Properties of Random Samples

For the next two weeks, I will discuss some of the concepts of random sample which we use very frequently. These are certainly not the central focus of this course, but it is extremely important for all of us to know these concepts. We have to use these ideas throughout this quarter. First we need to know what do we mean by a random sample.

Definition: The random variables $X_1, ..., X_n$ together is known as the random sample of size n from the population $f(x|\theta)$ if $X_1, ..., X_n$ are mutually independent, or the joint density of $X_1, ..., X_n$ is given by $\prod_{i=1}^n f(x_i|\theta)$. We will commonly write as $X_1, ..., X_n \overset{iid}{\sim} f$.

Example: $X_1, ..., X_n$ is a random sample from $exponential(\beta)$. What is $P(X_1 \leq a_1, ..., X_n \leq a_n)$.

Note that

$$P(X_1 \leq a_1, ..., X_n \leq a_n) = \prod_{i=1}^n P(X_i \leq a_i) = \prod_{i=1}^n P(X_i \leq a_i) = \prod_{i=1}^n \int \frac{1}{\beta} e^{-x/\beta} dx = \prod_{i=1}^n (1 - e^{-a_i/\beta}).$$

Remark: $X_1, ..., X_n$ are independent means $g_1(X_1), ..., g_n(X_n)$ are independent for any functions $g_1, ..., g_n$. This means if $X_1, ..., X_n$ is a random sample of size n, $g(X_1), ..., g(X_n)$ is also a random sample of size n for any function g.

Moral of the story is that in a random sample, the probability of any event related to X_i has nothing to do with X_j for $i \neq j$. There are some important advantages of dealing with random samples. By that I mean, some of the random variables derived from a random sample have closed form distributions. Let us see an example. For example, consider the random variable $\sum_{i=1}^n X_i$.

Example: $X_1, X_2, ..., X_n$ is a random sample from $Pois(\lambda)$. What is $P(X_1 + \cdots + X_n = a)$?

Note that

$$\begin{aligned}
P(X_1 + X_2 = m) &= \sum_{l=0}^{m} P(X_1 = l, X_2 = m - l) = \sum_{l=0}^{m} P(X_1 = l)P(X_2 = m - l) \\
&= \sum_{l=0}^{m} \frac{e^{-\lambda}\lambda^l}{l!} \frac{e^{-\lambda}\lambda^{m-l}}{(m-l)!} = \frac{e^{-2\lambda}(2\lambda)^m}{m!} \frac{1}{2^m} \sum_{l=0}^{m} \frac{m!}{l!(m-l)!} = \frac{e^{-2\lambda}(2\lambda)^m}{m!}
\end{aligned}$$

Therefore, $X_1 + X_2 \sim Pois(2\lambda)$. Using induction we can show $X_1 + \cdots + X_n \sim Pois(n\lambda)$.

Some Important definitions: $E[X^k] = \int x^k f(x|\theta)dx$, $Var(X) = E[X^2] - E[X]^2$, $Cov(X_i, X_j) = E[X_iX_j] - E[X_i]E[X_j]$. For any random sample $E[X_iX_j] = \int\int x_i x_j f(x_i, x_j|\theta)dx_i dx_j = \int x_i f(x_i|\theta) \left(\int x_j f(x_j|\theta)x_j\right) dx_i = E[X_i]E[X_j]$. Therefore, $Cov(X_i, X_j) = 0$. The reverse is not always true except for normal.

Moment generating function: What is the easiest way to find $E[X^k]$ for any k. There is a function known as moment generating function which is given by $M_X(t) = E[e^{tX}] = \int e^{tx} f(x|\theta)dx$. If MGF exists at a neighborhood of 0, then $E[X^k] = \frac{d^k}{dt^k} M_X(t)|_{t=0}$. For a random sample, $M_{\bar{X}}(t) = [M_{\bar{X}}(t/n)]^n$.

Example: Let $X \sim N(\mu, \sigma^2)$. Let us compute MGF of X. For every $t \in \mathbb{R}$,

$$\begin{aligned}
E[e^{tX}] &= \int \exp(tx) \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left[\frac{x^2}{\sigma^2} - 2x(\frac{\mu}{\sigma^2} + t) + \frac{\mu^2}{\sigma^2}\right]\right) dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}\left[x - \mu - t\sigma^2\right]^2\right) dx \exp\left(\frac{(\mu - t\sigma^2)^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) \\
&= \exp\left(\frac{(\mu + t\sigma^2)^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right) = \exp\left(t\mu + \frac{1}{2}t^2\sigma^2\right).
\end{aligned}$$

Note that MGF is exists in a range of t. For normal distribution, the range is entire $\mathbb{R}$. However, MGF might not be valid for the entire $\mathbb{R}$ for many other distribution.

Exercise: Let $X \sim Gamma(\alpha, \beta)$. Find the MGF of X.

Change of variable theorem: $X_1, ..., X_n$ random sample from a distribution $f(x|\theta)$. We would like to find the joint distribution of $(\psi_1(X_1, ..., X_n), ..., \psi_n(X_1, ..., X_n))$. Let $u_1 =$

$\psi_1(x_1,..,x_n)$,...,$u_n = \psi_n(x_1,...,x_n)$. Further $x_1 = H_1(u_1,...,u_n)$,...,$x_n = H_n(u_1,...,u_n)$. Then

$$f_U(u_1,...,u_n) = \left[\prod_{i=1}^{n} f(H_i(u_1,...,u_n)|\theta)\right] det\left(\left(\frac{\partial H_i(u_1,...,u_n)}{\partial u_j}\right)_{i,j=1}^{n}\right).$$

example (Box-Muller transformation): Let $U_1, U_2 \sim U(0,1)$. Show that $X_1 = \sqrt{-2\log(U_1)}cos(2\pi U_2)$ follows N(0,1). I will derive this in class. This will give you an idea about how to use the change of variable theorem.

Exercise: To be specified in the class.

Some important results on random sample

Result 1: $X_1,...,X_n$ be a random sample and $E[g(X_1)]$ and $Var(g(X_1))$ exist, then $E[\sum_{i=1}^n g(X_i)] = nE[g(X_1)]$, $Var(\sum_{i=1}^n g(X_i)) = nVar(g(X_1))$.

Result 2: If X and Y are independent random variables with pdf $f_X(x)$ and $f_Y(y)$ respectively, then the pdf of $Z = X+Y$ is $f_Z(z) = \int f_X(w)f_Y(z-w)dw$. Note that

$$P(Z \le z) = P(X+Y \le z) = \int_{-\infty}^{\infty} P(w+Y \le z)f_X(w)dw = \int_{-\infty}^{\infty} P(Y \le z-w)f_X(w)dw$$
$$= \int_{-\infty}^{\infty}\int_{-\infty}^{z-w} f_Y(y)f_X(w)dydw = \int_{-\infty}^{\infty}\int_{-\infty}^{z} f_Y(y-w)f_X(w)dydw = \int_{-\infty}^{z}\int_{-\infty}^{\infty} f_Y(y-w)f_X(w)dwdy.$$

Taking derivative w.r.t z on both sides $f_Z(z) = \int f_X(w)f_Y(z-w)dw$.

Result 4: If $Z \sim N(0,1)$, then $Z^2 \sim \chi_1^2$. If $X_i \sim \chi_1^2$ independently, then $\sum X_i \sim \chi_n^2$. (Note that the definition of χ_n^2 is $Gamma(\frac{n}{2}, \frac{1}{2})$).

$$P(Z^2 \le z) = P(-\sqrt{z} \le Z \le \sqrt{z}) = 2P(0 < Z \le \sqrt{z}) = 2\int_0^{\sqrt{z}} \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx.$$

let $w = x^2$, so that $dx = \frac{dw}{2\sqrt{w}}$. This implies the above integral is

$$P(Z^2 \leq z) = 2\int_0^z \frac{1}{2\sqrt{2w\pi}} \exp(-\frac{w}{2})dw = \int_0^z \frac{1}{\sqrt{2w\pi}} \exp(-\frac{w}{2})dw.$$

Recall the density of $Gamma(\alpha, \beta)$ is $f(x|\alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-\beta x}}{\Gamma(\alpha)}$, $0 < x < \infty$.

Take derivative on both sides w.r.t. z that implies density of Z is χ_1^2.

Result 3: Let $X_1, ..., X_n \sim N(\mu, \sigma^2)$ and let, $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$. Then

(a) $\bar{X}$ and S^2 are independent.

(b) $\bar{X} \sim N(\mu, \sigma^2/n)$.

(c) $(n-1)S^2/\sigma^2 \sim \chi_{n-1}^2$.

Some of the important distributions which you will frequently encounter

Students t distribution: When $X_1, ..., X_n \sim N(\mu, \sigma^2)$, if we know σ^2 then the quantity $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ can be used as a basis for inference on μ. We know the closed form distribution of that quantity. However when σ is unknown, one instead use the quantity $\frac{\bar{X}-\mu}{S/\sqrt{n}}$. It is very intuitive, S^2 is an unbiased estimator of σ^2. Now,

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/\sqrt{\sigma^2 n}}{\sqrt{(n-1)S^2}/\sqrt{n-1}} = \frac{N(0,1)}{\sqrt{\chi_{n-1}^2}/\sqrt{n-1}}.$$

We create a special class of distributions for handling such objects. In fact if $U \sim N(0,1)$, $V \sim \chi_p^2$ and U, V independent, then $U/\sqrt{V/p}$ follows a students t distribution with p degrees of freedom, denoted by t_p. By result 3, $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ follows a t_{n-1}. By the change of variable theorem, we can show that the density of t_p is

$$f(t) = \frac{\Gamma((p+1)/2)}{\Gamma(p/2)} \frac{1}{\sqrt{p\pi}} (1 + t^2/p)^{-(p+1)/2}, \ -\infty < t < \infty.$$

For $p = 1$ no moments exist for t, but for $p > 1$ $E[t_p] = 0$ and $Var(t_p) = \frac{p}{p-2}$ for $p > 2$.

F distribution

If $U \sim \chi^2_p$, $V \sim \chi^2_q$ and U, V are independent, then $\frac{U/p}{V/q}$ is said to follow an $F_{p,q}$ distribution. We will see the significance of distribution much later. But let us see some of the interesting facts about $F_{p,q}$ distribution.

(a) $X \sim F_{p,q}$ implies $1/X \sim F_{q,p}$. (b) $X \sim t_q$, then $X^2 \sim F_{1,q}$. (c) If $X \sim F_{p,q}$, then $(p/q)X/(1 + (p/q)X) \sim Beta(p/2, q/2)$.

Order Statistics: Suppose $X_1, ..., X_n$ be a random sample. The order statistics from the random sample is given by

$$X_{(1)} = \min_{1 \le i \le n} X_i, ..., X_{(n)} = \max_{1 \le i \le n} X_i.$$

$X_{(1)} \le X_{(2)} \le \cdots \le X_{(n)}$ are the order statistics from the random sample. The joint distribution of the order statistics is given by

$$f(X_{(1)}, ..., X_{(n)}|\theta) = n! f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Marginal density of the j-th order statistic

$$f_{X_{(j)}}(x) = \frac{n!}{((j-1)!(n-j)!} f_X(x)[F_X(x)]^{j-1}[1 - F_X(x)]^{n-j}.$$

example: $X_1, ..., X_n \sim \exp(\lambda)$. Then $f_X(x) = \frac{1}{\lambda}\exp(-x/\lambda)$ and $F_X(x) = 1 - \exp(-x/\lambda)$. Thus $f_{X_{(1)},...,X_{(n)}}(x_1, ..., x_n) = \frac{1}{\lambda^n}\exp(-\lambda\sum_{i=1}^n x_i)$, $x_1 < x_2 < \cdots < x_n$ and $f_{X_{(j)}}(x) = \frac{n!}{((j-1)!(n-j)!}\frac{1}{\lambda}\exp(-x/\lambda)[1-\exp(-x/\lambda)]^{j-1}[\exp(-x/\lambda)]^{n-j}$.

Joint density of $(X_{(i)}, X_{(j)})$ is given by

$$f_{X_{(i)},X_{(j)}}(x_1, x_2) = \frac{n!}{((i-1)!(j-i-1)!(n-j)!}f_X(x_1)f_X(x_2)[F_X(x_1)]^{i-1}[F_X(x_2) - F_X(x_1)]^{j-i-1}$$
$$[1 - F_X(x_2)]^{n-j},\ x_1 \leq x_2.$$

example: $X_1, ..., X_2 \overset{iid}{\sim} \exp(\lambda)$. Then

$$f_{X_{(i)},X_{(j)}}(x_1, x_2) = \frac{n!}{((i-1)!(j-i-1)!(n-j)!}[\frac{1}{\lambda^2}\exp(-(x_1+x_2)/\lambda)][1-\exp(-x_1/\lambda)]^{i-1}$$
$$[\exp(-x_1/\lambda) - \exp(-x_2/\lambda)]^{j-i-1}[\exp(-x_2/\lambda)]^{n-j},\ x_1 \leq x_2.$$

Some applications of order statistics.

- A electric device runs on 20 batteries and dies when 15th battery dies. If $X_1, ..., X_{20}$ are the random variables corresponding to lifetimes of 20 batteries, the lifetime of electric device is $X_{(15)}$.
- A policy of five family members are in an insurance policy which says that they will receive a a huge money when two people die. Here if $X_1, ..., X_5$ are life spans of 5 people, we are interested in $X_{(2)}$.

0.1 Some convergence concepts

We always receive a sample of size n. What if the sample size becomes infinite? We will talk about two concepts of convergence.

Convergence in Probability: A sequence $X_1, ...$ converges is probability to a random variable X if , for every $\epsilon > 0$ $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$. For example take a sequence $X_n \sim N(0, 1/n)$. Then $P(|X_n| > \epsilon) \leq \frac{E(X_n^2)}{\epsilon^2} = \frac{1}{n\epsilon^2} \to 0$.

There are two important properties for the convergence in probability.

Properties of convergence in probability: (a) X_n converges to X in probability implies $g(X_n)$ converges to $g(X)$ in probability, for any continuous fn. g.

(b) X_n converges to X and Y_n converges to Y in prob. means $X_n + Y_n$ converges to $X + Y$ in prob.

Convergence in distribution: A sequence of random variables $X_1, ...$ is said to converge in distribution to X, if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$, at all points where $F_X(x)$ is continuous. Convergence in probability implies convergence in distribution, reverse is not generally true except when convergence is happening on constants.

example: Let $X_1, ..., X_n$ be random sample from $U(0, 1)$, where does $n(1 - X_{(n)})$ converge in distribution as $n \to \infty$?

Note that $P(n(1-X_{(n)}) < t) = P(X_{(n)} > 1-\frac{t}{n}) = 1-P(X_{(n)} < 1-\frac{t}{n}) = 1-(1-\frac{t}{n})^n \to 1-e^{-t}$. Hence $n(1 - X_{(n)})$ converges in distribution to $\exp(1)$.

An important fact: X_n converges in probability implies X_n converges in distribution. The reverse is not true in general. For example, take $P(X = 0) = P(X = 1) = \frac{1}{2}$ and $X_n = X$ for all n. Then X and $1 - X$ have the same distribution. Thus X_n converges in distribution to $1 - X$. However, $P(|X_n - (1 - X)| > 1/2) = 1$ for all n. Therefore X_n doesn't converge in probability to $1 - X$.

Referring to the question in the class. Why the definition of convergence in distribution is limited to the continuity point of F_X. Let $X_n = \frac{1}{n}$ and $X = 0$. There is noting random in X_n and X and as a deterministic sequence X_n converges to X. Now we expect that when a deterministic sequence converges to a number, the sequence of random variables degenerate at this deterministic sequence should converge in distribution. Now

$$F_{X_n}(x) = \begin{cases} 0, \text{ if } x < \frac{1}{n} \\ 1 \;\; x \geq \frac{1}{n} \end{cases}$$

Thus

$$\lim_{n\to\infty} F_{X_n}(x) = \begin{cases} 0, \text{ if } x \leq 0 \\ 1 \;\; x > 0 \end{cases}$$

However,

$$F_X(x) = \begin{cases} 0, \text{ if } x < 0 \\ 1 \;\; x \geq 0 \end{cases}$$

In general X_n, Y_n converge in distribution to X, Y respectively in distribution does not mean $X_n + Y_n$ converges to $X + Y$. We need some additional condition provided by the following theorem.

An important result bridging two types of convergence (Slutsky Thoerem): If $X_n \to X$ in distribution and $Y_N \to a$ in probability, then (a)$Y_n X_n \to aX$ in distribution, (b) $Y_n + X_n \to Y + a$ in distribution.

Most Important applications of the two types of convergence

Weak law of large number: Let $X_1, ..., X_n$ be iid random variables with $EX_i = \mu$, $Var(X_i) = \sigma^2 < \infty$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Then, for every $\epsilon > 0$, $\lim_{n\to\infty} P(|\bar{X}_n - \mu| < \epsilon) = 1$.

Central limit theorem: Let $X_1, ..., X_n$ be a sequence of iid random variables whose mgf exists in a nbd. of 0. Let $EX_i = \mu$, $Var(X_i)\sigma^2 > 0$. Define $\bar{X}_n = (1/n)\sum_{i=1}^n X_i$. Let $G_n(x)$ denote the cdf of $\sqrt{n}(\bar{X}_n - \mu)/\sigma$. Then, for any x, $-\infty < x < \infty$, $\lim_{n\to\infty} G_n(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy$.

Central limit theorem is the single most important result in statistics. It talks about large sample behaviour of the mean of a random sample and also justifies popular usage of normal distribution in statistical world. What happens to functions of random variables. *Delta* *method* below is going to give that answer.

Delta Theorem: Let Y_n be a sequence of random variables that satisfies $\sqrt{n}(Y_n - \theta)$ converges in distribution to $N(0, \sigma^2)$. For a given function g and a specific value of θ, suppose that $g'(\theta)$ exists and is not 0. Then

$$\sqrt{n}[g(Y_n) - g(\theta)] \to N(0, \sigma^2[g'(\theta)]^2) \text{ in distribution.}$$

If $g'(\theta) = 0$ and $g''(\theta)$ exists and nonzero, then

$$n[g(Y_n) - g(\theta)] \to \sigma^2 \frac{g''(\theta)}{2} \chi_1^2 \text{ in distribution.}$$

example: CLT gives us $\sqrt{n}(\bar{X}_n - \theta) \to N(0, \sigma^2)$. What is the limiting distribution of $\sqrt{n}(\frac{1}{\bar{X}_n} - \frac{1}{\theta})$.

Exercise: 5.3, 5.4, 5.8, 5.13, 5.22, 5.23, 5.24, 5.44, 5.52 & 5.53 to check CLT.

Statistical Inferential Tools

Our subject is all about using a random sample to produce estimates of unknown parameters in the model. From random sample we create a number of summary measures to understand the behavior of the unknown distribution. For example, we calculate mean or variance to understand central tendency or dispersion of the unknown distribution. While calculating these statistics, we are essentially reducing our data. Question is how should we reduce data optimally? In the next few classes we are going to see some principles.

Sufficiency

Definition 1: Let $\boldsymbol{X} = (X_1, ..., X_n)$ and $\boldsymbol{X} \sim F(\boldsymbol{x} \,|\, \theta)$. $T(\boldsymbol{X})$ is known to be the *sufficient statistic* for θ if the conditional distribution of $\boldsymbol{X}|T(\boldsymbol{X})$ is independent of θ. Intuitively, $T(\boldsymbol{X})$ contains the "same information" about θ that $\boldsymbol{X}$ contains. There is no "additional information" which is required to make proper inference on θ.

Example: Let $X_1, X_2, X_3 \overset{iid}{\sim} \mathit{Bernoulli}(p)$. Density of the Bernoulli distribution is given by

$$f(X) = p^X(1-p)^{1-X}, \;\; X = 0, 1.$$

Claim: $T(X_1, X_2, X_3) = \sum_{i=1}^{3} X_i$ is the sufficient statistics for p.

Proof $P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T(X_1, X_2, X_3) = t) = 0$, if $\sum_{i=1}^{3} x_i \neq t$. If $\sum_{i=1}^{3} x_i = t$,

$$\begin{aligned}
&P(X_1 = x_1, X_2 = x_2, X_3 = x_3 | T(X_1, X_2, X_3) = t) \\
&= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3, T(X_1, X_2, X_3) = t)}{P(T(X_1, X_2, X_3) = t)} \\
&= \frac{P(X_1 = x_1, X_2 = x_2, X_3 = x_3)}{P(T(X_1, X_2, X_3) = t)} \\
&= \frac{P(X_1 = x_1)P(X_2 = x_2)P(X_3 = x_3)}{P(T(X_1, X_2, X_3) = t)} \quad [\text{As } X_1, X_2, X_3 \text{ are iid}] \\
&= \frac{p^{\sum_{i=1}^{3} x_i}(1-p)^{3-\sum_{i=1}^{3} x_i}}{\binom{3}{t}p^t(1-p)^{3-t}} \quad [X_1, X_2, X_3 \sim \mathit{Bernouilli}(p) \Rightarrow T(X_1, X_2, X_3) \sim \mathit{Bin}(3, p)] \\
&= \frac{p^t(1-p)^{3-t}}{\binom{3}{t}p^t(1-p)^{3-t}} = \frac{1}{\binom{3}{t}}.
\end{aligned}$$

Above is a rigorous proof the fact that $T(X_1, X_2, X_3) = \sum_{i=1}^{3} X_i$ is sufficient statistics for p. Let us examine that example with more details and try to make more intuition out of it. Let us see the probability of occurring different values $\mathcal{A}_1 = \{000\}, \mathcal{A}_2 = \{001, 010, 100\}, \mathcal{A}_3 = \{110, 011, 101\}, \mathcal{A}_4 = \{111\}$ are sets whose elements have the the same probability of occurrence. Note that, for every element of $\mathcal{A}_t$, $T(X_1, X_2, X_3) = t$. In other words, given any random sample $\boldsymbol{X} = (X_1, X_2, X_3)$ (more generally for $\boldsymbol{X} = (X_1, ..., X_n)$), it is enough to know $\sum X_i = T(\boldsymbol{X})$ to write down the likelihood of p. Therefore, only information on $T(\boldsymbol{X})$ is sufficient to infer on p as opposed to the entire sample, hence the name "sufficient statistics".

cases	probability
000	$(1-p)^3$
001	$(1-p)^2p$
010	$(1-p)p(1-p) = (1-p)^2p$
100	$p(1-p)^2$
110	$p^2(1-p)$
101	$p(1-p)p = p^2(1-p)$
011	$(1-p)p^2$
111	p^3

Table 1: Probabilities of random samples

This is a more formal way to look into it for a general distribution. Note that $P_\theta(X = x) = P(X = x|T(X) = T(x))P_\theta(T(X) = T(x))$. Therefore, only the distribution of $T(X)$ is contributing in the likelihood of θ. Hence $T(X)$ is sufficient.

Question: How to find out sufficient statistics in a general set up ?

Theorem (Factorization Theorem): Let $\boldsymbol{X}$ have joint p.d.f (or p.m.f) $f_\theta(\boldsymbol{X})$, where θ is the unknown parameter. A statistic $T(\boldsymbol{X})$ is sufficient statistic for θ if and only if $f_\theta(\boldsymbol{X})$ can be expressed as $f_\theta(\boldsymbol{X}) = g(T(\boldsymbol{X}), \theta)h(\boldsymbol{X})$, where $h(\boldsymbol{X})$ is a function of $\boldsymbol{X}$ which is independent of θ.

proof: We will see the proof in the discrete case only just to simplify things. Let us prove the "only if" part first.

$$P[\boldsymbol{X} = \boldsymbol{x}] = \sum_t P[\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t]P[T(\boldsymbol{X}) = t]$$

Now for only one t $P[\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t]$ is positive. Hence $P[\boldsymbol{X} = \boldsymbol{x}] = P[\boldsymbol{X} = \boldsymbol{x}|T(\boldsymbol{X}) = t]P[T(\boldsymbol{X}) = t] = h(\boldsymbol{x})g(T(\boldsymbol{x}), \theta)$. This proves the only if part. Now we will prove the "if part".

$$P[T(\boldsymbol{X}) = t] = \sum_{\boldsymbol{x}\mathcal{A}_t} f_\theta(\boldsymbol{x}) = \sum_{\boldsymbol{x}\mathcal{A}_t} g(T(\boldsymbol{x}), \theta)h(\boldsymbol{x}) = g(t, \theta)\sum_{\boldsymbol{x}\mathcal{A}_t} h(\boldsymbol{x}).$$

Thus

$$P[\boldsymbol{X}=\boldsymbol{x}|T(\boldsymbol{X})=t]=\begin{cases}\frac{g(t,\theta)h(\boldsymbol{x})}{g(t,\theta)\sum_{\boldsymbol{x}\mathcal{A}_t}h(\boldsymbol{x})}, \text{ if } \boldsymbol{x}\in\mathcal{A}_t\\ 0 \text{ o.w.}\end{cases}$$

Example 1: Recall the last example, $X_1,...,X_n \sim \mathit{Bernoulli}(p)$. Then

$$f_p(\boldsymbol{X})=\prod_{i=1}^n p^{X_i}(1-p)^{1-X_i}=p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}=\left(\frac{p}{1-p}\right)^{\sum_{i=1}^n X_i}(1-p)^n.$$

Therefore $h(\boldsymbol{X})=1$ and sufficient statistic is $T(\boldsymbol{X})=\sum_{i=1}^n X_i$.

Example 2: Suppose $X_1,...,X_n \sim \mathit{Poisson}(\lambda)$. Then

$$f_\lambda(\boldsymbol{X})=\prod_{i=1}^n\left[\frac{\exp(-\lambda)\lambda^{X_i}}{X_i}\right]=\frac{\exp(-n\lambda)\lambda^{\sum_{i=1}^n X_i}}{\prod_{i=1}^n X_i}.$$

Therefore $h(\boldsymbol{X})=\frac{1}{\prod_{i=1}^n X_i}$ and $T(\boldsymbol{X})=\sum_{i=1}^n X_i$ with $g(T(\boldsymbol{X}),\lambda)=\exp(-n\lambda)\lambda^{\sum_{i=1}^n X_i}$.

Example 3: Suppose $X_1,...,X_n \sim N(\mu,\sigma^2)$, μ is an unknown parameter, σ^2 known. Then

$$f_\mu(\boldsymbol{X})=\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\right)=\left[\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right)\right]\\ \times\exp\left(-\frac{n\mu^2-2\mu\sum_{i=1}^n X_i}{2\sigma^2}\right).$$

Hence $h(\boldsymbol{X})=\left[\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2\right)\right]$ and $T(\boldsymbol{X})=\sum_{i=1}^n X_i$.

Example 4: Suppose $X_1,...,X_n \sim N(\mu,\sigma^2)$, μ,σ^2 both unknown parameters. Then

$$f_{\mu,\sigma^2}(\boldsymbol{X})=\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n(X_i-\mu)^2\right)=\left[\frac{1}{(\sqrt{2\pi\sigma^2})^n}\exp\left(-\frac{\sum_{i=1}^n X_i^2}{2\sigma^2}+\frac{2\mu\sum_{i=1}^n X_i}{2\sigma^2}-\frac{n\mu^2}{2\sigma^2}\right)\right].$$

Therefore, $h(\boldsymbol{X})=1$ and $T(\boldsymbol{X})=(\sum_{i=1}^n X_i^2,\sum_{i=1}^n X_i)$.

Example 5: Suppose $X_1, ..., X_n \sim U(0, \theta)$. Then

$$f_\theta(\boldsymbol{X}) = \frac{1}{\theta^n} I(0 < X_1 < \theta, ..., 0 < X_n < \theta) = \frac{1}{\theta^n} I(X_{(n)} < \theta) I(X_{(1)} > 0),$$

where $X_{(n)}, X_{(1)}$ are biggest and smallest order statistics from $X_1, ..., X_n$. Therefore, $T(\boldsymbol{X}) = X_{(n)}$.

Example 6: Suppose $X_1, ..., X_n \sim U(\theta_1, \theta_2)$. Then

$$f_\theta(\boldsymbol{X}) = \frac{1}{(\theta_2 - \theta_1)^n} I(\theta_1 < X_1 < \theta_2, ..., \theta_1 < X_n < \theta_2) = \frac{1}{(\theta_2 - \theta_1)^n} I(X_{(n)} < \theta_2) I(X_{(1)} > \theta_1),$$

where $X_{(n)}, X_{(1)}$ are biggest and smallest order statistics from $X_1, ..., X_n$. Therefore, $T(\boldsymbol{X}) = (X_{(1)}, X_{(n)})$.

Some Important Facts:

(a) $T(\boldsymbol{X}) = (X_1, ..., X_n)$, i.e. the full sample is always sufficient for the unknown parameter.

(b) If $X_1, ..., X_n \overset{iid}{\sim} f_\theta(x)$ then, $f(\boldsymbol{X}) = \prod_{i=1}^{n} f_\theta(X_i) = \prod_{i=1}^{n} f_\theta(X_{(i)})$. This means order statistics $X_{(1)} \leq \cdots \leq X_{(n)}$ is always sufficient for θ. Of course this is not a big reduction, but with so little information you can't reduce sample much without losing any "information".

(c) Any one to one function of a sufficient statistics is also sufficient.

- In examples 1,2,3, $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$ is also sufficient, being a one-one function of $\sum_{i=1}^n X_i$.
- In example 4, $(\bar{X}, S^2) = K(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$, where $K(z_1, z_2) = (z_1/n, z_2/n - z_1^2/n^2)$ which is a one to one function. Therefore $(\bar{X}, S^2)$ is a sufficient statistics.

In general, you can create a lot of sufficient statistics for a problem. Let us go back to the Bernoulli example we started with. $X_1, X_2, X_3 \sim \textit{Bernoulli}(p)$. We have seen $\sum_{i=1}^{3} X_i$ is a sufficient statistic. We also know from (a) that the full sample is sufficient statistic. Note that

$$f_p(\boldsymbol{X}) = p^{\sum_{i=1}^{3} X_i}(1-p)^{3-\sum_{i=1}^{3} X_i} = p^{\sum_{i=1}^{2} X_i + X_3}(1-p)^{3-\sum_{i=1}^{2} X_i - X_3}.$$

Therefore $(\sum_{i=1}^{2} X_i, X_3)$ is a sufficient statistic. Also you will be able to find many other sufficient statistics. Any sufficient statistic is providing summary of the dataset that one can deal with without losing any information from the entire data. Therefore we are more interested in knowing the coarsest summary of the data without losing any information. Below is a concept that explains as to how far we can proceed in summarizing the data without losing any information contained in it.

Definition (Minimal Sufficiency): A statistic $T(\boldsymbol{X})$ is *minimal sufficient* if (a) it is sufficient, and (b) it is function of every other sufficient statistic.

Consider the good old example of Bernoulli. $T_1(\boldsymbol{X}) = (X_1, X_2, X_3)$, $T_2(\boldsymbol{X}) = (\sum_{i=1}^{2} X_i, X_3)$, $T_3(\boldsymbol{X}) = \sum_{i=1}^{3} X_i$ are all sufficient statistics foo p, as we have seen earlier. However T_2 is a function of T_1 and T_3 is a function of both T_1 and T_2. Further T_1 is one-dimensional and you can't make anything lower dimensional than that. So, T_1 has to be a minimal sufficient statistic for p.

Question: How to find *minimal sufficient* statistics in more general set ups.

Theorem (Minimal Sufficiency): Let $f_\theta(\boldsymbol{X})$ be the p.d.f (or, p.m.f) of $\boldsymbol{X}$. Suppose there exists a statistic T s.t. for any two realizations $\boldsymbol{x}$, $\boldsymbol{y}$ of the sample $T(\boldsymbol{x}) = T(\boldsymbol{y})$ if and only if $f_\theta(\boldsymbol{x}) = kf_\theta(\boldsymbol{y})$ where k is independent of θ, then T is a minimal sufficient statistic of θ.

Example 7: Lets look at our favorite example, $X_1, X_2, X_3 \sim Bernoulli(p)$. We have argued T_3 is minimal sufficient from a different angle. Now lets look at it in the light of this theorem.

$$\frac{f_p(\boldsymbol{x})}{f_p(\boldsymbol{y})} = \frac{p^{\sum_{i=1}^3 x_i}(1-p)^{3-\sum_{i=1}^3 x_i}}{p^{\sum_{i=1}^3 y_i}(1-p)^{3-\sum_{i=1}^3 y_i}} = \left(\frac{p}{1-p}\right)^{\sum_{i=1}^3 x_i - \sum_{i=1}^3 y_i}.$$

This ratio is constant if and only if $\sum_{i=1}^3 x_i = \sum_{i=1}^3 y_i$. Hence $T_3(\boldsymbol{X}) = \sum_{i=1}^3 X_i$ is the minimal sufficient statistic. Why $T_2(\boldsymbol{X}) = (\sum_{i=1}^2 X_i, X_3)$ is not the minimal sufficient. As $\left(\frac{p}{1-p}\right)^{\sum_{i=1}^3 x_i - \sum_{i=1}^3 y_i}$ can be a constant even if $\sum_{i=1}^2 x_i \neq \sum_{i=1}^2 y_i$.

Example 8: Suppose $X_1, ..., X_n \sim N(\mu, \sigma^2)$, μ, σ^2 both unknown parameters. Then

$$\begin{aligned}\frac{f_{\mu,\sigma^2}(\boldsymbol{x})}{f_{\mu,\sigma^2}(\boldsymbol{y})} &= \frac{\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i + n\mu^2\right]\right)}{\exp\left(-\frac{1}{2\sigma^2}\left[\sum_{i=1}^n y_i^2 - 2\mu\sum_{i=1}^n y_i + n\mu^2\right]\right)} \\ &= \exp\left(-\frac{1}{2\sigma^2}\left[(\sum_{i=1}^n x_i^2 - \sum_{i=1}^n y_i^2) - 2\mu(\sum_{i=1}^n x_i - \sum_{i=1}^n y_i)\right]\right).\end{aligned}$$

This ratio is constant if and only if $\sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$ and $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$. Therefore $T(\boldsymbol{X}) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is the minimal sufficient statistics.

Example 9: Suppose $X_1, ..., X_n \sim U(\theta, \theta+1)$, $-\infty < \theta < \infty$. We have seen the joint pdf is

$$\frac{f_\theta(\boldsymbol{x})}{f_\theta(\boldsymbol{y})} = \frac{I(\theta < x_1 < \theta+1, ..., \theta < x_n < \theta+1)}{I(\theta < y_1 < \theta+1, ..., \theta < y_n < \theta+1)} = \frac{I(x_{(1)} > \theta, x_{(n)} - 1 < \theta)}{I(y_{(1)} > \theta, y_{(n)} - 1 < \theta)}.$$

The ratio is constant if and only if $(x_{(1)}, x_{(n)}) = (y_{(1)}, y_{(n)})$. Hence the minimal sufficient statistics is $(X_{(1)}, X_{(n)})$.

Remark: Any one to one function of a minimal sufficient statistics is also minimal sufficient.

Minimal sufficient statistic is not unique.

Ancillary Statistics

In the previous subsection we see sufficient statistics which are summarization of the sample without losing any "information". Sufficient statistics are something which contain all information about θ. We are now going to introduce a different sort of statistics.

Definition (Ancillary Statistic): A statistics whose distribution does not depend on the unknown parameter θ is known as an *ancillary statistic*.

It seems to us that ancillary statistics has nothing to do with θ. Then why are we interested in it? We will see later that ancillary statistics sometimes can give information for inference about θ.

Location family ancillary statistics: Let $X_1, ..., X_n \sim F(x-\theta)$, $-\infty < \theta < \infty$. This implies $Z_i = X_i - \theta \sim F(x)$. Consider the distribution of $R = X_{(n)} - X_{(1)}$, the range statistic. Now

$$P_\theta(R \leq r) = P_\theta(X_{(n)} - X_{(1)} \leq r) = P_\theta(\max_i(Z_i + \theta) - \min_i(Z_i + \theta) \leq r) = P_\theta(Z_{(n)} - Z_{(1)} + \theta - \theta \leq r).$$

Last probability doesn't depend on θ. So R is an ancillary statistics for the location family.

Example 10: $X_1, ..., X_n \sim U(\theta, \theta+1)$. This implies $X_i - \theta \sim U(0,1)$. Thus $R = X_{(n)} - X_{(1)}$ is an ancillary statistics.

Example 11: $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ^2 known. This implies $X_i - \mu \sim N(0, \sigma^2)$. Thus $R = X_{(n)} - X_{(1)}$ is an ancillary statistics.

Scale family ancillary statistics: Let $X_1, ..., X_n \sim F(x/\sigma)$, $\sigma > 0$. Any statistic that depends on the sample only through the $n-1$ values X_1/X_n,...,X_{n-1}/X_n is an ancillary statistic.

Note that $Z_i = X_i/\sigma \sim F(x)$. Therefore the joint CDF of $X_1/X_n, ..., X_{n-1}/X_n$ is the same as the joint CDF of $Z_1/Z_n, ..., Z_{n-1}/Z_n$. Hence any function of X_1/X_n,...,X_{n-1}/X_n has distribution free of θ.

Example 12: $X_1, ..., X_n \sim N(0, \sigma^2)$, then $X_i/\sigma \sim N(0,1)$. Hence it is a scale family with ancillary statistics as above.

As was said earlier, ancillary statistics together with some other statistic provide important information about θ. For example, we have seen in example 9 that the minimal sufficient statistic is $(X_{(1)}, X_{(n)})$. By the property that any one to one function of a minimal sufficient statistic is also minimal sufficient means $(X_{(1)} - X_{(n)}, \frac{X_{(1)}+X_{(n)}}{2})$ is also minimal sufficient. However we have seen in this example $X_{(1)} - X_{(2)}$ is an ancillary statistic. Therefore, ancillary statistic although gives no information on θ alone can give information on θ together with some other statistic. Below we are going to give more insight on this phenomenon.

Example 13: Let X_1, X_2 be iid drawn from a distribution which has p.m.f

$$P(X = \theta) = P(X = \theta + 1) = P(X = \theta + 2) = \frac{1}{3},$$

where θ is an integer and unknown. Here also the minimal sufficient statistics is $(X_{(1)}, X_{(2)})$ and again by a one-one transformation $(X_{(1)} - X_{(n)}, \frac{X_{(1)}+X_{(n)}}{2})$ is minimal sufficient. Let me denote the minimal sufficient statistic by (r, m) and let m be an integer. Given only m, θ can be any of the three values $\theta = m, m-1, m-2$. However, if we additionally know $r = 2$ then it can be concluded that $X_{(1)} = \theta, X_{(2)} = \theta + 2$. Thus $m = \theta + 1 \Rightarrow \theta = m - 1$. Thus r also provides crucial information for the inference on θ.

This example also proves the fact that ancillary statistics, although contains no information about θ in itself, is not independent of the minimal sufficient statistics. We need some additional conditions to hold for a minimal sufficient statistic to be independent of ancillary statistics. A description of situations in which this occurs relies on the following definition.

Definition (Complete Statistic): Let $f_\theta(t)$ be a family of pdfs (or pmfs) for a statistic $T(\boldsymbol{X})$. The family of distributions is called *complete* if $E_\theta(g(T)) = 0$ for all θ implies $P_\theta(g(T) = 0) = 1$ for all θ. Equivalently, $T(\boldsymbol{X})$ is called a *complete statistic.*

Note that completeness is a stronger definition than minimal sufficiency. Indeed

Theorem: If a minimal sufficient statistic exists, then any complete sufficient statistic is also a minimal sufficient statistic.

Proof Let T be a complete sufficient statistic and S is minimal sufficient. S is a function of T as S is minimal sufficient. Now $E[T|S] = g(S) \Rightarrow E[(T - g(S))|S] = 0 \Rightarrow E[T - g(S)] = 0$. Given that S is a function of T, by completeness we have $T = g(S)$. Therefore T is minimal sufficient.

Notice that completeness is a property for a family of distributions, not of a particular distribution. Let us discuss a few examples of complete statistics. Later we will provide complete sufficient statistics for a broad class of distribution.

example: Suppose $T \sim Bin(n, p)$ and let g be a function s.t. $E_p[g(T)] = 0$. This implies for all p

$$0 = \sum_{k=0}^{n} g(k) \binom{n}{k} p^k (1-p)^{n-k} = (1-p)^n \sum_{k=0}^{n} g(k) \binom{n}{k} \left(\frac{p}{1-p}\right)^k.$$

Thus a polynomial $f(t) = \sum_{k=0}^{n} g(k) \binom{n}{k} t^k$ is identically zero for all t. This means every coefficient is zero, i.e. $g(k) = 0$ for all k. Hence $g = 0$.

example: $X_1, ..., X_n \overset{\sim}{iid} U(0, \theta)$, $0 < \theta < \infty$. Let $T(X_1, .., X_n) = \max_i X_i$ be a statistic. We will show it is a complete sufficient statistics for θ. Note that

$$\begin{aligned} P(T \leq t) &= P(X_1 < t, ..., X_n < t) = P(X_1 < t) \cdots P(X_n < t) = t^n \theta^{-n},\ 0 < t < \theta \\ &= 1 \text{ if } t > \theta \\ &= 0 \text{ if } t < 0. \end{aligned}$$

Therefore the density of T is given by $f(t|\theta) = nt^{n-1}\theta^{-n},\ 0 < t < \theta$. Suppose g be a fn. s.t. $E\theta[g(T)] = 0$ for all θ. Then

$$0 = \frac{d}{d\theta} E_\theta[g(T)] = \frac{d}{d\theta} \int_0^\theta g(t) n t^{n-1} \theta^{-n} dt = g(\theta) n \theta^{n-1} \theta^{-n}.$$

Since this is true for all θ, it implies that $g = 0$.

We are now in a position to discuss when a minimal sufficient statistic is independent of an ancillary statistic.

Basu's Theorem: If $T(\boldsymbol{X})$ is a complete and sufficient statistic, then $T(\boldsymbol{X})$ is independent of any ancillary statistic.

Proof (Only for the simple discrete case): Let $S(\boldsymbol{X})$ be any ancillary statistic. Then $P_\theta(S(\boldsymbol{X}) = s)$ does not depend on θ. Since $T(\boldsymbol{X})$ is a sufficient statistic hence $P_\theta(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) = P_\theta(\boldsymbol{X} \in \{\boldsymbol{x} : S(\boldsymbol{x}) = s\}|T(\boldsymbol{X}) = t)$ is independent of θ. Now

$$P_\theta(S(\boldsymbol{X}) = s) = \sum_t P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) P_\theta(T(\boldsymbol{X}) = t). \tag{1}$$

Furthermore since $P(S(\boldsymbol{X}) = s) = \sum_t P(S(\boldsymbol{X}) = s) P_\theta(T(\boldsymbol{X}) = t)$, using (1) we have for

$$g(t) = P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) = t) - P(S(\boldsymbol{X}) = s),$$

$E_\theta[g(T)] = 0$ for all θ. Now using completeness of T we obtain $P(S(\boldsymbol{X}) = s|T(\boldsymbol{X}) =$

$t) - P(S(\boldsymbol{X}) = s) = 0$. This proves that $T(\boldsymbol{X})$ and $S(\boldsymbol{X})$ are independent.

Basu's theorem sometimes turns out to be an extremely useful technique. Consider the following classic examples.

Example 13: Consider $X_1, ..., X_n \sim exp(\theta)$, need to find $E_\theta\left[\frac{X_n}{\sum_{i=1}^n X_i}\right]$. Note that $f_\theta(x) = \frac{1}{\theta}\exp(-x/\theta)$. Therefore $X/\theta \sim exp(1)$ implying that it is scale family. By a previous example, $g(\boldsymbol{x}) = \frac{X_n}{\sum_{i=1}^n X_i} = \frac{1}{\sum_{i=1}^n \frac{X_i}{X_n}}$ is an ancillary statistic. It is easy to show that $T(\boldsymbol{X}) = \sum_{i=1}^n X_i$ is a complete sufficient statistic. Therefore, $T(\boldsymbol{X})$ and $g(\boldsymbol{X})$ are independent. Thus

$$\theta = E_\theta[X_n] = E_\theta[g(\boldsymbol{X})T(\boldsymbol{X})] = E_\theta[g(\boldsymbol{X})]E_\theta[T(\boldsymbol{X})] = E_\theta[g(\boldsymbol{X})]n\theta.$$

Hence $E_\theta[g(\boldsymbol{X})] = n^{-1}$.

Exponential Family

A one parameter exponential family density is given by $f_\theta(x) = h(x)c(\theta)\exp\left(w(\theta)t(x)\right)$.

Exercise: Show how $Bin(p), Pois(\lambda)$ is a one parameter exponential family.

Now note that

$$\begin{aligned}
0 &= \frac{d}{d\theta}\int h(x)c(\theta)\exp\left(w(\theta)t(x)\right)d\theta \\
&= \int h(x)\left[c'(\theta)\exp\left(w(\theta)t(x)\right) + c(\theta)w'(\theta)t(x)\exp\left(w(\theta)t(x)\right)\right]d\theta \\
&= \frac{c'(\theta)}{c(\theta)} + w'(\theta)E[t(X)].
\end{aligned}$$

$E[t(X)] = -\frac{c'(\theta)}{w'(\theta)c(\theta)}$. Taking derivative one more time we can calculate $E[t(X)^2], Var(t(X))$.

Similarly one encounters multi-parameter exponential family. A multi-parameter exponential family has density

$$f_{\boldsymbol{\theta}}(x) = h(x)c(\boldsymbol{\theta})\exp\left(\sum_{i=1}^k w_i(\boldsymbol{\theta})t_i(x)\right).$$

Clearly by factorization theorem, $(\sum_{j=1}^n t_1(X_j), ..., \sum_{j=1}^n t_k(X_j))$ is sufficient and by the next theorem it is minimal sufficient.

Remark: It can also be shown that $(\sum_{j=1}^n t_1(X_j), ..., \sum_{j=1}^n t_k(X_j))$ is also complete sufficient statistic if $\{(w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \theta\}$ contains an open set in $\mathbb{R}^k$.

Result borrowed from the Fourier Transformation:

If $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(y_1, ..., y_k) \exp(t_1 y_1 + \cdots + t_k y_k) dy_1 \cdots dy_k = 0$ for $a_i < t_i < b_i$ for all $i = 1, ..., k$ then $g = 0$.

We are going to borrow this result to prove the remark. Note that $T(\boldsymbol{X}) = (\sum_{j=1}^n t_1(X_j), ..., \sum_{j=1}^n t_k(X_j))$ is a sufficient statistics for $\boldsymbol{\theta}$. Now $E[g(T(\boldsymbol{X}))] = 0$ for all $\boldsymbol{\theta}$ implies

$$\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(\sum_{j=1}^n t_1(X_j), ..., \sum_{j=1}^n t_k(X_j)) \exp(w_1(\boldsymbol{\theta}) \sum_{j=1}^n t_1(X_j) + \cdots + w_k(\boldsymbol{\theta}) \sum_{j=1}^n t_k(X_j)) = 0. \tag{2}$$

Now $\{(w_1(\boldsymbol{\theta}), ..., w_k(\boldsymbol{\theta})) : \boldsymbol{\theta} \in \theta\}$ contains an open set in $\mathbb{R}^k$ means it there exist intervals $[a_i, b_i]$ in every dimension so that $a_i < w_i(\boldsymbol{\theta}) < b_i$ for which (2) holds. By the previous result we have $g = 0$.

This if condition is important. For example if $X_1, ..., X_n \sim N(\theta, \theta^2)$. We can't apply the theorem here.

Likelihood Principle

The last topic of this chapter is another principle known as the "likelihood principle". Likelihood principle tells us that all the inferences on the parameter should be only based on the likelihood function. What is a likelihood function? Below we give definition of the likelihood function.

Definition (Likelihood function): Let $f_\theta(\boldsymbol{x})$ be the joint pdf or pmf of the sample $\boldsymbol{X} = (X_1, ..., X_n)$. Then given that $\boldsymbol{X} = \boldsymbol{x}$ is observed, the function of θ defined by

$L(\theta|\boldsymbol{x}) = f_\theta(\boldsymbol{x})$ is called the likelihood function.

Likelihood Principle: If $\boldsymbol{x}$ and $\boldsymbol{y}$ are two sample points such that $L(\theta|\boldsymbol{x})$ is proportional to $L(\theta|\boldsymbol{y})$, that is there exists a constant $C(\boldsymbol{x}, \boldsymbol{y})$ such that

$$L(\theta|\boldsymbol{x}) = C(\boldsymbol{x}, \boldsymbol{y})L(\theta|\boldsymbol{y}), \quad \text{for all } \theta,$$

then the conclusion drawn from $\boldsymbol{x}$ and $\boldsymbol{y}$ should be identical. Note that the constant $C(\boldsymbol{x}, \boldsymbol{y})$ may be different for different $(\boldsymbol{x}, \boldsymbol{y})$ pair, but it does not depend on θ.

Likelihood principle says inference must be fully based on the likelihood. If for two values θ_1, θ_2 of θ we have $L(\theta_2|\boldsymbol{x}) = 3L(\theta_1|\boldsymbol{x})$, then θ_2 is thrice "probable" as a value of θ. Further if likelihood principle is true then $L(\theta_2|\boldsymbol{y}) = 3L(\theta_2|\boldsymbol{y})$. Thus whether we observe $\boldsymbol{x}, \boldsymbol{y}$ we conclude that θ_2 is thrice more likely as a value of θ than θ_1. According to likelihood principle the most likely value of θ is the one that maximizes likelihood. This is how likelihood principle gives rise to the "maximum likelihood estimator".

However, likelihood principle is quite controversial and it contradicts frequentist inference in many example. I will show you a very popular one.

example: Let X be the number of success in twelve Bernoulli trial with success prob. θ. Then $X \sim Bin(12, \theta)$. Suppose we observe 3 successes. Then the likelihood of θ is

$$L(\theta|X = 3) = \binom{12}{3}\theta^3(1-\theta)^9.$$

Let Y be the number of trials required to have 3 successes. $Y \sim NegBin(3, \theta)$. The likelihood of θ here is

$$L(\theta|Y = 12) = \binom{11}{2}\theta^3(1-\theta)^9.$$

Since the two likelihoods are merely proportional to each other for all θ, therefore likelihood principle says we should have the same inference on θ. However, it has been shown that $H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta > \frac{1}{2}$ has p-value of 0.07 in the first case, while 0.03 in the second case. We will describe more when we study hypothesis testing. Therefore, with standard frequentist testing procedure, we draw two different conclusions. Therefore, frequentist procedure has contradiction with the likelihood principle.

Exercise: 6.2, 6.3, 6.5, 6.6, 6.9, 6.10, 6.13, 6.14, 6.16, 6.20, 6.22, 6.30.

1 Techniques to evaluate estimators

In the previous section we studied a few concepts on sufficiency, minimal sufficiency and completeness. Those are tools to evaluate "how good" is the data reduction achieved by an estimator and how much information is lost, if any. In this section, we will use these tools (and introduce some other) to create "optimal" point estimator. First we need a metric under which we can evaluate any estimator.

Definition (Mean Squared Error): If $\tau(\theta) \neq 0$ is a function of θ and $T(\boldsymbol{X})$ be an estimator used to estimate $\tau(\theta)$, then the mean squared error (MSE) of $T(\boldsymbol{X})$ is given by $E_\theta(T(\boldsymbol{X}) - \tau(\theta))^2$. Note that,

$$
\begin{aligned}
&E_\theta(T(\boldsymbol{X}) - \tau(\theta))^2 = E_\theta(T(\boldsymbol{X}) - E_\theta(T(\boldsymbol{X})) + E_\theta(T(\boldsymbol{X})) - \tau(\theta))^2 \\
&= E_\theta(T(\boldsymbol{X}) - E_\theta(T(\boldsymbol{X})))^2 + 2E_\theta((T(\boldsymbol{X}) - E_\theta(T(\boldsymbol{X})))(E_\theta(T(\boldsymbol{X})) - \tau(\theta))) + E_\theta(E_\theta(T(\boldsymbol{X})) - \tau(\theta))^2 \\
&= E_\theta(T(\boldsymbol{X}) - E_\theta(T(\boldsymbol{X})))^2 + E_\theta(E_\theta(T(\boldsymbol{X})) - \tau(\theta))^2 \\
&= Var_\theta(T(\boldsymbol{X})) + Bias_\theta(T(X))^2.
\end{aligned}
$$

Given any function of θ (say $\tau(\theta)$), we would ideally like to obtain an estimator $T(\boldsymbol{X})$ that has the lowest MSE, uniformly over all θ. However, this is not possible to achieve. Consider the estimator $T(\boldsymbol{X}) = 10$, which is a terrible as an estimator, but when $\theta = 10$, it gives $MSE = 0$. Therefore it is not possible to achieve an estimator which is uniformly best across

θ over all other estimators, in terms of MSE. We restrict the class of estimators among which we are going to find out estimator with the best MSE. Let

$$\mathcal{C}_\tau = \{T : E_\theta(T(\boldsymbol{X})) = \tau(\theta)\}$$

be a class of estimators. Clearly $T \in \mathcal{C}_\tau \Rightarrow Bias_\theta(T(\boldsymbol{X})) = 0$. We call the class $\mathcal{C}_\tau$ as the class of all *unbiased estimators* of $\tau(\theta)$. Our aim is to to find an estimator $T(\boldsymbol{X})$ of $\tau(\theta)$ which satisfies the property that given any other unbiased estimator $W(\boldsymbol{X})$ of $\tau(\theta)$, $MSE_\theta(W) \geq MSE_\theta(T)$ for all θ. Since, for unbiased estimators $MSE_\theta(T) = Var_\theta(T)$, it amounts to finding out an unbiased estimator T s.t $Var_\theta(W) \geq Var_\theta(T)$ for all θ. Such an estimator T is known as the *uniform minimum variance unbiased estimator* (**UMVUE**) of $\tau(\theta)$. We will see how to find UMVUE for different problems. While doing so, we are going to use concepts which have been introduced earlier. But first we should answer the question if such a UMVUE is unique.

Theorem (Uniqueness of UMVUE) If $T(\boldsymbol{X})$ is the best unbiased estimator of $\tau(\theta)$, then $T(\boldsymbol{X})$ is unique.

Proof: Suppose $W(\boldsymbol{X})$ be another best unbiased estimator and consider $T^*(\boldsymbol{X}) = \frac{T(\boldsymbol{X})+W(\boldsymbol{X})}{2}$. Note that $E[T^*(\boldsymbol{X})] = \tau(\theta)$, hence T^* is unbiased. Also

$$\begin{aligned} Var_\theta(T^*) &= Var_\theta(\frac{T+W}{2}) = \frac{1}{4}Var_\theta(T) + \frac{1}{4}Var_\theta(W) + \frac{1}{2}Cov_\theta(T, W) \\ &\leq Var_\theta(\frac{T+W}{2}) = \frac{1}{4}Var_\theta(T) + \frac{1}{4}Var_\theta(W) + \frac{1}{2}[Var_\theta(T)Var_\theta(W)]^{1/2} \end{aligned}$$
$$= Var_\theta(T),$$

where the second step follows from Cauchy-Schwartz inequality and last step follows from the fact that $Var_\theta(T) = Var_\theta(W)$ for all θ. If the inequality is strict, then it clearly gives a contradiction of the fact that T is UMVUE. If the inequality is an equality then $W = a(\theta)T + b(\theta)$, by the equality of Cauchy-Schwartz. Thus $Cov_\theta(T, W) = a(\theta)Var_\theta(T)$.

But, step 2 is an equality now, hence $Cov_\theta(T, W) = Var_\theta(T)$ implying that $a(\theta) = 1$. Now $E_\theta(T) = E_\theta(W)$ implies $b(\theta) = 0$. Hence $T = W$.

To see when an unbiased estimator is best unbiased, we want to see how can we improve upon a given unbiased estimator. Suppose $T(\boldsymbol{X})$ is an unbiased estimator of $\tau(\theta)$ and $U(\boldsymbol{X})$ is an unbiased estimator of 0, i.e. $E_\theta(T + aU) = \tau(\theta)$, this is also unbiased. Now

$$Var_\theta(T + aU) = Var_\theta(T) + 2aCov_\theta(T, U) + a^2 Var_\theta(U).$$

Now if for some θ_0, $Cov_{\theta_0}(T, U) < 0$, then we can make $2aCov_{\theta_0}(T, U) + a^2 Var_{\theta_0}(U) < 0$ by choosing $a \in (0, -2Cov_{\theta_0}(T, U)/Var_{\theta_0}(U))$. Hence $T + aU$ will be a better estimator at θ_0 and T cannot be UMVUE. Similarly we can show that if $Cov_{\theta_0}(T, U) > 0$ then also T cannot be best unbiased. In fact this observation characterizes an important property of UMVUE.

Theorem: $W(\boldsymbol{X})$ is the UMVUE for $\tau(\theta)$ if and only if W is uncorrelated with all unbiased estimators of 0.

proof: The above argument shows that if W is the UMVUE it must satisfy $Cov_\theta(W, U) = 0$ for all θ for all unbiased estimator U of 0. Now assume W is uncorrelated to all unbiased estimators of 0 and let W' be any other unbiased estimator of $\tau(\theta)$. This implies that W is uncorrelated to $W - W'$. Hence

$$Var_\theta(W) = Var_\theta(W') + Var_\theta(W - W').$$

Hence W is better than W'.

Note that this result is quite difficult to use in practice. However, it can be used as a negative result, i.e. if you like to show that some estimator is not UMVUE, just show that it is correlated to one unbiased estimator of 0.

Example: $X \sim U(\theta, \theta + 1)$. Then $E(X - \frac{1}{2}) = \theta$, i.e. $X - \frac{1}{2}$ is unbiased. If h is an unbiased estimator of 0, then $\int_\theta^{\theta+1} h(x)dx = 0 \Rightarrow h(\theta+1) - h(\theta) = 0$ for all θ. Now $h(x) = \sin(2\pi x)$ satisfies this and $Cov_\theta(X - \frac{1}{2}, \sin(2\pi X)) = -\frac{\cos(2\pi\theta)}{2\pi} \neq 0$.

The above results are all giving characterizations of UMVUE. Now we will move onto the task of constructing UMVUE in different problems.

Rao-Blackwell Theorem: Let W be any unbiased estimator of θ. Let T be a sufficient statistic for θ and $\phi(T) = E[W|T]$. Then

(i) $\phi(T)$ is an unbiased estimator of θ.

(ii) $Var(\phi(T)) \leq Var(W)$, with equality holding if and only if $\phi(T) = W$ with prob. 1.

Proof First of all $\phi(T)$ is a statistic (i.e. free of θ) as T is a sufficient statistic. Now, $E(\phi(T)) = E[E[W|T]] = E[W] = \theta$. So $\phi(T)$ is unbiased. Also $Var(W) = Var(E(W|T)) + E(Var(W|T)) = Var(\phi(T)) + E(Var(W|T)) \geq Var(\phi(T))$.

Example 14: $X_1, X_2, X_3 \sim Bernoulli(p)$. Lets start with any unbiased estimator, say $W = (X_1 + X_2)/2$. Clearly $E(W) = p$, i.e. W is unbiased. We know $T = \sum_{i=1}^3 X_i$ is a sufficient statistic for p. Then $\phi(T) = E[W|T] = T/3$ by symmetry. Now, $Var(W) = p(1-p)/2$, while $Var(\phi(T)) = p(1-p)/3$.

Given any unbiased estimator, Rao-Blackwell theorem provides a way to improve its MSE and we proceed towards achieving a UMVUE. But how much conditioning is needed? Is there any sufficient statistic with which conditioning provides UMVUE. Indeed it is achieved by a complete sufficient statistics as below.

Theorem (Lehman-Scheffe): Suppose T is complete and sufficient and there exists a function $\phi(T)$ of T s.t. $E[\phi(T)] = \psi(\theta)$. Then $\phi(T)$ is UMVUE for $\psi(\theta)$.

proof: Let T_1 be any other unbiased estimator of $\psi(\theta)$. Consider $\phi_1(T) = E[T_1|T]$, this is a statistic and by Rao-Blackwell we have $var(\phi_1(T)) \leq var(T_1)$. Now $E[\phi_1(T)] = E[\phi(T)] = \psi(\theta)$. By completeness of T, we have $\phi_1(T) = \phi(T)$ w.p. 1 for all θ. Hence $\phi(T)$ is the UMVUE.

The above theorem gives us a reasonably easy way to find a UMVUE for $\psi(\theta)$. We have two tasks, (a) find a complete sufficient statistics for θ. For exponential family we already know how to find that, (b) find an unbiased estimator of $\psi(\theta)$ as a function of the complete sufficient statistics. We will see some examples.

Example: Consider $X_1, ..., X_n \sim Bernoulli(p)$. We have already seen that $\sum_{i=1}^n X_i$ is a complete sufficient statistic. Therefore, $T = \sum_{i=1}^n X_i$ is UMVUE for p. What is the UMVUE for p^2? Note that $T = \sum_{i=1}^n X_i \sim Bin(n, p)$. Thus,

$$E[T(T-1)] = E[T^2] - E[T] = Var(T) + E[T]^2 - E[T] = np(1-p) + n^2p^2 - np = n(n-1)p^2$$

implying that $\frac{T(T-1)}{n(n-1)}$ is the UMVUE for p^2.

Example: Consider $X_1, ..., X_n \sim N(\mu, \sigma^2)$. We already know, $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is complete sufficient. $E(\sum_{i=1}^n X_i/n) = \mu$. Thus $\sum_{i=1}^n X_i/n$ is UMVUE FOR μ. Also $E[\sum_{i=1}^n X_i^2/n] = \mu^2 + \sigma^2$. Hence $\sum_{i=1}^n X_i^2/n$ is UMVUE for $\mu^2 + \sigma^2$.

There is also another technique to find out UMVUE for $\psi(\theta)$ using Lehman-Scheffe and Rao-Blackwell theorem. (a) First find out any unbiased estimator $H(\boldsymbol{X})$ of $\psi(\theta)$, (b) identify sufficient statistics for θ, (c) Compute $E[H(\boldsymbol{X})|T] = \phi(T)$. By Rao Blackwell theorem $\phi(T)$ is an unbiased estimator of $\psi(\theta)$ and a function of the complete sufficient statistics T. Therefore $\phi(T)$ is UMVUE for $\psi(\theta)$. Let us see an example.

Example: $X_1, ..., X_n \sim Pois(\lambda)$. What is the UMVUE of $P(X = 0) = e^{-\lambda}$?

Clearly $E[I(X_1 = 0)] = P(X_1 = 0) = e^{-\lambda}$. We already know $T = \sum X_i \sim Pois(n\lambda)$ is sufficient for λ. Now

$$E[I(X_1 = 0)|\sum_{i=1}^n X_i = t] = P(X_1 = 0|\sum_{i=1}^n X_i = t) = \frac{P(X_1 = 0, \sum_{i=2}^n X_i = t)}{P(\sum_{i=1}^n X_i = t)} = \frac{\frac{e^{-n\lambda}[(n-1)\lambda]^t}{t!}}{\frac{e^{-n\lambda}[n\lambda]^t}{t!}} = \left(1 - \frac{1}{n}\right)^t.$$

$\left(1 - \frac{1}{n}\right)^{\sum_{i=1}^n X_i}$ is the UMVUE for $e^{-\lambda}$.

Now we are going to see another result that gives us lower bound on the variance of any unbiased estimator. The theorem is popularly known as the **Cramer-Rao Inequality**. But before that, let us discuss a few concepts which are necessary.

Let $\lambda(x) = \log f(x|\theta)$. We call $u_\theta(x) = \frac{\partial \log(f_\theta(x))}{\partial\theta}$ = **score function**. Note that $E_\theta(u_\theta(x)) = 0$. This can be seen using the fact that

$$0 = \frac{\delta}{\delta\theta}\int f_\theta(x)dx = \int u_\theta(x)dx = E_\theta(u_\theta(X)) = 0.$$

We define, Fisher information as $I(\theta) = E[u_\theta(X)^2] = Var(u_\theta(X))$. Taking another derivative w.r.t θ we obtain $E[u_\theta(X)^2] = -E[u'_\theta(X)]$. This is true for scalar θ as

$$\begin{aligned}0 &= \frac{d}{d\theta}\int u_\theta f_\theta(x)dx = \int u'_\theta(x)f_\theta(x)dx + \int u_\theta(x)\frac{d}{d\theta}f_\theta(x)dx \\ &= E_\theta(u'_\theta(X)) + E_\theta(u_\theta(X)^2).\end{aligned}$$

Information for location family: If $X \sim f(x-\theta)$, $f(x) > 0$ for all x, then $I(\theta) = \int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)}dx$.

proof: Note that $u_\theta(x) = \frac{\delta}{\delta\theta}\log(f(x-\theta)) = -f'(x-\theta)$. Thus $I(\theta) = \int_{-\infty}^{\infty} u_\theta(x)^2 f(x-\theta)dx = \int_{-\infty}^{\infty} \frac{[f'(x-\theta)]^2}{f(x-\theta)}dx = \int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)}dx$.

Remark: When $X \sim \frac{1}{b}f\left(\frac{x-\theta}{b}\right)$, b known, $I(\theta) = \frac{1}{b^2}\int_{-\infty}^{\infty} \frac{[f'(x)]^2}{f(x)}dx$. The proof is done in a similar way.

Information for scale family: If $X \sim \frac{1}{\theta}f(x/\theta)$, then $I(\theta) = \frac{1}{\theta^2}\int \left[\frac{yf'(y)}{f(y)} + 1\right]^2 f(y)dy$.

proof: $u_\theta(X) = \frac{-1/\theta^2 f(x/\theta) - x/\theta^3 f'(x/\theta)}{\frac{1}{\theta}f(x/\theta)}$.

$$I(\theta) = \int_{-\infty}^{\infty} u_\theta(x)^2 \frac{1}{\theta} f(x/\theta)dx.$$

Let $y = x/\theta \Rightarrow dx = \theta dy$. Then

$$I(\theta) = \int_{-\infty}^{\infty} \frac{[-1/\theta^2 f(y) - y/\theta^2 f'(y)]^2}{f(y)^2} f(y)dy = \frac{1}{\theta^2} \int_{-\infty}^{\infty} \left[1 + \frac{yf'(y)}{f(y)}\right]^2 f(y)dy.$$

Information Inequality: Suppose $X \sim f_\theta(x)$ and $I(\theta) > 0$. Let $\delta(X)$ be any function of X with $E_\theta(\delta(X)^2) < \infty$, for which the derivative w.r.t θ of $E_\theta(\delta(X))$ exists and can be differentiated under the integral sign i.e. $\frac{d}{d\theta} E_\theta(\delta(X)) = \int \delta(x) \frac{d}{d\theta} f_\theta(x)dx = \int \delta(x)u_\theta(x)f_\theta(x)dx$. Then

$$var_\theta(\delta(X)) \geq \frac{\left[\frac{d}{d\theta} E_\theta(\delta(X))\right]^2}{I(\theta)}.$$

Proof: $cov_\theta(\delta(X), u_\theta(X))^2 \leq Var_\theta(u_\theta(X))Var_\theta(\delta(X))$, by Cauchy-Schwartz inequality. Now $cov_\theta(\delta(X), u_\theta(X)) = \int \delta(x)u_\theta(x)dx = \int \delta(x)u_\theta(x)f_\theta(x)dx = \frac{d}{d\theta} E_\theta(\delta(X))$. Also $Var_\theta(\delta(X)) \geq \frac{\left[\frac{d}{d\theta} E_\theta(\delta(X))\right]^2}{I(\theta)}$.

Suppose a random sample $X_1, ..., X_n \overset{iid}{\sim} f_\theta(x)$. The score function for a random sample is given by $u_\theta(\boldsymbol{X}) = \frac{d}{d\theta} \log[\prod_{i=1}^n f_\theta(X_i)] = \sum_{i=1}^n u_\theta(X_i)$. Also Fisher information contained in $X_1, ...X_n$, denoted by $I_n(\theta)$ is given by $I_n(\theta) = Var[u_\theta(\boldsymbol{X})] = Var[\sum_{i=1}^n u_\theta(X_i)] = nI(\theta)$.

Cramer-Rao Inequality: Let $X_1, ..., X_n$ be iid from a distribution with pdf or pmf $f(x|\theta)$. Let $T(\boldsymbol{X})$ be any unbiased estimator of s.t. $E[T(\boldsymbol{X})] = m(\theta)$. Assume that all the regularity conditions hold then, $Var(T(\boldsymbol{X})) \geq \frac{[m'(\theta)]^2}{nI(\theta)}$. When equality holds, $T(\boldsymbol{X})$ must be of the form $T(\boldsymbol{X}) = \frac{m'(\theta)}{nI(\theta)} \sum_{i=1}^n u_\theta(X_i) + m(\theta)$.

proof: Use Cauchy-Schwartz inequality on to obtain $Cov(T(\boldsymbol{X}), u_\theta(\boldsymbol{X}))^2 \leq Var[T(\boldsymbol{X})]Var[u_\theta(\boldsymbol{X})]$. Thus $Var[T(\boldsymbol{X})] \geq \frac{m'(\theta)}{nI(\theta)}$ with equality holding if and only if $T(\boldsymbol{X}) = a(\theta)\sum_{i=1}^n u_\theta(X_i) + b(\theta)$. Now $E(T(\boldsymbol{X})) = m(\theta)$ implies $b(\theta) = m(\theta)$. Also $Cov(T(\boldsymbol{X}), \sum_{i=1}^n u_\theta(X_i)) = m'(\theta)$ implies

$a(\theta) = \frac{m'(\theta)}{nI(\theta)}$.

Remark: It is very important that the regularity conditions hold. To show this use $U(0,\theta)$ case and show that the lower bound is not satisfied. Let $X_1, ..., X_n \sim U(0,\theta)$. Then $\frac{d}{d\theta}\log(f_\theta(x)) = -1/\theta$, $I(\theta) = 1/\theta^2$. So, the Cramer-Rao lower bound for the variance of any unbiased estimator of θ is θ^2/n. Note that $T(\boldsymbol{X}) = X_{(n)}$ has expectation $E[X_{(n)}) = \int_0^\theta \frac{ny^n}{\theta^n} = \frac{n}{n+1}\theta$. Thus $\frac{(n+1)}{n}X_{(n)}$ is an unbiased estimator of θ. Now $Var(\frac{(n+1)}{n}X_{(n)}) = \frac{(n+1)^2}{n^2}[\frac{n}{n+2}\theta^2 - (\frac{n}{n+1}\theta)^2] = \frac{\theta^2}{n(n+2)}$ which is lower than the Cramer-Rao inequality.

example: $X_1, ..., X_n \sim Pois(\lambda)$. $\log(f(x|\lambda)) = x\log(\lambda) - \lambda - \log(x!)$, $u_\lambda(x) = \frac{x}{\lambda} - 1$, $E[u_\lambda(X)^2] = \frac{1}{\lambda}$. Let $m(\lambda) = \lambda$. Let us see $T(\boldsymbol{X}) = \frac{\lambda}{n}\sum_{i=1}^n \left(\frac{X_i - \lambda}{\lambda}\right) + \lambda = \bar{X}$.

Bottomline is check this quantity and see if it is free of parameters. Then it has to be UMVUE. Otherwise find out in some other way as discussed before.

Multi-parameter case: When $X \sim f_{\boldsymbol{\theta}}(x)$ where $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ we define a score vector instead of a scalar score. The score vector is defined as $u_{\boldsymbol{\theta}}(x) = (\frac{\delta}{\delta\theta_1}f_{\boldsymbol{\theta}}(x), ..., \frac{\delta}{\delta\theta_k}f_{\boldsymbol{\theta}}(x))$. and the Fisher information matrix is given by $I(\boldsymbol{\theta}) = ((I_{ij}(\boldsymbol{\theta})))_{i,j=1}^k$, where $I_{ij}(\boldsymbol{\theta}) = E[\frac{\delta}{\delta\theta_i}\log f_{\boldsymbol{\theta}}(x)\frac{\delta}{\delta\theta_j}\log f_{\boldsymbol{\theta}}(x)]$.

Information matrix for the location-scale family: Let $X \sim \frac{1}{\theta_2}f(\frac{x-\theta_1}{\theta_2})$. It follows from the previous result that $I_{11}(\boldsymbol{\theta}) = \frac{1}{\theta_2^2}\int_{-\infty}^{\infty}\frac{[f'(x)]^2}{f(x)}dx$, $I_{22}(\boldsymbol{\theta}) = \frac{1}{\theta_2^2}\int\left[\frac{yf'(y)}{f(y)} + 1\right]^2 f(y)dy$. Using similar trick we can show that $I_{12}(\boldsymbol{\theta}) = \frac{1}{\theta_2^2}\int y\frac{[f'(y)]^2}{f(y)}dy$.

Example: $N(\mu, \sigma^2)$, $Gamma(\alpha, \beta)$.

Multi-parameter Information Inequality: Suppose that $I(\boldsymbol{\theta})$ is positive definite and $\alpha_i = \frac{\delta}{\delta\theta_i}E_{\boldsymbol{\theta}}(\delta(\boldsymbol{X}))$ exists and differentiation w.r.t θ_i can be done under integration w.r.t. x. Then $Var_\theta(\delta(\boldsymbol{X})) \geq \boldsymbol{\alpha}'I^{-1}(\boldsymbol{\theta})\boldsymbol{\alpha}$, where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_k)$.

2 Method for finding estimators

There a number of ways to estimate an unknown parameter or parameters. We will mainly discuss the following methods.

(i) Method of moments