8. Estimation

8.1. Point estimation

- Point estimation is the use of the value $\widehat{\theta}$ of a statistic $\widehat{\boldsymbol{\theta}} = g\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$ to make a guess for the unknown parameter vector $\theta \in \Theta \subset \mathbb{R}^K$ characterizing the distribution $P_{\tilde{X}}\left(\cdot; \theta\right)$ of the random vector (or sample) $\tilde{X} = \left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$.

- Note that, since the parameter θ is unknown, the estimator $\widehat{\boldsymbol{\theta}}$ is an statistic that cannot depend on θ but only on the sample $\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$.

- The statistic $\widehat{\boldsymbol{\theta}} = g\left(\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n\right)$ we use for estimation is called **estimator**. The value $\widehat{\theta} = g\left(x_1, x_2, \ldots, x_n\right)$ taken by an estimator for a particular sample value $X = \left(x_1, x_2, \ldots, x_n\right)$ is called an **estimate**.

- For instance, $\mathbf{s}^2 \neq \sigma^2$ in general. Questions: what is the relationship between $\mathbf{s}^2$ and σ^2?, how "close" are they?, does $\mathbf{s}_n^2$ converge to σ^2?, etc.

8.2. The mean square error of an estimator and the relative efficiency of estimators

- **Definition.** The mean square error (MSE) of an estimator $\widehat{\boldsymbol{\theta}}$ for the population parameter $\theta \in \mathbb{R}$ is

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}-\theta\right)^2\right].$$

- If $\theta \in \mathbb{R}^K$, then the MSE is

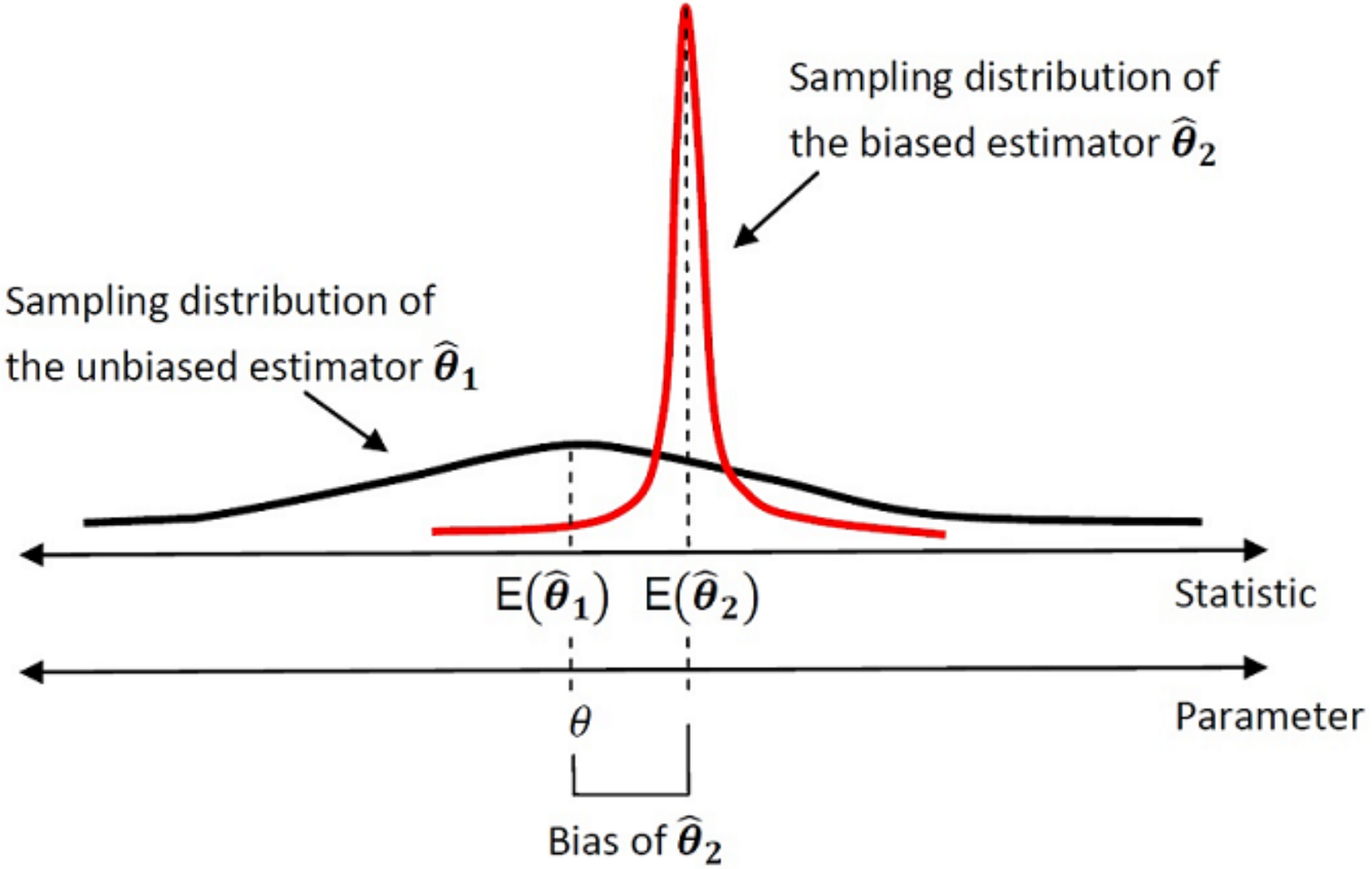
$$\mathrm{E}\left[\left\|\widehat{\boldsymbol{\theta}}-\theta\right\|^2\right].$$

- Note that

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}-\theta\right)^2\right]=\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}-\theta\right)+\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}}-\theta\right)\right]^2=\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)+\left[b_{\widehat{\boldsymbol{\theta}}}\left(\theta\right)\right]^2,$$

where $b_{\widehat{\boldsymbol{\theta}}}\left(\theta\right)=\mathrm{E}\left(\widehat{\boldsymbol{\theta}}-\theta\right)=\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right)-\theta$ is the bias of the estimator $\widehat{\boldsymbol{\theta}}$.

A biased estimator with small variance may be preferable to an unbiased estimator with large variance



- **Definition.** We say that $\widehat{\boldsymbol{\theta}}_1$ is a better (or a more efficient) estimator for $\theta \in \mathbb{R}$ than $\widehat{\boldsymbol{\theta}}_2$ if

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_1 - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_2 - \theta\right)^2\right].$$

- If the two previous estimators for θ were unbiased, $b_{\widehat{\boldsymbol{\theta}}_1}(\theta) = b_{\widehat{\boldsymbol{\theta}}_2}(\theta) = 0$, then $\widehat{\boldsymbol{\theta}}_1$ is better (or more efficient) than $\widehat{\boldsymbol{\theta}}_2$ if

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_1\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_2\right).$$

- **Definition.** We say that $\widehat{\boldsymbol{\theta}}^*$ is the best (or the most efficient) estimator in the class C of estimators for $\theta \in \mathbb{R}$ if

$$\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}^* - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right], \quad \text{for all } \widehat{\boldsymbol{\theta}} \in C.$$

8.3. The sample mean and sample variance as unbiased estimators

- Consider a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from a population $\tilde{x}$ with distribution $P_{\tilde{x}}$ having the mean μ and the finite variance σ^2.
- Sample mean:

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}_n = \frac{\sum_{i=1}^{n} \tilde{x}_i}{n}.$$

- Sample variance:

$$\mathbf{s^2} = \mathbf{s}_n^2 = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}_n\right)^2}{n-1}.$$

- Reminder:

(a)

$$\mathrm{E}\left(\bar{\mathbf{x}}\right) = \mu \quad \text{and} \quad \mathrm{Var}\left(\bar{\mathbf{x}}\right) = \frac{\sigma^2}{n}.$$

(b) Strong law of large numbers:

$$\bar{\mathbf{x}}_n \xrightarrow{a.s} \mu.$$

(c) Central limit theorem:

$$\tilde{z}_n \equiv \frac{\bar{\mathbf{x}}_n - \mathrm{E}(\bar{\mathbf{x}}_n)}{\sqrt{\mathrm{Var}\left(\bar{\mathbf{x}}_n\right)}} = \frac{\bar{\mathbf{x}}_n - \mu}{\sigma \big/ \sqrt{n}} \overset{a}{\sim} \mathrm{N}(0,1) \qquad (\text{or} \quad \tilde{z}_n \longrightarrow \mathrm{N}(0,1))$$

or, equivalently,

$$\sqrt{n}\left(\bar{\mathbf{x}}_n - \mu\right) \longrightarrow \mathrm{N}(0, \sigma^2).$$

(d) If the population is normal, then

$$\bar{\mathbf{x}}_n \sim \mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right) \iff \frac{\bar{\mathbf{x}}_n - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0,1) \iff \sqrt{n}\left(\bar{\mathbf{x}}_n - \mu\right) \sim \mathrm{N}(0, \sigma^2).$$

(e) $\mathbf{s}^2$ is an unbiased estimator for σ^2 :

$$\mathrm{E}\left(\mathbf{s}^2\right) = \sigma^2.$$

(f) If the population mean μ is known, then the statistic

$$\breve{\mathbf{s}}^2 = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \mu\right)^2}{n}$$

is an unbiased estimator for σ^2, $\mathrm{E}\left(\breve{\mathbf{s}}^2\right) = \mathrm{E}\left(\mathbf{s}^2\right) = \sigma^2$.

(g) If $\hat{\mathbf{s}}^2 = \dfrac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{n} = \dfrac{(n-1)\mathbf{s}^2}{n}$, then

$$\mathrm{E}\left(\hat{\mathbf{s}}^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2 = \mathrm{E}\left(\mathbf{s}^2\right).$$

However,

$$\lim_{n\to\infty} \mathrm{E}\left(\hat{\mathbf{s}}_n^2\right) = \mathrm{E}\left(\mathbf{s}_n^2\right) = \sigma^2.$$

(h) If the population is normal, then

$$\frac{(n-1)\mathbf{s}^2}{\sigma^2} \equiv \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}\right)^2}{\sigma^2} \sim \chi^2_{n-1}.$$

(i) If the population is normal, then

$$\mathrm{Var}\left(\mathbf{s}^2\right) = \frac{2\sigma^4}{n-1}, \quad \mathrm{Var}\left(\hat{\mathbf{s}}^2\right) = \frac{2(n-1)\sigma^4}{n^2}, \quad \text{and} \quad \mathrm{Var}\left(\breve{\mathbf{s}}^2\right) = \frac{2\sigma^4}{n}.$$

(j) If the population is normal, then

$$\frac{\bar{\mathbf{x}}_n - \mu}{\mathbf{s} \big/ \sqrt{n}} \sim \mathrm{t}_{n-1}$$

(k) If $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ are the variances of two independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively, then,

$$\frac{\mathbf{s_1^2} \big/ \sigma_1^2}{\mathbf{s_2^2} \big/ \sigma_2^2} = \frac{\sigma_2^2 \cdot \mathbf{s_1^2}}{\sigma_1^2 \cdot \mathbf{s_2^2}} \sim \mathrm{F}_{n_1-1, n_2-1}.$$

In particular, if the two population variances are equal, $\sigma_1^2 = \sigma_2^2$, then

$$\frac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \sim \mathrm{F}_{n_1-1, n_2-1}.$$

- Assume that the population is normal. Then,

$$\mathrm{E}\left(\mathbf{s^2}\right) = \sigma^2 \quad \text{so that } b_{\mathbf{s^2}}\left(\sigma^2\right) = 0,$$

while

$$\mathrm{E}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$$

so that the bias of $\hat{\mathbf{s}}^{\mathbf{2}}$ as an estimator for σ^2 is

$$b_{\hat{\mathbf{s}}^{\mathbf{2}}}\left(\sigma^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}.$$

- Moreover,

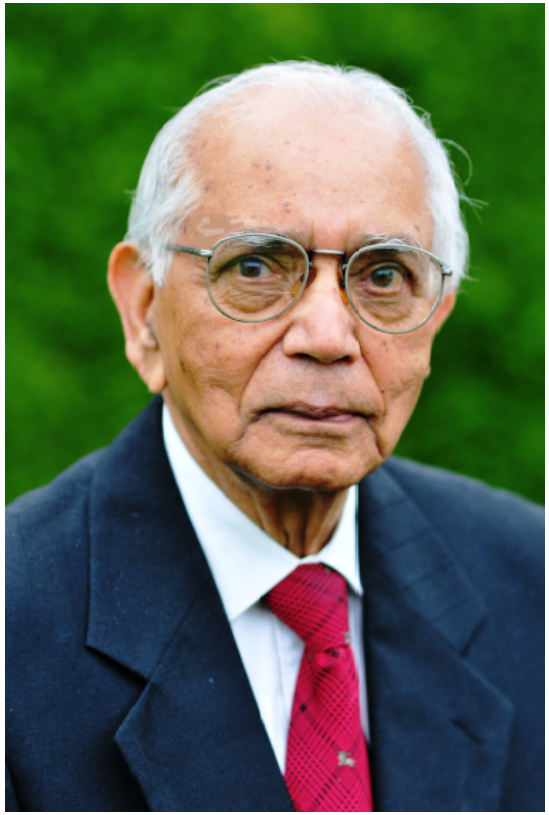
$$\mathrm{Var}\left(\mathbf{s^2}\right) = \frac{2\sigma^4}{n-1} \quad \text{and} \quad \mathrm{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) = \frac{2(n-1)\sigma^4}{n^2}.$$

- If the population is normal, $\mathbf{s^2}$ is an unbiased estimator for σ^2, while $\hat{\mathbf{s}}^{\mathbf{2}}$ is a biased estimator. However, $\mathrm{Var}\left(\hat{\mathbf{s}}^{\mathbf{2}}\right) < \mathrm{Var}\left(\mathbf{s^2}\right)$. In fact, it can be proved that $\hat{\mathbf{s}}^{\mathbf{2}}$ is an estimator for σ^2 more efficient than $\mathbf{s^2}$,

$$\mathrm{E}\left[\left(\hat{\mathbf{s}}^{\mathbf{2}} - \sigma^2\right)^2\right] < \mathrm{E}\left[\left(\mathbf{s^2} - \sigma^2\right)^2\right]. \qquad \text{(Exercise)}$$

8.4. The Cramér-Rao lower bound for unbiased estimators

Harald Cramér (1893 - 1985)



C. R. Rao (1920 - 2023)

- **Theorem (Cramér-Rao).** Let $\tilde{X}$ be a n-dimensional vector of random variables (not necessarily independent), the joint density of which is given by $h(X;\theta)$, where θ is a K-dimensional vector of parameters in some parameter space Θ. Let $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ be an unbiased estimator for θ with a finite covariance matrix $\text{Var}\left(\widehat{\boldsymbol{\theta}}\right)$. Furthermore, assume that $h(X;\theta)$ and $g(X)$ satisfy

- (a)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} h dX = \int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta} dX \leftarrow \text{This is a } K\text{-dimensional column vector}$$

- (b)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} ghdX = \int_{\mathbb{R}^n} g\frac{\partial h}{\partial\theta} dX \leftarrow \text{This is a } K\text{-dimensional column vector}$$

- (c)

$$\frac{\partial}{\partial \theta} \int_{\mathbb{R}^n} \frac{\partial h}{\partial \theta^T} dX = \int_{\mathbb{R}^n} \frac{\partial^2 h}{\partial \theta \partial \theta^T} dX \leftarrow \text{This is a } K \times K\text{-dimensional matrix}$$

- (d)

$$\text{Var}\left(\frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta} \right) \text{ is a positive definite } K \times K \text{ matrix.}$$

- Then, for all $\theta \in \Theta$,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-\text{E}\left[\frac{\partial^2 \ln h\left(\tilde{X}; \theta\right)}{\partial \theta \partial \theta^T} \right] \right)^{-1} \equiv CR \qquad \text{(CR-1)}$$

or, equivalently,

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(\text{E}\left[\frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta} \cdot \frac{\partial \ln h\left(\tilde{X}; \theta\right)}{\partial \theta^T} \right] \right)^{-1} \equiv CR. \qquad \text{(CR-2)}$$

- *Note:* $A \geqq B$ means that $A - B$ is a positive semi-definite matrix, which implies that the elements along the diagonal of $A - B$ are non-negative. Thus, the Cramér-Rao inequality provides a lower bound for $\text{Var}\left(\widehat{\boldsymbol{\theta}}_k\right)$, for all $k = 1, ..., K$.
- **Proof.** See the handout for the case $K = 1$.

- **About the assumptions:**
  - Assumption (d) is very mild.
  - Assumptions (a), (b) and (c) mean that the operations of differentiation and integration can be interchanged. This implies, among other things, that the domain of positive density of $\tilde{X}$ is independent of θ. Sometimes such a domain depends on the parameters (e.g. the uniform distribution).
- The matrix

$$I(\theta) = -\mathrm{E}\left[\frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial\theta\partial\theta^T}\right] = \mathrm{E}\left[\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta}\cdot\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta^T}\right]$$

  is called the Fisher information (matrix). Note that the Cramér-Rao lower bound (CR) is equal to $\left[I(\theta)\right]^{-1}$.
- Therefore, if $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator for θ and $\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) = CR$, then $\widehat{\boldsymbol{\theta}}$ is the best (or the most efficient) estimator within the class of unbiased estimators for θ. In this case, we say that the estimator $\widehat{\boldsymbol{\theta}}$ is an efficient estimator for θ.

- If the random vector $\tilde{X}$ were discrete instead of absolutely continuous, the result would be the same. In this case $h(X;\theta)$ would be the probability function of the random vector $\tilde{X}$ and we should change the integrals in both the statement and the proof by sums over the range of $\tilde{X}$. Obviously, this range should not depend on the parameters we are estimating in order to fulfil the assumptions (a), (b) and (c).

- **Corollary.** If in the previous theorem the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density (or probability function) $f(x;\theta)$, then

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{i}$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(n\mathrm{E}\left[\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta} \cdot \frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{ii}$$

- **Proof.** See the handout for the case $K = 1$.

- **Single parameter case:** $K = 1$ $(\theta \in \mathbb{R})$**.**
- **(a)** For $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, not necessarily a random sample, we have

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-\mathrm{E}\left[\frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial\theta^2}\right]\right)^{-1} \quad \text{(Theorem CR-1 for } K = 1\text{)}$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(\mathrm{E}\left[\left(\frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial\theta}\right)^2\right]\right)^{-1}. \quad \text{(Theorem CR-2 for } K = 1\text{)}$$

- **(b)** For the random sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$, we have

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f\left(\tilde{x};\theta\right)}{\partial\theta^2}\right]\right)^{-1} \quad \text{(Corollary (i) for } K = 1\text{)}$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(n\mathrm{E}\left[\left(\frac{\partial \ln f\left(\tilde{x};\theta\right)}{\partial\theta}\right)^2\right]\right)^{-1}. \quad \text{(Corollary (ii) for } K = 1\text{)}$$

8.5. Asymptotic properties of estimators: consistent estimators

- Let $\tilde{x}_n$ be a random variable with $\tilde{x}_n \sim P_n$, for $n = 1, 2, ...$ The limit of the expectation (or limiting expectation) of $\tilde{x}_n$ is

$$\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) = \lim_{n\to\infty} \int_{\mathbb{R}} x dP_n(x),$$

whereas the asymptotic expectation is

$$\mathrm{AE}\left(\tilde{x}_n\right) = \int_{\mathbb{R}} x d\widehat{P},$$

where $\tilde{x}_n \longrightarrow \widehat{P}$ as $n \longrightarrow \infty$. Thus, $\widehat{P}$ is the limiting distribution of $\left\{\tilde{x}_n\right\}_{n=1}^{\infty}$.

- **Theorem.** If $\mathrm{E}\left(\left|\tilde{x}_n\right|^r\right) < M$ for all n, where $M < \infty$ is a fixed bound, then

$$\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n^s\right) = \mathrm{AE}\left(\tilde{x}_n^s\right), \quad \text{for all } s < r.$$

  In particular, if $\mathrm{E}\left(\tilde{x}_n^2\right) < M$ for all n, where $M < \infty$ is a fixed bound, then $\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right) = \mathrm{AE}\left(\tilde{x}_n\right)$.

- Some statisticians refer to $\lim_{n\to\infty} \mathrm{E}\left(\tilde{x}_n\right)$ as the asymptotic expectation.

- Similarly, we can define the limit of the variance (or limiting variance) and the asymptotic variance as

$$\lim_{n\to\infty} \mathrm{Var}\left(\tilde{x}_n\right) = \lim_{n\to\infty} \mathrm{E}\left(\left[\tilde{x}_n - \mathrm{E}\left(\tilde{x}_n\right)\right]^2\right)$$

  and

$$\mathrm{AVar}\left(\tilde{x}_n\right) = \mathrm{AE}\left(\left[\tilde{x}_n - \mathrm{AE}\left(\tilde{x}_n\right)\right]^2\right),$$

  respectively.

- The subindex n in estimators will typically refer to the sample size.

- **Definition.** We say that the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is "unbiased in the limit" or "asymptotically unbiased" if

$$\lim_{n\to\infty} \mathrm{E}\left(\widehat{\boldsymbol{\theta}}_n\right) = \theta \qquad \text{or} \qquad \mathrm{AE}\left(\widehat{\boldsymbol{\theta}}_n\right) = \theta,$$

respectively.

Moreover, the "limiting bias" or the "asymptotic bias" of the estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is

$$\lim_{n\to\infty} \mathrm{E}\left(\widehat{\boldsymbol{\theta}}_n\right) - \theta \qquad \text{or} \qquad \mathrm{AE}\left(\widehat{\boldsymbol{\theta}}_n\right) - \theta,$$

respectively.

- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n^*$ is "the best (or the most efficient) in the limit" or "asymptotically the best (or the most efficient)" in the class C of estimators for $\theta \in \mathbb{R}$ if

$$\lim_{n\to\infty} \frac{\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n^* - \theta\right)^2\right]}{\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right]} \leq 1 \text{ or } \frac{\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n^* - \theta\right)^2\right]}{\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right]} \leq 1, \text{ for all } \widehat{\boldsymbol{\theta}}_n \in C,$$

  respectively.

- The function $\phi(n)$ is selected in order to make the previous relative asymptotic efficiency well defined in $\mathbb{R}$ (i.e., to make $\mathrm{AE}\left[\phi(n)\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] \neq 0$). Usually, we use $\phi(n) = n$.

- **Definition.** Let $\widehat{\theta}_n$ be an "unbiased in the limit" or an "asymptotically unbiased" estimator for $\theta \in \mathbb{R}$. Then, $\widehat{\theta}_n$ is an "efficient in the limit" or an "asymptotically efficient" estimator for θ if

$$\lim_{n\to\infty} \underbrace{\left(\frac{CR_n}{\text{Var}\left(\widehat{\theta}_n\right)} \right)}_{\leq 1 \text{ (Cramér-Rao)}} = 1 \quad \text{or} \quad \frac{CR^*}{\text{AVar}\left(\sqrt{\phi(n)}\,\widehat{\theta}_n\right)} = 1,$$

respectively, where CR^* is the following "adjusted" asymptotic Cramer-Rao lower bound (or inverse of the Fisher information):

$$CR^* = \left(-\text{AE}\left[\frac{1}{\phi(n)} \frac{\partial^2 \ln h\left(\tilde{X};\theta\right)}{\partial \theta^2} \right] \right)^{-1} =$$

$$\left(\text{AE}\left[\left(\frac{1}{\sqrt{\phi(n)}} \frac{\partial \ln h\left(\tilde{X};\theta\right)}{\partial \theta} \right)^2 \right] \right)^{-1}.$$

- If $\phi(n) = n$ and $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density function (probability function) $f(x;\theta)$, then

$$CR^* = \left(-\text{AE}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial \theta^2} \right] \right)^{-1} = \left(\text{AE}\left[\left(\frac{\partial \ln f(\tilde{x};\theta)}{\partial \theta} \right)^2 \right] \right)^{-1}.$$

- *Note:* An estimator may be the best (or the most efficient) in the limit (or asymptotically) in the class of unbiased estimators in the limit (or asymptotically) even if it is not efficient in the limit (or asymptotically). In other words, the Cramér-Rao lower bound does not need to be reached in the limit (or asymptotically) by an statistic that is the best (or the most efficient) unbiased estimator in the limit (or asymptotically).

- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is said to be (weakly) consistent if

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \theta \quad \left(\text{or } \operatorname*{plim}_{n\to\infty} \widehat{\boldsymbol{\theta}}_n = \theta \right),$$

  which is equivalent to $\widehat{\boldsymbol{\theta}}_n \xrightarrow{d} \theta$.
- **Definition.** An estimator $\widehat{\boldsymbol{\theta}}_n$ for $\theta \in \mathbb{R}$ is said to be strongly consistent if

$$\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s} \theta.$$

- **Theorem.** If $\lim_{n\to\infty} \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] = 0$, then $\widehat{\boldsymbol{\theta}}_n$ is a (weakly) consistent estimator for $\theta \in \mathbb{R}$.
- **Proof.** Obvious since $\widehat{\boldsymbol{\theta}}_n \xrightarrow{m} \theta$ implies that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{p} \theta$.
- Note that, since

$$\lim_{n\to\infty} \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}}_n - \theta\right)^2\right] = \lim_{n\to\infty} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_n\right) + \lim_{n\to\infty} \left[b_{\widehat{\boldsymbol{\theta}}_n}(\theta)\right]^2,$$

  then $\widehat{\boldsymbol{\theta}}_n \xrightarrow{m} \theta$ if and only if $\lim_{n\to\infty} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_n\right) = \lim_{n\to\infty} b_{\widehat{\boldsymbol{\theta}}_n}(\theta) = 0$.

- Let $\bar{\mathbf{x}}_n$ and $\mathbf{s}_n^2$ be the mean and variance of a random sample of size n.
- Since $\mathrm{E}(\bar{\mathbf{x}}_n) = \mu$ (or $b_{\bar{\mathbf{x}}_n}(\mu) = 0$) for all n, and $\mathrm{Var}(\bar{\mathbf{x}}_n) = \sigma^2 / n$ so that $\lim_{n\to\infty} \mathrm{Var}(\bar{\mathbf{x}}_n) = 0$, the sample mean $\bar{\mathbf{x}}_n$ is a (weakly) consistent estimator for the population mean μ.
- In fact, $\bar{\mathbf{x}}_n \xrightarrow{a.s} \mu$ because of any of the strong law of large numbers.
- Assume that the population is normal. Then $\mathrm{E}(\mathbf{s}_n^2) = \sigma^2$ $\left(\text{or } b_{\mathbf{s}_n^2}(\sigma^2) = 0\right)$ for all n, and $\mathrm{Var}(\mathbf{s}_n^2) = 2\sigma^4 / (n-1)$ so that $\lim_{n\to\infty} \mathrm{Var}(\mathbf{s}_n^2) = 0$. Hence, the sample variance $\mathbf{s}_n^2$ is a (weakly) consistent estimator for the population variance σ^2.

- Consider now the statistic

$$\hat{\mathbf{s}}_n^{\mathbf{2}} = \frac{\sum_{i=1}^{n} \left(\tilde{x}_i - \bar{\mathbf{x}}_n\right)^2}{n} = \frac{(n-1)\mathbf{s}_n^{\mathbf{2}}}{n}.$$

Then,

$$\mathrm{E}\left(\hat{\mathbf{s}}_n^{\mathbf{2}}\right) = \left(\frac{n-1}{n}\right)\sigma^2 \neq \sigma^2$$

and the bias is

$$b_{\hat{\mathbf{s}}_n^{\mathbf{2}}}\left(\sigma^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2 = -\frac{\sigma^2}{n} \longrightarrow 0, \text{ as } n \to \infty.$$

Moreover,

$$\mathrm{Var}\left(\hat{\mathbf{s}}_n^{\mathbf{2}}\right) = \frac{2(n-1)\sigma^4}{n^2} \to 0, \text{ as } n \to \infty.$$

- This means that $\hat{\mathbf{s}}_n^{\mathbf{2}} \xrightarrow{m} \sigma^2$ and, thus, $\hat{\mathbf{s}}_n^{\mathbf{2}}$ is a (weakly) consistent estimator for σ^2.

8.6. Sufficient estimators

- The statistic $\widehat{\boldsymbol{\theta}}$ is a sufficient estimator (or statistic) for the parameter θ if it uses all the information relevant for the estimation of the population parameter $\theta \in \Theta \subset \mathbb{R}^K$, that is, if all the knowledge we can gain about θ by actually specifying the individual sample values and their order can just as well be obtained by observing only the value of the statistic $\widehat{\boldsymbol{\theta}}$.

- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (a sample) whose joint distribution $P_{\tilde{X}}(\cdot; \theta)$ depends on the parameter θ. Then, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$, the conditional distribution of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$, $P_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right)$ should not depend on the value θ. Note that, if $P_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right)$ depends on θ, particular values of $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ yielding a particular value of $\widehat{\boldsymbol{\theta}}$ are more likely for some values of θ than for others, and the knowledge of these sample values will help in the estimation for θ.

- **Definition.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution $P_{\tilde{X}}(\cdot; \theta)$ depends on the parameter $\theta \in \Theta \subset \mathbb{R}^K$. The statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for θ if, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$, the conditional distribution of the sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$, $P_{\tilde{X}|\widehat{\theta}}\left(\cdot \left| \widehat{\theta}; \theta \right.\right) : \mathcal{B}(\mathbb{R}^n) \longrightarrow [0, 1]$, is independent of θ.

- *Note:* If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has a discrete (absolutely continuous) distribution, then sufficiency means that the conditional probability function (density) of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$,

$$f_{\tilde{X}|\widehat{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}; \theta \right.\right) = \frac{f_{(\tilde{X}, \widehat{\boldsymbol{\theta}})}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}{f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right)},$$

for all $\widehat{\theta}$ with $f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right) \neq 0$, does not depend on the parameter θ.

Note that

$$f_{\left(\tilde{X},\widehat{\theta}\right)}\left(\underbrace{x_1,x_2,...,x_n}_{X\in\mathbb{R}^n},\widehat{\theta};\theta\right)=\begin{cases} f_{\tilde{X}}\left(x_1,x_2,...,x_n;\theta\right) & \text{if } \widehat{\theta}=t\left(X\right), \\ 0 & \text{otherwise.}\end{cases} \quad (*)$$

and, for $\tilde{X}$ absolutely continuous,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\int_{\mathbb{R}^n} f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)dX=\int_A f_{\tilde{X}}\left(x_1,x_2,...,x_n;\theta\right)dX,$$

$$\text{with } A=\left\{X=(x_1,x_2,...,x_n)\in\mathbb{R}^n\,\middle|\,t(X)=\widehat{\theta}\right\},$$

whereas, if $\tilde{X}$ is discrete,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\sum_{X\in\prod_{i=1}^{n}\tilde{x}_i(\Omega)} f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)=\sum_{X\in B} f_{\tilde{X},}\left(x_1,x_2,...,x_n;\theta\right),$$

$$\text{with } B=\left\{X=(x_1,x_2,...,x_n)\in\prod_{i=1}^{n}\tilde{x}_i\left(\Omega\right)\,\middle|\,t(X)=\widehat{\theta}\right\}.$$

- **Factorization theorem (Fisher-Neyman).** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution is discrete (absolutely continuous) and depends on the parameter $\theta \in \Theta \subset \mathbb{R}^K$. Then, the statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the parameter θ if and only if the probability function (density) of $\tilde{X}$ can be factorized as follows:

$$f_{\tilde{X}}(x_1, x_2, ..., x_n; \theta) = \underbrace{h(x_1, x_2, ..., x_n)}_{\text{independent of } \theta} \cdot g\left(\widehat{\theta}, \theta\right),$$

where $\widehat{\theta} = t(x_1, x_2, ..., x_n)$.

- **Proof.** (Sufficiency $\Longrightarrow$ Factorization). If

$$f_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta \right.\right) = \frac{f_{\left(\tilde{X}, \widehat{\boldsymbol{\theta}}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}{f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right)}$$

is independent of θ, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}} = t\left(\tilde{X}\right)$, then

$$\underbrace{f_{\left(\tilde{X}, \widehat{\boldsymbol{\theta}}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}_{f_{\tilde{X}}(x_1, x_2, ..., x_n; \theta) \leftarrow \text{from } (*)}$$

$$= \underbrace{f_{\tilde{X}|\widehat{\boldsymbol{\theta}}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta \right.\right)}_{\text{independent of } \theta \rightarrow\ m\left(X, \widehat{\theta}\right) = m(X, t(X)) = h(X)} \times \underbrace{f_{\widehat{\boldsymbol{\theta}}}\left(\widehat{\theta}; \theta\right)}_{g\left(\widehat{\theta}, \theta\right)}.$$

- (Factorization $\Longrightarrow$ Sufficiency). If

$$f_{\tilde{X}}\left(x_1, x_2, ..., x_n; \theta\right) = h\left(x_1, x_2, ..., x_n\right) \cdot g\left(\widehat{\theta}, \theta\right),$$

then, for all for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}} = t\left(\tilde{X}\right)$,

$$f_{\tilde{X}|\widehat{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}\,; \theta\right.\right) = \frac{\overbrace{f_{\left(\tilde{X}, \widehat{\theta}\right)}\left(x_1, x_2, ..., x_n, \widehat{\theta}; \theta\right)}^{\leftarrow \text{from } (*)}}{f_{\widehat{\theta}}\left(\widehat{\theta}; \theta\right)} = \frac{h\left(x_1, x_2, ..., x_n\right) \cdot g\left(\widehat{\theta}, \theta\right)}{f_{\widehat{\theta}}\left(\widehat{\theta}; \theta\right)}. \quad (**)$$

If the random vector $\tilde{X}$ is absolutely continuous,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\int_{A}\underbrace{h\left(x_1,x_2,...,x_n\right)\cdot g\left(\widehat{\theta},\theta\right)}_{f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)}dX=g\left(\widehat{\theta},\theta\right)\cdot\underbrace{\int_{A}h\left(x_1,x_2,...,x_n\right)dX}_{c_1\left(\widehat{\theta}\right)},$$

where $A=\left\{X=\left(x_1,x_2,...,x_n\right)\in\mathbb{R}^n\,\middle|\,t(X)=\widehat{\theta}\right\}$ or, if $\tilde{X}$ is discrete,

$$f_{\widehat{\theta}}\left(\widehat{\theta};\theta\right)=\sum_{X\in B}\underbrace{h\left(x_1,x_2,...,x_n\right)\cdot g\left(\widehat{\theta},\theta\right)}_{f_{\left(\tilde{X},\widehat{\theta}\right)}\left(x_1,x_2,...,x_n,\widehat{\theta};\theta\right)}=g\left(\widehat{\theta},\theta\right)\cdot\underbrace{\sum_{X\in B}h\left(x_1,x_2,...,x_n\right)}_{c_2\left(\widehat{\theta}\right)},$$

where $B=\left\{X=\left(x_1,x_2,...,x_n\right)\in\prod_{i=1}^{n}\tilde{x}_i\left(\Omega\right)\,\middle|\,t(X)=\widehat{\theta}\right\}$ and $c_i\left(\widehat{\theta}\right)$, $i=1,2$, only depend on the value $\widehat{\theta}$ of the statistic. Therefore, from $(**)$ we get

$$f_{\tilde{X}\left|\widehat{\theta}\right.}\left(x_1,x_2,...,x_n\,\middle|\,\widehat{\theta}\,;\theta\right)=\frac{h\left(x_1,x_2,...,x_n\right)}{c_i\left(\widehat{\theta}\right)},\quad i=1,2,$$

which is independent of θ. *Q.E.D.*

8.7. The Rao-Blackwell theorem

C. R. Rao (1920 - 2023) - David Blackwell (1919 - 2010)

- **Theorem (Rao-Blackwell).** Let $\widehat{\boldsymbol{\theta}}$ be an estimator for the parameter θ. Suppose that $\tilde{t} = t\left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ is a sufficient statistic for θ and let $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$. Then,

$$\mathrm{E}\left[\left(\boldsymbol{\theta}^* - \theta\right)^2\right] \leq \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right]. \tag{1}$$

- **Proof.** Note that, from the sufficiency of $\tilde{t}$, $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$ does not depend on the parameter θ but only on the sample $\left(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\right)$ so that $\boldsymbol{\theta}^*$ is an estimator. Then,

$$\mathrm{E}\left[\left(\boldsymbol{\theta}^* - \theta\right)^2\right] = \mathrm{E}\left[\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\middle|\, \tilde{t}\right) - \theta\right]^2\right] = \mathrm{E}\left[\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}} - \theta\middle|\, \tilde{t}\right)\right]^2\right]$$

$$\leq \mathrm{E}\left[\mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\middle|\, \tilde{t}\right]\right] = \mathrm{E}\left[\left(\widehat{\boldsymbol{\theta}} - \theta\right)^2\right],$$

where the inequality follows from Jensen's inequality applied to a quadratic, and thus convex, function and the last equality follows from Adam's law. *Q.E.D.*

- The previous theorem tells us that an improvement in efficiency of an estimator can be obtained by taking its conditional expectation with respect to a sufficient statistic. This process of improvement is called Rao-Blackwellization. Thus, $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$ is the Rao-Blackwellized estimator.
- In other words, the theorem says that, if an estimator is not a function of a sufficient statistic, it can be improved in terms of MSE.
- **Corollary.** Let $\widehat{\boldsymbol{\theta}}$ be an unbiased estimator for the parameter θ. Suppose that $\tilde{t}$ is a sufficient statistic for θ and let $\boldsymbol{\theta}^* = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)$. Then, (a) $\boldsymbol{\theta}^*$ is an unbiased estimator for θ, and
(b) $\mathrm{Var}\left(\boldsymbol{\theta}^*\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)$.
- **Proof. (a)** $\mathrm{E}\left(\boldsymbol{\theta}^*\right) = \mathrm{E}\left[\mathrm{E}\left(\widehat{\boldsymbol{\theta}} \,|\tilde{t}\right)\right] = \mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right) = \theta$, where the second equality follows from Adam's law and the last from the unbiasedness of $\widehat{\boldsymbol{\theta}}$. **(b)** From the Rao-Blackwell theorem, we know that the inequality (1) holds. Since both $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^*$ are unbiased estimators for θ, then inequality (1) simply becomes $\mathrm{Var}\left(\boldsymbol{\theta}^*\right) \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)$. *Q.E.D.*

8.8. The method of moments

- **Definition.** The kth sample moment of the value $X = (x_1, x_2, ..., x_n)$ taken by a random sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is given by

$$m'_k(X) = \frac{\sum_{i=1}^{n} x_i^k}{n}.$$

- The population moments depend on the parameter $\theta \in \Theta \subset \mathbb{R}^K$, $\mu'_k(\theta)$, $k = 1, 2, ...$
- The Method of Moments (MM) consists of solving the following system of equations:

$$m'_k(X) = \mu'_k(\theta), \text{ for } k = 1, 2, ..., K,$$

for the K parameters of the population distribution. The solution is the estimate $\widehat{\theta}_{MM} = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_{MM} = t\left(\tilde{X}\right)$ is the Method of Moments estimator for θ.

- Intuition: Find the parameters of the distribution that match the empirical moments to the population moments of the distribution.
- We could generate sample moments of order larger than K,

$$m'_k(X) = \mu'_k(\theta), \text{ for } k = 1, 2, ..., J, \quad \text{where } J \geq K.$$

However, the previous system of J equations and K unknowns does not have solution generically when $J > K$.

- In this case, the Generalized Method of Moments (GMM) estimate is given by

$$\widehat{\theta}_{GMM} = \arg\min_{\theta \in \Theta} \left\{ \left[m'(X) - \mu'(\theta) \right]^{\mathsf{T}} W \left[m'(X) - \mu'(\theta) \right] \right\},$$

where $\left[m'(X) - \mu'(\theta) \right]^{\mathsf{T}} = \left(..., m'_j(X) - \mu'_j(\theta), ... \right)_{1 \times J}$ and $W_{J \times J}$ is a weighting matrix.

- Note that if $J = K$ and $W = \mathrm{I}$, then the GMM estimator coincides with the MM estimator.

8.9. Maximum likelihood estimation

- Let $X = (x_1, x_2, ..., x_n)$ be the value of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ having the density (probability function) $h(X; \theta)$ where θ is a K-dimensional vector of parameters in some parameter space Θ. The likelihood function $L : \Theta \longrightarrow \mathbb{R}$ is given by

$$L(\theta; X) \equiv h(X; \theta).$$

- Note that $h(X; \theta)$ is the value of the joint density (probability function) of the random vector $\tilde{X}$ at the observed sample value.
- The maximum likelihood (ML) estimate is given by

$$\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; X).$$

- The solution to the previous maximization problem is the estimate $\widehat{\theta}_{ML} = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_{ML} = t(\tilde{X})$ is the Maximum Likelihood estimator for θ.

- Intuition: Find the parameters of the distribution for which the observed data is most probable.
- If the parameter space Θ is compact and $L(\theta; X)$ is continuous on Θ, then $\widehat{\theta}_{ML}$ exists.
- Sometimes, by modifying the density function on a set with zero Lebesgue measure, we can make

$$\max_{\theta \in \Theta} L(\theta; X) = \sup_{\theta \in \Theta} L(\theta; X),$$

  where $\sup_{\theta \in \Theta} L(\theta; X)$ exists if the likelihood function $L(\cdot; X)$ is bounded and Θ is compact.
- **Invariance property.** If $\widehat{\theta}_{ML}$ is a ML estimator for θ and $g : \Theta \longrightarrow \Theta' \subset \mathbb{R}^K$ is a one-to-one correspondence (or bijection), then $g\left(\widehat{\theta}_{ML}\right)$ is a ML estimator for $g(\theta) \in \Theta'$.

- **Proof.** Let $L(\theta; X)$ be the likelihood function on the parameter space Θ, $\widehat{L}(\theta'; X)$ be the likelihood function on the parameter space Θ', $h(X; \theta)$ be the density (probability function) of the random vector $\tilde{X}$, and $\widehat{h}(X; \theta')$ be the density function (probability function) of $\tilde{X}$ under the reparameterization dictated by the function g, according to which $\theta' = g(\theta)$ and $\theta = g^{-1}(\theta')$ with $\theta \in \Theta$ and $\theta' \in \Theta'$. Then, since $h(X; \theta) = h(X; g^{-1}(\theta')) = \widehat{h}(X; \theta')$ and $L(\theta; X) = h(X; \theta)$, we have that

$$\widehat{L}(\theta'; X) = \widehat{h}(X; \theta') = h(X; g^{-1}(\theta')) = L(\underbrace{g^{-1}(\theta')}_{=\theta}; X).$$

Therefore,

$$\max_{\theta' \in \Theta'} \widehat{L}(\theta'; X) = \max_{g^{-1}(\theta') \in \Theta} L(g^{-1}(\theta'); X) = \max_{\theta \in \Theta} L(\theta; X).$$

The previous equality tells us that, if $\widehat{\theta}_{ML} = \arg\max_{\theta \in \Theta} L(\theta; X)$ and $\widehat{\theta}'_{ML} = \arg\max_{\theta' \in \Theta'} \widehat{L}(\theta'; X)$, then $\widehat{\theta}_{ML} = g^{-1}\left(\widehat{\theta}'_{ML}\right)$ or $\widehat{\theta}'_{ML} = g\left(\widehat{\theta}_{ML}\right)$.

Q.E.D.

- **Example:** Let $\widehat{\theta}_{ML}$ be the Maximum Likelihood estimator for the parameter θ of the exponential distribution. The variance of the exponential distribution is $\sigma^2 = g(\theta) = \theta^2$ and g is a one-to-one correspondence since $\theta > 0$, i.e., $\Theta = \mathbb{R}_{++}$. Therefore, $\left(\widehat{\theta}_{ML}\right)^2$ is the Maximum Likelihood estimator for the population variance σ^2.

- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random sample from a population with density (probability function) $f(x;\theta)$, then

$$L(\theta; X) = h(X;\theta) = \prod_{i=1}^{n} f(x_i;\theta).$$

- Moreover, in this case,

$$\widehat{\theta}_{ML} = \arg\max_{\theta\in\Theta} L(\theta; X) = \arg\max_{\theta\in\Theta} \ln L(\theta; X) = \arg\max_{\theta\in\Theta} \sum_{i=1}^{n} \ln f(x_i;\theta).$$

8.10. Bayesian estimation

- $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a random vector (sample) and $\tilde{w} = g(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a statistic, which may be a random vector.
- The random vector $\tilde{w} = g(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has the density (probability function) $f_{\tilde{w}|\boldsymbol{\theta}}(w|\theta)$, where the parameter θ is the realization of the random vector $\boldsymbol{\theta}: (\Omega, \mathcal{F}) \longrightarrow (\mathbb{R}^K, \mathcal{B}(\mathbb{R}^K))$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$.
- $f_{\tilde{w}|\boldsymbol{\theta}}(w|\theta)$ is also called the likelihood of $\tilde{w}$ for $\boldsymbol{\theta} = \theta$.
- The density (probability function) of $\boldsymbol{\theta}$ is $f_{\boldsymbol{\theta}}$ (this is called the prior density or prior probability function).

- The posterior density (probability function) of $\boldsymbol{\theta}$ given $\tilde{w} = w$ is

$$f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right) = \frac{f_{(\boldsymbol{\theta},\tilde{w})}\left(\theta, w\right)}{f_{\tilde{w}}\left(w\right)} = \frac{f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)}{f_{\tilde{w}}\left(w\right)}, \text{ for } f_{\tilde{w}}\left(w\right) \neq 0,$$

where $f_{\boldsymbol{\theta}}\left(\theta\right)$ is the prior density (probability function) of $\boldsymbol{\theta}$, $f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)$ is the likelihood of $\tilde{w}$ for $\boldsymbol{\theta} = \theta$, and

$$f_{\tilde{w}}\left(w\right) = \int_{\Theta} f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right) d\theta \tag{1}$$

or

$$f_{\tilde{w}}\left(w\right) = \sum_{\theta \in \Theta} f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right). \tag{2}$$

- Note that the formula for the posterior density/probability function,

$$f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right) = \frac{f_{\boldsymbol{\theta}}\left(\theta\right) f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)}{f_{\tilde{w}}\left(w\right)}, \text{ for } f_{\tilde{w}}\left(w\right) \neq 0,$$

also holds if $f_{\boldsymbol{\theta}}\left(\theta\right)$ is a density/probability function and the likelihood $f_{\tilde{w}|\boldsymbol{\theta}}\left(w\,|\theta\right)$ is a conditional probability function/density (see Exercise 13 of List 2). Then, $f_{\tilde{w}}\left(w\right)$ would be a probability function/density obtained through the formula (1)/(2).

- We define the loss function $L\left(\theta,\widehat{\theta}\right)$. For $K=1$, we usually use

$$L\left(\theta,\widehat{\theta}\right)=\left(\theta-\widehat{\theta}\right)^{2} \qquad \text{(quadratic loss function)}$$

or, for $K>1$,

$$L\left(\theta,\widehat{\theta}\right)=\left(\theta-\widehat{\theta}\right)^{\mathsf{T}} W\left(\theta-\widehat{\theta}\right),$$

where W is a $K\times K$ weighting matrix.

- Then, the Bayesian estimate is given by

$$\widehat{\theta}_{B}=\underset{\widehat{\theta}\in\Theta}{\arg\min}\,\mathrm{E}\left[L\left(\boldsymbol{\theta},\widehat{\theta}\right)\middle|\,\tilde{w}=w\right]=\underset{\widehat{\theta}\in\Theta}{\arg\min}\int_{\Theta}L\left(\theta,\widehat{\theta}\right)f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right)d\theta$$

or

$$\widehat{\theta}_{B}=\underset{\widehat{\theta}\in\Theta}{\arg\min}\,\mathrm{E}\left[L\left(\boldsymbol{\theta},\widehat{\theta}\right)\middle|\,\tilde{w}=w\right]=\underset{\widehat{\theta}\in\Theta}{\arg\min}\sum_{\theta\in\Theta}L\left(\theta,\widehat{\theta}\right)f_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right).$$

- In general,

$$\widehat{\theta}_B = \arg\min_{\widehat{\theta}\in\Theta} \mathrm{E}\left[L\left(\boldsymbol{\theta},\widehat{\theta}\right)\middle|\,\tilde{w}=w\right] = \arg\min_{\widehat{\theta}\in\Theta} \int_{\Theta} L\left(\theta,\widehat{\theta}\right) dP_{\boldsymbol{\theta}|\tilde{w}}\left(\theta\,|w\right).$$

- The solution is the estimate $\widehat{\theta}_B = h(w) = h\left(g(X)\right) = t(X)$ and the statistic $\widehat{\boldsymbol{\theta}}_B = t\left(\tilde{X}\right)$ is the Bayesian estimator for the random parameter $\boldsymbol{\theta}$ (or for the parameter value θ).
- Intuition: A Bayesian estimator minimizes the posterior expected value of the loss function.
- **Proposition.** Let $K = 1$. If $L\left(\theta,\widehat{\theta}\right) = \left(\theta-\widehat{\theta}\right)^2$, then $\widehat{\theta}_B = \mathrm{E}\left[\boldsymbol{\theta}\,|\tilde{w}=w\right]$.

- **Proof.** To solve

$$\underset{\widehat{\theta}\in\Theta\subset\mathbb{R}}{\arg\min}\ \mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right], \tag{1}$$

we compute the first derivative of $\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right]$ w.r.t. $\widehat{\theta}$ and equate it to zero,

$$-2\,\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)\middle|\,\tilde{w}=w\right]=0.$$

Solving for $\widehat{\theta}$ in the previous expression, we get

$$\widehat{\theta}_B=\mathrm{E}\left[\boldsymbol{\theta}\,|\tilde{w}=w\right]. \tag{2}$$

Note that $\mathrm{E}\left[\left(\boldsymbol{\theta}-\widehat{\theta}\right)^2\middle|\,\tilde{w}=w\right]$ is a strictly convex function of $\widehat{\theta}$.

To see this, we compute its second derivative w.r.t. $\widehat{\theta}$,

$$-2\,\mathrm{E}\left[-1|\,\tilde{w}=w\right]=(-2)(-1)=2>0.$$

Therefore, the value $\widehat{\theta}_B$ obtained in (2) is a solution for (1) and, hence, it is a Bayesian estimate for $\boldsymbol{\theta}$. *Q.E.D.*

- **Conjugate families of distributions.**

- There are cases where the prior distribution of the random vector $\boldsymbol{\theta}$ and the posterior distribution of $\boldsymbol{\theta}$ given any value of the statistic (or any value of the random sample) both belong to the same family of distributions. The prior and posterior are then called conjugate distributions for the corresponding likelihood.

- **Examples:**

- **1)** Prior: λ has the gamma distribution with the parameters α and β. Likelihood: $\tilde{x}$ has the Poisson distribution with the random parameter λ. Posterior of λ given the value of a random sample of size n from the population $\tilde{x}$, $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$: gamma distribution with the parameters $\alpha + \sum_{i=1}^{n} x_i = \alpha + n\bar{x}$, where $\bar{x}$ is the value of the sample mean, and $\frac{\beta}{1+n\beta}$. Thus, the family of gamma distributions is conjugate for the Poisson likelihood.

- **2)** Prior: $\boldsymbol{\mu}$ has the normal distribution with the mean μ_0 and the variance σ_0 (or precision $\tau_0 = 1/\sigma_0^2$). Likelihood: $\tilde{x}$ has the normal distribution with the random mean $\boldsymbol{\mu}$ and the known variance σ^2 (or precision $\tau = 1/\sigma^2$), which means that the mean $\bar{\mathbf{x}}$ of a random sample of size n from the population $\tilde{x}$ is normal with the random mean $\boldsymbol{\mu}$ and the known variance σ^2/n. Posterior of $\boldsymbol{\mu}$ given $\bar{\mathbf{x}} = \bar{x}$ (or given the sample value $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$): normal distribution with the mean μ_1 and the variance σ_1^2 (or precision $\tau_1 = 1/\sigma_1^2$), where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} = \left(\frac{\tau_0}{\tau_0 + n\tau}\right)\mu_0 + \underbrace{\left(\frac{n\tau}{\tau_0 + n\tau}\right)}_{=\left(\frac{\tau}{\tau_0+n\tau}\right)\sum_{i=1}^{n} x_i}\bar{x}$$

and

$$\tau_1 = \frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} = \tau_0 + n\tau.$$

Thus, the family of normal distributions is conjugate for the normal likelihood.

- **3)** Prior: $\boldsymbol{\theta}$ has the beta distribution with the parameters α and β. Likelihood: $\tilde{x}$ has the Bernoulli distribution with the random parameter $\boldsymbol{\theta}$. Posterior of $\boldsymbol{\theta}$ given the value of a random sample of size n from the population $\tilde{x}$, $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$: beta distribution with the parameters $\sum_{i=1}^{n} x_i + \alpha = n\bar{x} + \alpha$ and $n - \sum_{i=1}^{n} x_i + \beta = n(1 - \bar{x}) + \beta$, where $\bar{x}$ is the value of the sample mean (i.e., the percentage of successes). Thus, the family of beta distributions is conjugate for the Bernoulli likelihood.

  OR, equivalently,

- **3')** Prior: $\boldsymbol{\theta}$ has the beta distribution with the parameters α and β. Likelihood: $\tilde{y}$ has the binomial distribution with the parameters $\boldsymbol{\theta}$ and n, where n is known. Posterior of $\boldsymbol{\theta}$ given $\tilde{y} = y$: beta distribution with the parameters $y + \alpha$ and $n - y + \beta$. In fact, the family of beta distributions is conjugate for the binomial likelihood.

- **Sufficiency in Bayesian estimation.**
- **Definition.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution $P_{\tilde{X}|\boldsymbol{\theta}}(\cdot|\theta)$ depends on the random parameter $\boldsymbol{\theta}: (\Omega, \mathcal{F}) \longrightarrow \left(\mathbb{R}^K, \mathcal{B}\left(\mathbb{R}^K\right)\right)$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$. The statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the random parameter $\boldsymbol{\theta}$ if, for all values $\widehat{\theta}$ taken by $\widehat{\boldsymbol{\theta}}$ and all values θ taken by $\boldsymbol{\theta}$, the conditional distribution of the sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$ and $\boldsymbol{\theta} = \theta$, $P_{\tilde{X}|\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left(\cdot \left| \widehat{\theta}, \theta \right.\right) : \mathcal{B}(\mathbb{R}^n) \longrightarrow [0,1]$, is independent of θ.
- If $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ has a discrete (absolutely continuous) distribution, then sufficiency means that the conditional probability function (density) of $\tilde{X}$ given $\widehat{\boldsymbol{\theta}} = \widehat{\theta}$ and $\boldsymbol{\theta} = \theta$,

$$f_{\tilde{X}|\widehat{\boldsymbol{\theta}},\boldsymbol{\theta}}\left(x_1, x_2, ..., x_n \left| \widehat{\theta}, \theta \right.\right) = \frac{f_{(\tilde{X},\widehat{\boldsymbol{\theta}})}\left(x_1, x_2, ..., x_n, \widehat{\theta} \left| \theta \right.\right)}{f_{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}}\left(\widehat{\theta} \left| \theta \right.\right)},$$

for all $\widehat{\theta}$ and θ with $f_{\widehat{\boldsymbol{\theta}}|\boldsymbol{\theta}}\left(\widehat{\theta} \left| \theta \right.\right) \neq 0$, does not depend on θ.

- **Factorization theorem.** Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution is discrete (absolutely continuous) and depends on the random parameter $\boldsymbol{\theta} : (\Omega, \mathcal{F}) \longrightarrow \left(\mathbb{R}^K, \mathcal{B}\left(\mathbb{R}^K\right)\right)$ with $\boldsymbol{\theta}(\Omega) = \Theta \subset \mathbb{R}^K$. Then, the statistic $\widehat{\boldsymbol{\theta}} = t(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ is a sufficient estimator (or statistic) for the random parameter $\boldsymbol{\theta}$ if and only if, for all values θ taken by $\boldsymbol{\theta}$, the probability function (density) of $\tilde{X}$ can be factorized as follows:

$$f_{\tilde{X}}(x_1, x_2, ..., x_n | \theta) = h(x_1, x_2, ..., x_n) \cdot g\left(\widehat{\theta}, \theta\right),$$

where $\widehat{\theta} = t(x_1, x_2, ..., x_n)$.

8.11. Interval estimation: the pivotal method

- Let $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ be a collection of random variables (or sample) whose joint distribution depends on the parameter $\theta \in \Theta \subset \mathbb{R}$.
- Interval estimation is the use of a sample $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n)$ to calculate an interval $\left(a\left(\tilde{X}\right), b\left(\tilde{X}\right)\right)$ of probable values of the unknown parameter θ.
- **Pivotal method:** Find a real-valued random variable $\tilde{w} = g\left(\tilde{X}, \theta\right)$, called the pivotal variable (or pivot), such that its distribution is independent of the parameter θ we want to estimate.
- A $(1-\alpha)100\%$ confidence (Borel) set $\widehat{C}_{1-\alpha}$ for $g\left(\tilde{X}, \theta\right)$ is given by

$$P_{g\left(\tilde{X},\theta\right)}\left(\widehat{C}_{1-\alpha}\right) = P\left\{g\left(\tilde{X}, \theta\right) \in \widehat{C}_{1-\alpha}\right\} = 1-\alpha.$$

- If $\widehat{C}_{1-\alpha}$ is an interval (c, d), then a $(1-\alpha)100\%$ confidence interval for $g\left(\tilde{X}, \theta\right)$ is given by

$$P\left\{c < g\left(\tilde{X}, \theta\right) < d\right\} = 1 - \alpha. \qquad (*)$$

- Since there are many confidence intervals, we usually make

$$P\left\{g\left(\tilde{X}, \theta\right) \geq d\right\} = P\left\{g\left(\tilde{X}, \theta\right) \leq c\right\} = \frac{\alpha}{2}.$$

- Then, we solve for θ in $(*)$ to get a random $(1-\alpha)100\%$ confidence set $H\left(\tilde{X}, c, d\right)$ for θ,

$$P\left\{\theta \in H(\tilde{X}, c, d)\right\} = 1 - \alpha.$$

- If the set $H(\tilde{X}, c, d)$ is a random interval $\left(a\left(\tilde{X}\right), b\left(\tilde{X}\right)\right)$, then

$$P\left\{a\left(\tilde{X}\right) < \theta < b\left(\tilde{X}\right)\right\} = 1 - \alpha.$$

- This random interval is called the $(1-\alpha)100\%$ confidence interval for the parameter θ.

- Usually, the $(1-\alpha)100\%$ confidence interval for θ is presented like this:

$$a(X) < \theta < b(X),$$

where X is a particular value of the sample. Therefore, we are just providing a "realization" of the random $(1-\alpha)100\%$ confidence interval.

- **Examples:**
- Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample from a normal population with mean μ and variance σ^2
- **(1)** If σ^2 is known, the random variable $\dfrac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} \sim \mathrm{N}(0, 1)$. Therefore,

$$P\left\{-z_{\alpha/2} < \frac{\bar{\mathbf{x}} - \mu}{\sigma / \sqrt{n}} < z_{\alpha/2}\right\} = 1 - \alpha,$$

where $z_{\alpha/2} = 2.576$ if $1 - \alpha = 0.99$; $z_{\alpha/2} = 1.96$ if $1 - \alpha = 0.95$; and $z_{\alpha/2} = 1.645$ if $1 - \alpha = 0.9$. Then,

$$P\left\{\bar{\mathbf{x}} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{\mathbf{x}} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right\} = 1 - \alpha.$$

Thus, the $(1 - \alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},$$

where $\overline{x}$ is the value of the sample mean $\bar{\mathbf{x}}$.

- **(2)** If σ^2 is unknown, the random variable $\dfrac{\bar{\mathbf{x}} - \mu}{\mathbf{s}/\sqrt{n}} \sim \mathrm{t}_{n-1}$. Therefore,

$$P\left\{-t_{\alpha/2,n-1} < \frac{\bar{\mathbf{x}} - \mu}{\mathbf{s}/\sqrt{n}} < t_{\alpha/2,n-1}\right\} = 1 - \alpha.$$

Then,

$$P\left\{\bar{\mathbf{x}} - t_{\alpha/2,n-1} \cdot \frac{\mathbf{s}}{\sqrt{n}} < \mu < \bar{\mathbf{x}} + t_{\alpha/2,n-1} \cdot \frac{\mathbf{s}}{\sqrt{n}}\right\} = 1 - \alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}},$$

where $\overline{x}$ is the value of the sample mean $\bar{\mathbf{x}}$ and s is the value of the sample standard deviation $\left(\mathbf{s} = \left(\mathbf{s^2}\right)^{1/2}\right)$.

- **(3)** The random variable $\dfrac{(n-1)\mathbf{s^2}}{\sigma^2} \sim \chi^2_{n-1}$. Therefore,

$$P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)\mathbf{s^2}}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\frac{(n-1)\mathbf{s^2}}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)\mathbf{s^2}}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right\} = 1-\alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}},$$

where s^2 is the value of the sample variance $\mathbf{s^2}$.

- Let $\mathbf{s_1^2}$ and $\mathbf{s_2^2}$ be the variances of independent random samples of size n_1 and n_2 from two normal populations with variances σ_1^2 and σ_2^2, respectively.
- **(4)** The random variable $\dfrac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} \sim \mathrm{F}_{n_1-1,n_2-1}$. Therefore,

$$P\left\{F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\mathbf{s_1^2}/\sigma_1^2}{\mathbf{s_2^2}/\sigma_2^2} < F_{\alpha/2,n_1-1,n_2-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\frac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \cdot \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{\mathbf{s_1^2}}{\mathbf{s_2^2}} \cdot \frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}\right\} = 1-\alpha.$$

Thus, the $(1-\alpha)100\%$ confidence interval for σ_1^2/σ_2^2 is given by

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}}$$

or, since

$$F_{\alpha/2,n_2-1,n_1-1}=\frac{1}{F_{1-\frac{\alpha}{2},n_1-1,n_2-1}},$$ (see the handout)

the confidence interval becomes

$$\frac{s_1^2}{s_2^2}\cdot\frac{1}{F_{\alpha/2,n_1-1,n_2-1}}<\frac{\sigma_1^2}{\sigma_2^2}<\frac{s_1^2}{s_2^2}\cdot F_{\alpha/2,n_2-1,n_1-1},$$

where s_1 and s_2 are the values of the standard deviations of the two random samples.

8.11. Confidence intervals for means, for differences between means, for proportions, for differences between proportions, for variances, and for ratios of two variances

- See the handout.

Cramér-Rao lower bound

Theorem (Cramér-Rao). Let $\tilde{X}$ be a n-dimensional vector of random variables (not necessarily independent), the joint density of which is given by $h(X;\theta)$, where θ is a K-dimensional vector of parameters in some parameter space Θ. Let $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ be an unbiased estimator for θ with a finite covariance matrix $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$. Furthermore, assume that $h(X;\theta)$ and $g(X)$ satisfy

(a)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} h dX = \int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta} dX \;\leftarrow \text{This is a } K\text{-dimensional column vector}$$

(b)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} g h dX = \int_{\mathbb{R}^n} g\frac{\partial h}{\partial\theta} dX \;\leftarrow \text{This is a } K\text{-dimensional column vector}$$

(c)

$$\frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} \frac{\partial h}{\partial\theta^T} dX = \int_{\mathbb{R}^n} \frac{\partial^2 h}{\partial\theta\partial\theta^T} dX \;\leftarrow \text{This is a } K\times K\text{-dimensional matrix}$$

(d)

$$\mathrm{Var}\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right) \text{ is a positive definite } K\times K \text{ matrix.}$$

Then, for all $\theta\in\Theta$,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{CR-1}$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(\mathrm{E}\left[\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\cdot\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{CR-2}$$

Proof for the case $K = 1$. We have $1 = \int_{\mathbb{R}^n} h(X;\theta)dX$ for all $\theta\in\Theta$, where $X = (x_1, x_2, \ldots, x_n)\in\mathbb{R}^n$. Therefore, differentiating both sides with respect to θ and using assumption (a), we get

$$0 = \int_{\mathbb{R}^n} \frac{\partial h(X;\theta)}{\partial\theta} dX = \int_{\mathbb{R}^n} \frac{\partial \ln h(X;\theta)}{\partial\theta} h(X;\theta) dX \tag{1}$$

Define $\tilde{z} = \dfrac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}$, then (1) means that $\mathrm{E}(\tilde{z}) = 0$. Furthermore, $\mathrm{Var}(\tilde{z}) = \mathrm{E}(\tilde{z}^2) = \mathrm{E}\left[\left(\dfrac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}\right)^2\right]$. Since $\widehat{\boldsymbol{\theta}} = g\left(\tilde{X}\right)$ is an unbiased estimator for θ, we have

$$\theta = \mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right) \iff \theta = \int_{\mathbb{R}^n} \underbrace{g(X)}_{=\widehat{\theta}} h(X;\theta)dX.$$

Therefore, differentiating both sides with respect to θ and using assumption (b), we get

$$1 = \int_{\mathbb{R}^n} g(X)\frac{\partial h(X;\theta)}{\partial \theta}dX = \int_{\mathbb{R}^n} g(X)\frac{\partial \ln h(X;\theta)}{\partial \theta}h(X;\theta)dX. \tag{2}$$

Note that (2) means that $\mathrm{E}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) = 1$. Thus, we have

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) = \mathrm{E}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right) - \underbrace{\mathrm{E}\left(\widehat{\boldsymbol{\theta}}\right)}_{=\theta}\underbrace{\mathrm{E}(\tilde{z})}_{=0} = 1.$$

The Cauchy-Schwarz inequality implies that the absolute value of the correlation coefficient is smaller than 1 so that

$$\underbrace{\left[\mathrm{Cov}\left(\widehat{\boldsymbol{\theta}} \cdot \tilde{z}\right)\right]^2}_{=1} \leq \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \cdot \mathrm{Var}\left(\tilde{z}\right).$$

Note that $[\mathrm{Var}\left(\tilde{z}\right)]^{-1}$ is well defined if assumption (d) holds. Therefore, we obtain the inequality (CR-2) since

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq [\mathrm{Var}\left(\tilde{z}\right)]^{-1} = \left(\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial \theta}\right)^2\right]\right)^{-1}. \tag{3}$$

To get the inequality (CR-1) observe that

$$\frac{\partial \ln h(X;\theta)}{\partial \theta} = \frac{1}{h(X;\theta)}\frac{\partial h(X;\theta)}{\partial \theta} \tag{4}$$

so that

$$\left(\frac{\partial \ln h(X;\theta)}{\partial \theta}\right)^2 = \frac{1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2$$

and

$$\mathrm{E}\left[\left(\frac{\partial \ln h(X;\theta)}{\partial \theta}\right)^2\right] = \mathrm{E}\left[\frac{1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2\right]. \tag{5}$$

Differentiating (4) with respect to θ, we get

$$\frac{\partial^2 \ln h(X;\theta)}{\partial \theta^2} = \frac{-1}{[h(X;\theta)]^2}\left(\frac{\partial h(X;\theta)}{\partial \theta}\right)^2 + \frac{1}{h(X;\theta)}\frac{\partial^2 h(X;\theta)}{\partial \theta^2}.$$

Therefore, taking the expectation of the previous equation and using (5), we get

$$\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \mathrm{E}\left[\frac{1}{h(\tilde{X};\theta)}\frac{\partial^2 h(\tilde{X};\theta)}{\partial\theta^2}\right]$$

$$= -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \int_{\mathbb{R}^n} \frac{1}{h(X;\theta)}\frac{\partial^2 h(X;\theta)}{\partial\theta^2} h(X;\theta)dX$$

$$= -\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] + \int_{\mathbb{R}^n} \frac{\partial^2 h(X;\theta)}{\partial\theta^2} dX, \tag{6}$$

and using assumption (c), we get

$$\int_{\mathbb{R}^n} \frac{\partial^2 h(X;\theta)}{\partial\theta^2} dX = \frac{\partial}{\partial\theta} \underbrace{\int_{\mathbb{R}^n} \frac{\partial h(X;\theta)}{\partial\theta} dX}_{=0} = 0, \tag{7}$$

where the last equality follows from differentiating both sides of the first equality in (1) with respect to θ. Therefore, combining (6) and (7) we obtain

$$-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = \mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right]. \tag{8}$$

Plugging (8) into (3) we obtain the inequality (CR-1). *Q.E.D.*

Corollary. If in the previous theorem the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_n)$ is a random sample from a population $\tilde{x}$ having the density (or probability function) $f(x;\theta)$, then

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\tilde{x};\theta)}{\partial\theta\partial\theta^T}\right]\right)^{-1} \tag{i}$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geqq \left(n\mathrm{E}\left[\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta} \cdot \frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta^T}\right]\right)^{-1}. \tag{ii}$$

Proof for the case $K = 1$. Since $\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n$ are independent, we have

$$h(X;\theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta),$$

so that

$$\ln h(X;\theta) = \sum_{i=1}^{n} \ln f(x_i;\theta),$$

which implies that

$$\frac{\partial^2 \ln h(X;\theta)}{\partial\theta^2} = \sum_{i=1}^{n} \frac{\partial^2 \ln f(x_i;\theta)}{\partial\theta^2},$$

and, hence,

$$\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right] = n\mathrm{E}\left[\frac{\partial^2 \ln f(\widetilde{x};\theta)}{\partial\theta^2}\right],$$

which combined with (CR-1) proves inequality (i) in the statement of the Corollary.

Let us prove inequality (ii). We have that

$$\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] = \mathrm{E}\left[\left(\frac{\partial\left[\sum_{i=1}^n \ln f(\widetilde{x}_i;\theta)\right]}{\partial\theta}\right)^2\right] = \mathrm{E}\left[\left(\sum_{i=1}^n \frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2\right]$$

$$= \mathrm{E}\left[\sum_{i=1}^n \left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2 + \sum_{j\neq i}\sum_{i=1}^n \left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}\right)\right]$$

$$= \sum_{i=1}^n \left[\mathrm{E}\left(\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right)^2\right] + \sum_{j\neq i}\sum_{i=1}^n \mathrm{E}\left[\underbrace{\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\cdot\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}}_{\text{independent random variables}}\right]$$

$$= n\mathrm{E}\left[\left(\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right)^2\right] + \sum_{j\neq i}\sum_{i=1}^n \underbrace{\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right]}_{=0}\cdot\underbrace{\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x}_j;\theta)}{\partial\theta}\right]}_{=0}$$

since we know from (1) that

$$0 = \mathrm{E}\left[\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right] = \mathrm{E}\left[\sum_{i=1}^n \frac{\partial \ln f(\widetilde{x}_i;\theta)}{\partial\theta}\right] = n\mathrm{E}\left[\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right].$$

Therefore, we have proved that

$$\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right] = n\mathrm{E}\left[\left(\frac{\partial \ln f(\widetilde{x};\theta)}{\partial\theta}\right)^2\right], \tag{9}$$

which combined with (CR-2) proves inequality (ii) in the statement of the Corollary. *Q.E.D.*

Single parameter case: $K = 1$ ($\theta \in \mathbb{R}$)**.**

(a) For $\tilde{X} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$, not necessarily a random sample, we have

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-\mathrm{E}\left[\frac{\partial^2 \ln h(\tilde{X};\theta)}{\partial\theta^2}\right]\right)^{-1} \qquad \text{(Theorem CR-1 for } K = 1)$$

or, equivalently,

$$\mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(\mathrm{E}\left[\left(\frac{\partial \ln h(\tilde{X};\theta)}{\partial\theta}\right)^2\right]\right)^{-1}. \qquad \text{(Theorem CR-2 for } K = 1)$$

(b) For the random sample $\tilde{X} = (\widetilde{x}_1, \widetilde{x}_2, \ldots, \widetilde{x}_n)$, we have

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(-n\mathrm{E}\left[\frac{\partial^2 \ln f(\widetilde{x};\theta)}{\partial\theta^2}\right]\right)^{-1} \qquad \text{(Corollary (i) for } K = 1)$$

or, equivalently,

$$\operatorname{Var}\left(\widehat{\boldsymbol{\theta}}\right) \geq \left(n\mathrm{E}\left[\left(\frac{\partial \ln f(\tilde{x};\theta)}{\partial\theta}\right)^2\right]\right)^{-1}. \qquad \text{(Corollary (ii) for } K = 1)$$

Interval Estimation

Confidence intervals for means

$$\frac{\overline{\boldsymbol{x}}-\mu}{\sigma/\sqrt{n}} \sim \mathrm{N}(0,1) \Rightarrow P\left\{-z_{\alpha/2} < \frac{\overline{\boldsymbol{x}}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right\} = 1-\alpha,$$

$$\text{where } \int_{z_{\alpha/2}}^{\infty} n(x;0,1)dx = \frac{\alpha}{2}.$$

Then,

$$P\left\{\overline{\boldsymbol{x}} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{\boldsymbol{x}} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha.$$

Proposition 1. If $\overline{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, a $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}.$$

Remark 1. By virtue of the central limit theorem, this result can be also used for random samples from non-normal populations with the known variance σ^2, provided that n is sufficiently large ($n \geq 30$). If σ is unknown but $n \geq 30$, we can replace σ by the value s of the sample standard deviation $\boldsymbol{s} = (\boldsymbol{s^2})^{1/2}$.

– – – – – – – – – – – – – – –

$$\frac{\overline{\boldsymbol{x}}-\mu}{\boldsymbol{s}/\sqrt{n}} \sim \mathrm{t}_{n-1} \Rightarrow P\left\{-t_{\alpha/2,n-1} < \frac{\overline{\boldsymbol{x}}-\mu}{\boldsymbol{s}/\sqrt{n}} < t_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Then,

$$P\left\{\overline{\boldsymbol{x}} - t_{\alpha/2,n-1}\cdot\frac{\boldsymbol{s}}{\sqrt{n}} < \mu < \overline{\boldsymbol{x}} + t_{\alpha/2,n-1}\cdot\frac{\boldsymbol{s}}{\sqrt{n}}\right\} = 1-\alpha.$$

Proposition 2. If $\overline{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population with the unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for the population mean μ is given by

$$\overline{x} - t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2,n-1}\cdot\frac{s}{\sqrt{n}}.$$

Remark 2. Note that Proposition 2 applies to <u>small</u> samples ($n < 30$) from a normal population. For $n \geq 30$, the confidence interval of Proposition 2 and the one of Proposition 1 with σ replaced by s will generally yield nearly the same result.

Confidence intervals for differences between means

Let $\overline{\boldsymbol{x}}_1$ and $\overline{\boldsymbol{x}}_2$ be the means of independent random samples of size n_1 and n_2 from normal populations having means μ_1 and μ_2 and variances σ_1^2 and σ_2^2. Then, $\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}\right)$. It follows that

$$\frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1).$$

From Proposition 1 we get,

Proposition 3. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of independent random samples of size n_1 and n_2 from normal populations with the known variances σ_1^2 and σ_2^2, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

Remark 1 also applies to Proposition 3 for large samples ($n_1 \geq 30$ and $n_2 \geq 30$).

If we had two small independent random samples from normal populations having the same variance σ^2, which is unknown, then we must construct the "pooled" estimator

$$\boldsymbol{s}_p^{\boldsymbol{2}} = \frac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}} + (n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{n_1 + n_2 - 2}$$

which is, indeed, an unbiased estimator for σ^2. The independent random variables $\dfrac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}}}{\sigma^2}$ and $\dfrac{(n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{\sigma^2}$ have chi-square distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom and their sum

$$\frac{(n_1 - 1)\boldsymbol{s}_1^{\boldsymbol{2}}}{\sigma^2} + \frac{(n_2 - 1)\boldsymbol{s}_2^{\boldsymbol{2}}}{\sigma^2} = \frac{(n_1 + n_2 - 2)\boldsymbol{s}_p^{\boldsymbol{2}}}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Since

$$\frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \quad \text{and} \quad \frac{(n_1 + n_2 - 2)\boldsymbol{s}_p^{\boldsymbol{2}}}{\sigma^2}$$

are independent, it follows that

$$\frac{\dfrac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\sigma \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}}{\sqrt{\boldsymbol{s}_p^{\boldsymbol{2}}/\sigma^2}} = \frac{(\overline{\boldsymbol{x}}_1 - \overline{\boldsymbol{x}}_2) - (\mu_1 - \mu_2)}{\boldsymbol{s}_p \cdot \sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \sim \mathrm{t}_{n_1+n_2-2}.$$

From Proposition 2 we get,

Proposition 4. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of two independent random samples of size n_1 and n_2 from normal populations with unknown but equal variances, a $(1-\alpha)100\%$ confidence interval for $\mu_1 - \mu_2$ is given by

$$\begin{aligned}(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,n_1+n_2-2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} &< \mu_1 - \mu_2 \\ &< (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,n_1+n_2-2} \cdot s_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.\end{aligned}$$

Confidence intervals for proportions

Let $\tilde{x} \sim \mathrm{B}(n,\theta)$. Then $\dfrac{\tilde{x} - n\theta}{\sqrt{n\theta(1-\theta)}} \to \mathrm{N}(0,1)$ as $n \to \infty$.

Let $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{x}}{n}$ be the sample proportion. Then

$$\frac{\hat{\boldsymbol{\theta}} - \theta}{\sqrt{\dfrac{\theta(1-\theta)}{n}}} \to \mathrm{N}(0,1) \quad \text{as } n \to \infty.$$

Using Proposition 1 and Remark 1, we get,

Proposition 5. An approximate $(1-\alpha)100\%$ confidence interval (for large n) for the binomial parameter θ is given by

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

where $\hat{\theta} = \dfrac{x}{n}$.

Confidence intervals for differences between proportions

Using Proposition 3 we get,

Proposition 6. An approximate $(1-\alpha)100\%$ confidence interval (for large n) for the difference between two binomial parameters, $\theta_1 - \theta_2$, is given by

$$\begin{aligned}(\hat{\theta}_1 - \hat{\theta}_2) - z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}} &< \theta_1 - \theta_2 \\ &< (\hat{\theta}_1 - \hat{\theta}_2) + z_{\alpha/2}\sqrt{\frac{\hat{\theta}_1(1-\hat{\theta}_1)}{n_1} + \frac{\hat{\theta}_2(1-\hat{\theta}_2)}{n_2}}.\end{aligned}$$

where $\hat{\theta}_1 = \dfrac{x_1}{n_1}$ and $\hat{\theta}_2 = \dfrac{x_2}{n_2}$.

Confidence intervals for variances

If $\dfrac{(n-1)\boldsymbol{s^2}}{\sigma^2} \sim \chi^2_{n-1}$, then

$$P\left\{\chi^2_{1-\frac{\alpha}{2},n-1} < \frac{(n-1)\boldsymbol{s^2}}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right\} = 1-\alpha.$$

Proposition 7. If s^2 is the value of the variance of a random sample of size n from a normal population, a $(1-\alpha)100\%$ confidence interval for the population variance σ^2 is given by

$$\frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}.$$

Confidence intervals for ratios of two variances

If $\dfrac{\boldsymbol{s_1^2}/\sigma_1^2}{\boldsymbol{s_2^2}/\sigma_2^2} \sim \mathrm{F}_{n_1-1,n_2-1}$, then

$$P\left\{F_{1-\frac{\alpha}{2},n_1-1,n_2-1} < \frac{\boldsymbol{s_1^2}/\sigma_1^2}{\boldsymbol{s_2^2}/\sigma_2^2} < F_{\alpha/2,n_1-1,n_2-1}\right\} = 1-\alpha.$$

Lemma. $F_{1-\alpha,\nu_1,\nu_2} = \dfrac{1}{F_{\alpha,\nu_2,\nu_1}}$.

Proof. Exercise 8 of List 7 says that

$$\tilde{x} \sim \mathrm{F}_{\nu_1,\nu_2} \iff \tilde{y} \equiv \frac{1}{\tilde{x}} \sim \mathrm{F}_{\nu_2,\nu_1}.$$

Let $P\{\tilde{x} \geq F_{1-\alpha,\nu_1,\nu_2}\} = 1-\alpha$. Therefore,

$$P\left\{\frac{1}{\tilde{y}} \geq F_{1-\alpha,\nu_1,\nu_2}\right\} = 1-\alpha \iff P\left\{\tilde{y} \geq \frac{1}{F_{1-\alpha,\nu_1,\nu_2}}\right\} = \alpha.$$

Since $P\{\tilde{y} \geq F_{\alpha,\nu_2,\nu_1}\} = \alpha$, we arrive at the desired conclusion. *Q.E.D.*

Proposition 8. If s_1^2 and s_2^2 are the values of the variances of independent random samples of size n_1 and n_2 from two normal populations, a $(1-\alpha)100\%$ confidence interval for the ratio σ_1^2/σ_2^2 of the two population variances is given by

$$\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2,n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2,n_2-1,n_1-1}. \qquad \text{(Check it!)}$$

Exercises. Probability and Statistics. IDEA.
8. Estimation

1. Show that if $\hat{\boldsymbol{\theta}}$ is an unbiased estimator for θ and $\mathrm{Var}(\hat{\boldsymbol{\theta}})$ does not equal 0, then $\hat{\boldsymbol{\theta}}^{\mathbf{2}}$ is not an unbiased estimator for θ^2.

2. If $\tilde{x}_1$, $\tilde{x}_2$,..., and $\tilde{x}_n$ are independent Bernoulli random variables with the same parameter θ, show that the sample proportion $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{y}}{n}$, where $\tilde{y} = \sum_{i=1}^{n} \tilde{x}_i$, is a minimum variance unbiased estimator for the parameter θ. Note that $\tilde{y}$ is a binomial random variable with the parameters n and θ.

3. Let $\bar{\boldsymbol{x}}_{\mathbf{1}}$ be the mean of a random sample of size n from a normal population with the mean μ and the variance σ_1^2, and $\bar{\boldsymbol{x}}_{\mathbf{2}}$ be the mean of a random sample of size n from a normal population with the mean μ and the variance σ_2^2. If the two random samples are independent, show that

   (a) $w \cdot \bar{\boldsymbol{x}}_{\mathbf{1}} + (1-w) \cdot \bar{\boldsymbol{x}}_{\mathbf{2}}$, where $0 \leq w \leq 1$, is an unbiased estimator for μ;

   (b) the variance of the estimator is a minimum when $w = \dfrac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}$.

4. If $\bar{\boldsymbol{x}}_{\mathbf{1}}$ and $\bar{\boldsymbol{x}}_{\mathbf{2}}$ are the means of independent random samples of size n_1 and n_2 from a normal population with the mean μ and the variance σ^2, show that the variance of the unbiased estimator $w \cdot \bar{\boldsymbol{x}}_{\mathbf{1}} + (1-w) \cdot \bar{\boldsymbol{x}}_{\mathbf{2}}$ is a minimum when $w = \dfrac{n_1}{n_1 + n_2}$.

5. Show that the mean of a random sample of size n from an exponential population $\tilde{x}$ is a weakly consistent estimator for its parameter θ.

6. With reference to Exercise 5 show that the sample mean is a sufficient estimator for the exponential parameter θ. Show it in two ways: (a) without using the factorization theorem and (b) making use of the factorization theorem.

7. With reference to Exercise 2 show that the statistic $\hat{\boldsymbol{\theta}} = \dfrac{\tilde{y}}{n}$ is a sufficient estimator for θ. Show it in two ways: (a) without using the factorization theorem and (b) making use of the factorization theorem.

8. Use the method of moments to find an estimator for the parameter θ of the uniform density on the interval $(0, \theta)$.

9. If x_1, x_2,..., and x_n are the values of a random sample of size n from a population $\tilde{x}$ having the density

$$f(x;\theta) = \begin{cases} \dfrac{2(\theta - x)}{\theta^2} & \text{for } 0 < x < \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

   find an estimator for θ by the method of moments.

10. (a) Given a random sample of size n from a normal population with the mean μ and the variance σ^2, find joint maximum likelihood estimators for these two parameters.

    (b) Given a random sample of size n from a normal population with the known mean μ, find the maximum likelihood estimator for σ.

11. If x_1, x_2,..., and x_n are the values of a random sample of size n from a geometric population $\tilde{x}$, find an estimate of its parameter θ using

    (a) the method of moments;

    (b) the method of maximum likelihood.

12. Given a random sample of size n from a population $\tilde{x}$ having the density

$$f(x;\theta) = \begin{cases} e^{-(x-\theta)} & \text{for } x > \theta \\ \\ 0 & \text{elsewhere,} \end{cases}$$

    find an estimator for the parameter θ by the method of maximum likelihood.

13. Given independent random samples $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$, and $\{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_n\}$ from two normal populations having the means $\mu_1 = \alpha + \beta$ and $\mu_2 = \alpha - \beta$ and the common variance $\sigma^2 = 1$, find simultaneous maximum likelihood estimators for α and β.

14. (a) Prove that, if $\tilde{x}$ is a binomial random variable with the parameters $\boldsymbol{\theta}$ and n, where n is known, and the prior distribution of its parameter $\boldsymbol{\theta}$ is a beta distribution with parameters α and β, then the posterior distribution of $\boldsymbol{\theta}$ given $\tilde{x} = x$ is a beta distribution with the parameters $x + \alpha$ and $n - x + \beta$.

    (b) Under the same distributional assumptions as in (a), find the Bayesian estimate $\widehat{\theta}$ for the parameter value θ when the loss function is $L\left(\theta, \widehat{\theta}\right) = \left(\theta - \widehat{\theta}\right)^2$ and we observe one realization x of the random variable $\tilde{x}$.

    (c) Is the previous Bayesian estimator $\hat{\boldsymbol{\theta}}$ a strongly consistent estimator for the true parameter value $\theta \in (0,1)$, i.e., does $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \theta$ when $n \longrightarrow \infty$?

15. (a) Prove that, if $\bar{\boldsymbol{x}}$ is the mean of a random sample of size n from a normal population with the known variance σ^2, and the prior distribution of the mean $\boldsymbol{\mu}$ is a normal distribution with the mean μ_0 and the variance σ_0^2, then the posterior distribution of $\boldsymbol{\mu}$ given $\bar{\boldsymbol{x}} = \bar{x}$ is a normal distribution with the mean μ_1 and the variance σ_1^2, where

$$\mu_1 = \frac{n\bar{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \frac{1}{\sigma_1^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

    Write the previous formulas in terms of the precisions of the previous distributions, $\tau \equiv 1/\sigma^2$, $\tau_0 \equiv 1/\sigma_0^2$, and $\tau_1 \equiv 1/\sigma_1^2$.

(b) Under the same distributional assumptions as in (a), find the Bayesian estimate $\hat{\mu}$ for the value μ of the population mean when the loss function is $L(\mu, \hat{\mu}) = (\mu - \hat{\mu})^2$ and we observe the value $\bar{x}$ of the mean $\bar{\boldsymbol{x}}$ of a random sample of size n.

16. If a random sample of size $n = 20$ from a normal population with the variance $\sigma^2 = 225$ has the mean $\bar{x} = 64.3$, construct a 95% confidence interval for the population mean μ.

17. A paint manufacturer wants to determine the average drying time of a new interior wall paint. The drying time of a given area is normally distributed. If for 12 test areas of equal size he obtained a mean drying time of 66.3 minutes and a standard deviation of 8.4 minutes, construct a 95% confidence interval for the true mean μ.

18. Construct a 94% confidence interval for the actual difference between the average lifetimes of two kinds of light bulbs, given that a random sample of 40 light bulbs of one kind lasted on the average 418 hours of continuous use and 50 light bulbs of another kind lasted on the average 402 hours. The population standard deviations are known to be $\sigma_1 = 26$ and $\sigma_2 = 22$.

19. A study has been made to compare the nicotine contents of two brands of cigarettes. Ten cigarettes of Brand A had an average nicotine content of 3.1 milligrams with the standard deviation of 0.5 milligram, while eight cigarettes of Brand B had an average nicotine content of 2.7 milligrams with a standard deviation of 0.7 milligram. Assuming that the two sets of data are random samples from normal populations with equal variances, construct a 95% confidence interval for the true difference in the average nicotine content of the two brands of cigarettes.

20. If x is a value of a random variable having an exponential distribution, find k so that the interval from 0 to kx is a $(1 - \alpha)\,100\%$ confidence interval for the parameter θ.

21. If x_1 and x_2 are the values of a random sample of size 2 from a population having a uniform density on the interval $(0, \theta)$, find k so that

$$0 < \theta < k\,(x_1 + x_2)$$

is a $(1 - \alpha)\,100\%$ confidence interval for θ with $\alpha < 1/2$.

22. Measurements of the blood pressure of 25 elderly women have a mean of $\bar{x} = 140$ mm of mercury. If these data can be looked upon as a random sample from a normal population with $\sigma = 10$ mm of mercury, construct a 95% confidence interval for the population mean μ.

23. A study is being made to estimate the proportion of voters in a sizeable community who favor the construction of a nuclear plant. If it is found that only 140 of 400 voters selected at random favor the project, find a 95% confidence interval for the proportion of all voters in this community who favor the project.

24. If it is found that 132 of 200 voters in District A favor a given candidate for election to the United States Senate and 90 of 150 voters in District B favor the same candidate, find a 99% confidence interval for $\theta_1 - \theta_2$, the difference between the actual proportions of voters from the two districts who favor the candidate.

25. In 16 test runs the gasoline consumption of an experimental engine had a standard deviation of 2.2 gallons. Assume that the observed data can be looked upon as a random sample from a normal population. Construct a 99% confidence interval for σ^2, measuring the true variability of the gasoline consumption of this engine.

26. With reference to Exercise 19, find a 98% confidence interval for $\dfrac{\sigma_1^2}{\sigma_2^2}$.

27. Let $\tilde{X} = \{\tilde{x}_1, \tilde{x}_2, \tilde{x}_3\}$ be a random sample from a Bernoulli population $\tilde{x}$ with parameter θ. Recall that θ is the probability of success in a Bernoulli trial. Consider the statistic $\tilde{y} = \dfrac{1}{6}\left(\tilde{x}_1 + 2\tilde{x}_2 + 3\tilde{x}_3\right)$.

    (a) Find the probability function $f(x_1, x_2, x_3, y)$ of the random vector $(\tilde{x}_1, \tilde{x}_2, \tilde{x}_3, \tilde{y})$.

    (b) Find the probability function $f_{\tilde{y}}(y)$ of the random variable $\tilde{y}$.

    (c) Find the marginal probability function $h(x_2, y)$ of the random vector $(\tilde{x}_2, \tilde{y})$.

    (d) Find the covariance between the random variables $\tilde{x}_2$ and $\tilde{y}$, $\mathrm{Cov}\left(\tilde{x}_2, \tilde{y}\right)$.

    (e) Find the conditional probability function of the random vector $\tilde{X} = (\tilde{x}_1, \tilde{x}_2, \tilde{x}_3)$ given $\tilde{y} = \frac{1}{2}$ evaluated at $X = (1, 1, 0)$, $f_{\tilde{X}|\tilde{y}}\left(1, 1, 0 \,\middle|\, \frac{1}{2}\right)$. Is the statistic $\tilde{y}$ a sufficient estimator for θ?

    (f) Is the statistic $\tilde{y}$ an unbiased estimator for θ? Is the statistic $\tilde{y}$ an efficient estimator for θ? (i.e., is $\tilde{y}$ and unbiased estimator whose variance equals the Cramér-Rao lower bound?) Is the statistic $\tilde{y}$ a minimum variance unbiased estimator for θ? To answer this question compare the variance of $\tilde{y}$ with the variance of the sample mean. Is the Cramér-Rao lower bound reached by the sample mean?

28. Let $\tilde{x}_n$ be a random variable defined as follows:

$$\tilde{x}_n = \begin{cases} \tilde{z} & \text{with probability } (n-1)/n \\ \\ n & \text{with probability } 1/n, \end{cases}$$

    where $\tilde{z} \sim \mathrm{N}(0, 1)$. Compute $\operatorname{plim}_{n\to\infty} \tilde{x}_n$, $\lim_{n\to\infty} \mathrm{E}(\tilde{x}_n)$, and $\mathrm{AE}(\tilde{x}_n)$. Note that $\tilde{x}_n$ is a mixture.

29. Let $\tilde{y}_n$ be a random variable defined as follows:

$$\tilde{y}_n = \begin{cases} 0 & \text{with probability } (n-1)/n \\ \\ n^2 & \text{with probability } 1/n. \end{cases}$$

Compute $\underset{n\to\infty}{\text{plim}}\ \tilde{y}_n$, $\lim_{n\to\infty} \text{E}(\tilde{y}_n)$, and $\text{AE}(\tilde{y}_n)$.

30. We have obtained x "successes" in n identical and independent trials.

    (a) Find the value $\hat{\theta}_{ML}$ of the maximum likelihood estimator of the probability θ of success in each trial.

    (b) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ for θ is the most efficient (or the best) estimator in the class of unbiased estimators for θ.

    (c) Does the estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ converge in mean square to θ, $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{m} \theta$ as $n \to \infty$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{p} \theta$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{d} \theta$? Does $\hat{\boldsymbol{\theta}}_{\mathbf{ML}} \xrightarrow{a.s.} \theta$?

31. In the experiment of randomly and independently choosing n real numbers from the closed interval $[0, \beta]$, we have obtained the numbers $x_1, x_2, ..., x_n$. Let x^* be the largest of these numbers, $x^* = \max\{x_1, x_2, ..., x_n\}$.

    (a) Prove that x^* is the value of the maximum likelihood estimator for the parameter β.

    (b) Prove that the statistic $\tilde{x}^* = \max\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ is a weakly consistent estimator for β, that is, $\tilde{x}^* \xrightarrow{p} \beta$ as $n \to \infty$ (or $\underset{n\to\infty}{\text{plim}}\ \tilde{x}^* = \beta$).

    (c) Use the method of moments to find an estimator $\hat{\boldsymbol{\beta}}_{\mathbf{MM}}$ for the parameter β.

32. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$. Assume that the population distribution is normal with the density

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } x \in (-\infty, \infty)$$

    and that the population variance is known. Prove that the maximum likelihood estimator for the population mean is also a sufficient estimator. You can use the factorization theorem for your proof.

33. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ be a random sample of size n from a population $\tilde{x}$ having the Poisson distribution with a random parameter $\boldsymbol{\lambda}$. The prior distribution of $\boldsymbol{\lambda}$ is a gamma distribution with the parameters α and β.

    (a) Prove that the posterior distribution of $\boldsymbol{\lambda}$ given $(\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n) = (x_1, x_2, ..., x_n)$ is a gamma distribution with the parameters $\alpha + \sum_{i=1}^{n} x_i$ and $\dfrac{\beta}{1 + n\beta}$.

    (b) Find the Bayesian estimate $\hat{\lambda}_B$ for the parameter value λ when the loss function is quadratic,

$$L\left(\lambda, \widehat{\lambda}\right) = \left(\lambda - \widehat{\lambda}\right)^2,$$

    and we observe that $\bar{\boldsymbol{x}} = \bar{x}$, where $\bar{\boldsymbol{x}}$ is the mean of a random sample with size n and $\bar{x}$ is the realization of this sample mean.

    (c) Compute the following biases associated with the Bayesian estimator $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$:

(i) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \boldsymbol{\lambda} = \lambda\right]$, which is the bias when the true parameter value is λ, i.e., when $\boldsymbol{\lambda} = \lambda$. Note that $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \boldsymbol{\lambda} = \lambda\right] = \mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \lambda \middle| \boldsymbol{\lambda} = \lambda\right]$.

(ii) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda} \middle| \bar{\boldsymbol{x}} = \bar{x}\right]$, which is the posterior bias given a realization of the sample mean.

(iii) $\mathrm{E}\left[\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right]$, which is the prior bias.

(d) Answer the following questions about the mean square error associated with the Bayesian estimator $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$:

(i) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right]$, which is the mean square error when the true parameter value is λ. What is the limiting bias of $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$ as an estimator for the parameter value λ? Is $\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}}$ a weakly consistent estimator for λ? Note that $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right] = \mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \lambda\right)^2 \middle| \boldsymbol{\lambda} = \lambda\right]$.

(ii) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2 \middle| \bar{\boldsymbol{x}} = \bar{x}\right]$, which is the posterior mean square error given a realization of the sample mean.

(iii) Compute $\mathrm{E}\left[\left(\hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} - \boldsymbol{\lambda}\right)^2\right]$, which is the prior mean square error. Does $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\lambda}}_{\boldsymbol{B}} = \boldsymbol{\lambda}$?

34. (An application of Bayesian estimation) Imagine that you have a coin that might be unbalanced. Your prior distribution of the probability $\tilde{\theta}$ of getting a head when you flip once this coin is given by the uniform density on the interval $(0, 1)$. Then, you run the experiment of flipping the coin 7 times and you obtain 5 heads. What is the posterior distribution of the probability $\tilde{\theta}$ after observing the outcome of that experiment? Find the posterior expected probability of $\tilde{\theta}$ after observing the outcome of that experiment, i.e., compute the conditional expectation $\mathrm{E}\left(\tilde{\theta} \,|\tilde{x} = 5\right)$, where $\tilde{x}$ is the random number of heads when flipping the coin 7 times? Compare this posterior expectation with the prior (or unconditional) expectation.

35. Let $\{\tilde{x}_1, \tilde{x}_1, ..., \tilde{x}_n\}$ be a random sample of size n from a normal population $\tilde{x}$ with mean μ and variance σ^2. Consider the following two statistics:

(a) the sample variance,

$$\boldsymbol{s}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \overline{\boldsymbol{x}})^2}{n-1}$$

and

(*b*)

$$\hat{\boldsymbol{s}}^2 = \frac{\sum_{i=1}^{n} (\tilde{x}_i - \overline{\boldsymbol{x}})^2}{n},$$

where $\overline{\boldsymbol{x}}$ is the sample mean.

Is the statistic given in (a) a more efficient estimator for the population variance σ^2 than the statistic given in (b)?

36. Consider a random sample $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ of size n from a normal population $\tilde{x}$ with the known mean μ and the unknown variance $V > 0$.

(*a*) Find the Cramér-Rao lower bound on the variance of an unbiased estimator for the population variance V.

(*b*) Does the variance of the sample variance $\boldsymbol{s}_n^{\boldsymbol{2}}$ reach the Cramér-Rao lower bound? Is $\boldsymbol{s}_n^{\boldsymbol{2}}$ an efficient in the limit estimator for V?

Note: Recall that the parameter σ appearing in the normal density is the standard deviation of the population distribution and, thus, $\sigma \equiv V^{1/2} > 0$. Therefore, the normal density can be written as

$$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} = \frac{1}{(2\pi)^{1/2} \cdot V^{1/2}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{V}} \equiv f(x; V), \quad -\infty < x < \infty.$$

37. Let $\{\tilde{x}_i\}_{i=1}^n$ be a random sample of size n from a population $\tilde{x}$. Assume that the population distribution is Poisson with the parameter $\lambda > 0$.

(*a*) Find the the method of moments estimator for the parameter λ.

(*b*) Find the maximum likelihood estimator for the parameter λ.

(*c*) Prove that the sample mean is a sufficient estimator for the parameter λ. You can use the factorization theorem for your proof.

(*d*) Use the Cramér-Rao lower bound to check whether the sample mean is a minimum variance unbiased estimator for the parameter λ.

(*e*) The coefficient of skewness (or asymmetry) of a Poisson distribution is $\lambda^{-1/2}$. Find the maximum likelihood estimator for the coefficient of skewness of $\tilde{x}$.

38. Assume that the random variable $\tilde{x}$ has the **Laplace distribution**, whose density is:

$$f(x; \mu, \theta) = \frac{1}{2\theta} e^{-|x-\mu|/\theta}, \quad -\infty < x < \infty.$$

Note that in the exponent of the previous density we have the negative of the absolute value $|x - \mu|$ so that $-|x - \mu| = -(x - \mu)$ if $x \geq \mu$, whereas $-|x - \mu| = x - \mu$ if $x \leq \mu$.

(*a*) Draw the density of $\tilde{x}$ for the particular case where $\mu = -3$ and $\theta = 2$.

(*b*) Find the distribution function (cdf) of $\tilde{x}$. Is the cdf continuous? Is it differentiable everywhere? Is it twice differentiable everywhere?

(c) Find the moment generating function of $\tilde{x}$? Draw it for the particular case where $\mu = -3$ and $\theta = 2$.

(d) Find the mean and the variance σ^2 of $\tilde{x}$.

(e) Find the density of the random variable $\tilde{z} = |\tilde{x} - \mu|$. Find the mean of $\tilde{z}$.

(f) Let $\mu = -3$ and $\theta = 2$. Find the probability that $\tilde{x} \in (-4, 1)$ given that we know that the event $\tilde{x} \in (-5, -3)$ has occurred. *Note:* You can use for the computations of this part the distribution function you have found in part (b).

Assume for the rest of the exercise that you have a random sample $\{\tilde{x}_i\}_{i=1}^n$ of size n from the population $\tilde{x}$.

(g) Find the method of moments estimator for the parameter vector (μ, θ).

Assume from now on that the value of the parameter μ is known.

(h) Find the value of the Cramér-Rao lower bound for an unbiased estimator for the parameter θ of the distribution of $\tilde{x}$.

(i) Find the method of moments estimators $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}}$ and $\hat{\boldsymbol{\theta}}_{\mathbf{MM}}$ for the variance σ^2 of $\tilde{x}$ and for the parameter θ, respectively.

(j) Prove that the method of moments estimator $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}}$ is an unbiased estimator for σ^2. Is it consistent? (i.e., check if $\operatorname{plim}_{n\to\infty} \hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{MM}} = \sigma^2$ holds).

(k) Is $\hat{\boldsymbol{\theta}}_{\mathbf{MM}}$ an unbiased estimator for the parameter θ?

(l) Find the maximum likelihood estimators $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ and $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{ML}}$ for the parameter θ and for the variance σ^2 of $\tilde{x}$, respectively.

(m) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}_{\mathbf{ML}}$ is a biased estimator for the variance σ^2 of $\tilde{x}$. Is it unbiased in the limit?

(n) Prove that the maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\mathbf{ML}}$ is an unbiased estimator for θ. Is it efficient? Is it sufficient?

39. Let $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_n\}$ with $n > 1$ be a random sample from a Poisson population $\tilde{x}$ with parameter λ. Consider the following unbiased estimator for λ : $\hat{\boldsymbol{\lambda}} = \tilde{x}_1$. Improve this estimator in terms of mean square error using the Rao-Blackwell theorem and the result in part (c) of the previous Exercise 37.