{cases} \leq z^* & \Rightarrow \mu \leq 0 \\ > z^* & \Rightarrow \mu > 0 \end{cases}$$

One-sided test on the normal mean (2/5)

The is best conducted with $\mu_0 = 0$, for the following reasons.

1. Naturally, the probability to conduct a Type I error:

$$\mathbb{P}\left(\text{Type I error}\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - \mu_0}{\sigma} > z^* \,\middle|\, \text{H}_0 \text{ is true}\right) = \alpha$$

   depends on the actual value of μ_0 in the expression above. This is obviously maximized at $\mu_0 = 0$. The probability of rejecting the null hypothesis associated with that value of μ_0 is the *size* α of the test.

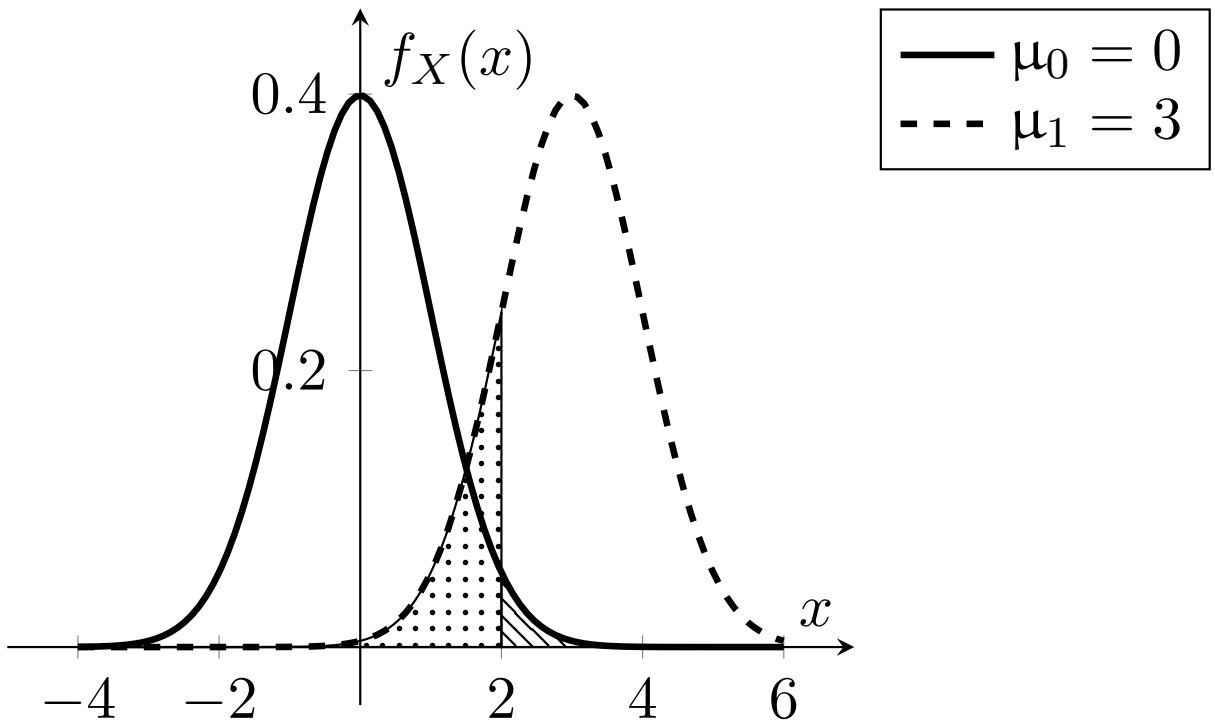
2. Similarly, the probability to conduct a Type II error:

$$\mathbb{P}\left(\text{Type II error}\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - \mu_1}{\sigma} \leq z^* \,\middle|\, \text{H}_1 \text{ is true}\right)$$

   is a function of the value of μ_1 *if the alternative hypothesis is true*. The researcher is ignorant about this value, but it is clear that regardless, the probability increases with z^*.

One-sided test on the normal mean (3/5)

- To illustrate, suppose that if H_1 is true it is $\mu_1 = 3$.
- The maximal Type I error probability attained under H_0 at $\mu_0 = 0$, as well as the Type II error probability for $\mu_1 = 3$, are graphically displayed below.
- For $z^* = 2$, these are the shaded (I) and dotted (II) areas.



One-sided test on the normal mean (4/5)

- Consider now the general case: $\mathrm{H}_0 : \mu \leq C$ and $\mathrm{H}_1 : \mu > C$.
- To use the standardized sample mean a test statistic with a given size α, one must solve the following equation in terms of the critical value z_α^*, where the subscript indicates size.

$$\mathbb{P}\left(\bar{X} > C + \frac{\sigma}{\sqrt{N}} z_\alpha^*\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - C}{\sigma} > z_\alpha^*\right) = \alpha$$

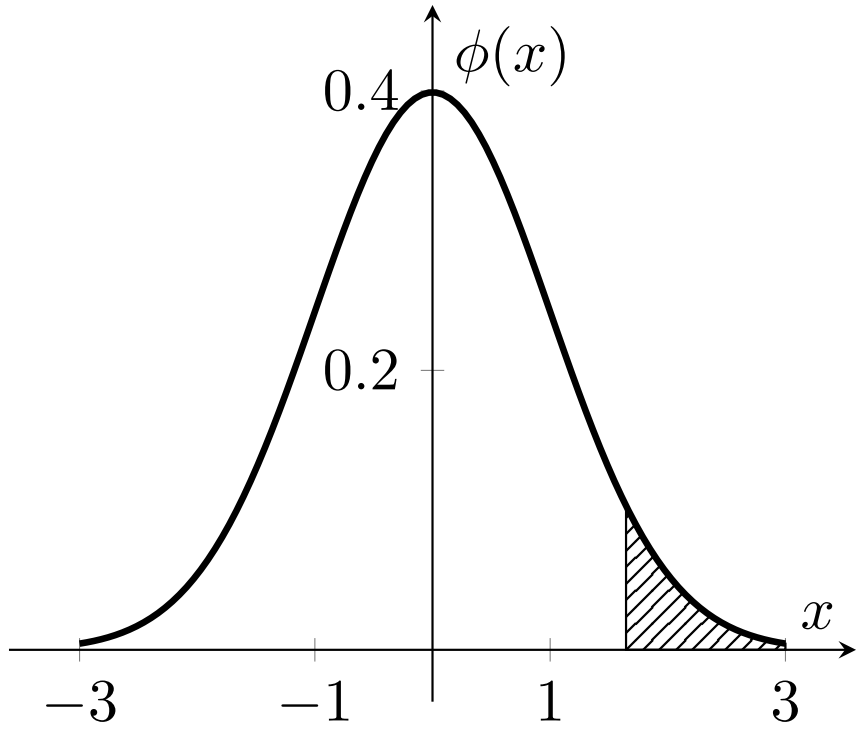
  This is evaluated against a standard normal c.d.f. $\Phi(z)$.

- This procedure implies that the null hypothesis is rejected if the observed realization of the sample mean is such that $\sqrt{N}\,(\bar{x} - C)\,/\sigma > z_\alpha^*$;
- The p-value is ultimately calculated (again evaluating it via the standard normal c.d.f.) as follows.

$$p(\bar{x}) = \mathbb{P}\left(\bar{X} \geq \bar{x}\right)$$

One-sided test on the normal mean (5/5)

- For $\alpha = 0.05$, the critical value is $z^*_{0.05} \approx 1.64$.
- The rejection region is represented by the shaded area in the figure below.



Two-sided test on the normal mean (1/3)

- Keep assuming σ^2; let however the test to be two-sided.

$$\mathrm{H}_0 : \mu = C \qquad\qquad \mathrm{H}_1 : \mu \neq C$$

- The researcher must now look for **two** *symmetric* **critical values**: $z^*_{\alpha/2} > 0$ and its mirror image $-z^*_{\alpha/2} < 0$.

- Intuitively, the researcher is agnostic about the sign of the deviation from C in case the alternative hypothesis is true.

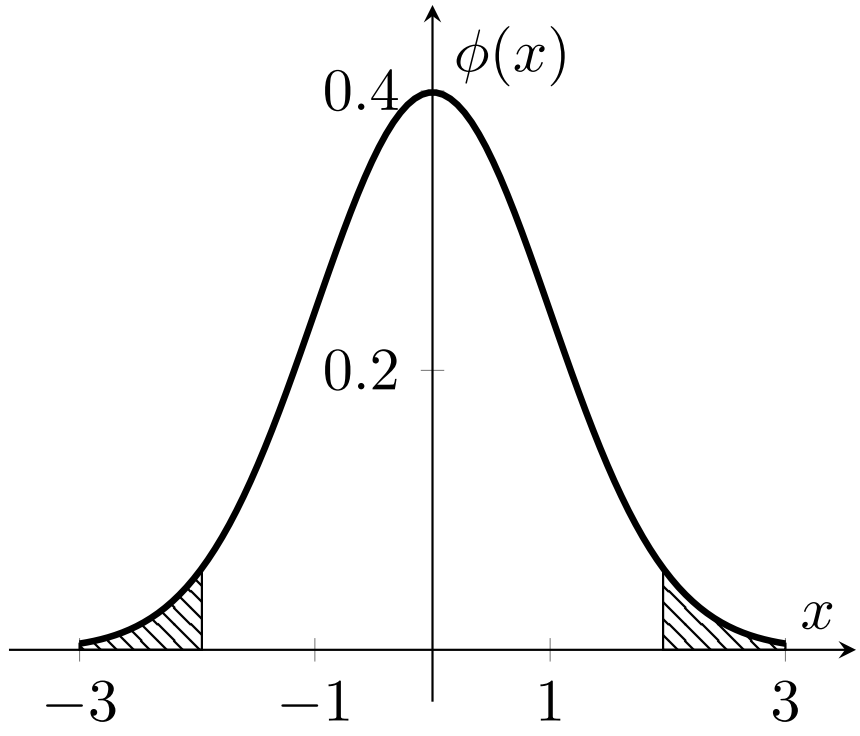
- Hence, given a level α the probabilities of *both* the Type I and the Type II errors are minimized when:

$$\mathbb{P}\left(\left|\bar{X} - C\right| > \frac{\sigma}{\sqrt{N}} z^*_{\alpha/2}\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\left|\bar{X} - C\right|}{\sigma} > z^*_{\alpha/2}\right) = \frac{\alpha}{2}$$

  and the null hypothesis is rejected if $\sqrt{N}\,|\bar{x} - C|\,/\sigma > z^*_{\alpha/2}$.

Two-sided test on the normal mean (2/3)

- For $\alpha = 0.05$, the critical value is $z^*_{0.025} \approx 1.96$.
- The rejection region is represented as the shaded area in the figure below. Notice the symmetry!



Two-sided test on the normal mean (3/3)

- Note how here, the p-value is calculated as *the sum of two symmetric probabilities.*

$$\begin{aligned} p\left(\bar{x}\right) &= \mathbb{P}\left(\bar{X} > \bar{x}\right) + \mathbb{P}\left(\bar{X} < -\bar{x}\right) \\ &= 2 \cdot \mathbb{P}\left(\bar{X} > |\bar{x}|\right) \\ &= 2 \cdot \mathbb{P}\left(\bar{X} > \bar{x}\right) \\ &= 2 \cdot \mathbb{P}\left(\bar{X} < -\bar{x}\right) \end{aligned}$$

- Two-sided tests about the mean of the normal distribution are perhaps the most common kinds of tests of hypotheses. A summary of critical values for two-sided tests is useful.

$$\begin{aligned} \alpha = 0.10 \quad &\Rightarrow \quad z^*_{0.050} \approx 1.64 \\ \alpha = 0.05 \quad &\Rightarrow \quad z^*_{0.025} \approx 1.96 \\ \alpha = 0.01 \quad &\Rightarrow \quad z^*_{0.005} \approx 2.33 \end{aligned}$$

Tests on the normal mean, unknown variance

- If σ^2 is unknown, the tests are based on the t-statistic and evaluated against the t-distribution with $N-1$ degrees for freedom (see Lecture 4).

- Conceptually though, little else changes. In one-sided tests, the critical value t^*_α is found as:

$$\mathbb{P}\left(\bar{X} > C + \frac{S}{\sqrt{N}} t^*_\alpha\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - C}{S} > t^*_\alpha\right) = \alpha$$

  and the test is rejected if $\sqrt{N}\,(\bar{x} - C)\,/s > t^*_\alpha$

- The p-value is calculated as the following function of both the *observed* sample mean and variance.

$$p\left(\bar{x}, s^2\right) = \mathbb{P}\left(\frac{\bar{X} - C}{S} > \frac{\bar{x} - C}{s}\right)$$

- The two-sided case is analogous.

Test on the normal variance (1/2)

- What if interest falls on the variance? Let the following null and alternative hypotheses.

$$\mathrm{H}_0 : \ 0 < \sigma^2 \leq C \qquad\qquad \mathrm{H}_1 : \ \sigma^2 > C$$

- In this case the test statistic is the rescaled sample variance $(N-1)\,S^2/C$; the critical value k^*_α for a test with size α is identified through the chi-squared distribution with $N-1$ degrees of freedom.

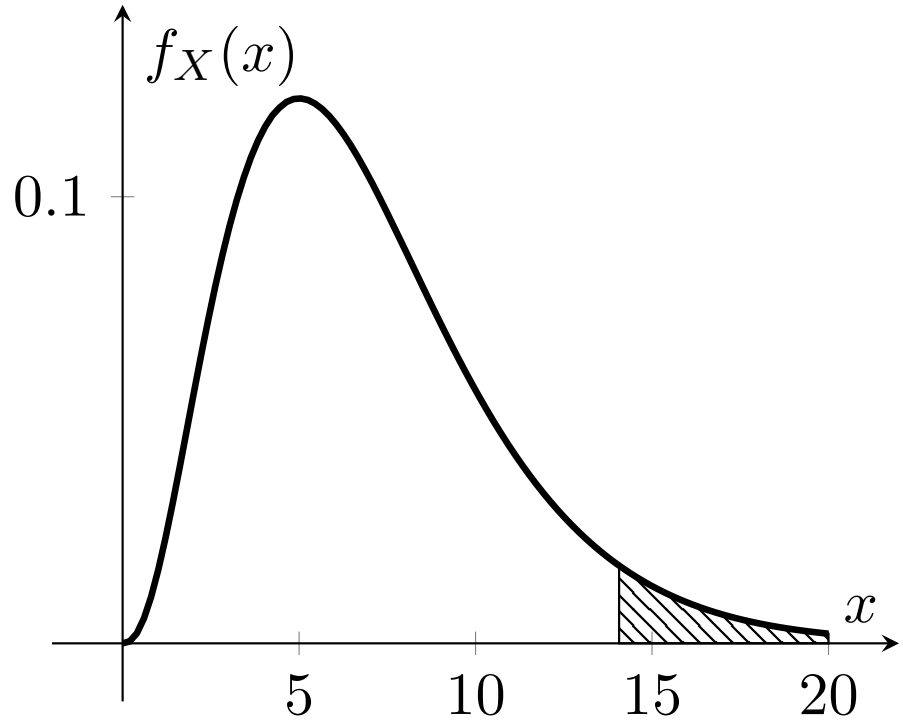
$$\mathbb{P}\left(S^2 > \frac{C}{N-1}\,k^*_\alpha\right) = \mathbb{P}\left((N-1)\,\frac{S^2}{C} > k^*_\alpha\right) = \alpha$$

- Hence, the null hypothesis is rejected if $(N-1)\,s^2/C > k^*_\alpha$, and the p-value is calculated as follows.

$$p\left(s^2\right) = \mathbb{P}\left(S^2 \geq s^2\right)$$

Test on the normal variance (2/2)

- See the figure for intuition: here $N = 8$, and $X \sim \chi^2_7$ is the random variable that models the rescaled sample variance.
- The higher the realization of $(N-1)\,S^2/C$, the less likely it is that H_0 is true.



Test for comparing two normal variances (1/2)

- Consider the comparison between two normal variances.

$$\mathrm{H}_0 : \ \frac{\sigma_X^2}{\sigma_Y^2} \leq C \qquad\qquad \mathrm{H}_1 : \ \frac{\sigma_X^2}{\sigma_Y^2} > C$$

- The relevant statistic here is the F-statistic (Lecture 4); the critical value k_α^* for a test of size α is therefore obtained by evaluating an F-distribution with paired $N_X - 1$ & $N_Y - 1$ degrees of freedom, as per the following expression.

$$\mathbb{P}\left(\frac{S_X^2}{S_Y^2} > C k_\alpha^*\right) = \mathbb{P}\left(\frac{S_X^2}{S_Y^2}\frac{1}{C} > k_\alpha^*\right) = \alpha$$

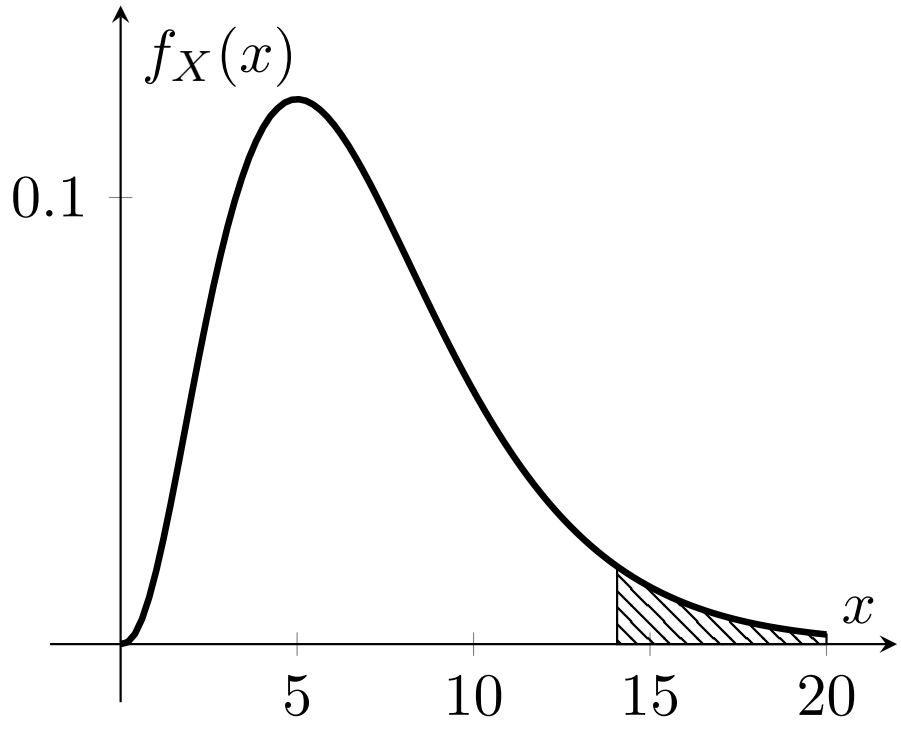
- Hence, the null hypothesis is rejected if $\left(s_X^2/s_Y^2\right)/C > k_\alpha^*$, and the p-value is calculated as follows.

$$p\left(s_X^2, s_Y^2\right) = \mathbb{P}\left(S_X^2/S_Y^2 > s_X^2/s_Y^2\right)$$

Test for comparing two normal variances (2/2)

- See the figure: in this case $N_X = N_Y = 12$ and $X \sim \mathcal{F}_{11,11}$ is the random variable that models the variance ratio.

- The higher the realization of the "empirical" variance ratio S_X^2/S_Y^2, the less likely H_0 is true.



Tests on the multivariate normal means (1/4)

- Consider some *composite* hypotheses about the means of K normally distributed random vector, for $k = 1, \ldots, K$.

$$\mathrm{H}_0 : \ \mu_k = C_k \qquad\qquad \mathrm{H}_1 : \ \mu_k \neq C_k$$

- This test is best expressed in vectorial form:

$$\mathrm{H}_0 : \ \boldsymbol{\mu} = \mathbf{c} \qquad\qquad \mathrm{H}_1 : \ \boldsymbol{\mu} \neq \mathbf{c}$$

  where $\boldsymbol{\mu}$ is the vector of means and $\mathbf{c} = (C_1, \ldots, C_K)^{\mathrm{T}}$.

- A straightforward test statistic in this environment is the u-statistic (Lecture 4):

$$\mathrm{u} = N \left(\bar{\boldsymbol{x}} - \mathbf{c}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{x}} - \mathbf{c}\right) \sim \chi^2_K$$

  it relates the sample mean $\boldsymbol{x}$ to the variance-covariance $\boldsymbol{\Sigma}$ of the random vector in the population.

Tests on the multivariate normal means (2/4)

- As in the univariate case however, $\boldsymbol{\Sigma}$ is generally unknown and must be replaced with its sample counterpart $\boldsymbol{S}$.

- The appropriate test statistic is thus Hotelling's t-squared:

$$\frac{N-K}{K\,(N-1)}\,\mathrm{t}^2 = \frac{(N-K)\,N}{K\,(N-1)}\,(\bar{\boldsymbol{x}}-\mathbf{c})^{\mathrm{T}}\,\boldsymbol{S}^{-1}\,(\bar{\boldsymbol{x}}-\mathbf{c})$$

  it follows the F-distribution with paired degrees of freedom K and $N-K$ (Lecture 4).

- Given a size α, the critical value k_α^* that leads to rejecting H_0 is determined as follows.

$$\mathbb{P}\left(\frac{N-K}{K\,(N-1)}\,\mathrm{t}^2 > k_\alpha^*\right) = \alpha$$

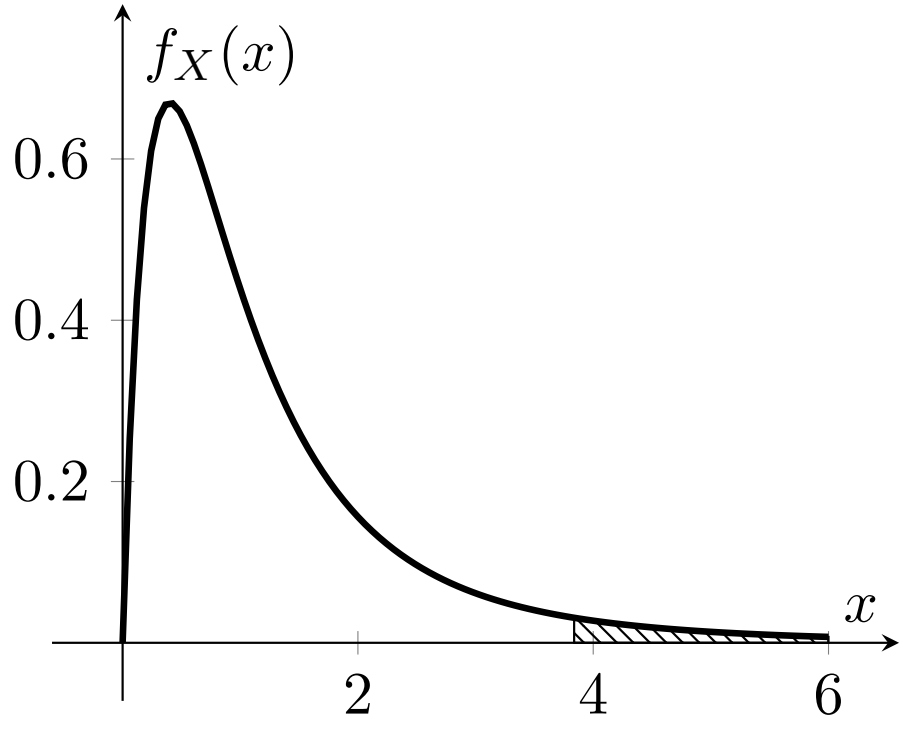
- The p-value here is calculated from the realized key sample statistic as follows.

$$p\,(\bar{\mathbf{x}},\mathbf{S}) = \mathbb{P}\left(\mathrm{t}^2 > N\,(\bar{\mathbf{x}}-\mathbf{c})^{\mathrm{T}}\,\mathbf{S}^{-1}\,(\bar{\mathbf{x}}-\mathbf{c})\right)$$

Tests on the multivariate normal means (3/4)

- See the figure again: here $N = 12$, $K = 4$ and $X \sim \mathcal{F}_{4,8}$ is the random variable that models Hotelling's t-squared.

- The higher the realization of Hotelling's t-squared, the less likely H_0. Note that negative deviations of $\bar{X}_k$ from C_k (for any k) contribute to the test statistic *positively*.



Tests on the multivariate normal means (4/4)

- The test's logic can extend further. Consider the following hypotheses for a bivariate normal random vector (X, Y).

$$\mathrm{H}_0 : \ \mu_X - \mu_Y = 0 \qquad \mathrm{H}_1 : \ \mu_X - \mu_Y \neq 0$$

- The appropriate test-statistic is:

$$\mathrm{t}^2 = \frac{N\left(\bar{X} - \bar{Y}\right)^2}{S_X^2 + S_Y^2 - 2S_{XY}} \sim \mathcal{F}_{1,N-1}$$

following the F-distribution with paired degrees of freedom 1 & $N-1$ (S_{XY} is the sample *covariance* between X & Y).

$$S_{XY} = \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$$

- This is equivalent to testing that the following transformed random variable has mean zero.

$$W = X - Y \sim \mathcal{N}\left(\mu_X - \mu_Y, \sigma_X^2 + \sigma_Y^2 - 2\rho_{XY}\sigma_X\sigma_Y\right)$$

Test on the exponential parameter λ (1/3)

- Consider any random sample drawn from an exponentially distributed random variable $X \sim \text{Exp}(\lambda)$.

- Suitable tests on the parameter λ are based, in samples of this sort, upon the sample mean $\bar{X}$.

- In such cases the sampling distribution of $\bar{X}$ is quite easy to identify. Note that:

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{1}{N}\,t\right)\right]^N = \left(1 - \frac{\lambda}{N}\,t\right)^N$$

and therefore:

$$\bar{X} \sim \Gamma\left(N, \frac{N}{\lambda}\right)$$

thus, $\bar{X}$ follows the Gamma distribution with the two given parameters.

Test on the exponential parameter λ (2/3)

- Consider hypotheses of the following sort.

$$\mathrm{H}_0 : \ 0 < \lambda \leq C \qquad\qquad \mathrm{H}_1 : \ \lambda > C$$

- Knowing the distribution of the sample mean, and given a test of size α, the critical value g_α^* is evaluated as:

$$\mathbb{P}\left(\bar{X} > C\left(\frac{g_\alpha^*}{\sqrt{N}} + 1\right)\right) = \mathbb{P}\left(\sqrt{N}\,\frac{\bar{X} - C}{C} > g_\alpha^*\right) = \alpha$$

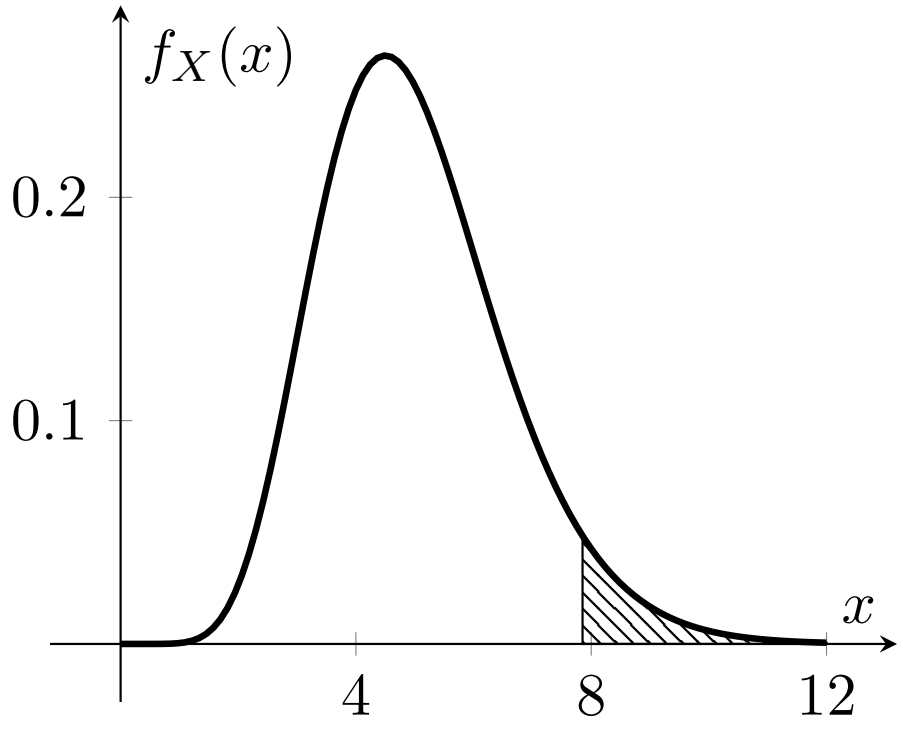
  and the null hypothesis is rejected if $\bar{x} > C g_\alpha^* / \sqrt{N} + C$.

- Recall that here $\mathbb{V}\mathrm{ar}\left[X\right] = \lambda^2$, hence the above formulae.

- Here the p-value is calculated similarly as in the one-sided normal test, that is $p\left(\bar{x}\right) = \mathbb{P}\left(\bar{X} \geq \bar{x}\right)$.

Test on the exponential parameter λ (3/3)

- In the figure, now $N = 10$, $C = 4$ and $X \sim \Gamma(10, 2)$ is the random variable that models the sample mean.
- Later in Lecture 6, it is discussed how the distribution of $\bar{X}$ approaches the normal in an *asymptotic* environment.



Interval Estimation

All point estimates are ultimately uncertain: hence, they are typically supplemented with other "likely" values of the parameter.

Definition 14

Interval estimators. Consider statistical inference about a *scalar* parameter θ. An *interval* estimator is a pair of statistics (L, U) that are functions of the sample: the "lower" bound statistic

$$L = L\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)$$

and the "upper" bound statistic

$$U = U\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)$$

such that $L \leq U$ and that if the values $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ are observed, the conclusion of the statistical inference is that θ falls within the interval defined by the two realized statistics, also called *confidence interval*.

$$\theta \in \left[L\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right), U\left(\mathbf{x}_1, \ldots, \mathbf{x}_N\right)\right]$$

Coverage probability

The larger the confidence interval, the higher the chances it contains the true parameter θ, but also the less informative it becomes.

Hence, the trick is to make a confidence interval as small as possible while keeping the chance it contains the true θ high.

Definition 15

Coverage probability. The *coverage* probability associated with an interval estimator is the probability that the associated confidence interval covers the true parameter, for a *given* parameter θ.

$$\text{Coverage Probability} = \mathbb{P}\left(L\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \leq \theta \leq U\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)\right)$$

Definition 16

Confidence coefficient. The confidence *coefficient* associated with an interval estimator is the infimum of all the confidence probabilities in the parameter space of θ (write it as Θ).

$$\text{Confidence Coefficient} = \inf_{\theta \in \Theta} \mathbb{P}\left(L\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \leq \theta \leq U\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)\right)$$

Inversion of test statistics (1/4)

- The problem can be reformulated as: choose L and U so as to maximize the coverage probability while also keeping the length of the interval fixed.

- Or conversely: choose the shortest $[L, U]$ interval while also keeping the coverage probability fixed.

- This problem is largely analogous to the trade-off between Type I and Type II errors in tests of hypothesis.

- In fact, all the leading methods for constructing confidence intervals are related to tests and based on the **inversion of test statistics**.

- This approach is first summarized in abstract terms, and it is then more conveniently illustrated via examples.

Inversion of test statistics (2/4)

The description of the method follows.

1. Start from a *two-sided* hypothesis about θ.

$$\mathrm{H}_0 : \ \theta = C \qquad\qquad \mathrm{H}_1 : \ \theta \neq C$$

2. Construct an *acceptance region* for C which is based on a test statistic T that is a function of C.

$$\mathbb{T}_0 = \left\{ T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2} \right] \right\}$$

   Here $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are suitable *critical values* associated with, respectively, the $(\alpha/2)$-th and $(1-\alpha/2)$-th quantiles of the distribution of the test statistic; if this is symmetric around zero it holds that $k^{**}_{1-\alpha/2} = -k^{*}_{\alpha/2}$.

(Continues after a discussion about notation.)

Inversion of test statistics (3/4)

Note how the notation here is somewhat counterintuitive, since $k^{**}_{1-\alpha/2}$ corresponds to the $(\alpha/2)$-th quantile:

$$\mathbb{P}\left(T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) > k^{**}_{1-\alpha/2}\right) = 1 - \frac{\alpha}{2}$$

and symmetrically $k^{*}_{\alpha/2}$ corresponds to the $(1-\alpha/2)$-th quantile:

$$\mathbb{P}\left(T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) > k^{*}_{\alpha/2}\right) = \frac{\alpha}{2}$$

The notation is chosen for the sake of consistency with the more general treatment of tests.

The acceptance region $\mathbb{T}_0$ here is associated with a size α which is defined in terms of the following probability.

$$\mathbb{P}\left(k^{**}_{1-\alpha/2} \leq T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C\right) \leq k^{*}_{\alpha/2}\right) = 1 - \alpha$$

This equals one minus the probability of a Type I error.

Inversion of test statistics (4/4)

(The description of the method continues.)

3. Derive two statistics I_1 & I_2 by *inverting* the function that defines the test statistic, $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; C)$, with respect to C, and by evaluating the inverse at the two critical values.

$$I_1 = T^{-1}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; k^{**}_{1-\alpha/2}\right)$$
$$I_2 = T^{-1}\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; k^{*}_{\alpha/2}\right)$$

4. The interval estimator is finally obtained as:

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \min\{I_1, I_2\}$$
$$U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \max\{I_1, I_2\}$$

where typically, $L = I_2$ and $U = I_1$. Note that the coverage probability associated with this interval estimator is $1 - \alpha$, since for any $C = \theta$ the procedure implies the following.

$$\mathbb{P}\left(L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \leq \theta \leq U(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)\right) = 1 - \alpha$$

Confidence interval for the normal mean

- In a random samples drawn from the normal distribution, a confidence interval for μ is based on two-sided tests. If the parameter σ^2 is known, the acceptance region is:

$$\mathbb{T}_{0,\mu} = \left\{ \sqrt{N}\,\frac{\bar{X} - C}{\sigma} \in \left[-z^*_{\alpha/2}, z^*_{\alpha/2} \right] \right\}$$

  since the null hypothesis is rejected if $\sqrt{N}\,|\bar{x} - C|\,/\sigma > z^*_{\alpha/2}$.

- Clearly the confidence interval for μ has $L = I_2$ and $U = I_1$.

$$\mu \in \left[\bar{X} - \frac{\sigma}{\sqrt{N}} z^*_{\alpha/2}, \bar{X} + \frac{\sigma}{\sqrt{N}} z^*_{\alpha/2} \right]$$

- If σ^2 is unknown, the procedure is based on the t-statistic:

$$\mu \in \left[\bar{X} - \frac{S}{\sqrt{N}} t^*_{\alpha/2}, \bar{X} + \frac{S}{\sqrt{N}} t^*_{\alpha/2} \right]$$

  where $t^*_{\alpha/2}$ obtains from the appropriate t-distribution.

Confidence interval for the normal variance (1/2)

- In a random samples drawn from the normal distribution, a confidence interval for σ^2 is also based on two-sided tests: a test of this sort would have the following acceptance region.

$$\mathbb{T}_{0,\sigma^2} = \left\{ (N-1)\,\frac{S^2}{C} \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2} \right] \right\}$$

  Here, $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are evaluated through the chi-squared distribution with $N-1$ degrees of freedom.

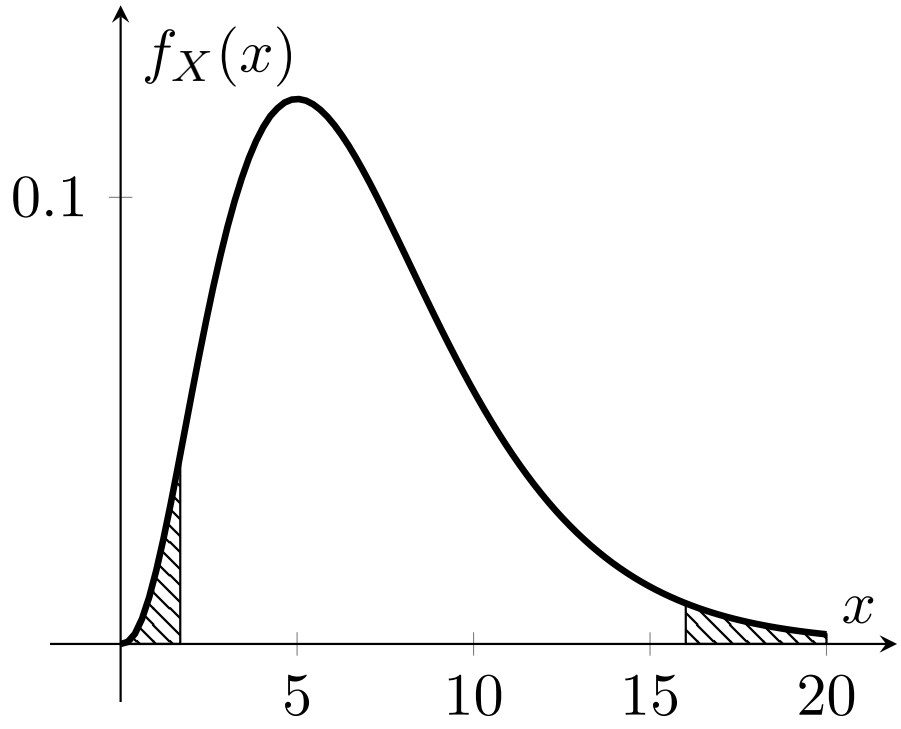
- Therefore, the confidence interval for σ^2 is:

$$\sigma^2 \in \left[(N-1)\,\frac{S^2}{k^{*}_{\alpha/2}}, (N-1)\,\frac{S^2}{k^{**}_{1-\alpha/2}} \right]$$

  and again, it is $L = I_2$ and $U = I_1$ because the test statistic is decreasing in the parameter of interest σ^2.

Confidence interval for the normal variance (2/2)

- Unlike in the previous example about the one-sided test for the normal variance, the test used now is two-sided.
- See the figure below to appreciate the difference: it displays the *two-sided rejection region* for the same N and C.



Confidence interval for the normal variance ratio

- In a similar manner, constructing a confidence interval for a *variance ratio* σ_X^2/σ_Y^2 from two normal samples requires a *two-sided* version of the test developed earlier. This test would have the following acceptance region.

$$\mathbb{T}_{0,\frac{\sigma_X^2}{\sigma_Y^2}} = \left\{ \frac{S_X^2}{S_Y^2}\frac{1}{C} \in \left[k^{**}_{1-\alpha/2}, k^{*}_{\alpha/2} \right] \right\}$$

  Here $k^{**}_{1-\alpha/2}$ and $k^{*}_{\alpha/2}$ are derived through the F-distribution with paired degrees of freedom $N_X - 1$ and $N_Y - 1$.

- Consequently, the confidence interval for the variance ratio here is:

$$\frac{\sigma_X^2}{\sigma_Y^2} \in \left[\frac{S_X^2}{S_Y^2}\frac{1}{k^{*}_{\alpha/2}}, \frac{S_X^2}{S_Y^2}\frac{1}{k^{**}_{1-\alpha/2}} \right]$$

  and it is once again $L = I_2$ and $U = I_1$.

Confidence interval for the exponential λ (1/2)

- Finally consider the case of parameter λ in sampling from the exponential distribution. Once again, it is necessary to start from a *two-sided* test; the acceptance region would be in this case as follows.

$$\mathbb{T}_{0,\lambda} = \left\{ \sqrt{N}\,\frac{\bar{X} - C}{C} \in \left[g^{**}_{1-\alpha/2}, g^{*}_{\alpha/2} \right] \right\}$$

  The quantiles $g^{**}_{1-\alpha/2}$ and $g^{*}_{\alpha/2}$ here obtain from the Gamma distribution with parameters $\alpha = N$ and $\beta = N/\lambda$

- The confidence interval for λ is thus:

$$\lambda \in \left[\frac{\bar{X}}{1 + N^{-1/2} g^{*}_{\alpha/2}}, \frac{\bar{X}}{1 + N^{-1/2} g^{**}_{1-\alpha/2}} \right]$$

  and one more time it is $L = I_2$ and $U = I_1$.

Confidence interval for the exponential λ (2/2)

- As usual, the interval is based on a two-sided test.
- See the figure below to best appreciate the difference with the one-sided case (using again the same N and C).