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: $E[X^k] = \int x^k f(x|\theta)dx$, $Var(X) = E[X^2] - E[X]^2$, $Cov(X_i, X_j) = E[X_i X_j] - E[X_i]E[X_j]$. For any random sample $E[X_i X_j] = \int\int x_i x_j f(x_i, x_j|\theta) dx_i dx_j = \int x_i f(x_i|\theta) \left(\int x_j f(x_j|\theta) x_j\right) dx_i = E[X_i]E[X_j]$. 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 $X_1, X_2, X_3 \overset{iid}{\sim} \textit{Bernoulli}(p)$. 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 $\sum_{i=1}^{3} x_i = t$,

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

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	$(1-p)^2 p$
010	$(1-p)p(1-p) = (1-p)^2 p$
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. $X_1, X_2, X_3 \sim \mathit{Bernoulli}(p)$. 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 $f_\theta(\boldsymbol{x}) = k f_\theta(\boldsymbol{y})$ 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

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

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

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

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

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

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

(ii) Method of maximum likelihood

(iii) Bayes and minimax estimators.

2.1 Method of Moments

Sometimes we don't know how to create estimators of a parameter and we need to depend on intuition. For example, estimating a parameter with its sample analogue can give good estimates. For example sample mean is a good estimator for the population mean. In general we create the method of moments estimator as follows.

$$m_j = \frac{1}{n}\sum_{i=1}^{n} X_i^j$$

Let $\mu_j' = E[X^j]$. Generally μ_j''s are functions of the unknown parameters $\theta_1, ..., \theta_k$. Therefore by solving k equations

$$m_j = \mu_j'(\theta_1, ..., \theta_k), j = 1, ..., k,$$

we have some estimates of $\theta_1, ..., \theta_k$. Using our previous techniques it might be shown that they are UMVUE sometimes. It might not be the case, but when you have nothing to start with, MOM gives you a fairly good starting point. Simplest example are binomial and normal where we have seen $\bar{X}$ is UMVUE for parameters.

Example: $X_1, ..., X_n \sim N(\mu, \sigma^2)$. MOM does the following

$$\bar{X} = \mu, \; \frac{1}{n}\sum_{i=1}^{n} X_i^2 = \mu^2 + \sigma^2 \Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

We already know they are UMVUE for μ and $\frac{(n-1)}{n}\sigma^2$. So here MOM estimator looks good. However, consider the situation when $X_1, ..., X_n \sim DE(\mu, \sigma^2)$. Even for this case MOM estimator remains the same, though it is not all the UMVUE. This is a disadvantage that the MOM estimator doesn't take care of the difference in distributions. Another important disadvantage is you are not using distribution of the random sample, just using some population moments. Sometimes this might produce estimators outside the range of the parameters.

Example: $X_1, ..., X_n \sim Bin(k, p)$. Then MOM estimators are obtained by

$$\bar{X} = kp, \frac{1}{n}\sum_{i=1}^{n} X_i^2 = kp(1-p) + k^2p^2 \Rightarrow k = \frac{\bar{X}}{\bar{X} - (1/n)\sum_{i=1}^{n}(X_i - \bar{X})^2}, p = \frac{\bar{X}}{k}.$$

Although we know estimate of k has to be an integer, there is no way to impose it in the method of moments.

Example: Another problem arises when one needs to calculate MOM estimator for a curved exponential family $N(\theta, \theta^2)$. MOM estimator might not exist in this case as there might not be any solution to the two equations $\bar{X} = \theta$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^2 = \theta^2 + \theta$.

2.2 Maximum Likelihood Estimators

Maximum likelihood estimator or MLE is the most popular technique for deriving estimators. It takes into account the actual distribution of the data and estimates parameters so as to maximize likelihood. Let $f_{\theta_1,...,\theta_k}(x)$ be the pdf of X. Then the likelihood $\boldsymbol{X}$ is $L(\theta_1, ..., \theta_k) = \prod_{i=1}^{n} f_{\theta_1,...,\theta_k}(X_i)$. MLE is the value of $\hat{\boldsymbol{\theta}} = (\theta_1, ..., \theta_l)$, denoted by $(\hat{\theta}_1, ..., \hat{\theta}_k)$, that maximizes the Likelihood $L(\theta_1, ..., \theta_k)$. MLE is found by solving $\frac{\partial L(\boldsymbol{\theta})}{\partial \theta_i} = 0$, $i = 1, ..., k$. Often we solve MLE by solving the score equation $\sum_{i=1}^{n} u_\theta(X_i) = 0$. Let $\hat{\theta}_1, ..., \hat{\theta}_k$ be the solutions of these equations. Then MLE should satisfy ${\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}}^2 |\boldsymbol{\theta} = \hat{\boldsymbol{\theta}} < 0$.

However, when MLE is in the boundary of the parameter space you have to be more careful.

Example: $X_1, ..., X_n \sim N(\theta, 1)$, $\theta > 0$. If $\bar{X} > 0$, then it is the MLE. Now when $\bar{X} < 0$, then

$$\exp(-\sum_{i=1}^{n}(X_i-\theta)^2/2) = \exp(-\sum_{i=1}^{n}(X_i-\bar{X})^2/2)\exp(-n(\bar{X}-\theta)^2/2).$$

The above expression is monotonically decreasing for $\theta > 0$, as $\bar{X} < 0$. Therefore MLE of $\theta = 0$. Besides this, MLE also satisfies another important property that makes it easy to interpret in many practical contexts.

Result: MLE, if exists, is always a function of a sufficient statistic. It is very clear from the factorization theorem $f_\theta(\boldsymbol{X}) = g(T(\boldsymbol{X}),\theta)h(\boldsymbol{X})$. However MLE need to be a function of the minimal sufficient statistic. Here is an example $X_1, ..., X_n \sim U(\theta - \frac{1}{,}\theta + 1)$. Now the likelihood of θ is

$$L(\theta) = \frac{1}{2^n}I[X_{(1)} > \theta - 1, X_{(n)} < \theta + 1] = \frac{1}{2^n}I[X_{(n)} - 1 < \theta < X_{(1)} + 1].$$

Here minimal sufficient statistics is $(X_{(1)}, X_{(n)})$. But MLE can be any value of θ between $X_{(n)} - 1$ and $X_{(1)} + 1$. So Let $\hat{\theta} = \frac{|\bar{X}|}{|\bar{X}|+1}(X_{(n)} - 1) + \frac{1}{|\bar{X}|+1}(X_{(1)} + 1)$ is an MLE which is not a function of minimal sufficient statistic.

Invariance Property of MLE: If $\hat{\theta}$ is the MLE of θ, then for any function $\tau(\theta)$, the MLE of $\tau(\theta)$ is $\tau(\hat{\theta})$.

There is another interesting result stating that MLE is "asymptotically the most efficient estimator". What do I mean by that? It means the following result.

Asymptotic Distribution of MLE: In smooth regular problems MLE $\hat{\theta}$ converges in probability to θ and $\sqrt{n}(\hat{\theta} - \theta)$ converges in distribution to $N(0, \frac{1}{I(\theta)})$ when $X_1, ..., X_n$ is a random sample from $f_\theta(x)$. Furthermore, by delta theorem, if $g'(\theta)$ exists and nonzero then $\sqrt{n}(g(\hat{\theta}) - g(\theta))$ converges in distribution to $N(0, \frac{g'(\theta)}{I(\theta)})$.

Remember $\frac{g'(\theta)}{I(\theta)}$ is the Cramer-Rao lower bound on the variance of any unbiased estimator of $g(\theta)$. It seems that asymptotically variance of $g(\hat{\theta})$ is achieving that variance and asymptotically it is unbiased too. Therefore, asymptotically MLE or any function of MLE is the "most efficient" estimator for its asymptotic expected value. "Asymptotic relative efficiency" of any other estimator is ratio of its asymptotic variance to the asymptotic variance of MLE.

Asymptotic Normality does not hold in non-regular problems: I will show an example of non-regularity.

example 1: $X_1, ...X_n \overset{iid}{\sim} U(0, \theta)$, $\theta > 0$. Here MLE of θ is $X_{(n)}$. In your assignment problem you have seen that $n(X_{(n)} - \theta)$ converges in distribution to $\exp(\theta)$.

2.3 Bayes and Minimax Estimator

Bayes and Minimax estimator is related to the Bayesian approach. You will learn them in detail in AMS 206. But, I am gonna give you a small introduction. You will learn that classical approaches face many different problems that Bayesians overcome. But, I will skip all those details. They are not the central focus of this course. Instead, I will show you how to compute Bayesian estimates. First recall what a frequentist is doing.

$$\text{Receive data } X_1, ..., X_n \rightarrow \text{Finds } \hat{\theta} \text{ as a fn. of } X_1, ..., X_n.$$

In Bayesian we do not see the parameter θ as fixed and unknown. Rather we think of it as a random variable with unknown distribution and our aim is to find out that unknown distribution. The setting is the following.

- $X_1, ..., X_n \sim f(x|\theta)$. Mainly thinking about the iid case, otherwise $(X_1, ..., X_n) \sim f(\boldsymbol{x}|\theta)$.
- Assume a prior distribution $\pi(\theta)$ for θ. Prior distribution can be anything. It depends on your subjective choice.
- Posterior distribution of $\theta | X_1, ..., X_n$ is given by $\pi(\theta | X_1, ..., X_n) = \frac{f(X_1,...,X_n|\theta)\pi(\theta)}{\int f(X_1,...,X_n|\theta)\pi(\theta)d\theta}$. When $X_1, ..., X_n$ is iid we have $f(X_1, ..., X_n|\theta) = \prod_{i=1}^n f(X_i|\theta)$.

- Bayes estimator is the posterior mean of θ, i.e. $E[\theta|X_1,...,X_n] = \int \theta\pi(\theta|X_1,...,X_n)d\theta$.

example: $X_1,...,X_n \sim Ber(p)$, prior distribution $p \sim Beta(\alpha,\beta)$. Then

$$\pi(p|X_1,...,X_n) \propto p^{\sum_{i=1}^n X_i}(i-p)^{n-\sum_{i=1}^n X_i}p^{\alpha-1}(1-p)^{\beta-1} = p^{\sum_{i=1}^n X_i+\alpha-1}(i-p)^{n-\sum_{i=1}^n X_i+\beta-1}.$$

Hence $p|X_1,...,X_n \sim Beta(\sum_{i=1}^n X_i+\alpha, n-\sum_{i=1}^n X_i+\beta)$.

example: $X_1,...,X_n \sim Poi(\lambda)$, prior distribution $\lambda \sim Gamma(\alpha,\beta)$. Then

$$\pi(p|X_1,...,X_n) \propto \lambda^{\sum_{i=1}^n X_i}e^{-n\lambda}\lambda^{\alpha-1}e^{-\lambda\beta} = \lambda^{\sum_{i=1}^n X_i+\alpha-1}e^{-\lambda(n+\beta)}.$$

Hence $\lambda|X_1,...,X_n \sim Gamma(\sum_{i=1}^n X_i+\alpha, n+\beta)$.

Have you noticed one thing? Prior and Posterior belong to the same class of distribution. This was deliberate. We wanted to chose the prior distribution so that calculation of the posterior is easier and this is one way to do it. Choosing prior in a way depending on the pdf of X so that prior and posterior distributions are the same. We call such prior distribution as the conjugate family. A formal definition is here

Definition (Conjugate family): Let $\mathcal{F}$ denote the class of pdfs or pmfs $f(x|\theta)$. A class Π of prior distributions is a conjugate family for $\mathcal{F}$ if the posterior distribution is in the class Π for all $f \in \mathcal{F}$, all priors in Π and all $x \in \mathcal{X}$.

In a wide range of applications it is simply not possible to find a conjugate family. One needs to employ some approximation algorithm to estimate posterior distributions in such cases. MCMC algorithm is the most popular of such. You will learn about them in AMS 206. Once you approximate the posterior distribution, you also have mean of that approximated posterior.

This is the basics of Bayesian statistics. Now we will learn how to create "good" estimators of θ using the "risk function" notion in Bayesian statistics. Let $\delta(\boldsymbol{X})$ be an estimator of θ. The loss in estimating θ by $\delta(\boldsymbol{X})$ is represented by a function given by a function $L(\theta,\delta)$. This function $L(\cdot,\cdot)$ is known as the loss function. $E_{\boldsymbol{X}|\theta}[L(\theta,\delta(\boldsymbol{X})] = R(\theta,\delta)$ is known as the risk function of δ. Recall that if $L(\theta, \delta(\boldsymbol{X})) = (\theta - \delta(\boldsymbol{X}))^2$, then $R(\theta, \delta) =$ MSE of $\delta(\boldsymbol{X})$. Given two estimators δ_1, δ_2, we say δ_1 is better than δ_2 if $R(\theta, \delta_1) \le R(\theta, \delta_2)$ for all θ, and the inequality is strict at least for one θ. When risk of two estimators intersect with each other we do not know which one to choose. At that time, we have to propose a summary measure from this entire risk function to choose one of them. Different summary measures give rise to different estimators. We will discuss two of them as following.

1. Minimize average risk $= E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta(\boldsymbol{X}))]$ which gives rise to the Bayes estimator.

2. Minimize supremum risk $= \sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta(\boldsymbol{X}))]$ which gives rise to the Minimax estimator.

2.3.1 Bayes Estimator

Note that $E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta(\boldsymbol{X}))] = E_{\boldsymbol{X}} E_{\theta|\boldsymbol{X}}[L(\theta, \delta(\boldsymbol{X}))]$. So, if we can find δ that minimizes $E_{\theta|\boldsymbol{X}}[L(\theta, \delta(\boldsymbol{X}))]$ for all $\boldsymbol{X}$, that is the solution of this problem. There are a few loss functions for which the solution is easy to find.

Squared Error Loss: Let $L(\theta, \delta(\boldsymbol{X})) = (\theta - \delta(\boldsymbol{X}))^2$. Then $E_{\theta|\boldsymbol{X}}[(\theta - \delta(\boldsymbol{X}))^2]$ is maximized by $E[\theta|\boldsymbol{X}]$, which is the posterior mean of θ, it is easy to see as $E_{\theta|\boldsymbol{X}}[(\theta - \delta(\boldsymbol{X}))^2] = E_{\theta|\boldsymbol{X}}[(\theta - E[\theta|\boldsymbol{X}])^2] + E_{\theta|\boldsymbol{X}}[(E[\theta|\boldsymbol{X}] - \delta(\boldsymbol{X}))^2]$. Sometimes for complicated loss function, people do it in the following way

$$\frac{d}{d\delta} E_{\theta|\boldsymbol{X}}[L(\theta, \delta(\boldsymbol{X}))] = 0,$$

and find δ. For example, in squared error loss $\frac{d}{d\delta} E_{\theta|\boldsymbol{X}}[(\theta - \delta)^2] = 0$ implies $E_{\theta|\boldsymbol{X}}[\theta] = \delta(\boldsymbol{X})$.

Example: In the previous examples, we have seen $p|X_1, ..., X_n \sim Beta(\sum_{i=1}^n X_i + \alpha, n - \sum_{i=1}^n X_i + \beta)$. Thus $E[p|\boldsymbol{X}] = \frac{\sum_{i=1}^n X_i + \alpha}{n+\alpha+\beta} = \bar{X}\frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{n+\alpha+\beta}$. This is clearly a biased estimator of θ.

Weighted Squared Error Loss: Let $L(\theta, \delta(\boldsymbol{X})) = w(\theta)(\theta - \delta(\boldsymbol{X}))^2$, where $w(\theta) > 0$ be some function of θ. Here to find the Bayes estimator we solve $\frac{d}{d\theta} E_{\theta|\boldsymbol{X}}[w(\theta)(\theta - \delta)^2] = 0$ that implies $E_{\theta|\boldsymbol{X}}[w(\theta)\theta] = \delta(\boldsymbol{X})E_{\theta|\boldsymbol{X}}[w(\theta)]$. Hence $\delta(\boldsymbol{X}) = \frac{E[w(\theta)\theta|\boldsymbol{X}]}{E[w(\theta)|\boldsymbol{X}]}$.

Note that the Bayes estimator is an "optimal" estimator in some sense. Can it be unbiased? The answer for the squared error loss is given as following.

Theorem 2.1 *No unbiased estimator $\delta(\boldsymbol{X})$ of θ can be a Bayes estimator unless $E_\theta E_{\boldsymbol{X}|\theta}[(\theta - \delta(\boldsymbol{X}))^2] = 0$.*

Proof Let $\delta(\boldsymbol{X})$ be an unbiased estimator which is also a Bayes estimator, so $E[\delta(\boldsymbol{X})|\theta] = \theta$ for all θ and $E[\theta|\boldsymbol{X}] = \delta(\boldsymbol{X})$. Note that

$$E_\theta E_{\boldsymbol{X}|\theta}[\delta(\boldsymbol{X})\theta] = E_\theta[\theta E_{\boldsymbol{X}|\theta}[\delta(\boldsymbol{X})]] = E_\theta[\theta^2] \tag{3}$$

$$E_{\boldsymbol{X}} E_{\theta|\boldsymbol{X}}[\delta(\boldsymbol{X})\theta] = E_{\boldsymbol{X}}[\delta(\boldsymbol{X})E[\theta|\boldsymbol{X}]] = E_{\boldsymbol{X}}[\delta(\boldsymbol{X})^2]. \tag{4}$$

Thus $E_\theta E_{\boldsymbol{X}|\theta}[\delta(\boldsymbol{X})^2] = E_\theta E_{\boldsymbol{X}|\theta}[\delta(\boldsymbol{X})\theta] = E_\theta E_{\boldsymbol{X}|\theta}[\theta^2]$. Hence

$$E_{\theta,\delta(\boldsymbol{X})}[(\theta - \delta(\boldsymbol{X}))^2] = E_{\theta,\delta(\boldsymbol{X})}[\theta^2 - 2\delta(\boldsymbol{X})\theta + \delta(\boldsymbol{X})^2] = 0.$$

This is an easy way to check if some estimator is a Bayes estimator for a parameter. For example, when $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ^2 known, $\bar{X}$ is the UMVUE, though it is not a Bayes estimator as $E_\mu E_{\boldsymbol{X}|\mu}[(\bar{X} - \mu)^2] = E_\mu[\sigma^2/n] = \sigma^2/n \neq 0$.

Remark: Also recall that the Bayes estimator is not unbiased (show that). In fact it is, in many cases, a convex combination of the prior mean and data mean.

2.3.2 Minimax Estimator

Minimax estimators minimize $\sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta(\boldsymbol{X}))]$. Thus minimax estimator is going to protect us in the worst case scenario. Identifying a minimax estimator is a hard task. However, sometimes we can rely on our intuitions to find a minimax estimator. The idea is that some parameter values are responsible for higher risk than others. If the prior distribution of θ gives high prior probability to those values, then maybe a Bayes estimator will be a minimax estimator. Let us formalize this intuition.

Definition (Least Favorable Distribution): A prior distribution is $\pi(\theta)$ on θ is known to be a least favorable prior if $E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi(\boldsymbol{X}))] \geq E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_{\pi'}(\boldsymbol{X}))]$ for all prior distribution π' on θ. Here δ_π and $\delta_{\pi'}$ are Bayes estimators w.r.t priors π and π' respectively. So, π is such a prior distribution that is just increasing the risk for the Bayes estimator.

Result: If $\pi(\theta)$ be a prior distribution for which $\int E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)]d\theta = \sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)]$, where δ_π is the Bayes estimator, then

(a) δ_π is minimax.

(b) π is least favorable.

Proof (a) For any estimator $\delta(\boldsymbol{X})$,

$$\sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta)] \geq E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta)] \geq E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)] = \sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)], \ \forall \, \theta.$$

Hence δ_π is minimax.

(b) Note that

$$\begin{aligned} E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)] &= \sup_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)] \geq \int E_{\boldsymbol{X}|\theta}[L(\theta, \delta_\pi)]\pi'(\theta)d\theta \geq E_{\boldsymbol{X}|\theta}[L(\theta, \delta_{\pi'})]\pi'(\theta)d\theta \\ &= E_\theta E_{\boldsymbol{X}|\theta}[L(\theta, \delta_{\pi'})], \end{aligned}$$

for any other prior distribution π'. Therefore, $]pi$ is least favorable.

This result gives us a way to find minimax estimator in some cases. This is the algorithm.

step 1: First find out Bayes estimator and calculate its risk function. For squared error loss that simply boils down to calculating its MSE.

Step 2: Find the prior distribution that will make this risk constant (free of the model parameter θ).

Step 3: Evaluate Bayes estimator at these prior parameters. This Bayes estimator will be minimax.

Example: $X_1, ..., X_n \sim Ber(p)$. Under squared error loss find the minimax estimator. With $p \sim Beta(\alpha, \beta)$, the Bayes estimator is given by $\delta_\pi(\boldsymbol{X}) = \frac{\sum_{i=1}^n X_i + \alpha}{\alpha+\beta+n}$. With a bit of calculation one can show that

$$E_{\boldsymbol{X}|p}((\delta(\boldsymbol{X}) - p)^2) = \frac{1}{(\alpha+\beta+n)^2}[\alpha^2 + \{n - 2\alpha(\alpha+\beta)\}p + \{(\alpha+\beta)^2 - n\}p^2].$$

This is constant as a function of p if $2\alpha(\alpha+\beta) = n$ and $\alpha+\beta = \sqrt{n}$. Solving these equations, we obtain $\alpha = \beta =$
$sqrtn/2$. Hence the minimax estimator is $\frac{\sum_{i=1}^n X_i + \frac{\sqrt{n}}{2}}{n+\sqrt{n}}$.

3 Testing of hypothesis

Statistical hypothesis testing is all about

- Beginning with a tentative idea about the unknown parameter.
- Want to test the validity of this tentative idea based on sample information. Existing tentative idea, status quo: H_0 (null hypothesis), new idea: H_1 (alternative hypothesis).
- We begin by assuming that the null hypothesis is true. Only when there is an overwhelming evidence contradicting null do we reject it in favor of alternative.

	H_0 is true	H_0 is false
Do not reject H_0	Correct	Type 2 error
Reject H_0	Type 1 error	Correct

Type 1 error $= P(reject\ H_0|H_0\ is\ true)$, Type 2 error $= P(do\ not\ reject\ H_0|H_0\ is\ false)$. Type 1 error is also known as the level of the test. While power of the test is defined by power$= 1-$ Type 2 error $= P(reject\ H_0|H_0\ is\ false)$.

Generality: Ideally we would like to minimize both type 1 and type 2 error. But it turns out that it is not possible to simultaneously minimize both of them. So, we fix level at a pre-specified value and find a test that maximizes power.

Parametric tests: $X_1, ..., X_n \sim f(x|\theta)$, we test $H_0 : \theta \in \mathbf{\Theta}_0$ vs. $H_1 : \theta \in \mathbf{\Theta}_1$. If $\mathbf{\Theta}_0$ is singleton we will call it a simple null hypothesis, o.w. we will call it a composite null hypothesis.

Let $\mathcal{R} = \{\boldsymbol{x} \in \mathcal{X} | \text{The null hypothesis is rejected for } \boldsymbol{x}\}$ be the *rejection region* or *critical region.*

$\phi(\boldsymbol{x})$ = prob. of rejecting H_0 when $\boldsymbol{x}$ is observed. The power function of a test is given by $\beta(\theta) = \int \phi(\boldsymbol{x}) f(\boldsymbol{x}|\theta)$, clearly for any level α test $\beta(\theta) \leq \alpha$, $\forall \theta \in \mathbf{\Theta}_0$. Let us first motivate with an example.

Find the best level 1/8th test. If you take $\mathcal{R} = \{0\}$, then *level* $= 1/8$, *power* $= 1/2$. It is the best. It seems like you need to include those points in the rejection rejection for which f_1/f_0 is higher. It is the most powerful test based on sample size 1.

X	0	1	2	3
f_0	1/8	1/8	1/4	1/2
f_1	1/2	1/4	1/8	1/8

Neyman-Pearson Lemma: Consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, where the pdf or pmf corresponding to θ_i is $f(x|\theta_i), i = 0, 1$, using a test with rejection region $\mathcal{R}$ that satisfies

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } f(\boldsymbol{x}|\theta_1) > kf(\boldsymbol{x}|\theta_0) \\ 0 \text{ if } f(\boldsymbol{x}|\theta_1) < kf(\boldsymbol{x}|\theta_0) \end{cases}$$

for some $k \geq 0$, and $\alpha = P_{\theta_0}(\boldsymbol{X} \in \mathcal{R})$. Then

(a) Any test that satisfies the above is the most powerful level α test.

(b) If there exists a test satisfying the above, then every MP level α test is a size α test and every MP level α test satisfies the above except perhaps on a set A satisfying $P_{\theta_0}(\boldsymbol{x} \in A) = P_{\theta_1}(\boldsymbol{X} \in A) = 0$.

Let $\phi'(\boldsymbol{x})$ be the test function of any level α test. Consider the function $[\phi(\boldsymbol{x})-\phi'(\boldsymbol{x})][f(\boldsymbol{x}|\theta_1)-kf(\boldsymbol{x}|\theta_0)] \geq 0$.

$$0 \leq \int [\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})][f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)] = \beta(\theta_1) - \beta'(\theta_1) - k(\beta(\theta_0) - \beta'(\theta_0))$$

$\beta(\theta_0) \geq \beta'(\theta_0)$, thus $\beta(\theta_1) > \beta'(\theta_1)$. To prove statement (b), note that if $\phi'(\boldsymbol{x})$ is a MP level α test. Now from the previous equation, $\beta(\theta_1) = \beta'(\theta_1)$, thus $\beta'(\theta_0) \geq \beta(\theta_0) = \alpha$. It is level α means size α. Nonnegative integrand $[\phi(\boldsymbol{x}) - \phi'(\boldsymbol{x})][f(\boldsymbol{x}|\theta_1) - kf(\boldsymbol{x}|\theta_0)]$ has to be zero except for a set satisfying...

Remember Factorization theorem, which states $f(\boldsymbol{x}|\theta) = g(T,\theta)h(\boldsymbol{x})$, where T is the sufficient statistic. Using this result we find the Neyman-Pearson most powerful test of level α as

$$\phi(t) = \begin{cases} 1 \text{ if } g(t,\theta_1) > kg(t,\theta_0) \\ 0 \text{ if } g(t,\theta_1) < kg(t,\theta_0) \end{cases}$$

for some $k \geq 0$, where $\alpha = P_{\theta_0}(\phi(T) = 1)$.

example: Let $X_1, X_2 \sim Ber(\theta)$. Want to test $H_0 : \theta = \frac{1}{2}$ vs. $H_1 : \theta = \frac{3}{4}$. We know the sufficient statistics here is $T = \sum_{i=1}^{2} X_i \sim Binomial(2,\theta)$. The three likelihood ratios are given below

$$\frac{f(0|\theta = \frac{3}{4})}{f(0|\theta = \frac{3}{4})} = \frac{1}{4}, \frac{f(1|\theta = \frac{3}{4})}{f(1|\theta = \frac{3}{4})} = \frac{3}{4}, \frac{f(2|\theta = \frac{3}{4})}{f(2|\theta = \frac{3}{4})} = \frac{9}{4}.$$

If we choose $\frac{3}{4} < k < \frac{9}{4}$, Neyman-Pearson lemma gives us that the test is MP-level $\alpha = P(T = 2|\theta = \frac{1}{2}) = \frac{1}{4}$. If we choose $\frac{1}{4} < k < \frac{3}{4}$, we have the MP level $\alpha = P(T = 1,2|\theta = \frac{1}{2}) = \frac{3}{4}$ test. Choosing $k < \frac{1}{4}$ or $k > \frac{9}{4}$ yields MP level $\alpha = 1$ or $\alpha = 0$ respectively.

example: $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ^2 known. Find MP level α test for $H_0 : \mu = \mu_0$ vs $\mu = \mu_1$. We know $\bar{X} \sim N(\mu, \sigma^2/n)$, sample mean is sufficient for μ. Now $\frac{g(\bar{X},\mu_1)}{g(\bar{X},\mu_0)} > k$ implies $\bar{X} < \frac{(2\sigma^2\log(k))/n - \mu_0^2 + \mu_1^2}{2(\mu_1 - \mu_0)}$. Finding a level α test boils down to finding k. k is determined by the equation $P(\bar{X} < \frac{(2\sigma^2 \log(k))/n - \mu_0^2 + \mu_1^2}{2(\mu_1 - \mu_0)}) = \alpha$. Solve for k.

Neyman-Pearson lemma provides us the most powerful test of level α for testing a point null vs point alternative. Now we will see more general sets $\mathbf{\Theta}_0$ and $\mathbf{\Theta}_1$ and how to carry out testing there. For point null vs. point alternative, our goal was to find out MP level α test. Now we will try to find out test which is uniformly most powerful (UMP) for a composite hypothesis. Here is the definition.

Definition: Let $\mathcal{C}$ be a class of tests for testing $H_0 : \theta \in \mathbf{\Theta}_0$ vs. $H_1 : \theta \in \mathbf{\Theta}_0^c$. A test in class $\mathcal{C}$, with power function $\beta(\theta)$ is a uniformly most powerful (UMP) class $\mathcal{C}$ test if $\beta(\theta) \geq \beta_1(\theta)$ for every $\theta \in \mathbf{\Theta}_0^c$ and every $\beta_1(\theta)$ that is a power function of a test in class $\mathcal{C}$.

In our context we want UMP level α test for composite hypothesis. It is not always easy to derive such a test for general $\mathbf{\Theta}_0$. But for some specific $\mathbf{\Theta}_0$, we can construct such tests. But to do so, we need the family of distributions $\{f(\cdot|\theta) : \theta \in \mathbf{\Theta}\}$ to satisfy the following property.

Monotone Likelihood Ratio: The family $f(\cdot|\theta)$ is said to have monotone likelihood ratio in $T(\boldsymbol{x})$ if $\frac{f(\boldsymbol{x}|\theta_1)}{f(\boldsymbol{x}|\theta_0)}$ is an increasing function of $T(\boldsymbol{x})$ whenever $\theta_1 > \theta_0$.

Example: $X_1, ..., X_n \sim N(\theta, 1)$. Let $\theta_1 > \theta_0$. Then

$$\frac{f(\boldsymbol{x}|\theta_1)}{f(\boldsymbol{x}|\theta_0)} = \exp\left(\sum_{i=1}^{n} X_i(\theta_1 - \theta_0) - \frac{n}{2}(\theta_1^2 - \theta_0^2)\right),$$

showing that $N(\theta, 1)$ family has MLR in $\sum_{i=1}^n X_i$.

Suppose a parametric family $f(\cdot|\theta)$ has MLR in $T(\boldsymbol{x})$. Let $\theta_1 > \theta_0$, we already know that the MP level α test for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$ is

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } T(\boldsymbol{x}) > b \\ 0 \text{ if } T(\boldsymbol{x}) < b \end{cases}$$

s.t. $P_{\theta_0}(T(\boldsymbol{x}) > b) = \alpha$. This test does not depend on the value of θ_1. So, this is the uniformly most powerful test of level α for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$.

Remark: It can be shown that the power function $\beta(\theta) = E_\theta(\phi)$, is non-decreasing. To prove this, let $\theta_1 > \theta_2$. Clearly, the power function $\beta(\theta) = P_\theta(T > t)$. Note that

$$\frac{d}{dt}\left[P_{\theta_1}(T \le t) - P_{\theta_2}(T \le t)\right] = f_{\theta_1}(t) - f_{\theta_2}(t) = f_{\theta_2}(t)\left(\frac{f_{\theta_1}(t)}{f_{\theta_2}(t)} - 1\right).$$

R.H.S is increasing as a function of t means, it can only change sign from negative to positive. Therefore, any internal extremum is a minimum. Therefore, the function in the bracket in L.H.S is maximized at ∞ or $-\infty$. At both these points, value of the function is 0. Thus $P_{\theta_1}(T \le t) - P_{\theta_2}(T \le t) < 0$. Hence the power function $\beta(\theta)$ is nondecreasing.

This means that if $\theta \le \theta_0$, $\beta(\theta) \le \beta(\theta_0) \le \alpha$. Therefore, the above test is also a UMP level α test for testing $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > theta_0$.

Remark: One parameter exponential family has MLR in $\sum_{i=1}^n t(X_i)$. Therefore, there exists a UMP test of level α for $H_0 : \theta \le \theta_0$ vs. $H_1 : \theta > \theta_0$.

Refer to the earlier normal example.

Remark: It is easy to see that the UMP level α test for testing $H_0 : \theta \ge \theta_0$ vs. $H_1 : \theta < \theta_0$ is given by

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } T(\boldsymbol{x}) < b \\ 0 \text{ if } T(\boldsymbol{x}) > b \end{cases}$$

s.t. $P_{\theta_0}(T(\boldsymbol{x}) < b) = \alpha$.

proof: Proof is very similar to the previous one. Start with $H_0 : \theta = \theta_0$ vs. $H_1 : \theta = \theta_1$, $\theta_1 < \theta_0$. Since $f_\theta(\cdot)$ has MLR in $T(\boldsymbol{X})$, means $\frac{f_{\theta_0}(\cdot)}{f_{\theta_1}(\cdot)}$ is a nondecreasing function of $T(\boldsymbol{X})$. Hence $\frac{f_{\theta_1}(\cdot)}{f_{\theta_0}(\cdot)}$ is a nondecreasing function of $-T(\boldsymbol{X})$. Thus, the NP test can be written as

$$\phi(\boldsymbol{x}) = \begin{cases} 1 \text{ if } T(\boldsymbol{x}) < b \\ 0 \text{ if } T(\boldsymbol{x}) > b \end{cases}$$

s.t. $P_{\theta_0}(T(\boldsymbol{x}) < b) = \alpha$. Since this test function is free of θ_1, this is a UMP test for testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta < \theta_0$. By our previous proof the test function for this test is $\beta(\theta) = P_\theta(T(\boldsymbol{X}) < b)$ which is a nonincreasing function of θ. Hence $\beta(\theta_0) \geq \beta(\theta)$ for all $\theta \geq \theta_0$. Hence it is a UMP test for $H_0 : \theta \geq \theta_0$ vs. $H_1 : \theta < \theta_0$.

p-value: We have seen that in the NP lemma $k = k(\alpha)$ is determined by using the fact that probability of $f_{\theta_1}(\boldsymbol{X})/f_{\theta_0}(\boldsymbol{X}) > k$ is α. Note that, in many such cases the rejection region $\mathcal{R}_\alpha$ is nested in a way that $\mathcal{R}_\alpha \subset \mathcal{R}_{\alpha'}$ for $\alpha < \alpha'$. When this is the case, it is a good practice to see not only whether the hypothesis is accepted or rejected, but also determine the smallest level at which the hypothesis is rejected. It is known as the p-value. More formally, it is $p(\boldsymbol{X}) = \inf\{\alpha : \boldsymbol{X} \in \mathcal{R}_\alpha\}$.

Example: Consider testing $H_0 : \mu = 0$ vs. $H_1 : \mu > 0$, where $X_1, ..., X_n \overset{iid}{\sim} N(\mu, 1)$. Thus the rejection region for the UMP test is $\mathcal{R}_\alpha = \{\bar{X} : \bar{X} > \frac{1}{\sqrt{n}} z_{1-\alpha}\} = \{\bar{X} : 1 - \Phi(\sqrt{n}\bar{X}) < \alpha\}$. Thus the infimum over all α where the last inequality holds for a given $\bar{X}$ is $p(\boldsymbol{X}) = 1 - \Phi(\sqrt{n}\bar{X})$.

Remark: For the one parameter exponential family, there does not exist a UMP test of level α that tests $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$.

Proof in the general case requires quite a bit of notations. I will show it for normal distribution to develop the intuition. Suppose $X_1, ..., X_n \sim N(\theta, \sigma^2)$, σ^2 known. For any $\theta_1 < \theta_0$ the UMP-level α test for testing θ_0 vs. θ_1 is given by a test that rejects if $\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0$. It has the highest possible power for any $\theta_1 < \theta_0$. Any other level α test will have the same rejection region by Neyman-Pearson except for a set with probability measure zero. Now consider the test that rejects H_0 if $\bar{X} > \sigma z_\alpha/\sqrt{n} + \theta_0$. For any $\theta_2 > \theta_0$,

$$P_{\theta_2}(\bar{X} > \sigma z_\alpha/\sqrt{n} + \theta_0) = P_{\theta_2}(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} > z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}) > P(Z > z_\alpha) = P(Z < -z_\alpha)$$
$$> P_{\theta_2}(\frac{\bar{X} - \theta_2}{\sigma/\sqrt{n}} < -z_\alpha + \frac{\theta_0 - \theta_2}{\sigma/\sqrt{n}}) = P_{\theta_2}(\bar{X} < -\sigma z_\alpha/\sqrt{n} + \theta_0).$$

Therefore the latter test has more power than the former at θ_2. Hence the former test is not UMP.

Remark: However outside exponential family such a proposition is not true. For example consider testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$ for $X_1, ..., X_n \sim U(0, \theta)$. I will assign this as a homework problem.

Note that the problem here is for $H_1 : \theta > \theta_0$ we have a UMP test ϕ_1 and for $H_1 : \theta < \theta_0$ we have a UMP test ϕ_2. They have power curves like the one in the picture. However we want a power curve something like this. We will restrict the class of tests among which we are going to find out the most powerful test. Give a pictorial depiction of what we want.

Unbiasedness

A test ϕ is unbiased level α for $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ if (i) $\beta(\theta) \leq \alpha$ for $\theta \in \Theta_0$ (ii) $\beta(\theta) \geq \alpha$ for $\theta \in \Theta_1$.

If the power function $\beta(\theta)$ is continuous and if the test T is unbiased, then $\beta(\theta) = \alpha$ for all θ belonging to the boundary of $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$.

Result: If the densities are s.t. all tests have continuous power functions, then if there exists a UMP test T among the tests satisfying $E(T') = \alpha$ whenever θ belongs to the boundary of $\boldsymbol{\Theta}_0$ and $\boldsymbol{\Theta}_1$, then T is UMPU level α test.

Application to one parameter exponential family: We already know that the one parameter exponential family is given by $p_\theta(x) = c(\theta)h(x)\exp(w(\theta)T(x))$. Let us reparametrize this family and take $w(\theta) = \eta$. W.r.t the new parametrization, entire density can be written as $p_\eta(x) = c_1(\eta)h(x)\exp(\eta T(x))$. With such a parametrization, the following is true.

UMPU test: To test $H_0 : \eta = \eta_0$ vs. $H_1 : \eta \neq \eta_0$, the test function $\phi(T(\boldsymbol{X}))$ with

$$\phi(T(\boldsymbol{X})) = \begin{cases} 1 \text{ if } T(\boldsymbol{X}) < c_1 \text{ or } T(\boldsymbol{X}) > c_2 \\ 0 \text{ o.w.} \end{cases}$$

where $E_{\eta_0}(\phi(T(\boldsymbol{X}))) = \alpha$, $E_{\eta_0}(T(\boldsymbol{X})\phi(T(\boldsymbol{X}))) = \alpha E_{\eta_0}[T(\boldsymbol{X})]$, is the UMPU level α test.

With little calculation you will be able to see that if $T(\boldsymbol{X})$ is symmetrically distributed about a under H_0, then $E_{\eta_0}[\phi(T(\boldsymbol{X}))] = \alpha$, $c_1 + c_2 = 2a$ determine c_1, c_2.

Example (UMPU test): Suppose $X_1, ..., X_n \sim N(\theta, \sigma^2)$. To test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$. We use the previous result. Note that, $\eta = \theta$ and $T(\boldsymbol{X}) = \bar{X}$. Now the distribution of $\sqrt{n}(\bar{X} - \theta_0)/\sigma$ is symmetric around 0. Thus $c_1 + c_2 = 0$ and $P_{\theta_0}(\sqrt{n}(\bar{X} - \theta_0)/\sigma > c_2) + P_{\theta_0}(\sqrt{n}(\bar{X} - \theta_0)/\sigma < c_1) = \alpha$. Thus we have to solve the two equations

$$\begin{aligned} c_1 + c_2 &= 0 \\ 1 - \Phi(c_2) + \Phi(c_1) &= \alpha. \end{aligned}$$

$c_1 = -c_2 = -z_{\alpha/2}$.

Consider the test that rejects if $|\bar{X} - \theta_0| > \sigma z_{\alpha/2}/\sqrt{n}$. This is clearly an unbiased test as $P_{\theta_0}(|\bar{X} - \theta_0| > \sigma z_{\alpha/2}/\sqrt{n}) = P(Z > z_{\alpha/2}) + P(Z < -z_{\alpha/2}) = \alpha$ and $P_{\theta'}(|\bar{X} - \theta_0| > \sigma z_{\alpha/2}/\sqrt{n}) = P_{\theta'}(\bar{X} - \theta_0 > \sigma z_{\alpha/2}/\sqrt{n}) + P_{\theta'}(\bar{X} - \theta_0 < -\sigma z_{\alpha/2}/\sqrt{n}) = P_{\theta'}(\bar{X} - \theta' > \sigma z_{\alpha/2}/\sqrt{n} + \theta_0 - \theta') + P_{\theta'}(\bar{X} - \theta' < -\sigma z_{\alpha/2}/\sqrt{n} + \theta_0 - \theta') = 1 - \Phi(z_{\alpha/2} + \frac{\theta_0 - \theta'}{\sigma/\sqrt{n}}) + \Phi(-z_{\alpha/2} + \frac{\theta_0 - \theta'}{\sigma/\sqrt{n}})$.

Likelihood Ratio Test: The aim is to test $H_0 : \theta \in \boldsymbol{\Theta}_0$ vs. $H_1 : \theta \in \boldsymbol{\Theta}_1$, $\boldsymbol{\Theta}_1 = \boldsymbol{\Theta} - \boldsymbol{\Theta}_0$ based on samples $X_1, ..., X_n$. Let $f_\theta(\boldsymbol{X})$ be the likelihood at θ. The likelihood ratio test statistic is given by

$$\lambda = \frac{\sup\limits_{\theta \in \boldsymbol{\Theta}_0} f_\theta(\boldsymbol{X})}{\sup\limits_{\theta \in \boldsymbol{\Theta}} f_\theta(\boldsymbol{X})}.$$

The decision is to reject when $\lambda < c$, where c is chosen to satisfy the level condition, i.e. $\sup\limits_{\theta \in \boldsymbol{\Theta}_0} P_\theta(\lambda < c) \leq \alpha$. Since $0 \leq \lambda \leq 1$, so is c.

Example: Let $X_1, ..., X_n$ be random sample from a location shifted exponential density

$$f_\theta(x) = \begin{cases} e^{-(x-\theta)} & \text{if } x \geq \theta \\ 0 \text{ o.w.} \end{cases}$$

Consider testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. The likelihood function is

$$L(\theta) = \begin{cases} e^{-\sum X_i + n\theta)} & \text{if } X_{(1)} \geq \theta \\ 0 \text{ o.w.} \end{cases}$$

Now this is an increasing function of θ in the region $\theta \leq X_{(1)}$. Clearly, the denominator in the likelihood ratio is maximum at $X_{(1)}$. If $X_{(1)} \leq \theta_0$, the numerator is also maximum at $X_{(1)}$. Otherwise it is maximum at θ_0. Therefore the likelihood ratio test statistic (LRT) is

$$\lambda = \begin{cases} 1 \text{ if } X_{(1)} \leq \theta_0 \\ e^{-n(X_{(1)} - \theta_0)} \text{ o.w.} \end{cases}$$

The test is rejected if $\lambda < c$ or $X_{(1)} \geq \theta_0 - \frac{\log(c)}{n}$. Here LRT test is dependent on the data only through a sufficient statistic and this is not automatic. In fact

Result: If $T(\boldsymbol{X})$ is a sufficient statistic for θ and $\lambda^*(T(\boldsymbol{X}))$ and $\lambda(\boldsymbol{X})$ are the LRT test statistics based on $T(\boldsymbol{X})$ and $\boldsymbol{X}$, respectively, then $\lambda^*(T(\boldsymbol{X})) = \lambda(\boldsymbol{X})$ for every $\boldsymbol{X}$.

The proof of this result follows directly from the factorization theorem. Note that $f_\theta(\boldsymbol{X}) = g_\theta(T(\boldsymbol{X}))h(\boldsymbol{X})$. Now

$$\lambda(\boldsymbol{X}) = \frac{\sup\limits_{\theta \in \boldsymbol{\Theta}_0} f_\theta(\boldsymbol{X})}{\sup\limits_{\theta \in \boldsymbol{\Theta}} f_\theta(\boldsymbol{X})} = \frac{\sup\limits_{\theta \in \boldsymbol{\Theta}_0} g_\theta((T(\boldsymbol{X}))}{\sup\limits_{\theta \in \boldsymbol{\Theta}} g_\theta(T(\boldsymbol{X}))} = \lambda^*(T(\boldsymbol{X})).$$

Union-Intersection and Intersection-Union test

The union-intersection method of test construction might be useful when the null hypothesis is conveniently expressed as an intersection,

$$H_0 : \theta \in \bigcap_{\gamma \in \boldsymbol{\Gamma}} \boldsymbol{\Theta}_\gamma,$$

where $\mathbf{\Gamma}$ is an indexing set. Lets say the tests are available for each of the problems of testing $H_{0\gamma} : \theta \in \mathbf{\Theta}_\gamma$ vs. $H_{1\gamma} : \theta \in \mathbf{\Theta}_\gamma^c$. Further assume that the rejection region for each test is $\mathcal{R}_\gamma = \{\boldsymbol{X} : T_\gamma(\boldsymbol{X}) \in S_\gamma\}$. Then the rejection region for the union-intersection test is $\bigcup_{\gamma \in \mathbf{\Gamma}} \mathcal{R}_\gamma$.

The rational is that if any one of the $H_{0\gamma}$ is rejected, then H_0 is rejected.

The intersection-union test is constructed when the null hypothesis is expressed as $H_0 : \theta \in \bigcup_{\gamma \in \mathbf{\Gamma}} \mathbf{\Theta}_\gamma$. Suppose for each test the rejection region is $\mathcal{R}_\gamma = \{\boldsymbol{X} : T_\gamma(\boldsymbol{X}) \in S_\gamma\}$. Then the rejection region for the intersection-union test is $\bigcap_{\gamma \in \mathbf{\Gamma}} \mathcal{R}_\gamma$.

Note that intersection-union or union-intersection tests are constructed in such a way that it is difficult to evaluate size of these tests. However, we can find certain bounds on their size in various examples. These bounds are important to ensure that they are of level α.

Theorem: Consider a UIT test for testing $H_0 : \mathbf{\Theta}_0$ vs. $H_1 : \theta \in \mathbf{\Theta}_0^c$, where $\mathbf{\Theta}_0 = \bigcap_{\gamma \in \mathbf{\Gamma}} \mathbf{\Theta}_\gamma$. Let $\lambda_\gamma(\boldsymbol{X})$ be the LRT statistics for testing $H_{0\gamma}$ vs. $H_{1\gamma}$. Let $\lambda(\boldsymbol{X})$ be the LRT statistic for testing $H_0 : \theta \in \mathbf{\Theta}_0$ vs. $H_1 : \theta \in \mathbf{\Theta}_0^c$. Define $T(\boldsymbol{X}) = \inf_{\gamma \in \mathbf{\Gamma}} \lambda_\gamma(\boldsymbol{X})$ so that the UIT rejection region $\bigcup \mathcal{R}_\gamma = \{\boldsymbol{X} : T(\boldsymbol{X}) < c\}$, $\mathcal{R}_\gamma = \{\boldsymbol{X} : \lambda_\gamma(\boldsymbol{X}) < c\}$. Then

1. $T(\boldsymbol{X}) > \lambda(\boldsymbol{X})$ for every $\boldsymbol{X}$.
2. If $\beta_T(\theta)$ and $\beta_\lambda(\theta)$ are power functions based on the tests T and λ respectively, then $\beta_T(\theta) \leq \beta_\lambda(\theta)$, for every θ.

proof: Note that, $\lambda(\boldsymbol{X}) \geq \lambda_\gamma(\boldsymbol{X})$, for any γ, as numerator is maximized over a bigger set. Thus $\lambda(\boldsymbol{X}) \geq T(\boldsymbol{X})$.

Also $\beta_T(\theta) = P_\theta(T(\boldsymbol{X}) < c) \leq P_\theta(\lambda(\boldsymbol{X}) < c) = \beta_\lambda(\theta)$.

Therefore, LRT is uniformly most powerful than UIT. The usefulness of UIT lies in the fact that when LRT rejects H_0, by looking at various tests in UIT you might get additional information. Note that this result is true only for UITs constructed through likelihood ratio tests. If LRT $\lambda(\boldsymbol{X})$ is difficult to compute for the entire region, you don't know how to keep your UIT test at level α. For intersection-union or IUT test such problems do not arise. Here is the result that says so.

Theorem: Let α_γ be the size of the test of $H_{0\gamma}$ with rejection region $\mathcal{R}_\gamma$. Then the IUT test with rejection region $\bigcap\limits_{\gamma\in\boldsymbol{\Gamma}} \mathcal{R}_\gamma$ rejects null hypothesis H_0 at level $\alpha = \sup\limits_{\gamma\in\boldsymbol{\Gamma}} \alpha_\gamma$.
proof: Let $\theta \in \boldsymbol{\Theta}_0$, then $\theta \in \boldsymbol{\Theta}_\gamma$ for some γ and $P_\theta(\boldsymbol{X} \in \bigcap\limits_{\gamma\in\boldsymbol{\Gamma}} \mathcal{R}_\gamma) \leq P_\theta(\boldsymbol{X} \in \mathcal{R}_\gamma, \gamma \in \boldsymbol{\Gamma}) = \alpha_\gamma \leq \alpha$.

Therefore, for IUT test one can start with any tests and ensure level α.

Example: Let $X_1, ..., X_n$ be a random sample from a $N(\mu, \sigma^2)$ population. Consider testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. $H_0 : \{\mu : \mu \leq \mu_0\} \cap \{\mu : \mu \geq \mu_0\}$. The LRT for $H_{01} : \mu \leq \mu_0$ vs. $H_{11} : \mu > \mu_0$ is reject if $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \geq t_1$. Similarly, the LRT of $H_{02} : \mu \geq \mu_0$ vs. $H_{12} : \mu < \mu_0$ if $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \leq t_2$. Then the union intersection test is to reject $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \geq t_1$ or $\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \leq t_2$. If $t_2 = -t_1 \geq 0$, the union intersection is the same as the LRT test. This is also the two sided t-test.

Bayesian Test

In Bayesian testing, entire testing can be formulated based on the posterior distribution. Suppose we want to test $H_0 : \theta \in \boldsymbol{\Theta}_0$ vs. $H_1 : \theta \in \boldsymbol{\Theta}_0^c$. In Bayesian testing, we compute $P(\theta \in \boldsymbol{\Theta}_0|\boldsymbol{X})$ and $P(\theta \in \boldsymbol{\Theta}_0^c|\boldsymbol{X})$ and choose the hypothesis having higher posterior probability, i.e. choose H_0 if $P(\theta \in \boldsymbol{\Theta}_0|\boldsymbol{X}) \geq 0.5$.

Example: Let $X_1, ..., X_n \sim N(\theta, \sigma^2)$ and the prior distribution for θ is $N(\mu, \tau^2)$, σ^2, tau^2, μ are known. Consider testing $H_0 : \theta \leq \theta_0$ vs. $H_1 : \theta > \theta_0$. Note that $\theta|\bar{X} \sim N(\frac{n\tau^2\bar{X}+\sigma^2\mu}{n\tau^2+\sigma^2}, \frac{\sigma^2\tau^2}{n\tau^2+\sigma^2})$. We do not reject H_0 is and only if $P(\theta \leq \theta_0|\boldsymbol{X}) \geq 0.5$. Since the posterior distribution is symmetric, this is true if and only if $\frac{n\tau^2\bar{X}+\sigma^2\mu}{n\tau^2+\sigma^2} \leq \theta_0$. This implies $\bar{X} \leq \theta_0 + \sigma^2(\theta_0 - \mu)/n\tau^2$.

4 Interval Estimation

Until now, we have seen

- How to provide point estimate for an unknown parameter.
- What is the optimal way to do it.
- How to test various hypotheses on the parameter.

In this section, we are going to talk about interval estimates. What do we mean by interval estimates? Here is the definition.

Definition (Interval Estimator): An interval estimate of a real valued parameter θ is any pair of functions $L(X_1, ..., X_n)$ and $U(X_1, ..., X_n)$ of a sample that satisfy $L(\boldsymbol{X}) \leq U(\boldsymbol{X})$ for all $\boldsymbol{X}$. The random interval $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is called an interval estimator. Note that the interval can also be open/half open and $L(\boldsymbol{X})$, $U(\boldsymbol{X})$ can be $-\infty, \infty$ respectively.

Example: For a sample $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ^2 known. Now an interval estimate of μ is $(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})$, which means that we will assert that μ lives in the interval. Obviously a point estimate is more precise than an interval estimator. Then what do we gain by moving to an imprecise estimator from a precise estimator? We gain confidence. What do I mean by that?

We known that for our normal example point estimate of μ is $\bar{X}$. We have seen this estimator is UMVUE. But $P_\mu(\bar{X} = \mu) = 0$. However, $P_\mu(\mu \in (\bar{X} - z_{\alpha/2}\sigma/\sqrt{n}, \bar{X} + z_{\alpha/2}\sigma/\sqrt{n})) = 1 - \alpha$. Thus with the imprecise interval estimator we can say with $100(1-\alpha)\%$ confidence that μ lies in the interval. $(1-\alpha)$ in this case is known as the coverage probability of the interval. A formal definition of the *coverage probability* is given below.

Definition: For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the coverage probability of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the probability that the random interval covers the parameter θ. In symbol, It is denoted by $P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$.

This means if we go on constructing intervals with different samples many times, roughly $P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$ proportion of time θ will lie in the interval.

Definition: For an interval estimator $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ of a parameter θ, the *confidence coefficient* of $[L(\boldsymbol{X}), U(\boldsymbol{X})]$ is the infimum of the coverage probabilities, $inf_\theta P_\theta(\theta \in [L(\boldsymbol{X}), U(\boldsymbol{X})])$. Interval estimators together with the confidence coefficient is sometimes known as the *confidence interval.*

4.1 Methods of Finding Interval Estimators

In this section we will describe two different methods of finding interval estimators, though both of them are not very different from each other.

4.1.1 Inverting a test statistics

Think about the old example of $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ^2 known and let $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. We known an UMPU level α test is that which accepts H_0 if $|\bar{X} - \mu_0| \leq \sigma z_{\alpha/2}/\sqrt{n}$, or

$$P_{\mu_0}(\bar{X} - \sigma z_{\alpha/2}/\sqrt{n} \leq \mu_0 \leq \bar{X} + \sigma z_{\alpha/2}/\sqrt{n}) = 1 - \alpha.$$

However this statement is true for all μ. Hence a $100(1-\alpha)\%$ confidence interval of μ is given by $[\bar{X} - \sigma z_{\alpha/2}/\sqrt{n}, \bar{X} + \sigma z_{\alpha/2}/\sqrt{n}]$.

Result: For each $\theta_0 \in \boldsymbol{\Theta}$. Let $A(\theta_0)$ be the acceptance region of a level α test of $H_0 : \theta = \theta_0$. For each $\boldsymbol{X} \in \mathcal{X}$, define a set $C(\boldsymbol{X})$ in the parameter space by $C(\boldsymbol{X}) = \{\theta_0 : \boldsymbol{X} \in A(\theta_0)\}$. Then the random set $C(\boldsymbol{X})$ is the $(1-\alpha)$ confidence set.

We have seen an illustration of the above result in the normal case. We will see another way to do it.

Inverting an LRT: Let $X_1, ..., X_n \sim \mathit{Exponential}(\lambda)$. We have to find $(1-\alpha)$ confidence interval for λ. Let us try to invert an LRT test of $H_0 : \lambda = \lambda_0$ vs. $H_1 : \lambda \neq \lambda_0$. The LRT statistics in this case is given by

$$L = \frac{\frac{1}{\lambda_0^n} e^{-\sum X_i/\lambda_0}}{\sup_\lambda \frac{1}{\lambda^n} e^{-\sum X_i/\lambda}} = \left(\frac{e\sum X_i}{n\lambda_0}\right)^n e^{-\sum X_i/\lambda_0}.$$

For fixed λ_0, the acceptance region is given by

$$A(\lambda_0) = \{\boldsymbol{X} : \left(\frac{\sum X_i}{\lambda_0}\right)^n e^{-\sum X_i/\lambda_0} \geq k^*\}.$$

Inverting the LRT we obtain the $(1-\alpha)$ confidence set as $C(\boldsymbol{X}) = \{\lambda : \left(\frac{\sum X_i}{\lambda}\right)^n e^{-\sum X_i/\lambda} \geq k^*\}$. Now $C(\boldsymbol{X})$ depends only through $\sum X_i$. So the confidence interval can be expressed in the form $C(\sum X_i) = \{\lambda : L(\sum X_i) \leq \lambda \leq U(\sum X_i)\}$. Now note that $\sum X_i/\lambda \sim$ $Gamma(2,1)$. Hence $P(a < \sum X_i/\lambda < b) = 1-\alpha$. So the $(1-\alpha)$ confidence interval is $(\sum X_i/b, \sum X_i/a)$.

Normal one sided confidence bound: Remember testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$. The UMP α test rejects alternative when $\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha$. So a $(1-\alpha)$ confidence interval for μ is $[\bar{x} - Z_\alpha \sigma/\sqrt{n}, \infty)$.

4.1.2 Pivotal Quantities

In the last example we saw that $\sum X_i/\lambda$ is a quantity whose distribution is indenpendent of the parameter. We used that to create a confidence interval. Actually this is a specific example to a very general case. We call such quantities as *pivotal quantities* whose distribution do not depend on the parameters.

Definition (pivotal quantity): A random variable $Q(\boldsymbol{X}, \theta) = Q(X_1, ..., X_n; \theta)$ is a *pivotal quantity (or pivot)* if the distribution of $Q(\boldsymbol{X}, \theta)$ is independent of all parameters.

There are some class of distributions where constructing pivotal quantities are fairly easy. They are the following.

(i) *Location Family:* When $X_i \sim f(x-\mu)$, $\bar{X} - \mu$ is an easily conceivable pivotal quantity.

(ii) *Scale Family:* When $X_i \frac{1}{\sigma} \sim f(\frac{x}{\sigma})$, $\frac{\bar{X}}{\sigma}$ is an easily conceivable pivotal quantity.

(iii) *Location-scale Family:* When $X_i \frac{1}{\sigma} \sim f(\frac{x-\mu}{\sigma})$, $\frac{\bar{X}-\mu}{S}$ is an easily conceivable pivotal quantity. Here S is the standard deviation of $X_1, ..., X_n$.

Of course for a specific family of distributions, there can be multiple pivotal quantities.

Examples: $N(\mu, \sigma^2)$ is a location scale family, $Exponential(\lambda)$ is a scale family, $Gamma(\alpha, \beta)$, α known, is a scale family.

As you might have already guessed that once we have a pivotal quantity $Q(\boldsymbol{X}, \theta)$, it is really easy to come up with a confidence interval for θ. If $Q(\boldsymbol{X}, \theta)$ is the pivotal quantity, there can be found a, b s.t.

$$P_\theta(a < Q(\boldsymbol{X}, \theta) < b) \geq 1 - \alpha.$$

Then for each θ_0, $A(\theta_0) = \{\boldsymbol{X} : a < Q(\boldsymbol{X}, \theta) < b\}$ is the acceptance region for a level α test. Therefore by a previous result we have $C(\boldsymbol{X}) = \{\theta : a < Q(\boldsymbol{X}, \theta) < b\}$ is the $(1-\alpha)$ confidence set for θ. If θ is a real valued parameter and if for each $\boldsymbol{X}$, $Q(\boldsymbol{X}, \theta)$ is a monotone function of θ, then $C(\boldsymbol{X})$ will be an interval.

Example: $X_1, ..., X_n \sim N(\mu, \sigma^2)$, μ is known. We know $T = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. T is a pivotal quantity so that $P(\chi^2_{n-1,\alpha/2} < T < \chi^2_{n-1,1-\alpha/2}) = 1 - \alpha$, which gives us a $(1-\alpha)$ confidence interval.

Poisson interval estimator: Let $X_1, ..., X_n$ be a random sample with Poisson distribution having parameter λ. Need to find a $100(1-\alpha)\%$ confidence interval for λ. Now $Y = \sum_{i=1}^n X_i \sim Pois(n\lambda)$. Note that a set of λ that contains $(1-\alpha)$ probability is $\{\lambda : F_\lambda(y) \leq 1 - \alpha/2, F_\lambda(y) \geq \alpha/2\}$, y is the observed value of $\sum X_i$. If $F_\lambda(\cdot)$ is monotone as a function of λ, this set is an interval. To find upper bound and lower bound for this interval, we solve the following equations

$$\sum_{k=0}^{y} e^{-n\lambda}\frac{(n\lambda)^k}{k!} = \alpha/2, \quad \sum_{k=y}^{\infty} e^{-n\lambda}\frac{(n\lambda)^k}{k!} = \alpha/2.$$

Luckily, $\sum_{k=0}^{y} e^{-n\lambda}\frac{(n\lambda)^k}{k!} = P(\chi^2_{2(y+1)} > 2n\lambda)$, thus the solution to the above equation is $\lambda = 1/2n\chi^2_{2(y+1),\alpha/2}$. Similarly the lower bound can be found out.

4.1.3 Bayesian Intervals

Problems with frequentist intervals: Frequentist interval is essentially a a random interval and coverage of the frequentist interval is a statement based on repeated experiments. This might lead to some anomalies.

example 1: $X_1, ..., X_n \overset{iid}{\sim} U(\theta - \frac{1}{2}, \theta + \frac{1}{2})$. Clearly this is a location family and hence $\bar{X} - \theta$ is a pivotal quantity. Therefore, the 95% confidence interval is $[\bar{X} - C, \bar{X} + C]$ where $P(-C < \bar{X} - \theta < C) = 0.95$. This is a 95% confidence interval for always. Now assume we observed $X_1 = 1, X_2 = 2$, then we know with 100% confidence that $\theta = 1.5$. But you can't determine it from frequentist viewpoint.

In the classical statistics when we obtain a $(1-\alpha)$ confidence interval say $a \leq \theta \leq b$, we are tempted to say that the probability is $(1-\alpha)$ for the parameter to stay in the interval [a,b]. However in classical statistics the interval is a random quantity while the parameter is fixed. Therefore such a conclusion does not hold. What holds is that *if we construct the interval for many times with different random samples, about* $(1-\alpha)$ *fraction of the time it will contain the true parameter.* However θ is a fixed quantity and it should either be or not be in the interval. This is in some sense leads to conceptual ambiguity.

In contrast in the Bayesian set up θ is a random variable and the observed data are fixed. So conceptually $\theta \in [a, b]$ should occur with a prob. not with 0 or 1. In the Bayesian set both hypothesis testing and interval estimation are incredibly unambiguous and easy. Suppose

$X_1, ..., X_n \sim f_\theta(x)$, $\theta \sim \pi$, then the posterior distribution of $\theta|X_1, ..., X_n$ is given by

$$\pi(\theta|X_1, ..., X_n) \propto \prod f_\theta(X_i)\pi(\theta).$$

Once we have the posterior, to carry out a hypothesis testing of $H_0 : \theta \in \mathbf{\Theta}_0$ vs. $H_1 : \theta \in \mathbf{\Theta}_1$ we only need to see if $\int_{\mathbf{\Theta}_0} \pi(\theta|X_1, ..., X_n)d\theta > 0.5$. If so then null is not rejected, o.w. null is rejected. Note that the cut-off might change, but the principle is the same and it saves us from different case consideration and finding out some UMP test. Similarly we define the idea of credible set in Bayesian statistics.

Definition: For $0 < \alpha < 1$ a $100(1-\alpha)\%$ credible set C is a set s.t. $P(\theta \in C|\boldsymbol{X}) = 1 - \alpha$. When a credible set is an interval we call it a credible interval. We always want to work with a credible interval rather than a general credible set.

Note that unlike frequentist confidence interval, in Bayesian statistics θ is a random variable. Therefore, the probability that θ belongs to some set is between 0 and 1. Ideally you can have thousands of $100(1-\alpha)\%$ credible intervals. Which one to choose among them. In interval estimation, you will always look for choosing an interval which has same coverage but smaller length. For that one needs highest posterior density (HPD) intervals.

Result: Suppose the posterior density of θ is unimodal. Then the HPD interval for θ is the interval $C = \{\theta : \pi(\theta|\boldsymbol{X}) \geq k\}$, where k is chosen such that $\pi(C|\boldsymbol{X}) = 1 - \alpha$. This interval is obviously the smallest interval among all $100(1-\alpha)\%$ credible intervals.

Example: Recall an earlier example where $X_1, ..., X_n \sim Ber(p)$, $p \sim Beta(a, b)$ then $p|X_1, ..., X_n \sim Beta(\sum_{i=1}^n X_i + a, n - \sum_{i=1}^n X_i + b)$. Let a, b be chosen cut-offs from the Beta distribution so that the interval contains prob. $(1-\alpha)$, then that interval becomes a $(1-\alpha)$ credible interval for p.

Another significant advantage of Bayesian credible interval is that it is very easy to accurately estimate even with most complicated examples. This is what you are going to do.

- You will run MCMC to draw sufficient number of posterior samples.

- Find an interval so that the empirical probability of this interval is $(1-\alpha)$.

5 Method of Evaluating Intervals

In general while constructing an interval estimate we want the intervals to have smallest possible size with largest possible coverage. However the requirements are conflicting. Hence, we fix the coverage and try to minimize length of the confidence interval.

Remember the example where we had $X_1, ..., X_n \sim N(\mu, \sigma^2)$, σ known. A $(1-\alpha)$ confidence interval can be of the form $\{\mu : \bar{X} - b\sigma/\sqrt{n} \leq \mu \leq \bar{X} - a\sigma/\sqrt{n}\}$ where $P(a \leq Z \leq b) = 1-\alpha$, $Z \sim N(0,1)$. We have always chosen $b = -a = z_{\alpha/2}$. Was there any reason? The answer is yes. In fact

Result: Let $f(x)$ be a unimodal pdf. If the interval $[a,b]$ satisfies
(1) $\int_a^b f(x)dx = 1-\alpha$ (2) $f(a) = f(b) > 0$ (3) $a \leq x^* \leq b$, where x^* is the mode a mode of $f(x)$. Then $[a,b]$ is the shortest interval among all that satisfy (1).

Proof Let $[a', b']$ be any interval with $b' - a' < b - a$. We will show that this implies $\int_{a'}^{b'} f(x)dx < 1-\alpha$. The result will only be proved for $a' \leq a$, the proof being similar for $a' > a$. Also for $a' \leq a$, we need to consider two cases, $b' \leq a$ and $b' > a$.

case 1: If $b' \leq a$, then $a' \leq b' \leq a \leq x^*$. Thus

$$\int_{a'}^{b'} f(x)dx \leq f(b')(b'-a') \leq f(a)(b'-a') < f(a)(b-a) \leq \int_a^b f(x)dx$$

Thus, case 1 is done.

case 2: If $b' > a'$, then $a' \leq a < b' < b$, o.w. $[a', b']$ will be contained in $[a,b]$. Now

$$\int_{a'}^{b'} f(x)dx = \int_a^b f(x)dx + \left[\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx\right].$$

$$\int_{a'}^{a} f(x)dx \leq f(a)(a-a'), \int_{b'}^{b} f(x)dx \geq f(b)(b-b').$$

Thus

$$\int_{a'}^{a} f(x)dx - \int_{b'}^{b} f(x)dx \leq f(a)(a-a') - f(b)(b-b')$$
$$= f(a)[(a-a')-(b-b')] = f(a)[(b'-a')-(b-a)].$$

The last expression is negative as we have assumed that $b'-a' < b-a$.

We have previously seen examples of normal and t-tests as examples of this theorem. You can mainly use this result for pivoting with location families, otherwise you have to be careful to use this result. Now I am going to show another way to find the above result. This is a more general way and will be useful in other cases, as we will see.

Our objective is to minimize $b-a$ subject to $\int_a^b f(x)dx = 1-\alpha$. This is a constrained optimization and also note that the optimized value of b is a function of a, denote it by $b(a)$. Clearly we have to minimize two equations

$$\frac{db}{da} - 1 = 0, \quad \frac{db}{da} f(b) - f(a) = 0.$$

They together give $f(b) = f(a)$.

Example: Suppose $X \sim Gamma(k, \beta)$. The quantity $Y = X/\beta$ is a pivot, with $Y \sim Gamma(k, 1)$. Therefore confidence interval of β can be found choosing cut-offs so that $P(a \leq Y \leq b) = 1-\alpha$. We can't use the previous theorem blindly here. Because the interval of β is of the form $\{\beta : \frac{X}{b} \leq \beta \leq \frac{X}{a}\}$. Therefore, the length of the interval is $X(1/a - 1/b)$, not $b-a$. Now the task is to maximize $1/a - 1/b$ subject to $\int_a^b f(x)dx = 1-\alpha$. Taking derivative w.r.t. a, we have two equations $-\frac{1}{a^2} + \frac{db}{da}\frac{1}{b^2} = 0$, $\frac{db}{da} f(b) - f(a) = 0$. This gives $b^2 f(b) = a^2 f(a)$.

There is one interesting question still unanswered. Is there any connection between optimality of hypothesis testing and optimality of finding confidence set. Now we are going to define another intriguing concept that connects them. Note that the probability of true coverage for an interval $C(\boldsymbol{X})$ is given by $P_\theta(\theta \in C(\boldsymbol{X}))$, i.e probability of covering the true parameter. The probability of *false* coverage is the function of θ, θ' s.t.

$P_\theta(\theta' \in C(\boldsymbol{X})),\ \theta \neq \theta',\ \text{ if } C(\boldsymbol{X}) = [L(\boldsymbol{X}), U(\boldsymbol{X})]$

$P_\theta(\theta' \in C(\boldsymbol{X})),\ \theta < \theta',\ \text{ if } C(\boldsymbol{X}) = (-\infty, U(\boldsymbol{X})]$

$P_\theta(\theta' \in C(\boldsymbol{X})),\ \theta > \theta',\ \text{ if } C(\boldsymbol{X}) = [L(\boldsymbol{X}), \infty),$

it is the probability of covering θ' when the true parameter is θ. Note that false coverage will be big if we unnecessarily cover unimportant θ' and it would potentially increase the length of the interval without increasing coverage. Therefore our aim should be in reducing false coverage. Here is a theorem that connects an acceptance region of a UMP test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$ to the purpose of reducing false coverage.

Result: Let $X_1, ..., X_n \sim f_\theta(x)$, where θ is a real-valued parameter. For each $\theta_0 \in \boldsymbol{\Theta}$, let $A^*(\theta_0)$ be the UMP level α acceptance region of a test of $H_0 : \theta = \theta_0$ vs. $H_1 : \theta > \theta_0$. Let $C^*(\boldsymbol{X})$ be the $(1-\alpha)$ confidence set formed by inverting the UMP acceptance regions. Then for any $(1-\alpha)$ confidence set C,

$$P_\theta(\theta' \in C^*(\boldsymbol{X})) \leq P_\theta(\theta' \in C(\boldsymbol{X})), \text{ for all } \theta' < \theta.$$

Proof Let $\theta' < \theta$ be any value. Let $A(\theta')$ be the acceptance region of the level α test of $H_0 : \theta = \theta'$ obtained by inverting C. Since $A^*(\theta')$ is the UMP acceptance region for testing $H_0 : \theta = \theta'$ vs. $H_1 : \theta > \theta'$. Since $\theta > \theta'$, we have

$$P_\theta(\theta' \in C^*(\boldsymbol{X})) = P_\theta(\boldsymbol{X} \in A^*(\theta')) \leq P_\theta(\boldsymbol{X} \in A(\theta')) = P_\theta(\theta' \in C(\boldsymbol{X})).$$

We have previously seen that for the normal problem $C(\bar{X}) = \{\mu : \mu \geq \bar{X} - z_\alpha \sigma/\sqrt{n}\}$ is the acceptance region for the UMP test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$. Therefore this interval reduces false coverage probability.

Definition: A $(1-\alpha)$ confidence set $C(\boldsymbol{X})$ is unbiased if $P_\theta(\theta' \in C(\boldsymbol{X})) \le 1-\alpha$ for all $\theta \neq \theta'$.

An unbiased confidence set ensures that false coverage is never more than the minimum probability of true coverage. An unbiased confidence set is obtained by inverting an biased test. I will conclude this section by stating a result that connects our intuition of length of $C(\boldsymbol{X})$ and probability of false coverage. This is a theorem by Pratt, but Ghosh independently proved it at about the same time.

Result: Let $X \sim f_\theta(x)$, θ is a real valued parameter. Let $C(X) = [L(X), U(X)]$ be the confidence interval for θ. If $L(x)$ and $U(x)$ are both increasing functions of x, then for any θ'

$$E_{\theta'}(Length[C(X)]) = \int_{\theta\neq\theta'} P_{\theta'}(\theta \in C(X))d\theta.$$

Proof From the definition

$$\begin{aligned}
E_{\theta^*}(Length[C(X)]) &= \int_{\mathcal{X}} Length[C(X)] f_{\theta^*}(x)dx \\
&= \int_{\mathcal{X}} [U(x)-L(x)] f_{\theta^*}(x)dx \\
&= \int_{\mathcal{X}} [\int_{L(x)}^{U(x)} d\theta] f_{\theta^*}(x)dx \\
&= \int_{\boldsymbol{\Theta}} [\int_{U^{-1}(\theta)}^{L^{-1}(\theta)} f_{\theta^*}(x)]d\theta \\
&= \int_{\boldsymbol{\Theta}} P_{\theta^*}(U^{-1}(\theta) \le X \le L^{-1}(\theta))d\theta \\
&= \int P_{\theta^*}(\theta \in C(X))d\theta = \int_{\theta\neq\theta^*} P_{\theta^*}(\theta \in C(X))d\theta
\end{aligned}$$