Statistics

Ivan Zhytkevych

April 21, 2022

Contents

Distributions

0.1 Bernoulli

$$P(X = 1) = p$$
$$P(X = 0) = 1 - p = q$$
$$E(X) = p$$
$$E\left[X^2\right] = P(X = 1) \cdot 1^2 + P(X = 0) \cdot 0^2 = p \cdot 1 + q \cdot 0 = p = E[X]$$
$$\mathcal{D}[X] = E[X^2] - E[X] = p - p^2 = p(1 - p) = pq$$

0.2 Binomial

$$P(\xi = k) = C_n^k p^k q^{n-k}, \quad k = 0, 1, 2, 3, \ldots, n \quad p \in [0, 1], \quad q = 1 - p \quad n \in \mathbb{N}$$
$$E[X] = np$$
$$\mathcal{D}[X] = np(1 - p)$$

0.3 Poisson

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$
$$\lambda = E[X] = \mathcal{D}[X]$$

0.4 Hypergeometric

$$P(X=k)=\frac{C_D^k\cdot C_{N-D}^{n-k}}{C_N^n}$$
$$E[X]=\frac{nD}{N}$$
$$\mathcal{D}[X]=\frac{n(D/N)(1-D/N)(N-n)}{N-1}$$

0.5 Continuos Uniform

$$f_X(x)=\begin{cases}\frac{1}{b-a} & x\in[a,b]\\ 0 & x\in[a,b]\end{cases}$$
$$P(X\le x)=\begin{cases}0 & x<a\\ \frac{x-a}{b-a} & a\le x\le b\\ 1 & x\ge b\end{cases}$$
$$E[X]=\frac{a+b}{2}$$
$$\mathcal{D}[X]=\frac{(b-a)^2}{12}$$

0.6 Normal

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$
$$E[X]=\mu$$
$$\mathcal{D}[X]=\sigma^2$$

0.7 Exponential

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$F(x,\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

$$E[X] = \frac{1}{\lambda}$$

$$\mathcal{D}[X] = \frac{1}{\lambda^2}$$

0.8 Cauchy

$$F(x, x_0, \gamma) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}$$

Definition 1 (Convergence in distribution)**.** *A sequence of random variables $X_1, X_2, \ldots, X_n$ converges in distribution, or* ***converge weakly,*** *or* ***converge in law*** *$X_n \rightsquigarrow X$ ($X_n \xrightarrow{d} X$) to a random variable X if:*

$$\forall x \in \mathbb{R} : F \in C[x] \Rightarrow \lim_{n\to\infty} F_n(x) = F(x).$$

Definition 2 (Convergence in probability)**.** *A sequence of random variables $\{X_n\}$ converges in probability $X_n \xrightarrow{p} X$ towards the random variable X if:*

$$\forall \varepsilon > 0 \quad \lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0.$$

Definition 3 (Almost surely convergence)**.** *A sequnce of random variables $\{X_n\}$ converges almost surely, or* ***almost everywhere,*** *or* ***with probability 1,*** *or* ***strongly*** *$X_n \xrightarrow{a.s.} X$ towards X means that:*

$$P\left(\lim_{n\to\infty} X_n = X\right) = 1.$$

1 Sample and sample characteristics

1.1 Sample

Definition 4. *Vector $\vec{X} = (x_1, \ldots, x_n)$, where $x_i \in P(\xi)$ are independent equally distributed random values (i.e.d. - independent equally distributed) is called a sample of volume n with distribution $P(\xi)$ (from general totality (з генеральної сукупності) $P(\xi)$).*

Remark. $F_{\vec{X}}(y_1, \ldots, y_n) = P(x_1 \le y_1, \ldots, x_n \le y_n) = \prod_{i=1}^{n} P(x_i \le y_i) = \prod_{i=1}^{n} F_\xi(y_i)$*, where $F_\xi(x) = P(\xi \le x)$ distribution function ξ.*

$\mathcal{F} = \{F_\xi\}$ we define a class of allowable ditribution functions for random value ξ.

$\mathcal{F}\{F(x,\theta), \theta \in \Theta\}$, Θ - a set of all allowable values for θ.

Example. $P(\xi)$ *normal distribution with known dispersion σ^2 but unknown expectation θ. Then our parametric model is:*

$\mathcal{F} = \{F(x,\theta), \theta \in \Theta = (-\infty, \infty)\},$ *where $F(x,\theta)$ has density of distribution*

$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, x \in \mathbb{R}.$$

Example. $P(\xi)$ *has Puasson distribution with unknown parameter θ. Then the parametric model is:*

$$\mathcal{F} = \{F(x,\theta),\ \theta \in \Theta = (0, \infty)\}.$$

$$F(x,\theta) = P(\xi = x) = \frac{\theta^x}{x!} e^{-\theta}, x = 0, 1, 2, \ldots.$$

Definition 5. *Measurable function from sampling (and only from sample) is called statistics.*

$$T_n(\vec{X}) \text{ - statistics .}$$

Example.

$$x_1, \dots, x_n \text{ - \textit{i.e.d. random values} .}$$
$$T(x_1, \dots, x_n) = x_1.$$
$$T(x_1, \dots, x_n) = \frac{1}{n} \sum_{i=1}^{n} x_i.$$
$$T(x_1, \dots, x_n) = \min(x_1, \dots, x_n).$$

Example.

$$x_i \sim Poiss(\theta), \theta \text{ - \textit{unknown parameter} .}$$

$T(x_1, \dots, x_n) = \frac{x_1}{\theta}$ *- is not a statistics function as long as it depends on unknown parame*

1.2 Variation series of the sample

Suppose $\vec{X} = (X_1, \dots, X_n)$ a sample, $\vec{x} = (x_1, \dots, x_n)$ a realization of the sample.

Let

$$x_{(1)} = min(x_1, \dots, x_n)$$
$$x_{(2)} \text{ - second by range}$$
$$\dots$$
$$x_{(n)} = max(x_1, \dots, x_n).$$

In probability and statistics, a realization, observation, or observed value, of a random variable is the value that is actually observed (what actually happened).

Let $X_{(k)}$ to be a random value that for every realization $\vec{x}$ of sample $\vec{X}$ is $x_{(k)}$. Then the series

$$R = \left(X_{(1)}, X_{(2)}, \dots, X_{(n)}\right).$$

is a variation series of the sample.

$X_{(k)}$ - is kth ordinal statistics.

Remark. *Ordinal statistics $X_{(1)}, \dots, X_{(n)}$ are neither independent nor equally distributed.*

Let's find $F_{X_{(1)}}, F_{X_{(k)}}, F_{X_{(n)}}$:

$$F_{X_{(1)}}(y) = P(X_{(1)} \leq y) = P(\min(X_1, \ldots, X_n) \leq y) =$$
$$= 1 - P(\min(X_1, \ldots, X_n) > y) = 1 - P(X_1 > y, \ldots, X_n > y) =$$
$$= 1 - \prod_{i=1}^{n} P(X_i > y) = 1 - (1 - F(y))^n ;$$

$$F_{X_{(n)}}(y) = P(\max(X_1, \ldots, X_n) \leq y) =$$
$$= P(X_1 \leq y, \ldots, X_n \leq y) = [F(y)]^n .$$

$$F_{X_{(k)}}(y) = P\left(X_{(k)} \leq y\right) =$$

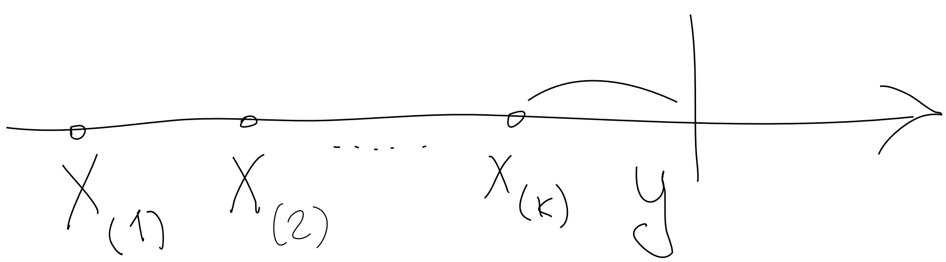

Figure 1: Distribution of k-th ordinal statistics

$$= P(\text{ at least } k \text{ elements do not exceed } y) =$$
$$= \sum_{m=k}^{n} C_n^m \left[F(y)\right]^m (1 - F(y))^{n-m} .$$

Proposition 1.1 (joint distribution of variation series). *Let* $\vec{X} = (X_1, \dots, X_n)$ *- a sample and* X_i *has density* $f(x)$*. Then:*

$$f_{(X_{(1)},\dots,X_{(n)})}(y_1, \dots, y_n) = n! f(y_1) \dots f(y_n) \times \mathbb{1}(y_1 \leq y_2 \dots \leq y_n).$$

Proof. Consider distribution function of variation series:

$$F_{(X_{(1)},\dots,X_{(n)})}(y_1, y_2, \dots, y_n) = P\left(X_{(1)} \leq y_1, \dots, X_{(n)} \leq y_n\right).$$

Consider that $y_1 > y_2$. Then $X_{(2)} \leq y_2 \Rightarrow X_{(1)} \leq y_1$. $\left(X_{(1)} \leq X_{(2)} \leq y_2 < y_1\right)$.

$$\left\{X_{(2)} \leq y_2\right\} \cap \left\{X_{(1)} \leq y_1\right\} = \left\{X_{(2)} \leq y_2\right\}.$$

That's why:

$$F_{\left(X_{(1)},\dots,X_{(n)}\right)}(y_1, \dots, y_n) = P\left(X_{(2)} \leq y_2, \dots, X_{(n)} \leq y_n\right).$$

Because right side does not depend from y_1 then

$$f_R(y_1, \dots, y_n) = 0$$

in case of non-fulfillment of the condition of orderliness: $y_1 \leq y_2 \leq \dots \leq y_n$.

Let $\Gamma = \{(y_1, \dots, y_n) \in \mathbb{R}^n : y_1 \leq \dots, y_n\}$.

$$\forall A \subset \Gamma : P\left(\left(X_{(1)}, \dots, X_{(n)}\right) \in A\right) = \int_A f_R(y_1, \dots, y_n) dy_1 \dots dy_n.$$

On the other side:

$$P\left(\left(X_{(1)}, \dots, X_{(n)}\right) \in A\right) = \sum_{\sigma \in S_n} P\left(\left(X_{\sigma(1)}, \dots, X_{\sigma(n)}\right) \in A\right) =$$

$$= n! \cdot P\left((X_1, \dots, X_n) \in A\right) = n! \cdot \int_A f_{\vec{X}}\left(y_1, \dots, y_n\right) dy_1 \dots dy_n$$

Hence:

$$f_r(y_1, \dots, y_n) = n! f_{\vec{X}}(y_1, \dots, y_n) = n! f(y_1) \cdot \ldots \cdot f(y_n).$$

At last:

$$f_R(y_1, \dots, y_n) = n! \prod_{i=1}^{n} f(y_i) \cdot \mathbb{1}(y_1 \leq \dots \leq y_n).$$

$\square$

1.3 Empirical distribution function

Let $\vec{X} = (x_1, \ldots, x_n)$ - a sample.

Consider:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left(X_i \leq x\right).$$

$F_n(x)$ is called **empirical distirbution function**.

$F_n(x)$ - random function: for every $x \in \mathbb{R}$ takes values $0, \frac{1}{n}, \frac{2}{n}, \ldots, 1$.

Herewith:

$$P\left(F_n(x) = \frac{k}{n}\right) = C_n^k \left[F(x)\right]^k (1 - F(x))^{n-k}, \quad k = \overline{0, n}.$$

$$n \cdot F_n(x) \sim \text{Bin}(n, F(x)).$$

Hence:

1.

$$M\left[n \cdot F_n(x)\right] = n \cdot F(x)$$
$$\Downarrow$$
$$MF_n(x) = F(x).$$

2.

$$\mathcal{D}(n \cdot F_n(x)) = n \cdot F(x)(1 - F(x))$$
$$\Downarrow$$
$$\mathcal{D}F_x(x) = \frac{1}{n} F(x)(1 - F(x)).$$

Using law of large numbers:

$$F_n(x) = \frac{\mathbb{1}(X_1 \leq x) + \ldots + \mathbb{1}(X_n \leq x)}{n} \xrightarrow[n \to \infty]{P} M\mathbb{1}(X_1 \leq x) = P(X_1 \leq x) = F(x).$$

Using central limit theorem (ЦГТ):

$$\frac{n \cdot F_n(x) - nF(x)}{\sqrt{nF(x)(1 - F(x))}} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 1)$$

$$\sqrt{n} \cdot \frac{F_n(x) - F(x)}{\sqrt{F(x)(1 - F(x))}} \xrightarrow{d} \mathcal{N}(0, 1)$$

1.4 Hystogram and frequency range

Suppose that $X_1, \dots, X_n$ = a sample;

$X_i \sim \xi$; ξ has continuous density $f(x)$ (unknown).

Let $\mathcal{I}_1, \dots, \mathcal{I}_m$ - some division of the area $\mathcal{I}$ of possible values of ξ:

Let $\nu_r = \sum_{j=1}^n \mathbb{1}(X_j \in \mathcal{I}_r)$ - number of elements of the sample that are in $\mathcal{I}_r$.

Then by the Law of Large Numbers:

$$\frac{\nu_r}{n} = \frac{\sum_{j=1}^n \mathbb{1}(X_j \in \mathcal{I}_r)}{n} \underset{n\to\infty}{\overset{P}{\to}} M\mathbb{1}(X_1 \in \mathcal{I}_r) = P(X_1 \in \mathcal{I}_r) = \int_{\mathcal{I}_r} f(x)dx.$$

Because f is continuos then by the theorem about mean (теорема про середнє):

$$\int_{\mathcal{I}_r} f(x)dx = |\mathcal{I}_r| \cdot f(x_r)$$

where x_r - is inner point of the interval $\mathcal{I}_r$, $|\mathcal{I}_r|$ - length of the interval.

We can consider, that (n is big and $|\mathcal{I}_r|$ is small)

$$\frac{\nu_r}{n \cdot |\mathcal{I}_r|} \approx f(x_r)$$

where x_r - middle of $\mathcal{I}_r$.

Definition 6. *Piecewise constant function*

$$f_n(x) = \frac{\nu_r}{n \cdot |\mathcal{I}_r|} \mathbb{1}(x \in \mathcal{I}_r), \quad r = \overline{1, m}$$

is called a hystogram.

Within large n and small enough division the hystogram $f_n(x)$ is an approximation of true density $f(x)$.

Example. *Height of $n = 500$ students was measured; The results are shown in view of interval statistical series:*

145-150	150-155	155-160	160-165	165-170	170-175	175-180	180-185
1	2	28	90	169	132	55	23

$$|\mathcal{I}_r| = 5; \qquad n = 500; \qquad f_n(x) = \frac{\nu_r}{2500}\mathbb{1}(x \in \mathcal{I}_r).$$

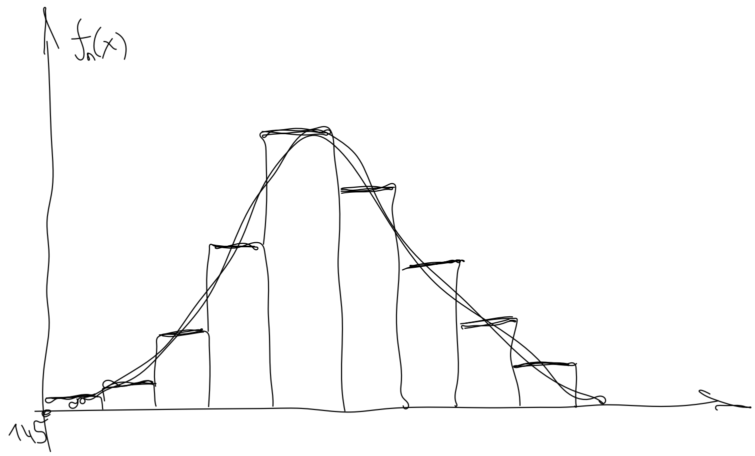

Figure 2: Students height hystogram

Definition 7. *The frequency polygon is a polyline that connects the midpoints of the segments in the histogram.*

1.5 Sample mean

Definition 8. *Statistics* $\overline{X} = \frac{X_1+\ldots+X_n}{n}$ *is called sample mean (вибіркове середнє) or selective first moment (вибірковий перший момент).*

Properties:

1. $M\overline{X} = \frac{1}{n}\sum_{i=1}^{n} MX_i = MX_1 = m$
2. $\mathcal{D}\overline{X} = \frac{1}{n^2}\sum_{i=1}^{n} \mathcal{D}X_i = \frac{1}{n}\mathcal{D}X_1 = \frac{\sigma^2}{n}$
3. Using Law of Large Numbers:

$$\overline{X} = \frac{X_1 + \ldots + X_n}{n} \xrightarrow[n\to\infty]{P} MX_1 = m.$$

4. Using central limit theorem (центральна гранична теорема):

$$\frac{n \cdot \overline{X} - n \cdot M\overline{X}}{\sqrt{\mathcal{D}[n\overline{X}]}} \xrightarrow{d} \mathcal{N}(0,1), \quad n \to \infty$$

$$\frac{n\left(\overline{X} - m\right)}{n\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{d} \mathcal{N}(0,1)$$

$$\sqrt{n \cdot \frac{\overline{X} - m}{\sigma}} \xrightarrow{d} \mathcal{N}(0,1), \quad n \to \infty.$$

1.6 Sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)$$

where $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

$$S^2 = \overline{X^2} - \left(\overline{X}\right)^2$$

where $\overline{X}^2 = \frac{X_1^2 + \ldots + X_n^2}{n}$ is second sample moment.

Using Law of Large Numbers:

$$S^2 = \overline{X^2} - \left(\overline{X}\right)^2 \xrightarrow[n\to\infty]{P} MX_1^2 - (MX_1)^2 = \mathcal{D}X_1.$$

$$S_o^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2.$$

S_0^2 - is **unbiased sample dispersion**.

Then

$$MS_o^2 = \sigma^2 \tag{1}$$

Indeed :

$$\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n}\left(\left(X_i - \overline{X}\right) + \left(\overline{X} - m\right)\right)^2 =$$

$$= \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + 2\left(\overline{X} - m\right) \cdot \overbrace{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)}^{0} + n \cdot \left(\overline{X} - m\right)^2 =$$

$$= \sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 + n\left(\overline{X} - m\right)^2 =$$

$$= (n-1)\, S_o^2 + n \cdot \left(\overline{X} - m\right)^2 .$$

Let's take *expectation* on the left and right side:

$$M\sum_{i=1}^{n}(X_i - m)^2 = \sum_{i=1}^{n} M\,(X_i - m)^2 = \sum_{i=1}^{n}\mathcal{D}X_i = n \cdot \sigma^2$$

$$M\left(\overline{X} - m\right)^2 = M\left(\overline{X} - M\left(\overline{X}\right)\right)^2 = \mathcal{D}\overline{X} = \frac{\sigma^2}{n}$$

$$n \cdot \sigma^2 = (n-1) \cdot MS_o^2 + n \cdot \frac{\sigma^2}{n}$$

$$\Downarrow$$

$$MS_o^2 = \sigma^2.$$

2 Lection 2. Point extimates and their properties

Let $X_1, X_2, \ldots X_n$ is a sample from parametric family $\mathcal{F} = \{F(x,\theta), \theta \in \Theta\}$ where θ is unknown parameter.

Problem: find statistics $T_n = T\left(\vec{X}\right)$, values of which, by defined realization of a sample, are taken as approximated value of θ(значення якої при заданій реалізації $\vec{x}$ вибірки $\vec{X}$ приймається за наближене значення θ). Then T_n is called a point estimate of evaluation (точкова оцінка) θ.

Definition 9. *Statistics*

$$T_n = T(X_1, X_2, \ldots, X_n).$$

is called meaningful evaluation (змістовна оцінка) of θ if

$$T_n \xrightarrow{p} \theta, \quad n \to \infty.$$

$$\forall \varepsilon > 0 \quad P_\theta\left(|T_n - \theta| > \varepsilon\right) \to 0, \quad n \to \infty.$$

Remark. $P_\theta, M_\theta, D_\theta$ *means that respective values are evaluated with an assumption that* $X_i \sim F(x, \theta)$.

Definition 10. *Statistics T_n is called unbiased evaluation of θ if*

$$M_\theta T_n = \theta \quad \forall \theta \in \Theta.$$

Definition 11. *Statistics T_n is called asymptotically unbiased evaluations if*

$$M_\theta T_n \to \theta, \;\; n \to \infty \;\; \forall \theta \in \Theta.$$

Example. $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ *- meaningful unbiased statistics for $\theta = MX_1$*.

Example.

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \overline{X^2} - \left(\overline{X}\right)^2$$

Using Law of Large Numbers:

$$S^2 \xrightarrow{p} MX_1^2 - (MX_1)^2 = \mathcal{D}X_1$$

$$MS^2 = M\left(\frac{n-1}{n} MS_0^2\right) \quad S_0^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

That's why

$$MS^2 = \frac{n-1}{n} MS_0^2 = \frac{n-1}{n} \cdot \mathcal{D}X_1 \to \mathcal{D}X_1$$

S^2 *- meaningful asymptotic unbiased for $\theta = \mathcal{D}X_1$*

Corollary 2.1 (About meaningfullness and unbias of sampling moments)**.** *Let g - borel function that $Mg(X_1) < \infty$. Then the evaluation*

$$\overline{g(X)} = \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

is meaningful unbiased evaluation for $\Theta = Mg(X_1)$.

Example (Unbiased evaluation does not exist)**.**

$$X_1 \sim \text{Poiss}(\theta), \quad g(\theta) = \frac{1}{\theta}.$$

Assume that $\exists T(X)$ unbiased estimate for $\frac{1}{\theta}$:

$$M_\theta T(X) = \frac{1}{\theta} \quad \forall \theta > 0.$$

this means

$$\sum_{k=0}^{\infty} T(k) \cdot \frac{\theta^k}{k!} e^{-\theta} = \frac{1}{\theta} \quad \forall \theta > 0.$$

$$\sum_{k=0}^{\infty} T(k) \cdot \frac{\theta^{k+1}}{k!} = e^{\theta} \quad \forall \theta > 0.$$

$$\sum_{k=0}^{\infty} T(k) \cdot \frac{\theta^{k+1}}{k!} = \sum_{m=0}^{\infty} \frac{\theta^m}{m!} \quad \forall \theta > 0.$$

The last equality is unpossible: theres no such function T that does not depend on θ in such way that the last equality is true.

Remark. *If T_1 and T_2 are unbiased evaluations for θ then*

$$T = C_1 T_1 + C_2 T_2, \quad C_1 + C_2 = 1$$

is unbiased too.

Example.

$$x_1, \ldots, x_{2n} \sim B(p),\ p \text{ is unknown }.$$

$$\hat{p_1} = \frac{1}{2n} \sum_{i=1}^{2n} x_i.$$

is unbiased and measningful evaluations for p.

$$\hat{p_2} = \frac{1}{n} \sum_{i=1}^{n} X_{2i}.$$

is unbiased and measningful too.

Which one is better?

2.1 Root mean square approach to comparing estimates

Definition 12 (RMS)**.** *Value*

$$M_\theta(T(\vec{X}) - \theta)^2$$

is called a root ***mean square evaluation*** *of* T.

Definition 13. *The evaluation* T_1 *is better in root mean square than evaluation* T_2 *if*

$$\forall \theta \in \Theta : M_\theta(T_1 - \theta)^2 \leq M_\theta \left(T_2 - \theta\right)^2.$$

and at least for one θ*:*

$$M_\theta \left(T_1 - \theta\right)^2 < M_\theta \left(T_2 - \theta\right)^2.$$

Example (continuation)**.**

$$\hat{p_1} = \frac{1}{2n} \sum_{i=1}^{2n} X_i \quad \hat{p_2} = \frac{1}{n} \sum_{i=1}^{n} X_{2i}.$$

$$M_\theta \left(\hat{p_1} - p\right)^2 = M_\theta \left(\hat{p_1} - M\hat{p_1}\right)^2 = \mathcal{D}\hat{p_1} = \frac{1}{2n}\mathcal{D}X_1.$$

$$M_\theta \left(\hat{p_2} - p\right)^2 = \mathcal{D}\hat{p_2} = \frac{1}{n}\mathcal{D}X_1.$$

$$\frac{1}{2n} < \frac{1}{n} \Rightarrow \ \hat{p_1} \text{ is better in rms }.$$

Example.

$$X_1, \ldots, X_n \sim U(0, \theta), \quad \theta \textit{ unknown } .$$

$$T_1 = 2 \cdot \overline{X}; \quad T_2 = X_{(n)}.$$

for T_1:

$$MT_1 = 2 \cdot \frac{\theta}{2} = \theta$$

is unbiased.

$$M_\theta (T_1 - \theta)^2 = \mathcal{D}T_1 = 4 \cdot \mathcal{D}(\overline{X}) = \frac{4}{n} \cdot \mathcal{D}X_1 = \frac{4}{n} \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n}.$$

For T_2:

$$F_{X_{(n)}}(y) = [F_{X_i}(y)]^n = \begin{cases} 1 & y \geq 0 \\ \left(\frac{y}{\theta}\right)^n & y \in [0, \theta] \\ 0 & y < 0 \end{cases}.$$

$$f_{X_{(n)}}(y) = n \cdot \frac{y^{n-1}}{\theta^n} \cdot \mathbb{1}(y \in [0, \theta]).$$

Herewith

$$MX_{(n)} = \frac{n}{\theta^n} \int_0^\theta y^n dy = \frac{n}{n+1} \cdot \theta \underset{n \to \infty}{\to} \theta.$$

$X_{(n)}$ is asymptotically unbiased

$$MX_{(n)}^2 = \frac{n}{\theta^n} \int_0^\theta y^{n+1} dy = \frac{n}{n+2} \theta^2.$$

Hence

$$M_\theta \left(X_{(n)} - \theta\right)^2 = \frac{n}{n+2}\theta^2 - \frac{2n}{n+1}\theta^2 + \theta^2 = \frac{2}{(n+1)(n+2)}\theta^2.$$

Within $n = 1, n = 2$ then RMS are equal. It follows that no is better. Within $n \geq 3$:

$$\frac{2}{(n+1)(n+2)} < \frac{1}{3n}$$

and $X_{(n)}$ is better in RMS.

Definition 14. *Value* $b(\theta) = M_\theta T(\vec{X}) - \theta$ *is called biased evaluation* $T\left(\vec{X}\right)$.

Suppose K_b is a class of biased evaluations $b = b(\theta)$. K_0 is a class of unbiased evaluations.

Remark. *If* $\forall T \in K_b$*:*

$$M_\theta \left(T(X) - \theta\right)^2 = \mathcal{D}_\theta T(X) + \left(b(\theta)\right)^2.$$

Definition 15. *Evaluation* $T^* \in K_b$ *is called an optimal in this class if it is not worse (better) than any other evaluation from this class.*

$$\forall T \in K_b, \forall \theta \in \Theta \;\; \mathcal{D}_\theta T^* \leq \mathcal{D}_\theta T.$$

Theorem 2.2 (about unity of optimal evaluation). *Let* T_1 *and* T_2 *are two optimal evaluations. Then*

$$T_1 = T_2 \textit{ almost certainly}.$$

Proof. We got:

$$M_\theta T_1 = \theta, \quad M_\theta T_2 = \theta.$$
$$\mathcal{D}_\theta T_1 = \mathcal{D}_\theta T_2 = \sigma^2.$$

$$\text{Consider } T = \frac{1}{2}\left(T_1 + T_2\right)$$
$$T \in K_0 : M_\theta T = \theta \text{ besides}$$
$$\mathcal{D}_\theta T \geq \sigma^2$$
$$\mathcal{D}_\theta T = \mathcal{D}_\theta \left(\frac{1}{2}T_1 + \frac{1}{2}T_2\right) = \frac{1}{4}\mathcal{D}T_1 + \frac{1}{4}\mathcal{D}T_2 + \frac{1}{2}\operatorname{cov}(T_1, T_2) =$$
$$= \frac{1}{2}\sigma^2 + \frac{1}{2}\operatorname{cov}(T_1, T_2)$$
$$\text{Causy-Bunyakovskiy: } |\operatorname{cov}(T_1, T_2)| = |\operatorname{cov}(T_1 - \theta, T_2 - \theta)| \leq$$
$$\leq \sqrt{\mathcal{D}_\theta T_1 \cdot \mathcal{D}_\theta T_2} = \sigma^2$$
$$\mathcal{D}_\theta T \leq \sigma^2$$

This means that in Causy-Bnyakovskiy inequality turn into equality, and this means that
$$T_1 \overset{\text{a.s.}}{=} kT_2 + a.$$
Using conditions:
$$a = 0, k = 1.$$
So
$$T_1 \underset{\text{a.c.}}{=} T_2.$$
$\square$

2.2 Sample contribution and Fisher information quantity

$$X_1, \dots, X_n \sim \mathcal{F}_\theta = \{F(x, \theta), \theta \in \Theta\}.$$
$f(x, \theta)$ - corresponding distribution densit X_i ; $f(x, \theta) = P(X_i = x)$ - in discrete case.

Definition 16.
$$\mathcal{L}(\vec{x}, \theta) = f(x_1, \theta) \cdot \ldots \cdot f(x_n, \theta)$$
that is considered as function θ within fixed $\vec{x}$, is called a likelihood function.

Consider further that
$$\mathcal{L}(x, \vec{\theta}) > 0 \quad \forall \vec{x} \in X \quad \forall \theta \in \Theta.$$
$$\mathcal{L}(\vec{x}, \theta) \text{ differentiable by } \theta\ .$$
The model is regular: the order of differential by θ and integration by $\vec{X}$ may be swapped.

Example (not regular model)**.**
$$X_1, X_2, \dots, X_n \sim U(0, \theta).$$
$$f(x, \theta) = \frac{1}{\theta}\mathbb{1}(x \in (0, \theta)).$$

$$\int_0^\theta f(x,\theta)dx = 1.$$

Then

$$\frac{\partial}{\partial\theta}\int_0^\theta \frac{1}{\theta}dx = 0.$$

But

$$\int_0^\theta \frac{\partial}{\partial\theta}\left(\frac{1}{\theta}\right)dx = \int_0^\theta -\frac{1}{\theta^2}dx = -\frac{1}{\theta} \neq 0.$$

Definition 17. *Random value*

$$U(\vec{X},\theta) = \frac{\partial}{\partial\theta}\ln\mathcal{L}(\vec{X},\theta) = \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\ln f(X_i,\theta).$$

is called sample contribution. And addition

$$\frac{\partial}{\partial\theta}\ln f(X_i,\theta).$$

is called a contribution of i-th observation.

Corollary 2.3 (about expectation of U)**.** *For regular model*

$$M_\theta U(\vec{X},\theta) = 0 \quad \forall\theta\in\Theta.$$

Proof.

$$L(\vec{X},\theta) = f(x_1,\theta)\ldots f(x_n,\theta).$$

L is compatible (сумісна) density function $\vec{X}$.

$$\int_{\mathbb{R}} L(\vec{X},\theta)dx = 1.$$

Differentiate by θ

$$0 = \frac{\partial}{\partial\theta}\int_{\mathbb{R}^n} L(\vec{X},\theta)d\vec{x} = \int_{\mathbb{R}^n}\frac{\partial}{\partial\theta}L(\vec{X},\theta)d\vec{x} =$$

$$= \int_{\mathbb{R}^n}\left[\frac{\partial}{\partial\theta}\ln L(\vec{x},\theta)\right]\cdot L(\vec{x},\theta)d\vec{x} = M_\theta\left[\frac{\partial}{\partial\theta}\ln L(\vec{x},\theta)\right] = M_\theta U(\vec{X},\theta).$$

□

Definition 18. *Value*

$$I_n(\theta) = M_\theta U^2(X,\theta) = \mathcal{D}_\theta U(X,\theta)$$

is called a number of Fisher information in sample $\vec{X}$. Value

$$i(\theta) = M_\theta \left(\frac{\partial}{\partial\theta} \ln f(X_1,\theta)\right)^2.$$

is a number of information in one observation.

As long as $X_1, \dots X_n$ are independent equaly distributed values:

$$I_n(\theta) = n \cdot i(\theta).$$

This means that number of information is increasing proprtionally to the grows of sample volume.

Corollary 2.4. *If $f(x,\theta)$ is differentiable twice by θ then*

$$i(\theta) = -M_\theta \left[\frac{\partial^2}{\partial\theta^2} \ln f(X_1,\theta)\right].$$

Proof. By the latest proof:

$$0 = M_\theta U(\vec{X},\theta) = \int_{\mathbb{R}^n} \left[\frac{\partial}{\partial\theta} \ln \mathcal{L}(\vec{x},\theta)\right] \mathcal{L}(\vec{x},\theta) d\vec{x}.$$

Within $n = 1$:

$$0 = \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta} \ln f(x,\theta)\right] f(x,\theta) dx.$$

Differentiate by θ:

$$0 = \int_{\mathbb{R}} \left[\frac{\partial^2}{\partial\theta^2} \ln f(x,\theta)\right] f(x,\theta)dx + \int_{\mathbb{R}} \left[\frac{\partial}{\partial\theta} \ln f(x,\theta)\right] \cdot \frac{\partial}{\partial\theta} f(x,\theta)dx =$$

$$= M_\theta \left[\frac{\partial^2}{\partial\theta^2} \ln f(x,\theta)\right] \cdot \left[\frac{\partial}{\partial\theta} \ln f(x,\theta)\right] \cdot f(x,\theta)dx =$$

$$= M_\theta \left[\frac{\partial^2}{\partial\theta^2} \ln f(x,\theta)\right] + M_\theta \left(\frac{\partial}{\partial\theta} \ln f(x,\theta)\right)^2 =$$

$$= M_\theta \left[\frac{\partial^2}{\partial\theta^2} \ln f(x,\theta)\right] + i(\theta).$$

Herewith:

$$i(\theta) = -M_\theta \left[\frac{\partial^2}{\partial \theta^2} \ln f(x, \theta) \right].$$

$\square$

Example (Calculation of information quantity)**.**

$$X_1, \dots, X_n \sim \text{Poiss}(\theta).$$

$$\begin{gathered}
f(x,\theta) = \frac{\theta^x}{x!} e^{-\theta} \quad x = 0, 1, 2, \dots \\
\ln f(x,\theta) = -\theta + x \ln \theta - \ln x! \\
\frac{\partial^2}{\partial \theta^2} \ln f(x,\theta) = \frac{\partial}{\partial \theta} (-1 + \frac{x}{\theta}) = -\frac{x}{\theta^2} \\
i(\theta) = -M_\theta \left(-\frac{x}{\theta^2} \right) = \frac{M_\theta X}{\theta^2} = \frac{\theta}{\theta^2} = \frac{1}{\theta}
\end{gathered}$$

Example.

$$\begin{gathered}
X_i \sim \mathcal{N}(\theta, \sigma^2) \quad \sigma^2 \textit{ known} \\
f(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\theta)^2}{2\sigma^2}} \\
\ln f(x,\theta) = C - \frac{(x-\theta)^2}{2\sigma^2} \\
\frac{\partial^2}{\partial \theta^2} \ln f(x,\theta) = \frac{\partial}{\partial \theta} \left(\frac{1}{\sigma^2} (x - \theta) \right) = -\frac{1}{\sigma^2} \\
i(\theta) = -M_\theta \left(-\frac{1}{\sigma^2} \right) = \frac{1}{\sigma^2}
\end{gathered}$$

Example.

$$X_i \sim \mathcal{N}(a, \theta^2); \quad a - \text{ known}$$
$$f(x,\theta) = \frac{1}{\sqrt{2\pi}\cdot\theta} e^{-\frac{(x-a)^2}{2\theta^2}}$$
$$\ln f(x,\theta) = C - \ln\theta - \frac{(x-a)^2}{2\theta^2};$$
$$\frac{\partial^2}{\partial\theta^2}\ln f(x,\theta) = \frac{\partial}{\partial\theta}\left(-\frac{1}{\theta} + \frac{(x-a)^2}{\theta^3}\right) =$$
$$\frac{1}{\theta^2} - \frac{3\,(x-a)^2}{\theta^4}$$
$$i(\theta) = -M_\theta\left(\frac{1}{theta^2} - \frac{3\,(x-a)^2}{\theta^4}\right) = -\frac{1}{\theta^2} + \frac{3\theta^2}{\theta^4} = \frac{2}{\theta^2}.$$

2.3 Cramér–Rao bound

Theorem 2.5. *Let* $X_1, \ldots, X_n \in \mathcal{F}_\theta = \{F(x,\theta), \theta \in \Theta\}$ *a sample from regular parametrized model;*

Let $g(\theta)$ *- differentiable function;* K_0^g *- a class of unbiased evaluations for* $g(\theta)$*. Then*

$$\forall T \in K_0^g \quad \mathcal{D}_\theta T \geq \frac{[g'(\theta)]^2}{ni(\theta)} \tag{2}$$

moreover the equation is true only if

$$T(\vec{X}) - g(\theta) = a(\theta)U(X,\theta)$$

for some function $a(\theta)$*;* U *- sample contribution.*

Notice, that within $g(\theta) = \theta$ *:*

$$\mathcal{D}_\theta T \geq \frac{1}{ni(\theta)} \quad \forall T \in K_0.$$

Proof. As far as $T \in K_0^g$ then

$$M_\theta T = g(\theta) \quad \forall \theta \in \Theta$$
$$\Updownarrow$$
$$\int_{\mathbb{R}^n} T(\vec{x}) L(\vec{x}, \theta) d\vec{x} = g(\theta), \quad \forall \theta \in \Theta.$$

Differentiate by θ:

$$\begin{gathered} g'(\theta) = \int_{\mathbb{R}^n} T(\vec{x}) \cdot \frac{\partial}{\partial \theta} \mathcal{L}(\vec{x}, \theta) d\vec{x} = \\ = \int_{\mathbb{R}^n} T(\vec{x}) \left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\vec{x}, \theta) \right] \cdot \mathcal{L}(\vec{x}, \theta) d\vec{x} = \\ = M_\theta \left[T(\vec{x}) \cdot \frac{\partial}{\partial \theta} \ln \mathcal{L}(\vec{x}, \theta) \right] = \\ = M_\theta \left[T(\vec{x}) \cdot U(\vec{x}, \theta) \right] = \\ = \langle M_\theta U(\vec{x}, \theta) = 0 \rangle = \\ = M_\theta \left[(T(\vec{x}) - g(\theta)) \cdot U(\vec{x}, \theta) \right] \leq \langle \text{ Causy-Bunyakovskiy inequality } \rangle \leq \\ \leq \sqrt{\mathcal{D}_\theta T(\vec{x}) \cdot \mathcal{D}_\theta U(\vec{x}, \theta)} = \sqrt{\mathcal{D}_\theta T(\vec{x}) \cdot (ni(\theta))} \end{gathered}$$

Herewith:

$$[g'(\theta)]^2 \leq \mathcal{D}_\theta T(\vec{x}) \cdot I_n(\theta)$$
$$\Downarrow$$
$$\mathcal{D}_\theta T(\vec{x}) \geq \frac{[g'(\theta)]^2}{I_n(\theta)}.$$

Equality in inequality of Causy-Bunyakovskiy is true if and only if:

$$T(\vec{x}) - g(\theta) = a(\theta) \cdot U(\vec{x}), \theta).$$

□

Definition 19. *Evaluation $T^* \in K_0^g$ is called **effective** (by Cramér–Rao) if*

$$\mathcal{D}_\theta T^* = \frac{[g'(\theta)]^2}{ni(\theta)}.$$

Criterion of effectiveness:

$$T^* = g(\theta) + a(\theta)U(\vec{x}, \theta).$$

Effective evaluation is optimal and therefore the only one.

Example.

$$X_1, \ldots, X_n \sim \text{Poiss}(\theta); \quad \theta = MX_1$$
$$T(\vec{X}) = \overline{X}.$$
$$i(\theta) = \frac{1}{\theta}$$
$$\mathcal{D}_\theta T = \frac{1}{n}\mathcal{D}X_1 = \frac{\theta}{n} = \frac{1}{ni(\theta)} \Rightarrow T = \overline{X} \quad \text{— \textit{effective for} } \theta = MX_1$$

Example.

$$X_i \sim \mathcal{N}(\theta, \sigma^2) \quad \theta = MX_1$$
$$T(\vec{X}) = \overline{X} \quad i(\theta) = \frac{1}{\sigma^2}$$
$$\mathcal{D}(\overline{X}) = \frac{1}{n}\mathcal{D}X_1 = \frac{\sigma^2}{n} = \frac{1}{ni(\theta)} \Rightarrow \overline{X} \text{ — \textit{effective} }.$$

Example.

$$X_i \sim \mathcal{N}(a, \theta^2) \quad \theta^2 = \mathcal{D}X_1$$
$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - a)^2$$
$$MS^2 = \mathcal{D}X_1 = \theta^2$$
$$g(\theta) = \theta^2; \quad g'(\theta) = 2\theta$$
$$i(\theta) = \frac{2}{\theta^2}; \quad \frac{(g'(\theta))^2}{ni(\theta)} = \frac{2\theta^4}{n};$$

$$\mathcal{D}S^2 = \frac{1}{n}\mathcal{D}\left(X_1 - a\right)^2 = \frac{1}{n}\left(M\left(X_1 - a\right)^4 - \left(M(X_1 - a)^2\right)^2\right) =$$
$$= \frac{1}{n}\left((4-1)!!\left(\mathcal{D}X_1\right)^2 - \left(\mathcal{D}X_1\right)^2\right) =$$
$$= \frac{1}{n}\left(3\theta^4 - \theta^4\right) = \frac{1}{n}2\theta^4;$$
$$\frac{\left[g'(\theta)\right]^2}{ni(\theta)} = \frac{4\theta^4}{2n} = \frac{2\theta^4}{n};$$
$$\mathcal{D}S^2 = \frac{\left[g'(\theta)\right]^2}{ni(\theta)} \Rightarrow S^2 - \textit{effective}.$$

3 Lecture 3. Exponential model. Methods of point evaluation contruction

Let $X_1, \ldots, X_n$ a sample from regular parametric model $\mathcal{F} = \{F(x,\theta), \theta \in \Theta\}$; $f(x,\theta)$ — probability density function (or $P(X_1 = x)$ in descrete variant). $g(\theta)$ — some differentiable parametric function; K_0^g — class of unbiased evaluations for $g(\theta)$.

Then:

$$\forall T \in K_0^g : \mathcal{D}_\theta T \geq \frac{\left[g'(\theta)\right]^2}{n \cdot i(\theta)} \tag{3}$$

where

$$i(\theta) = M_\theta\left(\frac{\partial}{\partial\theta}\ln f(x,\theta)\right)^2$$

is quantity of Fisher information. Besides:

$$\mathcal{D}_\theta T = \frac{\left[g'(\theta)\right]^2}{ni(\theta)} \iff T - g(\theta) = a(\theta)U(\vec{X},\theta)$$

where $U(\vec{X},\theta) = \frac{\partial}{\partial\theta}\ln\mathcal{L}(\vec{X},\theta)$ — sample contribution.

The statistics for which the equation 3 is true is called effective evaluation.

3.1 Exponential model

Definition 20. *Model $\mathcal{F} = \{F(x,\theta), \theta \in \Theta\}$ is called exponential if function $f(x,\theta)$ is like:*

$$f(x,\theta) = \exp A(\theta) \cdot B(x) + C(\theta) + D(x)$$

or

$$\ln f(x,\theta) = A(\theta) \cdot B(x) + C(\theta) + D(x).$$

Example (Bin is exponential model)**.**

$$X_i \sim \text{Bin}(k,\theta);$$

$$f(x,\theta) = P(X = x) = C_k^x \theta^x (1-\theta)^{k-x} \quad x = \overline{0,k}$$

$$\ln f(x,\theta) = \ln C_k^x + x \ln\theta + (k-x)\ln(1-\theta) =$$
$$= \underbrace{\ln k! + \ln \frac{1}{x!\,(k-x)!}}_{D(x)} + \underbrace{x \ln \frac{\theta}{1-\theta}}_{B(x)\cdot A(\theta)} + \underbrace{k \ln(1-\theta)}_{C(\theta)}$$

So model $\text{Bin}(k,\theta)$ *is exponential.*

Remark. *Models* $\mathcal{N}(\theta,\sigma^2)$*,* $\mathcal{N}(a,\theta^2)$*,* $\text{Poiss}(\theta)$*,* $\Gamma(\theta,\lambda)$ *are* ***exponential*** *models*

Example. *Causy distribution is not an exponential model. Indeed:*

$$f(x,\theta) = \frac{1}{\pi} \cdot \frac{1}{1+(x-\theta)^2}, \quad \theta \in \mathbb{R},\ x \in \mathbb{R}.$$

$$\ln f(x,\theta) = C - \ln\left(1 + (x-\theta)^2\right).$$

Assume that

$$\ln f(x,\theta) = A(\theta)B(x) + C(\theta) + D(x)$$

then

$$\frac{\partial^2}{\partial\theta\partial x} \ln f(x,\theta) = A'(\theta)B'(x) \equiv a(\theta)b(x).$$

On the other side:

$$\frac{\partial^2}{\partial\theta\partial x}\ln f(x,\theta) = 2\cdot\frac{1-(x-\theta)^2}{\left(1+(x-\theta)^2\right)^2}.$$

Therefore:

$$a(\theta)\cdot b(x) = 2\cdot\frac{1-(x-\theta)^2}{\left(1+(x-\theta)^2\right)^2}.$$

But the right side cannot be represented as $a(\theta)b(x)$ *(separate the variables).*

Lets compute sample contribution for a regular exponential model:

$$U(\vec{X},\theta) = \frac{\partial}{\partial\theta}\ln\mathcal{L}(\vec{X},\theta) =$$
$$= \frac{\partial}{\partial\theta}\ln\prod_{i=1}^{n}\exp A(\theta)B(X_i) + C(\theta) + D(X_i) =$$
$$= \frac{\partial}{\partial\theta}\ln\exp A(\theta)\cdot\sum_{i=1}^{n}B(X_i) + nC(\theta) + \sum_{i=1}^{n}D(X_i) =$$
$$= \frac{\partial}{\partial\theta}\left(A(\theta)\cdot n\cdot\overline{B(X)} + nC(\theta) + \overline{D(X)}\cdot n\right) =$$
$$= A'(\theta)\cdot n\cdot\overline{B(X)} + n\cdot C'(\theta) = nA'(\theta)\left[\overline{B(X)} + \frac{C'(\theta)}{A'(\theta)}\right]$$

where $\overline{B(X)} = \frac{1}{n}\sum_{i=1}^{n}B(X_i)$. Suppose:

$$a(\theta) = \frac{1}{nA'(\theta)} \quad T(\vec{X}) = \overline{B(x)} \quad g(\theta) = -\frac{C'(\theta)}{A'(\theta)}.$$

then:

$$T(x) - g(\theta) = a(\theta)\cdot U(\vec{X},\theta).$$

By Cramer-Rao criterion: $T(x) = \overline{B(x)}$ is an effective evaluation for $g(\theta) = -\frac{C'(\theta)}{A'(\theta)}$.