Sampling Distributions

Sofia Olhede

EPFL

October 4, 2020

Sampling Distributions III

- We recall that t_{n-1} denotes the t distribution on $n-1$ degrees of freedom. The t distribution on k degrees of freedom takes the form

$$f_X(x,k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(k/2)\sqrt{k\pi}}\left(1+\frac{x^2}{k}\right)^{-\frac{k+1}{2}}, \quad x \in \mathbb{R}.$$

- Let us also assume $k > 2$ then the mean and variance of $X \sim t_k$ are

$$\mathbb{E}(X) = 0, \quad \mathbb{V}\mathrm{ar}(X) = \frac{k}{k-2}.$$

- Theorem: let $Y_1 \sim \chi^2_{d_1}$ and let $Y_2 \sim \chi^2_{d_2}$ be independent. Then

$$\frac{Y_1/d_1}{Y_2/d_2} \sim F_{d_1,d_2}.$$

Sampling Distributions IV

- A random variable follows the Fisher F distribution with integer parameters d_1 and d_2, written as $X \sim F_{d_1,d_2}$ if

$$f_X(x; d_1, d_2) = \frac{1}{B(d_1/2, d_2/2)}\left(\frac{d_1}{d_2}\right)^{d_1/2} x^{d_1/2-1}\left(1+\frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}, \quad x \geq 0.$$

The mean and variance in this case are for $d_2 > 4$

$$\mathbb{E}(X) = \frac{d_2}{d_2-2}, \quad \mathbb{V}\mathrm{ar}(X) = \frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-4)(d_2-2)^2}.$$

Sampling Distributions V

- Theorem: (Sampling from an Exponential Family).
  Let $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} f$ where
  $$f(y) = \exp\left\{ \sum_{j=1}^{k} \phi_j T_j(y) - \gamma(\phi_1, \cdots, \phi_k) + S(y) \right\}, \quad y \in \mathcal{Y},\ \phi \in \Phi.$$
  This is a density of a k parameter exponential family form. If Φ is open then

  (1) The minimal sufficient statistic for ϕ is $\boldsymbol{\tau}$ where
  $$\tau_j = \sum_{i=1}^{n} T_j(y_i).$$
  (2) The function γ is infinitely differentiable in all k of its variables and
  $$\mathbb{E}(\boldsymbol{\tau}) = n\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \quad \text{and} \quad \mathbb{C}\text{ov}(\boldsymbol{\tau}) = n\nabla^2_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}).$$

Sampling Distributions VI

- Unfortunately the sampling distribution of T is not always available in closed form.
- It may become easier to work with an approximation valid for large n.
- These approximations will be understood as a form of convergence of F.
- Definition: Convergence in Distribution (Weak Convergence). Let $\{F_n\}_{n\geq 1}$ be a sequence of distribution functions and let G be a distribution function on $\mathbb{R}$. We say that F_n converges weakly or in distribution to G and write $F_n \stackrel{\mathcal{L}}{\rightarrow} G$ whenever

$$F_n(y) \stackrel{n\rightarrow\infty}{\rightarrow} G(y),$$

  for all y constituting continuity points of G.
- A stronger notion of convergence corresponds to convergence in probability.

Sampling Distributions VI

- Definition (convergence in probability): When a sequence of random variables satisfies $\Pr\{\|Y_n - Y\| > \epsilon\} \to 0$ for all $\epsilon > 0$ and a given (random variable) Y, then we say that Y_n converges in probability to Y, and write $Y_n \xrightarrow{P} Y$.
- $\xrightarrow{\mathcal{L}}$ relates distribution functions. It says that the probabilistic behaviour of a sequence Y_n becomes more and more alike that of the limit Y.
- $\xrightarrow{P}$ relates random variables. It says that the actual realisations of Y_n can be progressively approximated with high probability by those of Y.
- Theorem: (a) $Y_n \xrightarrow{P} Y \Rightarrow Y_n \xrightarrow{\mathcal{L}} Y$.

  (b) $Y_n \xrightarrow{\mathcal{L}} c \Rightarrow Y_n \xrightarrow{p} c$ for $c \in \mathbb{R}$.
- Theorem (The Continuous Mapping Theorem)

  Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on the range of Y. Then

  $$\text{(a) } Y_n \xrightarrow{p} Y \Rightarrow g(Y_n) \xrightarrow{p} g(Y),$$

  $$\text{(b) } Y_n \xrightarrow{\mathcal{L}} Y \Rightarrow g(Y_n) \xrightarrow{\mathcal{L}} g(Y).$$

Limiting Distributions I

- Theorem (Slutsky's theorem): Let $X_n \xrightarrow{\mathcal{L}} X$ and let $Y_n \xrightarrow{\mathcal{L}} c$ where $c \in \mathbb{R}$. Then

  (a) $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$.

  (b) $X_n Y_n \xrightarrow{\mathcal{L}} Xc$.

- Theorem (General version of Slutsky's theorem): Let $g: \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be continuous and assume that $X_n \xrightarrow{\mathcal{L}} X$ and let $Y_n \xrightarrow{\mathcal{L}} c$ where $c \in \mathbb{R}$. Then $g(X_n, Y_n) \xrightarrow{\mathcal{L}} g(X, c)$ with increasing n.
- By applying the continuous mapping theorem and by applying Slutsky's theorem we can obtain new approximations. But they require a starting point.
- Theorem (Law of Large Numbers): let Y_n be independent random variables with $\mathbb{E}\, Y_k = \mu$ and $\mathbb{E}\,|Y_k| < \infty$ for all k. Then $n^{-1}(Y_1 + \cdots + Y_n \xrightarrow{p} \mu$.

Limiting Distributions II

- Theorem (Central Limit Theorem). Let $\{Y_n\}$ be a sequence of iid random variables with mean μ and variance σ^2 which is assumed finite. Then

$$\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu)) \xrightarrow{\mathcal{L}} N(0,\sigma^2).$$

- Theorem (Delta method): Let $Z_n = a_n(X_n-\theta) \xrightarrow{\mathcal{L}} Z$ where $a_n \in \mathbb{R}^+$ and $\theta \in \mathbb{R}$ for all n and assume $a_n \to \infty$. Let $g()$ be continuously differentiable at θ. Then

$$a_n\{g(X_n)-g(\theta)\} \xrightarrow{\mathcal{L}} g'(\theta)Z.$$

- Proof: Taylor expansion around θ gives

$$g(X_n) = g(\theta) + g'(\theta^*)\{X_n-\theta\}, \quad \|\theta-\theta^*\| < \|\theta - X_n\|.$$

(This is the Lagrange form of the remainder).

Limiting Distributions III

- Thus we may deduce

$$\|\theta - \theta^*\| < \|\theta - X_n\| = a_n^{-1}|a_n(\theta - X_n)| = a_n^{-1}|Z_n| \xrightarrow{P} 0. \tag{1}$$

The latter equation uses Slutsky's theorem. Thus $\theta^* \xrightarrow{P} 0$. By the continuous mapping theorem $g'(\theta^*) \xrightarrow{P} g'(\theta)$.
Therefore it follows

$$\begin{aligned} a_n\{g(X_n) - g(\theta)\} &= a_n\{g(\theta) + g'(\theta^*)(X_n - \theta) - g(\theta)\} \qquad (2) \\ &= g'(\theta^*)a_n(X_n - \theta) \xrightarrow{\mathcal{L}} g'(\theta)Z. \qquad (3) \end{aligned}$$

$\square$

Limiting Distributions IV

- We can apply these methods to derive sampling distributions.
- Corollary: Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} f$ where

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \quad x \in \mathcal{X},$$

with $\phi \in \Phi \subset \mathbb{R}$ and

$$\overline{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i) = n^{-1}\tau(X_1, \ldots, X_n).$$

If Φ is open then γ is infinitely differentiable and so

$$\sqrt{n}\big(\overline{T}_n - \gamma'(\phi)\big) \xrightarrow{\mathcal{L}} N(0, \gamma''(\phi)).$$

Limiting Distributions V

- Theorem (Weighted sum CLT): Let $\{W_n\}$ be an iid sequence of real random variables, with common mean 0 and variance 1. Let $\{\gamma_n\}$ be a sequence of real constants. Then

$$\sup_{1 \le j \le n} \frac{\gamma_j^2}{\sum_{i=1}^n \gamma_i^2} \to 0 \Longrightarrow \frac{1}{\sqrt{\sum_{i=1}^n \gamma_i^2}} \sum_{j=1}^n \gamma_j W_j \xrightarrow{\mathcal{L}} N(0,1).$$

- For joint convergence we need to consider random vectors.

Limiting Distributions VI

- Definitions. Let $\{\mathrm{Y}_n\}$ be a sequence of random vectors of $\mathbb{R}^d$ and Y a random vector of $\mathbb{R}^d$ with $\mathrm{Y}_n = \left(Y_n^{(1)} \quad \dots \quad Y_n^{(d)}\right)^T$ and $\mathrm{Y} = \left(Y^{(1)} \quad \dots \quad Y^{(d)}\right)^T$. Also define the CDFs F_{Y_n} and F_{Y}. We say that Y_n converges in distribution to Y as $n \to \infty$ (and write $\mathrm{Y}_n \xrightarrow{\mathcal{L}} \mathrm{Y}$) if for every continuity point of F_{Y} we have

$$F_{\mathrm{Y}_n}(\mathrm{y}) \overset{n\to\infty}{\longrightarrow} F_{\mathrm{Y}}(\mathrm{y}).$$

- Theorem (Cramér–Wold Device). Let $\{\mathrm{Y}_n\}$ be a sequence of random variables of $\mathbb{R}^d$ and let Y be a random vector of $\mathbb{R}^d$. Then

$$\mathrm{Y}_n \xrightarrow{\mathcal{L}} \mathrm{Y} \Leftrightarrow \mathrm{u}^T\mathrm{Y}_n \xrightarrow{\mathcal{L}} \mathrm{u}^T\mathrm{Y},\ \forall \mathrm{u} \in \mathbb{R}^d.$$

Limiting Distributions VII

- Continuous mapping theorem and Slutsky's lemma generalize to the vector case.
- In either case

(a) Continuous mapping: if $g : \mathbb{R}^p \to \mathbb{R}^d$ is continuous on the range of U and if $\mathrm{U}_n \xrightarrow{\mathcal{L}} \mathrm{U}$ in $\mathbb{R}^p$ then $g(\mathrm{U}_n) \xrightarrow{\mathcal{L}} g(\mathrm{U})$ in $\mathbb{R}^d$.
(b) Slutsky: if $g : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}^d$ is continuous and $\mathrm{U}_n \xrightarrow{\mathcal{L}} \mathrm{U}$ in $\mathbb{R}^p$ as well as $\mathrm{W}_n \xrightarrow{\mathcal{L}} \mathrm{u} \in \mathbb{R}^q$ for deterministic u then $g(\mathrm{U}_n, \mathrm{W}_n) \xrightarrow{\mathcal{L}} g(\mathrm{U}, \mathrm{u})$.

Limiting Distributions VIII

- Convergence in probability also easily generalizes to the vector case:
- Definition (Convergence in Probability (vectors)): Given a sequence of random vectors $\{Y_n\}$ in $\mathbb{R}^d$ satisfies $\Pr\{\|Y_n - Y\| > \epsilon\} \overset{n\to\infty}{\longrightarrow} 0$ for any $\epsilon > 0$ and a given random vector Y we say that Y_n converges in probability to Y and write $Y_n \overset{p}{\to} Y$.
- Theorem (Multivariate law of large numbers). Let $\{Y_n\}$ be iid random vectors with expectation $\boldsymbol{\mu}$ and if $\mathbb{E}\,\|Y_n\| < \infty$, for all $k \geq 1$,

$$\frac{1}{n}\sum_{k=1}^{n} Y_k \overset{p}{\to} \boldsymbol{\mu}.$$

- Theorem (Multivariate CLT). Let $\{X_n\}$ be a sequence of random vectors in $\mathbb{R}^d$ with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Omega}$ and define $\overline{X}_n$ as the mean of the vectors $X_1, \ldots X_n$. Then

$$\sqrt{n}(\overline{X}_n - \boldsymbol{\mu}) \overset{\mathcal{L}}{\to} Z \sim \mathcal{N}_d(0, \boldsymbol{\Omega}).$$

Limiting Distributions VIII

- Theorem: (Delta Method, vector case). Set $\mathrm{Z}_n \equiv a_n(\mathrm{X}_n - \mathrm{u}) \xrightarrow{\mathcal{L}} \mathrm{Z}$ in $\mathbb{R}^d$ where $a_n \in \mathbb{R}$, $\mathrm{u} \in \mathbb{R}^d$ and $a_n \to \infty$. Let $g : \mathbb{R}^d \to \mathbb{R}^p$ be continuously differentiable at u. Then

$$a_n\{g(\mathrm{X}_n) - g(\mathrm{u})\} \xrightarrow{\mathcal{L}} J_g(\mathrm{u})\mathrm{Z},$$

where $J_g(\mathrm{y})$ is the $p \times d$ Jacobian matrix of g,

$$J_g(\mathrm{y}) = \begin{pmatrix} \frac{\partial}{\partial x_1} g_1(\mathrm{y}) & \dots & \frac{\partial}{\partial x_d} g_1(\mathrm{y}) \\ \dots & \dots & \dots \\ \frac{\partial}{\partial x_1} g_p(\mathrm{y}) & \dots & \frac{\partial}{\partial x_d} g_p(\mathrm{y}) \end{pmatrix}. \tag{4}$$

Estimation I

- What is estimation ($\equiv$ "learning" in machine learning)?
- Imagine you assume Y is distributed according to $F(y_1, \ldots, y_n; \theta)$ where $y \in \mathcal{Y}^n$.
- Assume you know the form of $F(y_1, \ldots, y_n; \theta)$ but not the value of θ.
- Guessing θ on having observed $y_1, \ldots, y_n$ is estimation.
- Point estimation corresponds to producing a decent estimator of θ given we have observed $y_1, \ldots, y_n$.
- Or more formally we define a point estimator as a statistic with codomain Θ or $T : \mathcal{Y}^n \mapsto \Theta$.
- We usually write this as $\widehat{\theta}(Y_1, \ldots, Y_n)$. θ is deterministic and $\widehat{\theta}$ is stochastic.

Estimation II

- Not that whenever we realise a different set of $Y_1, \ldots, Y_n$ then we realise a different $\widehat{\theta}(Y_1, \ldots, Y_n)$.
- How do we design an estimator $\widehat{\theta}(Y_1, \ldots, Y_n)$?
- A good estimator would normally produce a value of $\widehat{\theta}(Y_1, \ldots, Y_n)$ near θ.
- We usually address this in terms of the mean and variance of $\widehat{\theta}(Y_1, \ldots, Y_n)$.
- The first interpretation of this is that "on average" we get the right value from $\widehat{\theta}(Y_1, \ldots, Y_n)$, and the spread of values obtained is not significant.
- This brings us round to the notion of the mean square error.

Estimation III

- Definition (mean square error): assume that $\widehat{\theta}$ is an estimator of the parameter θ corresponding to the model $F(y;\theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$. The mean square error of $\widehat{\theta}$ is then defined as

$$\mathrm{MSE}\{\widehat{\theta},\theta\} = \mathbb{E}\left[\|\widehat{\theta}-\theta\|^2\right] \tag{5}$$

- Lemma: (Mean Square Error Decomposition) The mean square error admits the decomposition

$$\mathbb{E}\left[\|\widehat{\theta}-\theta\|^2\right] = \|\,\mathbb{E}\,\widehat{\theta}-\theta\|^2 + \mathbb{E}\,\|\widehat{\theta}-\mathbb{E}\,\widehat{\theta}\|^2.$$

  This decomposition can colloquially be described as the mean square error of an estimator, is the bias square plus its variance.

Estimation IV

- The proof of this result is straightforward and just corresponds to

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\theta}-\theta\|^2\right] &= \mathbb{E}\left[\|\widehat{\theta}-\mathbb{E}\,\widehat{\theta}+\mathbb{E}\,\widehat{\theta}-\theta\|^2\right]\\ &= \mathbb{E}\left[\|\widehat{\theta}-\mathbb{E}\,\widehat{\theta}\|^2\right]+0+0+\left[\|\,\mathbb{E}\,\widehat{\theta}-\theta\|^2\right]. \qquad (6)\end{aligned}$$

- Why is the MSE important? Basically it transpires that the concentration of $\widehat{\theta}$ around θ can always be phrased in terms of the MSE.
- Lemma: Let $\widehat{\theta}$ be an estimator of $\theta \in \mathbb{R}^p$. For any $\epsilon > 0$

$$\mathrm{Pr}\left\{\|\widehat{\theta}-\theta\| > \epsilon\right\} \leq \frac{\mathrm{MSE}\{\widehat{\theta},\theta\}}{\epsilon^2}.$$

From this relationship we can note that $\mathrm{MSE}\{\widehat{\theta},\theta\} \to 0 \Rightarrow \widehat{\theta} \xrightarrow{p} \theta$.

Measures of Performance

- When an estimator has this property, then it is called consistent for θ.
- Definition: An estimator $\widehat{\theta}_n$ is consistent for parameter θ in terms of sample size n if $\widehat{\theta}_n \xrightarrow{p} \theta$.
- So once the MSE vanishes, consistency follows, but the converse is more tricky.
- Can we then always design a consistent estimator? Not necessarily!
- Definition: (Identifiability). A given model $F(y;\theta)$ is identifiable if given any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$ if it holds that

$$\theta_1 \neq \theta_2 \Rightarrow F(y;\theta_1) \neq F(y;\theta_2).$$

- If the model is not identifiable then you can get the same model with different values of the parameters.

Measures of Performance

- As an example consider $N(\mu_1 + \mu_2, \sigma^2)$. From observations we can only estimate $\mu_1 + \mu_2$.
- We will always assume identifiability unless we explicitly specify otherwise.
- We can use $\mathrm{MSE}\{\widehat{\theta}, \theta\}$ to determine which estimators are "better".
- But what is the best MSE given a problem? This is a difficult problem.
- A easier problem might be, among unbiased estimators (bias zero), can we make the MSE arbitrarily small?

Measures of Performance

- Assume that $Y_1, \ldots, Y_n$ have a joint density or probability mass function $f(\boldsymbol{y};\theta)$ depending on an unknown $\theta \in \mathbb{R}$.
- Let us try to find a function $\Lambda(\theta)$ such that
$$\mathbb{V}\mathrm{ar}\{\widehat{\theta}\} \geq \Lambda(\theta)\,\forall\theta \in \Theta,$$
with the choice of any $\widehat{\theta}$ such that $\mathbb{E}\{\widehat{\theta}\} = \theta$.
- To be able to study this question, we shall start with studying expectations with respect to y.
- We start from (with some regularity conditions thrown in):
$$\begin{aligned}\mathscr{Y}(\theta) &\equiv \mathbb{E}\{S(\mathrm{Y})\} \\ &= \int_{\mathcal{Y}^n} S(\mathrm{y})f(\mathrm{y};\theta)\,d\mathrm{y} \\ \frac{\partial}{\partial\theta}\mathscr{Y}(\theta) &= \frac{\partial}{\partial\theta}\int_{\mathcal{Y}^n} S(\mathrm{y})f(\mathrm{y};\theta)\,d\mathrm{y} = \int_{\mathcal{Y}^n} S(\mathrm{y})\frac{\partial}{\partial\theta}f(\mathrm{y};\theta)\,d\mathrm{y}.\end{aligned}$$

Measures of Performance

- Now with some further manipulation we note that

$$\begin{aligned}\frac{\partial}{\partial\theta}\mathscr{Y}(\theta) \equiv \frac{\partial}{\partial\theta}\mathbb{E}\{S(\mathrm{Y})\} &= \int_{\mathcal{Y}^n} S(\mathrm{y})\frac{\partial}{\partial\theta} f(\mathrm{y};\theta)\,dy \\ &= \int_{\mathcal{Y}^n} S(\mathrm{y})\frac{f(\mathrm{y};\theta)}{f(\mathrm{y};\theta)}\frac{\partial}{\partial\theta} f(\mathrm{y};\theta)\,dy \\ &= \int_{\mathcal{Y}^n} S(\mathrm{y}) f(\mathrm{y};\theta)\frac{\partial}{\partial\theta}\log(f(\mathrm{y};\theta))\,dy. \end{aligned} \tag{7}$$

- Now we can use this to derive further results, taking $U(\mathrm{y};\theta) = \frac{\partial}{\partial\theta}\log(f(\mathrm{y};\theta))$ and $S(\mathrm{y}) = 1$, we get

$$\begin{aligned}\mathbb{E}\{U\} &= \int_{\mathcal{Y}^n} f(\mathrm{y};\theta)\frac{\partial}{\partial\theta}\log(f(\mathrm{y};\theta))\,dy \\ &= \frac{\partial}{\partial\theta}\int_{\mathcal{Y}^n} f(\mathrm{y};\theta)\,dy = \frac{\partial}{\partial\theta}1 = 0. \end{aligned} \tag{8}$$

Measures of Performance

- Given $\mathbb{E}\{U\} = 0$ it follows that $\mathbb{V}\text{ar}\{U\} = \mathbb{E}\{U^2\}$. We note

$$\mathbb{V}\text{ar}\{U\} = \mathbb{E}\left\{ \left(\frac{\partial}{\partial\theta} \log(f(\mathrm{Y};\theta) \right)^2 \right\}. \tag{9}$$

- If we instead take $S(\mathrm{Y}) = \widehat{\theta}(\mathrm{Y})$ then we arrive at

$$\begin{aligned} \mathbb{C}\text{ov}\{\widehat{\theta}(\mathrm{Y}), U(\mathrm{Y})\} &= \mathbb{E}\{\widehat{\theta}(\mathrm{Y})U(\mathrm{Y})\} - \mathbb{E}\{\widehat{\theta}(\mathrm{Y})\}\,\mathbb{E}\{U(\mathrm{Y})\} \\ &= \mathbb{E}\{\widehat{\theta}(\mathrm{Y})U(\mathrm{Y})\}. \end{aligned} \tag{10}$$

Furthermore, we note that

$$\begin{aligned} \mathbb{E}\{\widehat{\theta}(\mathrm{Y})U(\mathrm{Y})\} &= \int_{\mathcal{Y}^n} \widehat{\theta}(\mathrm{y}) f(\mathrm{y};\theta) \frac{\partial}{\partial\theta} \log(f(\mathrm{y};\theta))\, d\mathrm{y} \\ &= \frac{\partial}{\partial\theta} \int_{\mathcal{Y}^n} \widehat{\theta}(\mathrm{y}) f(\mathrm{y};\theta)\, d\mathrm{y} \overset{\text{unbiased}}{=} \frac{\partial}{\partial\theta}\theta = 1. \end{aligned} \tag{11}$$

- What can we learn from this?

Measures of Performance

- The Cauchy–Schwarz inequality states that if we have any two random variables A and B then it follows that

$$\mathbb{V}\text{ar}\{A\}\,\mathbb{V}\text{ar}\{B\} \geq \mathbb{C}\text{ov}^2\{A, B\}. \tag{12}$$

Taking $A = \widehat{\theta}(\mathrm{Y})$ and $B = U(\mathrm{Y})$ we therefore arrive at:

$$\begin{aligned} \mathbb{V}\text{ar}\left\{\widehat{\theta}(\mathrm{Y})\right\}\mathbb{V}\text{ar}\{U(\mathrm{Y})\} &\geq \mathbb{C}\text{ov}^2\left\{\widehat{\theta}(\mathrm{Y}), U(\mathrm{Y})\right\} \\ \Rightarrow \mathbb{V}\text{ar}\left\{\widehat{\theta}(\mathrm{Y})\right\} &\geq \frac{1^2}{\mathbb{V}\text{ar}\{U(\mathrm{Y})\}} \\ &= \frac{1}{\mathbb{E}\left\{\left(\frac{\partial}{\partial\theta}\log(f(\mathrm{Y};\theta)\right)^2\right\}}. \end{aligned} \tag{13}$$

- This yields the Cramér–Rao lower bound on the variance of an unbiased estimator.

Measures of Performance

- We define the Fisher information to be

$$\mathcal{I}_n(\theta) = \mathbb{E}\left\{ \left(\frac{\partial}{\partial\theta} \log(f(\mathrm{Y};\theta) \right)^2 \right\}.$$

- Then the Cramér–Rao lower bound on the variance of unbiased estimator $\widehat{\theta}(\mathrm{y})$ states that

$$\mathbb{V}\mathrm{ar}\left\{\widehat{\theta}(\mathrm{Y})\right\} \geq \frac{1}{\mathcal{I}_n(\theta)}.$$

- If $\mathrm{Y} = (Y_1, \ldots, Y_n)$ has n iid entries then $f(\mathrm{y};\theta) = f(y_1;\theta) f(y_2;\theta) \ldots f(y_n;\theta)$ and so we find:

$$\mathcal{I}_n(\theta) = n \cdot \mathcal{I}_1(\theta).$$

- Unless the density is very peculiar further simplifications can be made.

Measures of Performance

- If we play around further using Fubini's theorem to exchange the order of integration and differentiation then we arrive at

$$\mathcal{I}_n(\theta) = \mathbb{E}\left\{-\left(\frac{\partial^2}{\partial\theta^2}\log(f(\mathsf{Y};\theta)\right)\right\}.$$

- The real reason why this hold, and why we have a negative sign, will become apparent later in the course.
- The next mathematical question is–can we achieve the Cramér–Rao bound?
- OK let us start backwards!
- Assume that

$$\mathrm{var}\left\{\widehat{\theta}\right\} = \frac{1}{\mathcal{I}_n(\theta)}.$$

Measures of Performance

- Then the bound becomes an equality. We then have

$$\text{var}\left\{\widehat{\theta}\right\} = \frac{\mathbb{C}\text{ov}^2\left\{\widehat{\theta}, \frac{\partial}{\partial\theta}\log(f(\mathsf{Y};\theta)\right\}}{\mathbb{V}\text{ar}\left\{\frac{\partial}{\partial\theta}\log(f(\mathsf{Y};\theta)\right\}}.$$

- This is true if and only if $\widehat{\theta}$ is a linear function of $\frac{\partial}{\partial\theta}\log(f(\mathsf{Y};\theta)$ or

$$\widehat{\theta} = a\frac{\partial}{\partial\theta}\log(f(\mathsf{Y};\theta) + b,$$

  for some constant a and b.

- Solving this equation yields that we can achieve this if and only if the density (frequency) of Y is a one-parameter exponential family with sufficient statistic $\widehat{\theta}$.

Measures of Performance

- Sufficiency is key to this result.
- Theorem (Rao-Blackwell Theorem): Let $\widehat{\theta}$ be an unbiased estimator of θ that has finite variance. Assume T is sufficient for θ. In this case $\widehat{\theta}^* = \mathbb{E}\{\widehat{\theta}|T\}$ is an unbiased estimator and

$$\mathbb{V}\text{ar}\{\widehat{\theta}^*\} \leq \mathbb{V}\text{ar}\{\widehat{\theta}\}.$$

- Equality is attained if and only if $\Pr\{\widehat{\theta}^* = \widehat{\theta}\} = 1$.
- Getting rid of irrelevant information improves estimation performance.
- The new estimator $\widehat{\theta}^*$ is called a "Rao–Blackwellised" version of $\widehat{\theta}$.