Notes on PSTAT 207

Haosheng Zhou

Sept, 2022

Contents

Moments and Generating Functions

Moments of Distribution

Theorem 1. *If $\mathbb{E}|X|^k < \infty$ for some $k > 0$, then $n^k \mathbb{P}(|X| > n) \to 0$ $(n \to \infty)$.*

Proof.

$$\mathbb{E}|X|^k = \int_{\mathbb{R}} |x|^k \, dF(x) < \infty < \infty \tag{1}$$

$$\forall \varepsilon > 0, \exists N > 0, \int_{|x| \geq N} |x|^k \, dF(x) < \varepsilon \tag{2}$$

with a simple Markov estimation

$$\int_{|x| \geq N} |x|^k \, dF(x) \geq N^k \mathbb{P}(|X| \geq N) \tag{3}$$

set $\varepsilon \to 0, N \to \infty$ to get the conclusion.

□

Remark. *This theorem shows that the existence of moments can infer the behavior of tail probability.*

However, the converse is not necessarily true. Consider $\mathbb{P}(X = n) = \frac{C}{n^2 \log n}$ $(n = 2, 3, ...)$ with C appropriately picked such that it's a probability distribution ($\sum_n \frac{1}{n^2 \log n} < \infty$). Note that $\mathbb{E}|X| = \sum_n \frac{C}{n \log n} = \infty$ but

$$n\mathbb{P}(|X| > n) \sim Cn \int_n^\infty \frac{1}{x^2 \log x} \, dx \sim \frac{C}{\log n} \to 0 (n \to \infty) \tag{4}$$

*To go from tail probability to the existence of $\mathbb{E}|X|^k$, **moment condition** is needed, i.e.*

$$\exists \delta > 0, n^{k+\delta} \mathbb{P}(|X| > n) \to 0 \ (n \to \infty) \tag{5}$$

This moment condition will be proved below and the upper counterexample shows us that if $\delta = 0$, the logarithm factor may have dominant effect, and that's why the converse fails.

Theorem 2. *If $\exists \alpha > \beta > 0, n^\alpha \mathbb{P}(|X| > n) \to 0$ $(n \to \infty)$, then $\mathbb{E}|X|^\beta < \infty$.*

Proof.

$$\mathbb{E}|X|^\beta = \beta \int_0^\infty y^{\beta - 1} \mathbb{P}(|X| > y) \, dy \tag{6}$$

$$= \beta \int_0^N + \beta \int_N^\infty \tag{7}$$

with the first term to be finite. The second term is also finite since $\beta - \alpha - 1 < -1$

$$\int_N^\infty y^{\beta-1}\mathbb{P}\left(|X| > y\right)\, dy = \int_N^\infty y^{\beta-\alpha-1}y^\alpha \mathbb{P}\left(|X| > y\right)\, dy \tag{8}$$

$$\leq \varepsilon \int_N^\infty y^{\beta-\alpha-1}\, dy < \infty \tag{9}$$

□

Remark. *Note that there exists r.v. that does not have any positive powered moments. Consider*

$$f(x) = \frac{1}{2|x|\log^2|x|}\ (|x| > e) \tag{10}$$

$$\mathbb{E}|X|^\alpha = \int_e^\infty \frac{|x|^{\alpha-1}}{\log^2 x}\, dx = \infty \tag{11}$$

This counterexample has $\forall c > 1, \frac{\mathbb{P}(|X|>cx)}{\mathbb{P}(|X|>x)} \to 1\ (x \to \infty)$*, which is the reason of failure since the survival function is varying too slowly (a heavy tail). To verify,*

$$\mathbb{P}\left(|X| > x\right) = \int_x^\infty \frac{1}{t\log^2 t}\, dt = \frac{1}{\log x} \tag{12}$$

$$\frac{\mathbb{P}\left(|X| > cx\right)}{\mathbb{P}\left(|X| > x\right)} = \frac{\log x}{\log x + \log c} \to 1\ (x \to \infty) \tag{13}$$

If positive function L *on* $(0, +\infty)$ *satisfies* $\frac{L(cx)}{L(x)} \to 1\ (x \to \infty)$*, then it's called* ***slowly varying****. If* $\mathbb{P}\left(|X| \geq x\right)$ *is slowly varying, then* $\forall \alpha > 0, x^\alpha \mathbb{P}\left(|X| > x\right) \to \infty\ (x \to \infty)$*, so* $\forall \alpha > 0, \mathbb{E}|X|^\alpha = \infty$*, giving a r.v. with no finite moments. Opposite to that, if the two-sided survival function decays quickly enough, then all moments exist.*

Theorem 3. *If* $\forall c > 1, \frac{\mathbb{P}(|X|>cx)}{\mathbb{P}(|X|>x)} \to 0\ (x \to \infty)$*, then all moments exist.*

Proof.

$$\forall c > 1, \forall \varepsilon > 0, \exists x_0, \forall x > x_0, \frac{\mathbb{P}\left(|X| > cx\right)}{\mathbb{P}\left(|X| > x\right)} < \varepsilon \tag{14}$$

$$\forall \alpha > 0, \mathbb{E}|X|^\alpha = \alpha \int_0^\infty y^{\alpha-1}\mathbb{P}\left(|X| > y\right)\, dy = \alpha \int_0^{x_0} + \alpha \int_{x_0}^\infty \tag{15}$$

it's clear that the first term is finite, let's only prove that the second term w.r.t. the tail is also finite.

$$\int_{x_0}^{\infty} y^{\alpha-1}\mathbb{P}\left(|X|>y\right)\,dy = \int_{x_0}^{cx_0} + \int_{cx_0}^{c^2x_0} + ... \tag{16}$$

$$\leq \int_{x_0}^{cx_0} y^{\alpha-1}\mathbb{P}\left(|X|>y\right)\,dy + \int_{cx_0}^{c^2x_0} y^{\alpha-1}\varepsilon\mathbb{P}\left(|X|>\frac{y}{c}\right)\,dy \tag{17}$$

$$+ \int_{c^2x_0}^{c^3x_0} y^{\alpha-1}\varepsilon^2\mathbb{P}\left(|X|>\frac{y}{c^2}\right)\,dy + ... \tag{18}$$

$$= \int_{x_0}^{cx_0} y^{\alpha-1}\mathbb{P}\left(|X|>y\right)\,dy + c\varepsilon\int_{x_0}^{cx_0} y^{\alpha-1}\mathbb{P}\left(|X|>u\right)\,du \tag{19}$$

$$+ c^2\varepsilon^2\int_{x_0}^{cx_0} y^{\alpha-1}\mathbb{P}\left(|X|>u\right)\,du + ... \tag{20}$$

$$= \frac{1}{1-c\varepsilon}\int_{x_0}^{cx_0} y^{\alpha-1}\mathbb{P}\left(|X|>y\right)\,dy < \infty \tag{21}$$

for any fixed $c>1$ and ε small enough.

□

Generating functions

Probability generating function is defined for non-negative discrete random variable with

$$p(s) = \mathbb{E}s^X = \sum_{k=0}^{\infty} p_k s^k \tag{22}$$

since $\sum_k p_k = 1$, this power series is absolute convergence for $|s| \leq 1$. By the property of power series, we can interchange differentiation and infinite sum to get:

$$p'(s) = \sum_{k=0}^{\infty} k p_k s^{k-1} \tag{23}$$

$$p''(s) = \sum_{k=0}^{\infty} k(k-1) p_k s^{k-2} \tag{24}$$

$$\mathbb{E}X = p'(1), \mathbb{E}X^2 = p'(1) + p''(1) \tag{25}$$

$$Var(X) = p'(1) + p''(1) - (p'(1))^2 \tag{26}$$

the information of moments if they exist.

Moment generating function is defined as

$$M(t) = \mathbb{E}e^{tX} \tag{27}$$

however, it may only exist in a certain interval but not on the whole $\mathbb{R}$. To calculate the moments,

$$M'(t) = \mathbb{E}Xe^{tX} \tag{28}$$
$$M'(0) = \mathbb{E}X \tag{29}$$
$$M''(t) = \mathbb{E}X^2e^{tX} \tag{30}$$
$$M''(0) = \mathbb{E}X^2 \tag{31}$$

Note that MGF and the distribution function has a 1-on-1 correspondence, which means that in order to prove two random variables have the same distribution, calculating the MGF and comparing suffice. (Laplacian transform)

One question is that can the moments of a distribution uniquely determine the distribution? The answer is NO generally. The classical counterexample is given by the lognormal distribution.

$$f(x) = \frac{1}{\sqrt{2\pi}x}e^{-\frac{\log^2 x}{2}} \ (x > 0) \tag{32}$$
$$f_\varepsilon(x) = f(x)[1 + \varepsilon \sin(2\pi \log x)] \ (x > 0, |\varepsilon| < 1) \tag{33}$$

It's easy to see that

$$\int_0^\infty f(x)\sin(2\pi\log x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{u^2}{2}}\sin(2\pi u)\,du = 0 \tag{34}$$

so f_ε really is a density. However, all its moments are the same as lognormal since

$$\int_0^\infty x^n f(x)\sin(2\pi\log x)\,dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{nu-\frac{u^2}{2}}\sin(2\pi u)\,du \ (y = u - n) \tag{35}$$
$$= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}y^2+\frac{n^2}{2}}\sin(2\pi y)\,dy \ (y = u - n) = 0 \tag{36}$$

Consider **factorial moments**

$$m_{[k]} = \mathbb{E}X(X-1)...(X-k+1) \tag{37}$$

where $m_{[1]} = \mathbb{E}X, m_{[2]} = \mathbb{E}X^2 - \mathbb{E}X, ...$

If $X \sim P(\lambda)$ (the factorial moments are consistent with the factorial in Poisson)

$$m_{[k]} = \sum_{j=0}^\infty \frac{\lambda^j}{j!}e^{-\lambda}j(j-1)...(j-k+1) \tag{38}$$
$$= \sum_{j=k}^\infty \frac{\lambda^j}{(j-k)!}e^{-\lambda} \ (l = j-k) = \lambda^k\sum_{l=0}^\infty \frac{\lambda^l}{l!}e^{-\lambda} = \lambda^k \tag{39}$$

Note that there's a theorem for Poisson r.v. saying that if the factorial moments are known the same as Poisson

then the r.v. have to be a Poisson r.v.

Actually, **if the Maclaurin series of moment generating function is absolutely convergent, then moments can uniquely determine distribution**. Which is to say, if $\exists C, \forall n, \mathbb{E}|X|^n \leq C^n$

$$\sum_{n=0}^{\infty} \frac{|\mathbb{E}X^n|}{n!} t^n \leq e^{Ct} < \infty \tag{40}$$

so if the moments are uniformly bounded by the exponential of a constant, then moments can uniquely determine the distribution.

Cumulant Generating Function

Defined as

$$K_X(t) = \log M_X(t) = \log \mathbb{E}e^{tX} \tag{41}$$

take derivative to see

$$K'_X(t) = \frac{\mathbb{E}Xe^{tX}}{\mathbb{E}e^{tX}} \tag{42}$$

$$K'_X(0) = \mathbb{E}X \tag{43}$$

$$K''_X(t) = \frac{\mathbb{E}X^2e^{tX} \cdot \mathbb{E}e^{tX} - (\mathbb{E}Xe^{tX})^2}{(\mathbb{E}e^{tX})^2} \tag{44}$$

$$K''_X(0) = Var(X) \tag{45}$$

Multi-dimensional Distributions and Order Statistics

Multinomial Distribution

Models the n times independent and identical repetition of an experiment that has k possible outcomes. For each time of experiment, has k possible outcomes following the probability $p = (p_1, ..., p_k)$ with $p_1 + ... + p_k = 1$.

$X = (X_1, ..., X_k)$ is a random vector with X_i denoting the number of outcomes that gives outcome i within the n total experiments. Then such $X \sim multi(n, p)$ with probability mass

$$\mathbb{P}(X = x) = \binom{n}{x_1, ..., x_k} p_1^{x_1} ... p_k^{x_k}, \ (x_1 + ... + x_k = n) \tag{46}$$

can consider the MGF (use multinomial thm)

$$\begin{aligned} M_{X_1,...,X_{k-1}}(t_1, ..., t_{k-1}) &= \mathbb{E}e^{t_1 X_1 + ... + t_{k-1} X_{k-1}} && (47)\\ &= \left(p_1 e^{t_1} + ... + p_{k-1} e^{t_{k-1}} + p_k\right)^n && (48) \end{aligned}$$

set $t_2 = 0, ..., t_{k-1} = 0$ to get

$$\begin{aligned} M_{X_1}(t) &= M_{X_1,...,X_{k-1}}(t, 0, ..., 0) && (49)\\ &= \left(p_1 e^t + p_2 + ... + p_k\right)^n && (50) \end{aligned}$$

this is telling us that the marginal of multinomial is still multinomial and $X_1 \sim B(n, p_1)$ (since only two possible outcomes, becomes binomial). Similarly,

$$\begin{aligned} M_{X_1,X_2}(t_1, t_2) &= M_{X_1,...,X_{k-1}}(t_1, t_2, 0, ..., 0) && (51)\\ &= \left(p_1 e^{t_1} + p_2 e^{t_2} + p_3 + ... + p_k\right)^n && (52) \end{aligned}$$

so $(X_1, X_2) \sim multi(n, [p_1, p_2, 1 - p_1 - p_2])$ is a natural result.

To get the covariance between any two components in a multinomial distributed random vector, one can calculate $cov(X_1, X_2)$ as follows

$$\begin{aligned} Var(X_1 + X_2) &= n(p_1 + p_2)(1 - p_1 - p_2) && (53)\\ Var(X_1) &= np_1(1 - p_1) && (54)\\ Var(X_2) &= np_2(1 - p_2) && (55) \end{aligned}$$

since $X_1 + X_2 \sim B(n, p_1 + p_2)$ by MGF. Consider the decomposition of variance that

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2) + 2cov(X_1, X_2) \tag{56}$$

$$cov(X_1, X_2) = \frac{1}{2}(n(p_1 + p_2)(1 - p_1 - p_2) - np_1(1 - p_1) - np_2(1 - p_2)) \tag{57}$$

$$= -np_1p_2 \tag{58}$$

note that $cov(X_1, X_2) < 0$ is natural since if X_1 increases, that will account for the space of X_2 so X_2 will decrease in general.

Another way to see the same result is to set $A_{i,1}$ as the event that the i-th experiment gives outcome 1 and $A_{i,2}$ as the event that the i-th experiment gives outcome 2, then

$$X_1 = \sum_{k=1}^{n} \mathbb{I}_{A_{k,1}} \tag{59}$$

$$X_2 = \sum_{k=1}^{n} \mathbb{I}_{A_{k,2}} \tag{60}$$

$$cov(X_1, X_2) = \sum_{i,j=1}^{n} cov(\mathbb{I}_{A_{i,1}}, \mathbb{I}_{A_{j,2}}) \tag{61}$$

$$= \sum_{i=1}^{n} cov(\mathbb{I}_{A_{i,1}}, \mathbb{I}_{A_{i,2}}) + 2\sum_{i<j} cov(\mathbb{I}_{A_{i,1}}, \mathbb{I}_{A_{j,2}}) \tag{62}$$

$$= \sum_{i=1}^{n} -p_1p_2 \tag{63}$$

$$= -np_1p_2 \tag{64}$$

since $A_{i,1}, A_{j,2}$ are independent for $i \neq j$.

Multivariate Gaussian Distribution

For n-dimensional Gaussian random vector X with mean vector $\mu \in \mathbb{R}^n$ and covariance matrix $\Sigma \in \mathbb{R}^{n\times n}$, the MGF is

$$M_X(t) = e^{\mu^T t + \frac{1}{2}t^T \Sigma t} \ (t \in \mathbb{R}^n) \tag{65}$$

the transformations of multivariate Gaussian can be derived based on MGF. For example,

$$M_{X_1}(t) = M_X(t, 0, ..., 0) = e^{\mu_1 t + \frac{1}{2}\sigma_1^2 t^2} \tag{66}$$

and the two dimensional marginal has

$$M_{X_1,X_2}(t_1,t_2) = M_X(t_1,t_2,0,...,0) = e^{\mu_1 t_1+\mu_2 t_2+\frac{1}{2}t^{(2)^T}\Sigma^{(2)}t^{(2)}} \tag{67}$$

where the column vector $t^{(2)} = [t_1, t_2]$ and $\Sigma^{(2)} = \begin{bmatrix} \sigma_1^2 & cov(X_1,X_2) \\ cov(X_1,X_2) & \sigma_2^2 \end{bmatrix}$. This is showing that the marginal of multi-dimensional Gaussian is still Gaussian.

Similarly, any linear transformation of multi-dim Gaussian random vector is still multi-dim Gaussian, for example, for $A \in \mathbb{R}^{n\times n}, b \in \mathbb{R}^n$, let's consider $AX + b$:

$$\begin{aligned} M_{AX+b}(t) &= \mathbb{E}e^{t^T AX+t^T b} && (68)\\ &= e^{t^T b}\cdot M_X(A^T t) && (69)\\ &= e^{t^T b}\cdot e^{\mu^T A^T t+\frac{1}{2}t^T A\Sigma A^T t} && (70)\\ &= e^{(A\mu+b)^T t+\frac{1}{2}t^T A\Sigma A^T t} && (71) \end{aligned}$$

so $AX + b \sim N(A\mu + b, A\Sigma A^T)$.

Transformation of Random Variables

When there are random vectors $X, Y \in \mathbb{R}^n$ and $Y = f(X), f : \mathbb{R}^n \to \mathbb{R}^n$, and we know the density of X, to get the density of Y, just need to realize that density is actually a Radon-Nikodym derivative and thus

$$f_Y(y_1,...,y_n)\,dy_1...\,dy_n \sim f_X(x_1,...,x_n)\,dx_1...\,dx_n \tag{72}$$

so the approach is saying that

$$f_Y(y_1,...,y_n) = f_X(x_1(y),...,x_n(y))\cdot\left|\frac{\partial(x_1,...,x_n)}{\partial(y_1,...,y_n)}\right| \tag{73}$$

this works when the determinant of the Jacobian is not zero, showing us that the transformation of random variable is actually just a change of variables in the density.

Order Statistics

$(X_{(1)},...,X_{(n)})$ is the order statistics of $(X_1,...,X_n)$, where $X_{(i)}$ is the i-th smallest value among $X_1,...,X_n$ *i.i.d.* random variables.

In discrete case, directly apply the definition of order statistic to compute its distribution. If each X_i has probability mass $p_k = \mathbb{P}(X_i = k)$, then

$$\mathbb{P}\left(X_{(r)} \le x\right) = \sum_{i=0}^{n-r}\binom{n}{i}[\mathbb{P}(X_1 > x)]^i[\mathbb{P}(X_1 \le x)]^{n-i} = \sum_{i=r}^{n}\binom{n}{i}[\mathbb{P}(X_1 \le x)]^i[\mathbb{P}(X_1 > x)]^{n-i} \tag{74}$$

so

$$\mathbb{P}\left(X_{(r)}=x\right)=\mathbb{P}\left(X_{(r)}\leq x\right)-\mathbb{P}\left(X_{(r)}<x\right) \tag{75}$$

For continuous r.v. with density, the joint density of order statistics is

$$f(x_{(1)},...,x_{(n)})=n!\cdot f(x_{(1)})...f(x_{(n)})\ (x_{(1)}<...<x_{(n)}) \tag{76}$$

as a result, the marginal distribution of $X_{(r)}$ can be figured out using integration

$$f_{X_{(r)}}(x_r)=n!f(x_r)\cdot\int_{x_1<...<x_{r-1}<x_r<x_{r+1}<...<x_n} f(x_1)...f(x_{r-1})f(x_{r+1})...f(x_n)\,dx_1...\,dx_{r-1}\,dx_{r+1}...\,dx_n \tag{77}$$

$$=n!f(x_r)\cdot\frac{[1-F(x_r)]^{n-r}}{(n-r)!}\frac{[F(x_r)]^{r-1}}{(r-1)!} \tag{78}$$

where the calculations are followed by the fact that

$$\int_{x_1<...<x_r} f(x_1)...f(x_{r-1})\,dx_1...\,dx_{r-1}=\int_{-\infty}^{x_r} f(x_1)\,dx_1\int_{x_1}^{x_r} f(x_2)\,dx_2...\int_{x_{r-2}}^{x_r} f(x_{r-1})\,dx_{r-1} \tag{79}$$

$$=\int_{-\infty}^{x_r} f(x_1)\,dx_1\int_{x_1}^{x_r} f(x_2)\,dx_2...\int_{x_{r-3}}^{x_r} f(x_{r-2})[F(x_r)-F(x_{r-2})]\,dx_{r-2} \tag{80}$$

$$=\cdots \tag{81}$$

$$=\frac{[F(x_r)]^{r-1}}{(r-1)!} \tag{82}$$

Remark. *A good way to understand the marginal density of order statistics is to use the* ***binomial*** *distribution and the likelihood interpretation.*

$f_{X_{(r)}}(x_r)$ is the likelihood of $X_{(r)}$ taking value x_r, this means that there should be $r-1$ values less than x_r and $n-r$ values larger than x_r. View "less than x_r" and "larger than x_r" as two different bins, then we are actually throwing each i.i.d. observation into one of the two bins. So the likelihood should be

$$nf(x_r)\cdot\binom{n-1}{r-1}[F(x_r)]^{r-1}[1-F(x_r)]^{n-r} \tag{83}$$

the explanation is that we ***select 1 r.v. from the*** n ***r.v. to be*** $X_{(r)}$ ***taking value*** x_r ***with likelihood*** $f(x_r)$***, from the remaining*** $n-1$ ***r.v. select*** $r-1$ ***to be less than*** x_r***, each with likelihood*** $F(x_r)$***; the remaining*** $n-r$ ***r.v. to be larger than*** x_r***, each with likelihood*** $1-F(x_r)$*.*

Use the same reasoning, it will be easy to write out the joint density of $(X_{(j)},X_{(k)})$ $(j<k)$ using the multinomial

interpretation. The likelihood that $X_{(j)} = x_j, X_{(k)} = x_k$ is equal to

$$n(n-1)f(x_j)f(x_k) \cdot \binom{n-2}{j-1, k-j-1} [F(x_j)]^{j-1} [F(x_k) - F(x_j)]^{k-j-1} [1 - F(x_k)]^{n-k} \tag{84}$$

with support $x_j < x_k$. First select 2 out of n to be $X_{(j)}, X_{(k)}$ (it's a permutation since order matters) taking x_j, x_k respectively, from the remaining $n-2$ r.v., split it into 3 bins "less than x_j", "between x_j and x_k", and "larger than x_k". After multiplying the multinomial coefficient to put $j-1$ in the first bin, $k-j-1$ in the second bin, the likelihood of being in the first bin is $F(x_j)$, the likelihood of being in the second bin is $F(x_k) - F(x_j)$ and the likelihood of being in the third bin is $1 - F(x_k)$.

Tolerance Limits

Consider $\mathbb{P}\left(X_{(1)} < X < X_{(n)}\right)$, the probability of having the observation between the maximum and the minimum of the *i.i.d.* samples of size n. One might notice that

$$\mathbb{P}\left(X_{(1)} < X < X_{(n)}\right) = \mathbb{P}\left(X < X_{(n)}\right) - \mathbb{P}\left(X < X_{(1)}\right) = F(X_{(n)}) - F(X_{(1)}) = U_{(n)} - U_{(1)} \tag{85}$$

where $U_{(i)}$ is the order statistics of the *i.i.d.* $U(0,1)$ r.v. since $F(X) \sim U(0,1)$.

Note that the joint density of $(U_{(1)}, U_{(n)})$ is

$$g(u_1, u_n) = n(n-1)f(u_n)f(u_1) \cdot [F(u_n) - F(u_1)]^{n-2} \tag{86}$$

$$= n(n-1) \cdot (u_n - u_1)^{n-2} \ (0 < u_1 < u_n < 1) \tag{87}$$

consider the transformation

$$\begin{cases} R = U_{(n)} - U_{(1)} \\ S = U_{(1)} \end{cases} \tag{88}$$

and apply the Jacobian to find

$$\left| \frac{\partial(u_1, u_n)}{\partial(r, s)} \right| = 1 \tag{89}$$

$$f_{(R,S)}(r, s) = g(s, s+r) \cdot \left| \frac{\partial(u_1, u_n)}{\partial(r, s)} \right| \tag{90}$$

$$= n(n-1)r^{n-2} \ (0 < s < 1, 0 < r < 1 - s) \tag{91}$$

so the marginal is

$$f_R(r) = \int_0^{1-r} n(n-1)r^{n-2}\,ds = n(n-1)r^{n-2}(1-r)\ r \in (0,1) \tag{92}$$

$$f_S(s) = \int_0^{1-s} n(n-1)r^{n-2}\,dr = n(1-s)^{n-1}\ s \in (0,1) \tag{93}$$

so $R \sim Beta(n-1,2)$ is the difference of the maximum and the minimum. We can find sample size n such that $\mathbb{P}(R \geq 1-\beta) = 1-\alpha$ to ensure that under confidence level $1-\alpha$ the difference is large enough to cover a large percentage of the distribution.

Sampling Distribution and Hypothesis Testing

Sampling Distribution

For $X_1, ..., X_n \sim N(\mu, \sigma^2)$ *i.i.d.*, define

$$\overline{X} = \frac{\sum_{i=1}^n X_i}{n} \tag{94}$$

$$S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1} \tag{95}$$

as sample mean and sample variance, then obvious that $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$, for the distribution of S^2 and the relationship between sample mean and variance, we have

Theorem 4.

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{96}$$

and $\overline{X}$ *and* S^2 *are independent.*

Proof. Take $Z \sim N(0, I_n)$ as standard Gaussian random vector of length n. Then the joint density is

$$f_Z(z) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2} z^T z} \tag{97}$$

and the MGF is

$$M_Z(t) = e^{-\frac{1}{2} t^T t} \tag{98}$$

consider applying orthogonal matrix $A_{n\times n}$ as an action on Z, then it's obvious that

$$M_{AZ}(t) = \mathbb{E} e^{t^T AZ} = M_Z(A^T t) = e^{-\frac{1}{2} t^T A A^T t} = e^{-\frac{1}{2} t^T t} \tag{99}$$

so $AZ \sim N(0, I_n)$ is still standard Gaussian random vector.

Now we assume WLOG that $X \sim N(0, I_n)$ and we specify an A that we want, we pick A with the first row as $\left[\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}}\right]$ and all other entries are arbitrarily picked, such matrix always exists. Notice that the first component of $Y = AX$ is just $Y_1 = \sqrt{n} \cdot \overline{X}$ and

$$n\overline{X}^2 + Y_2^2 ... + Y_n^2 = \sum_{i=1}^n X_i^2 \tag{100}$$

since orthogonal transformation preserves the length of the vector. It's obvious that according to the variance

decomposition,

$$\frac{\sum_{i=2}^{n} Y_i^2}{n} = \frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2 = \frac{n-1}{n} S^2 \tag{101}$$

it's easy to see that

$$(n-1)S^2 = \sum_{i=2}^{n} Y_i^2 \sim \chi_{n-1}^2 \tag{102}$$

since $Y_1, ..., Y_n$ *i.i.d.* and it's obvious that $\overline{X}$ is independent of S^2.

For general $X \sim N(\mu\vec{1}, \sigma^2 I_n)$, just consider $\frac{X-\mu\vec{1}}{\sigma} \sim N(0, I_n)$ and the proof is immediate.

□

When $X \sim N(0,1)$ and $Y \sim \chi_k^2$ are independent, $T = \frac{X}{\sqrt{\frac{Y}{k}}}$ is defined to have **t-distribution** with degree of freedom k. To calculate the density of t-distribution, consider the transformation

$$(X, Y) \to (T, U) \tag{103}$$

with $U = Y$ to apply the Jacobian and then get the marginal distribution of T. (add U to make life easier)

Now that

$$f_{T,U}(t,u) = f_{X,Y}(x(t,u), y(t,u)) \cdot \left| \frac{\partial(x,y)}{\partial(t,u)} \right| \tag{104}$$

with

$$\begin{cases} x = t\sqrt{\frac{u}{k}} \\ y = u \end{cases} \tag{105}$$

so

$$\left| \frac{\partial(x,y)}{\partial(t,u)} \right| = \left| \det \begin{bmatrix} \sqrt{\frac{u}{k}} & t\frac{1}{2\sqrt{ku}} \\ 0 & 1 \end{bmatrix} \right| = \sqrt{\frac{u}{k}} \tag{106}$$

and plug in to get

$$f_{T,U}(t,u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2 u}{2k}} \cdot \frac{\left(\frac{1}{2}\right)^{\frac{k}{2}}}{\Gamma(\frac{k}{2})} u^{\frac{k}{2}-1} e^{-\frac{u}{2}} \cdot \sqrt{\frac{u}{k}} \tag{107}$$

$$\propto e^{-(\frac{1}{2}+\frac{t^2}{2k})u} u^{\frac{k-1}{2}} \ (u > 0, t \in \mathbb{R}) \tag{108}$$

to integrate u, we get the marginal of T that

$$f_T(t) \propto \int_0^\infty e^{-(\frac{1}{2}+\frac{t^2}{2k})u} u^{\frac{k-1}{2}}\, du \tag{109}$$

$$= \frac{\Gamma(\frac{k+1}{2})}{(\frac{1}{2}+\frac{t^2}{2k})^{\frac{k+1}{2}}} \tag{110}$$

$$\propto \left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}} \quad (t \in \mathbb{R}) \tag{111}$$

Remark. *In particular,* ***when*** $k = 1$*,*

$$f_T(t) \propto \frac{1}{1+t^2} \tag{112}$$

$$f_T(t) = \frac{1}{\pi}\frac{1}{1+t^2} \tag{113}$$

a ***Cauchy distribution*** *with heavy tail! One may see from the definition that this is telling us that for* $Z_1, Z_2 \sim N(0,1)$ *independent,*

$$\frac{Z_1}{Z_2} \sim Cauchy \tag{114}$$

When $k \to \infty$*, notice that*

$$f_T(t) \propto e^{-\frac{1}{2}t^2} \tag{115}$$

so this is just ***standard Gaussian****!*

When $X \sim \chi^2_m, Y \sim \chi^2_n$ are independent, define $F = \frac{\frac{X}{m}}{\frac{Y}{n}}$ to follow the **F-distribution** with degree of freedom m, n. Consider the transformation

$$(X, Y) \to (F, U) \tag{116}$$

with $U = Y$ to apply the Jacobian and then get the marginal distribution of F. (add U to make life easier)

Now that

$$f_{F,U}(f,u) = f_{X,Y}(x(f,u), y(f,u)) \cdot \left|\frac{\partial(x,y)}{\partial(f,u)}\right| \tag{117}$$

with

$$\begin{cases} x = \frac{muf}{n} \\ y = u \end{cases} \tag{118}$$

so

$$\left|\frac{\partial(x,y)}{\partial(f,u)}\right| = \left|\det\begin{bmatrix}\frac{mu}{n} & \frac{mf}{n}\\ 0 & 1\end{bmatrix}\right| = \frac{mu}{n} \tag{119}$$

and plug in to get

$$f_{F,U}(f,u) = \frac{\left(\frac{1}{2}\right)^{\frac{m}{2}}}{\Gamma(\frac{m}{2})}\left(\frac{muf}{n}\right)^{\frac{m}{2}-1} e^{-\left(\frac{muf}{2n}\right)} \cdot \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma(\frac{n}{2})} u^{\frac{n}{2}-1} e^{-\frac{u}{2}} \cdot \frac{mu}{n} \tag{120}$$

$$\propto u^{\frac{m+n}{2}-1} f^{\frac{m}{2}-1} e^{-\frac{u}{2}-\frac{muf}{2n}} \ (f>0, u>0) \tag{121}$$

to integrate u, we get the marginal of F that

$$f_F(f) \propto \int_0^\infty u^{\frac{m+n}{2}-1} f^{\frac{m}{2}-1} e^{-\frac{u}{2}-\frac{muf}{2n}}\, du \tag{122}$$

$$\propto f^{\frac{m}{2}-1}\left(1+\frac{m}{n}f\right)^{-\frac{m+n}{2}} \ (f>0) \tag{123}$$

Remark. *Notice that if $T \sim t(n)$, then it's actually a quotient $T \overset{d}{=} \frac{X}{\sqrt{\frac{Y}{n}}}$ where $X \sim N(0,1), Y \sim \chi^2_n$ are independent. As a result, $T^2 \overset{d}{=} \frac{X^2}{\frac{Y^2}{n}} \sim F(1,n)$. So* ***the square of t-distribution with d.f. n is the F-distribution with d.f.*** *$1, n$.*

Notice the ***symmetricity*** *of F-distribution that for $F \sim F(m,n)$, $\frac{1}{F} \sim F(n,m)$. This is easy to see from the definition. In particular, when $F \sim F(n,n), F \overset{d}{=} \frac{1}{F}$. One can use this property to figure out the left tail probability of F-distribution using only the right tail probability of F-distribution (percentiles) by interchanging two d.f.*

If $F \sim F(m,n)$, then $\beta = \frac{1}{1+\frac{m}{n}F} \sim Beta\left(\frac{n}{2}, \frac{m}{2}\right)$. To see this fact,

$$f_\beta(b) \propto \left(\frac{1}{b}-1\right)^{\frac{m}{2}-1} b^{\frac{m+n}{2}} b^{-2} = (1-b)^{\frac{m}{2}-1} b^{\frac{n}{2}-1} \tag{124}$$

Application of Sampling Distribution

Now let's assume that the population follows Gaussian $N(\mu, \sigma^2)$ and we can observe *i.i.d.* sample values $X_1, X_2, ..., X_n$. It's easy to see that $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ **the (population) normalized sample mean follows standard Gaussian**. However, this requires us to know σ is we want to build the confidence interval of μ, which is always infeasible.

By replacing the population variance with the sample variance, one would see that

$$\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{\frac{(n-1)S^2}{\sigma^2}}{n-1}}} \tag{125}$$

and notice that

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1} \tag{126}$$

S is independent of $\overline{X}$, one immediately conclude that $\frac{\overline{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$. So **the (sample) normalized sample mean follows t-distribution with d.f.** $n-1$.

Consider two independent populations $X_1, ..., X_m \sim N(\mu_x, \sigma_x^2), Y_1, ..., Y_n \sim N(\mu_y, \sigma_y^2)$ with sample means $\overline{X}, \overline{Y}$ and sample variances S_x^2, S_y^2 for each population. One can find that $\overline{X}, S_x^2, \overline{Y}, S_y^2$ are independent. That's why we consider

$$\frac{\frac{S_x^2}{S_y^2}}{\frac{\sigma_x^2}{\sigma_y^2}} = \frac{\frac{(m-1)\frac{S_x^2}{\sigma_x^2}}{m-1}}{\frac{(n-1)\frac{S_y^2}{\sigma_y^2}}{n-1}} \sim F(m-1, n-1) \tag{127}$$

to find that **the quotient of sample variance over the quotient of population variance follows F-distribution with d.f.** $m-1, n-1$.

Another observation follows that

$$\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{m} + \frac{\sigma_y^2}{n}\right) \tag{128}$$

$$(m-1)\frac{S_x^2}{\sigma_x^2} + (n-1)\frac{S_y^2}{\sigma_y^2} \sim \chi^2_{m+n-2} \tag{129}$$

by independency, and the two parts are still independent. As a result,

$$\frac{\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{m}+\frac{\sigma_y^2}{n}}}}{\sqrt{\frac{(m-1)\frac{S_x^2}{\sigma_x^2}+(n-1)\frac{S_y^2}{\sigma_y^2}}{m+n-2}}} \sim t(m+n-2) \tag{130}$$

which is **the t-test on the difference of population mean for two independent populations**. However, in this formula one still has to know about the true population variance σ_x, σ_y so it's not that useful. If one plugs in $\sigma_x^2 = \sigma_y^2 = C^2$ **(assume equal population variance)**, one would find that

$$\frac{(\overline{X}-\overline{Y})-(\mu_x-\mu_y)}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}} \sim t(m+n-2) \tag{131}$$

$$S_p^2 = \frac{\sum_{i=1}^m (X_i-\overline{X})^2 + \sum_{j=1}^n (Y_i-\overline{Y})^2}{m+n-2} \tag{132}$$

where S_p is called **the pooled variance**. This test does not require us to know about population variance but requires one to test whether the variance of two populations are the same. However, we have already stated the

equal variance test conducted by taking the quotient and applying F-test above.

Delta Method

Theorem 5. ***(Delta Method)*** *Let* $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \sigma^2)$ $(n \to \infty)$ *and* $g \in C^1$ *in a neighborhood of* θ*, then*

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2) \ (n \to \infty) \tag{133}$$

Delta method tells us that the C^1 transformation of asymptotically Gaussian random variable series is still asymptotically Gaussian. The proof is a simple application of Taylor expansion.

An example for Delta method is that $X_1, X_2, ... \overset{i.i.d.}{\sim} B(1, p)$ and an estimate for p is formed as $\hat{p}_n = \overline{X}_n$ so by CLT

$$\sqrt{n}(\hat{p}_n - p) \xrightarrow{d} N(0, p(1-p)) \ (n \to \infty) \tag{134}$$

as a result, by Delta method,

$$\sqrt{n}(g(\hat{p}_n) - g(p)) \xrightarrow{d} N(0, [g'(p)]^2 p(1-p)) \ (n \to \infty) \tag{135}$$

so it's possible to find the g such that the asymptotic variance after transformation is constant, i.e.

$$[g'(p)]^2 p(1-p) = c \tag{136}$$

one can solve out to see that $g(p) = \arcsin\sqrt{p}$ suffices. This is called **variance stabilizing transformation**.

Consider *i.i.d.* sample $X_1, X_2, ...$ and ξ_p as the p-quantile for the distribution of X_1. Now $X_{(\lfloor np \rfloor + 1)}$ is the sample estimate of ξ_p (it's actually an order statistic) so when $p = \frac{1}{2}$ it's just the median of the sample. One can prove the following asymptotic estimate for sample quantile that

Theorem 6. ***(Asymptotic Normality of Quantile)*** *For i.i.d. sample* $X_1, X_2, ...,$

$$\sqrt{n}(X_{(\lfloor np \rfloor + 1)} - \xi_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right) \tag{137}$$

where f *is the density of* X_1*.*

For example, let's take $X_1 \sim N(\mu, \sigma^2)$ so

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} N(0, \sigma^2) \ (n \to \infty) \tag{138}$$

however, the theorem above tells us that

$$\sqrt{n}(M_n - \mu) \xrightarrow{d} N\left(0, \frac{\pi}{2}\sigma^2\right) \ (n \to \infty) \tag{139}$$

where M_n is the sample median given n observations. It's interesting to see that **for large Gaussian sample, sample median has larger asymptotic variance than sample mean**.

Statistical Inference

Settings

Now $X = (X_1, ..., X_n)$ denotes n *i.i.d.* samples, each with likelihood $p_\theta(x), \theta \in \Theta$ where θ is the parameter and Θ is the space of all admissible parameter values. Denote $x = (x_1, ..., x_n) \in \mathscr{X}$ as a real vector (realization of X taking values in $\mathscr{X}$). A **statistic** is $T = T(X_1, ..., X_n)$, a random variable with $T(x) = t \in \mathscr{T}$ taking values in $\mathscr{T}$. The **orbit** of T is defined as $A_t = \{x : T(x) = t\} = T^{-1}(\{t\})$ is the set of realizations such that the statistic takes value t.

For example, consider $X_1, ..., X_n \sim B(1, \theta)$ so $p_\theta(x) = \theta^x(1-\theta)^{1-x}$ $(x \in \{0, 1\})$ with $\mathscr{X} = \{x = (x_1, ..., x_n) : x_i \in \{0, 1\}\}$. Now consider the statistic

$$T = \sum_{i=1}^{n} X_i \sim B(n, \theta) \tag{140}$$

so $p_\theta(T = t) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$. The orbits are $A_0 = \{x \in \mathscr{X} : \sum_{i=1}^n X_i \sim B(n, \theta) = 0\} = \{(0, 0, ..., 0)\}$, $A_1 = \{(1, 0, ..., 0), ..., (0, 0, ..., 1)\}$ etc, so there are altogether $\binom{n}{0} + \binom{n}{1} + ... + \binom{n}{n} = 2^n$ orbits.

Sufficiency

A statistic T is **sufficient** for the parameter family $\mathscr{P} = \{p_\theta, \theta \in \Theta\}$ if $X|_T$ is independent of θ, i.e. if T is known, no more information of θ is required to know the distribution of the samples X, the information contained in T is sufficient to find out the distribution of X. In other words, given that the sample X belongs to a certain orbit A_t, the sample is not relevant to θ.

For example, consider $X_1, ..., X_n \sim B(1, \theta)$ and $T = \sum_{i=1}^n X_i \sim B(n, \theta)$. Now that

$$\mathbb{P}(X = x | T = t) = \frac{\mathbb{P}(X = x, T = t)}{\mathbb{P}(T = t)} \tag{141}$$

$$= \frac{\mathbb{P}(X = x, \sum_{i=1}^n X_i = t)}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} \tag{142}$$

$$= \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \frac{\theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} & \sum_{i=1}^n x_i = t \end{cases} \tag{143}$$

$$= \begin{cases} 0 & \sum_{i=1}^n x_i \neq t \\ \frac{1}{\binom{n}{t}} & \sum_{i=1}^n x_i = t \end{cases} \tag{144}$$

which has nothing to do with θ, so such T is a sufficient statistic.

Another example, $X_1, ..., X_n \sim N(\theta, 1)$, $T = \overline{X} \sim N\left(\theta, \frac{1}{n}\right)$ is sufficient. To see this, consider

$$p_\theta(X = x|T = t) = \frac{p(X = x, T = t)}{p(T = t)} \tag{145}$$

$$= \begin{cases} 0 & \overline{x} \neq t \\ \frac{\prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{(x_i-\theta)^2}{2}}}{\frac{1}{\sqrt{\frac{2\pi}{n}}} e^{-\frac{(t-\theta)^2}{\frac{2}{n}}}} & \overline{x} = t \end{cases} \tag{146}$$

$$= \begin{cases} 0 & \overline{x} \neq t \\ C(n) \cdot e^{\frac{n}{2}(t-\theta)^2 - \sum_{i=1}^n \frac{(x_i-\theta)^2}{2}} & \overline{x} = t \end{cases} \tag{147}$$

where $\frac{n}{2}(t-\theta)^2 - \sum_{i=1}^n \frac{(x_i-\theta)^2}{2} = \frac{n}{2}(t-\theta)^2 - \sum_{i=1}^n \frac{(x_i-t+t-\theta)^2}{2} = -\frac{1}{2}\sum_{i=1}^n (x_i - t)^2 + \sum_{i=1}^n (x_i - t)(t-\theta)$ and notice that here $\overline{x} = t$ so it's equal to $-\frac{1}{2}\sum_{i=1}^n (x_i - t)^2$, independent of θ.

Theorem 7. ***(Factorization Theorem)*** *The joint likelihood $p_\theta(x) = \prod_{i=1}^n p_\theta(x_i)$, then $T = T(x)$ is sufficient statistic if and only if*

$$p_\theta(x) = g_\theta(T(x)) \cdot h(x) \tag{148}$$

where g_θ is a function of x only through T and θ while h is a function only of x.

Proof. WLOG, assume discrete r.v. (for continuous r.v. we would require $T(x)$ to be continuous in x), if there is sufficiency

$$p_\theta(X = x) = p_\theta(X = x|T(X) = t) \cdot p_\theta(T(X) = t) \tag{149}$$

the first term is independent of θ so it's $h(x)$ and the second term is a function of $T(x), \theta$ so it's $g_\theta(T(x))$, factorization is true.

Conversely, if factorization holds, consider decomposition

$$p_\theta(T(X) = t) = \sum_{x:T(x)=t} p_\theta(X = x) \tag{150}$$

$$= \sum_{x:T(x)=t} g_\theta(T(x)) \cdot h(x) \tag{151}$$

$$= g_\theta(t) \cdot \sum_{x:T(x)=t} h(x) \tag{152}$$

so for x such that $T(x) = t$ (nontrivial case), $p_\theta(X = x|T(X) = t) = \frac{p_\theta(X=x, T(X)=t)}{p_\theta(T(X)=t)} = \frac{g_\theta(t)h(x)}{g_\theta(t)\sum_{x:T(x)=t} h(x)} = \frac{h(x)}{\sum_{x:T(x)=t} h(x)}$ independent of θ. □

Remark. *The key point in the proof is to realize that $p_\theta(T(X) = t) = \sum_{x:T(x)=t} p_\theta(X = x)$, so the orbits of the*

statistic induces ***a partition of the sample space****, connecting the likelihood of the statistic and the likelihood of the sample.*

The factorization theorem enables us to judge sufficiency easily. For example, for the first Bernoulli example above,

$$p_\theta(x) = \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} \tag{153}$$

where $g_\theta(T(x)) = \theta^{T(x)}(1-\theta)^{n-T(x)}, h(x) = 1$, so T is sufficient.

For the second Gaussian example above,

$$\begin{aligned} p_\theta(x) &= (2\pi)^{-\frac{n}{2}} e^{-\sum_{i=1}^n \frac{(x_i-\theta)^2}{2}} &(154)\\ &= (2\pi)^{-\frac{n}{2}} e^{-\sum_{i=1}^n \frac{x_i^2-2x_i\theta+\theta^2}{2}} &(155)\\ &= (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} \cdot e^{\theta\sum_{i=1}^n x_i - \frac{n}{2}\theta^2} &(156) \end{aligned}$$

where $g_\theta(T(x)) = e^{n\theta T(x) - \frac{n}{2}\theta^2}, h(x) = (2\pi)^{-\frac{n}{2}} e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}$, so T is sufficient.

Let's look at more interesting examples. Consider $X_1, ..., X_n \sim U(0,\theta)$ *i.i.d.*, so the joint likelihood is $p_\theta(x) = \mathbb{I}_{x_1\in(0,\theta)}\cdot ... \cdot \mathbb{I}_{x_n\in(0,\theta)} = \mathbb{I}_{\min\{x_1,...,x_n\}>0, \max\{x_1,...,x_n\}<\theta}$ so by setting $h(x) = \mathbb{I}_{\min\{x_1,...,x_n\}>0}, g_\theta(T(x)) = \mathbb{I}_{\max\{x_1,...,x_n\}<\theta}$, we see that $T(X) = X_{(n)}$ is a sufficient statistic.

Consider $X_1, ..., X_n \sim N(\mu, \sigma^2)$ *i.i.d.* with both parameters μ, σ^2 unknown, so the joint likelihood is $p_\theta(x) = (2\pi)^{-\frac{n}{2}}\sigma^{-n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2}$ and let's look at the exponential term

$$e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2} = e^{-\frac{n\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^n x_i^2 - 2\mu\sum_{i=1}^n x_i)} \tag{157}$$

so it's quite obvious that the sufficient statistics will be formed as $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ by the factorization theorem. Note that $T(X) = (\overline{X}, S^2)$ is also a sufficient statistic since $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a function of $(\overline{X}, S^2)$.

Remark. *Sufficient statistic is* ***not unique****! Actually* ***any one-to-one function of the sufficient statistic is still sufficient****. By factorization theorem, $T' = \psi(T)$ for bijection ψ, so $p_\theta(x) = g_\theta(T(x))\cdot h(x) = g_\theta \circ \psi^{-1}(T'(x))\cdot h(x)$ so T' is still sufficient.*

Actually from the proof we can see that for any function ψ and sufficient statistic T such that $T = \psi(T')$, T' is always sufficient. So ***if a statistic after the action of a function becomes a sufficient statistic, it must also be sufficient****. (The action of the function reduces the information contained in the statistic, so if after the information is reduced we still have sufficient information, the original statistic must also contain sufficient information) This is telling us some trivial conclusions like if $T = \sum_{i=1}^n X_i$ is sufficient statistics, then $T' = (\sum_{i=1}^m X_i, \sum_{i=m+1}^n X_i), T'' = (X_1, ..., X_n)$ are also sufficient.*

Minimal Sufficient Statistic

A sufficient statistic is a **minimal sufficient statistic** if it is a function of all sufficient statistics, i.e. it contains the minimal amount of information required to get rid of the dependence on the parameter.

Remark. *There is an information theory perspective on sufficiency. Recall that mutual information is defined as*

$$I(X;Y) = D_{KL}(\mathbb{P}_{X,Y}||\mathbb{P}_X \times \mathbb{P}_Y) \tag{158}$$

for r.v. X, Y to measure the 'distance' between the joint probability measure and the product measure (as if they are independent r.v.). Data processing inequality tells us that if $X \to Y \to Z$ is Markov chain, then

$$I(X;Y) \geq I(X;Z) \tag{159}$$

with the equality to be true iff $I(X;Y|Z) = 0$.

Now consider the Markov chain $\theta \to X \to T(X)$, so $I(\theta;X) \geq I(\theta;T(X))$ meaning that there can only be decrease in mutual information if we replace the sample with the statistic T. Then sufficiency of T is actually defined as T such that

$$I(\theta;T(X)) = I(\theta,X) \tag{160}$$

which is equivalent to saying $I(\theta;X|T(X)) = 0$. In other words, given $T(X)$, θ is independent of X (the meaning of mutual information to be 0).

Remark. *The 'minimal' in minimal sufficient statistic refers to the **minimal amount of information contained**. By data processing inequality, we know that applying any functions on sample cannot increase the amount of information. Similarly, a function of sufficient statistic cannot contain more information than a sufficient statistic, i.e. with equal or less amount of information. Now minimal sufficient statistic is a function of all sufficient statistic, so it contains the 'minimal' amount of information among all sufficient statistic but itself is still sufficient. From this point of view, minimal sufficient statistic can be viewed as the solution to a min-max variational problem.*

Theorem 8. ***(Minimal Sufficient Statistics)*** *Consider statistic T such that $\forall x, y \in A_t$ in the same orbit and*

$$T(x) = T(y) \iff \frac{p_\theta(x)}{p_\theta(y)} \text{ independent of } \theta \tag{161}$$

then T is minimal sufficient statistic for θ.

Proof. WLOG assume $\forall x \in \mathscr{X}, \forall \theta, p_\theta(x) > 0$ so there's no concern about the likelihood ratio blowing up. First, let's show that T is sufficient.

For each non-empty orbit A_t, $\exists x_t \in A_t$ fixed (for each t, x_t is always a fixed representative in A_t). So $\forall x \in \mathscr{X}, x_{T(x)}$ is in the same orbit as x because if $x \in A_t$ then $x_{T(x)} = x_t \in A_t$. Now since $x, x_{T(x)}$ are always in the same orbit, we know that $T(x) = T(x_{T(x)})$ so $\frac{p_\theta(x)}{p_\theta(x_{T(x)})} = h(x)$ is independent of θ. Note that for the function

$p_\theta(x_{T(x)})$, let's assume that $T(x) = t$, then $x_{T(x)} = x_t$ so $p_\theta(x_{T(x)}) = p_\theta(x_t)$. By defining $g_\theta(t) = p_\theta(x_t)$, we see that $p_\theta(x_{T(x)}) = g_\theta(T(x))$.

Now $p_\theta(x) = h(x) \cdot g_\theta(T(x))$ so by Factorization theorem, T is sufficient.

Next, we prove that T is minimal. Let T' be any other sufficient statistic, so by factorization theorem, $\exists g'_\theta, h'$ such that $p_\theta(x) = g'_\theta(T'(x)) \cdot h'(x)$. Pick $\forall x, y \in \mathscr{X}$ such that $T'(x) = T'(y)$, so

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{g'_\theta(T'(x)) \cdot h'(x)}{g'_\theta(T'(y)) \cdot h'(y)} = \frac{h'(x)}{h'(y)} \tag{162}$$

is independent of θ, so $T(x) = T(y)$. Now we have proved that $\forall x, y \in \mathscr{X}$, if $T'(x) = T'(y)$, then $T(x) = T(y)$. So T must be a function of T'. □

The example is for $X_1, ..., X_n \sim B(1, \theta)$, calculate

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i}}{\theta^{\sum_i y_i}(1-\theta)^{n-\sum_i y_i}} = \theta^{\sum_i x_i - \sum_i y_i}(1-\theta)^{-\sum_i x_i + \sum_i y_i} \tag{163}$$

to find it's independent of θ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i$, so $T(X) = \sum_{i=1}^n X_i$ is a minimal sufficient statistic.

Another example is for $X_1, ..., X_n \sim N(\mu, \sigma^2)$ (both parameters unknown), calculate

$$\frac{p_{\mu,\sigma}(x)}{p_{\mu,\sigma}(y)} = e^{-\frac{1}{2\sigma^2}[\sum_i (y_i^2 - x_i^2) + 2\mu \sum_i (x_i - y_i)]} \tag{164}$$

to find it's independent of θ iff $\sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \sum_{i=1}^n x_i^2 = \sum_{i=1}^n y_i^2$, so $T(X) = (\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is minimal sufficient statistic.

Another example is for $X_1, ..., X_n \sim U(0, \theta)$, calculate

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\mathbb{I}_{\min\{x_1,...,x_n\}>0, \max\{x_1,...,x_n\}<\theta}}{\mathbb{I}_{\min\{y_1,...,y_n\}>0, \max\{y_1,...,y_n\}<\theta}} \tag{165}$$

to find it's independent of θ iff $x_{(n)} = y_{(n)}$, so $T = X_{(n)}$ is sufficient statistic.

Remark. *Be careful with* ***the support of the distribution*** *since it may contain dependence on the parameter!*

Let's consider a slightly different example $X_1, ..., X_n \sim U(\theta, \theta+1)$ where two endpoints of the uniform distribution are both unknown. Calculate

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\mathbb{I}_{\min\{x_1,...,x_n\}>\theta, \max\{x_1,...,x_n\}<\theta+1}}{\mathbb{I}_{\min\{y_1,...,y_n\}>\theta, \max\{y_1,...,y_n\}<\theta+1}} \tag{166}$$

is independent of θ iff $x_{(1)} = y_{(1)}, x_{(n)} = y_{(n)}$. ***So*** *$T = (X_{(1)}, X_{(n)})$* ***is the minimal sufficient statistic while*** *$T = X_{(n)}$* ***is not sufficient any longer!***

Remark. *The reason why we care about sufficiency is due to the* ***sufficiency principle*** *that if T is the sufficient statistic and x, y are sample points such that $T(x) = T(y)$ (in the same orbit), then the inference of θ should be the*

same regardless of whether x or y is observed as the realization of the sample. This is because given the value of the sufficient statistic, the inference on parameter θ has nothing to do with the sample any longer.

Remark. *In order to argue that in the determination of minimal sufficient statistic, the likelihood ratio $\frac{p_\theta(x)}{p_\theta(y)}$ is independent of θ if and only if $T(x) = T(y)$, we have ways to make it rigorous. If $T(x) = T(y)$, then $\frac{p_\theta(x)}{p_\theta(y)}$ is independent of θ is always easy to prove. For the converse direction, we can start from noticing that if $\frac{p_\theta(x)}{p_\theta(y)}$ is independent of θ, then its partial derivative w.r.t. θ is constantly 0.*

More Examples on Delta Method and Sufficiency

Now there's $Y_1, ..., Y_n \sim \mathcal{E}(\theta), X_1, ..., X_n \sim \mathcal{E}(\delta\theta)$ to be independent observations, the estimator of δ is formed as $\hat{\delta} = \frac{\overline{Y}}{\overline{X}}$, the ratio of the sample mean of Y and X. We want to see if this estimator is asymptotically normal and get its asymptotic variance.

It's obvious that we shall consider $\hat{\delta} = g(\overline{X}, \overline{Y})$ where

$$g(x, y) = \frac{y}{x} \tag{167}$$

and by CLT, since $\mathbb{E}\overline{X} = \frac{1}{\delta\theta}, Var(\overline{X}) = \frac{1}{n\delta^2\theta^2}$ we know that by CLT

$$\begin{cases} \frac{\overline{X} - \frac{1}{\delta\theta}}{\frac{1}{\sqrt{n}\delta\theta}} \xrightarrow{d} N(0,1) \ (n \to \infty) \\ \frac{\overline{Y} - \frac{1}{\theta}}{\frac{1}{\sqrt{n}\theta}} \xrightarrow{d} N(0,1) \ (n \to \infty) \end{cases} \tag{168}$$

so both $\overline{X}, \overline{Y}$ are asymptotically normal. Note that $g : \mathbb{R}^2 \to \mathbb{R}$ is a multivariate function, we have to apply the multivariate Delta method that

Theorem 9. ***(Multivariate Delta Method)*** *If $Y_n = (Y_{n1}, ..., Y_{nk}) \in \mathbb{R}^k$ is a random vector that is asymptotically normal*

$$\sqrt{n}(Y_n - \mu) \xrightarrow{d} N(0, \Sigma) \ (n \to \infty) \tag{169}$$

and $g : \mathbb{R}^k \to \mathbb{R}$ is the transformation and ∇g is a column vector, so

$$\sqrt{n}(g(Y_n) - g(\mu)) \sim N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu)) \tag{170}$$

is still asymptotically normal.

Now let's write

$$\sqrt{n}[(\overline{X}, \overline{Y})^T - \mu] \xrightarrow{d} N(0, \Sigma) \ (n \to \infty) \tag{171}$$

where

$$\mu = \begin{bmatrix} \frac{1}{\delta\theta} \\ \frac{1}{\theta} \end{bmatrix}, \Sigma = \begin{bmatrix} \frac{1}{\delta^2\theta^2} & 0 \\ 0 & \frac{1}{\theta^2} \end{bmatrix} \tag{172}$$

so by Delta method, we know that $\sqrt{n}[g(\overline{X}, \overline{Y}) - g(\mu)]$ is still asymptotically normal with

$$\nabla g(x, y) = \begin{bmatrix} -\frac{y}{x^2} \\ \frac{1}{x} \end{bmatrix} \tag{173}$$

and the asymptotic variance is

$$\nabla g(\mu)^T \Sigma \nabla g(\mu) = \begin{bmatrix} -\delta^2\theta \\ \delta\theta \end{bmatrix}^T \begin{bmatrix} \frac{1}{\delta^2\theta^2} & 0 \\ 0 & \frac{1}{\theta^2} \end{bmatrix} \begin{bmatrix} -\delta^2\theta \\ \delta\theta \end{bmatrix} = 2\delta^2 \tag{174}$$

Remark. *The Delta Method is still true for $g : \mathbb{R}^n \to \mathbb{R}^k$. In this case, just replace the gradient by the Jacobian matrix $J_{n\times k}$ so the asymptotic covariance matrix of the transformed random vector is $J^T\Sigma J$.*

Another example is that the samples $X_1, ..., X_n$ come from distribution $N(\theta, \theta^2)$ with the estimator $S_n = \frac{\sum_{i=1}^n X_i^2}{2n}$. Firstly, we can show that this estimator is asymptotically normal. By CLT, since $\mathbb{E}X_1^2 = 2\theta^2, Var(X_1^2) = \mathbb{E}X_1^4 - (\mathbb{E}X_1^2)^2 = 6\theta^4$,

$$2S_n = \frac{\sum_{i=1}^n X_i^2}{n}, \sqrt{n}\frac{2S_n - 2\theta^2}{\sqrt{6}\theta^2} = \sqrt{n}\frac{S_n - \theta^2}{\frac{\sqrt{6}}{2}\theta^2} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{175}$$

so obviously $\sqrt{n}(S_n - \theta^2) \xrightarrow{d} N(0, \frac{3}{2}\theta^4) \ (n \to \infty)$. If we set $W_n = \frac{1}{S_n}$, then $g(x) = \frac{1}{x}, g'(x) = -\frac{1}{x^2}$ so by Delta method

$$\sqrt{n}\left(W_n - \frac{1}{\theta^2}\right) \xrightarrow{d} N\left(0, \frac{3}{2}\theta^{-4}\right) \ (n \to \infty) \tag{176}$$

since $g'(\theta^2) = -\theta^{-4}, [g'(\theta^2)]^2\frac{3}{2}\theta^4 = \frac{3}{2}\theta^{-4}$.

For the sufficiency, consider this example with n *i.i.d.* observations of a discrete distribution on $\{0, 1, 2, 3, 4\}$ with

$$p_\theta(X = k) = \frac{\theta^k(1 - \theta^5)}{(1 + \theta)^{k+4}} \ (k = 0, 1, ..., 4) \tag{177}$$

and we want to find a minimal sufficient statistic for θ. Notice that the likelihood ratio

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\theta^{\sum_i (x_i - y_i)}}{(1 + \theta)^{\sum_i (x_i - y_i)}} \tag{178}$$

is independent of θ iff $\sum_i x_i = \sum_i y_i$, so $T(X) = \sum_{i=1}^n X_i$ is a minimal sufficient statistic.

Completeness

A statistic or its family of distributions $\left\{p_\theta^T(t), \theta \in \Theta\right\}$ is called (bounded) **complete** if for any (bounded) Borel function g, $\forall \theta \in \Theta, \mathbb{E}g(T) = 0$ always implies that $g(T) = 0$ *a.s.*. This means that **for complete statistic T, the only way to estimate zero function with an unbiased estimator $g(T)$ is to use the zero function**.

Remark. *Completeness implies bounded completeness but the converse is not true. A counter example is that $X \sim p_\theta$*

$$p_\theta(X = j) = \begin{cases} \theta & j = -1 \\ (1-\theta)^2\theta^j & j = 0, 1, ... \end{cases} \tag{179}$$

where $\theta \in (0,1)$. For any bounded g, if $\forall \theta \in (0,1), \mathbb{E}g(X) = 0$, then notice that

$$\mathbb{E}g(X) = \theta g(-1) + \sum_{j=0}^{\infty}(1-\theta)^2\theta^j g(j) \tag{180}$$

so $\forall \theta \in (0,1), \sum_{j=0}^{\infty}\theta^j g(j) = -(1-\theta)^{-2}\theta g(-1)$. Notice that we can expand the RHS and compare coefficient to see that

$$\forall \theta \in (0,1), \sum_{j=0}^{\infty}\theta^j g(j) = -(1-\theta)^{-2}\theta g(-1) = -\theta g(-1)\sum_{j=0}^{\infty} j\theta^{j-1} = \sum_{j=0}^{\infty} -g(-1)j\theta^j \tag{181}$$

since two power series are equal in an open set, the coefficients must match

$$\forall j = 1, 2, ..., g(j) = -g(-1)j \tag{182}$$

since g is bounded, we must have $g(-1) = 0$ and $\forall j \in \{-1, 0, ...\}, g(j) = 0$, so $g(X) = 0$ a.s. ***it's bounded complete****. However, if we consider the completeness of X where the boundedness of g is cancelled, we can take $g(-1)$ to be 1 such that $\forall j = 1, 2, ..., g(j) = -j$ is not constantly zero on $\{-1, 0, 1, ...\}$, so $g(X)$ is not almost surely 0 and* ***it's not complete****.*

Example: $X_1, ..., X_n \sim U(0, \theta)$, so $T = X_{(n)}$ is minimal sufficient. If for any Borel $g : \mathbb{R}^n \to \mathbb{R}, \forall \theta > 0, \mathbb{E}g(T) = 0$, then since T has density $f(t) = \frac{nt^{n-1}}{\theta^{n-1}}$ $(0 < t < \theta)$, we know that $\forall \theta > 0, \int_0^\theta \frac{nt^{n-1}}{\theta^{n-1}} g(t)\, dt = 0$ so $\forall t \in (0, \theta), g(t)t^{n-1} = 0$ *a.e.*, $g(t) = 0$ *a.e.*. So $g(T) = 0$ *a.s.*, this statistic is complete.

Example: $X_1, ..., X_n \sim B(1, \theta)$ and $T = \sum_{i=1}^n X_i \sim B(n, \theta)$ is minimal sufficient. If for any Borel $g : \mathbb{R}^n \to \mathbb{R}$, $\forall \theta \in (0,1), \mathbb{E}g(T) = 0$, then $\forall \theta \in (0,1), \sum_{k=0}^n \binom{n}{k}\theta^k(1-\theta)^{n-k}g(k) = 0$ so

$$\forall \theta \in (0,1), \sum_{k=0}^{n}\binom{n}{k}\left(\frac{\theta}{1-\theta}\right)^k g(k) = 0 \tag{183}$$

which is a power series in $\frac{\theta}{1-\theta}$, so $\forall k = 0, 1, ..., n, \binom{n}{k}g(k) = 0, g(k) = 0$. It's easy to see that $g(T) = 0$ *a.s.* so T is complete.

Theorem 10. ***(Completeness and Sufficiency)*** *A complete sufficient statistic is minimal sufficient.*

Proof. Let T be complete and sufficient and T^* be minimal sufficient, then $\exists \psi, T^* = \psi(T)$. Consider $\forall \theta, \mathbb{E}[T - \mathbb{E}(T|T^*)] = 0$ where $\mathbb{E}(T|T^*)$ does not depend on θ and

$$g(T) = T - \mathbb{E}(T|T^*) \tag{184}$$

so $\mathbb{E}g(T) = 0$ and by completeness $g(T) = 0$ $a.s.$ so $T = \mathbb{E}(T|T^*)$ $a.s.$ and this proves that T is a function of T^* and T must also be minimal sufficient. □

Remark. *We are taking $\mathbb{E}(T|T^*)$ since it's the best approximation of T as a function of T^* under mean square error and the law of iterated expectation can combine perfectly with completeness. We are proving that such approximation actually gives the perfect T so the information contained in T cannot be less than T^*.*

Remark. *The converse is not true. Consider the counter example*

$$X_1, ..., X_n \sim N(\theta, \theta^2) \tag{185}$$

where $T(X) = (\overline{X}, \sum_{i=1}^n X_i^2)$ is the minimal sufficient statistic by computing likelihood ratio. However, consider

$$g(x, y) = \frac{n}{n+1}x^2 - \frac{1}{2n}y \not\equiv 0 \tag{186}$$

to find that

$$\forall \theta, \mathbb{E}g(T) = 0 \tag{187}$$

since $\mathbb{E}\overline{X}^2 = \frac{2n\theta^2 + n(n-1)\theta^2}{n^2} = \frac{n+1}{n}\theta^2, \mathbb{E}\sum_{i=1}^n X_i^2 = 2n\theta^2$ so it's ***a minimal sufficient statistic that is not complete****.*

Another counter example is that $X_1, ..., X_n \sim U\left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ with ***minimal sufficient statistic*** *$T(X) = (X_{(1)}, X_{(n)})$* ***is not complete****. To see this, if X_1 has density f and CDF F, we find that $X_{(1)}$ has density $f_{X_{(1)}}(x_1) = n[1 - F(x_1)]^{n-1} f(x_1) = n\left(\frac{1}{2} + \theta - x_1\right)^{n-1}$ $x_1 \in \left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$ and $X_{(n)}$ has density $f_{X_{(n)}}(x_n) = n[F(x_n)]^{n-1} f(x_n) = n\left(\frac{1}{2} - \theta + x_n\right)^{n-1}$ $x_n \in \left(\theta - \frac{1}{2}, \theta + \frac{1}{2}\right)$. As a result,*

$$\mathbb{E}X_{(1)} = n\int_{\theta - \frac{1}{2}}^{\theta + \frac{1}{2}} x\left(\frac{1}{2} + \theta - x\right)^{n-1} dx \quad \left(u = x - \theta + \frac{1}{2}\right) \tag{188}$$

$$= n\int_0^1 \left(u + \theta - \frac{1}{2}\right)(1-u)^{n-1} du \tag{189}$$

$$= n\left[B(2, n) + \left(\theta - \frac{1}{2}\right)B(1, n)\right] \tag{190}$$

$$= \frac{1}{n+1} + \theta - \frac{1}{2} \tag{191}$$

similarly,

$$\mathbb{E}X_{(n)} = n\int_{\theta-\frac{1}{2}}^{\theta+\frac{1}{2}} x\left(\frac{1}{2}-\theta+x\right)^{n-1}dx \quad \left(u = x-\theta+\frac{1}{2}\right) \tag{192}$$

$$= n\int_0^1 \left(u+\theta-\frac{1}{2}\right)u^{n-1}\,du \tag{193}$$

$$= n\left[B(n+1,1)+\left(\theta-\frac{1}{2}\right)B(n,1)\right] \tag{194}$$

$$= \frac{n}{n+1}+\theta-\frac{1}{2} \tag{195}$$

so

$$\mathbb{E}\left[X_{(n)}-X_{(1)}-\frac{n-1}{n+1}\right] = 0 \tag{196}$$

and we see that $g(x,y) = y-x-\frac{n-1}{n+1} \not\equiv 0$ *on* $\left(\theta-\frac{1}{2},\theta+\frac{1}{2}\right)^2$ *and that* $\mathbb{E}g(T) = 0$*, so this minimal sufficient statistic is not complete.*

From the examples above, we find out that in order to prove a sufficient statistic is not complete, we just need to try to find the moment of a function of the statistic such that the parameters can cancel out. Conversely, for complete statistic T*, the moments of any function of* T *always cannot be independent of the parameter, this is another way to* ***describe the amount of information a statistic holds****.*

Example: $X_1,...,X_n \sim C(\theta,1)$ follows Cauchy distribution, meaning that $p_\theta(x) = \frac{1}{\pi}\frac{1}{1+(x-\theta)^2}, x\in\mathbb{R}$, to find the minimal sufficient statistic, consider likelihood ratio

$$\frac{p_\theta(x)}{p_\theta(y)} = \frac{\prod_{i=1}^n[1+(y_i-\theta)^2]}{\prod_{i=1}^n[1+(x_i-\theta)^2]} \tag{197}$$

is independent of θ iff $\forall i\in\{1,2,...,n\}, x_{(i)} = y_{(i)}$, so $T(X) = (X_{(1)},...,X_{(n)})$ is the minimal sufficient statistic.

Remark. *One can never state that* $T(X) = \prod_{i=1}^n[1+(X_i-\theta)^2]$ *is the minimal sufficient statistic because the statistic is a function of the sample but contains no information about the parameter* θ*!*

Remark. *Note that* ***for*** *i.i.d.* ***samples,*** $T(X) = (X_{(1)},...,X_{(n)})$ ***is always sufficient and non-trivial*** *(actually it's still sufficient as long as* ***exchangeability*** *is maintained). To see why it's non-trivial, note that if* $T(X)$ *is known,* $(X_1,...,X_n)$ *is still unknown since any permutation of* $X_{(1)},...,X_{(n)}$ *is a possible realization of samples.*

On the other hand, notice that

$$p_\theta(x) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n f(x_{(i)}) \tag{198}$$

where f *is the density of* X_1*. By factorization theorem,* $T(X) = (X_{(1)},...,X_{(n)})$ *is sufficient.*

Remark. *Notice that for i.i.d. samples, all order statistics is always sufficient but not necessarily minimal sufficient.*

However, there are cases where this is minimal sufficient. For example, when the samples come from the Cauchy distribution $C(\theta, 1)$ or the logistic distribution with density $\frac{e^{-(x-\theta)}}{[1+e^{-(x-\theta)}]^2}$.

Generally, we can prove that for i.i.d. samples from a location family, i.e. a population with density $f(x-\theta)$ (where θ is the location parameter), if the distribution is ***not in the exponential family****, it's* ***often*** *the minimal sufficient statistic for θ.*

Note that exercise 6.8 in Casella Berger is wrong without additional conditions! *The counterexample for this conclusion is $X_1, ..., X_n \sim N(\theta, 1)$ which still forms a location family but is also in the exponential family.*

Consider

$$p_\theta(x) = (2\pi)^{-\frac{1}{2}} e^{-\frac{1}{2}(x-\theta)^2} \tag{199}$$

$$= (2\pi)^{-\frac{1}{2}} e^{-\frac{\theta^2}{2}} e^{-\frac{1}{2}x^2} e^{\theta x} \tag{200}$$

where $c(\theta) = (2\pi)^{-\frac{1}{2}} e^{-\frac{\theta^2}{2}}, h(x) = e^{-\frac{1}{2}x^2}, Q(\theta) = \theta, T(x) = x$. Verify the OSC to find that $\overline{X}$ is a complete sufficient statistic.

Exponential Family

If X has density or probability mass of form

$$p_\theta(x) = c(\theta) \cdot e^{Q(\theta)T(x)} \cdot h(x) \tag{201}$$

then it's in the **one-parameter exponential family**.

Remark. *Notice that since $h(x)$ and $c(\theta)$ are separate functions,* ***the support of the probability distribution can NOT depend on parameter θ!***

Example: $X \sim B(1,\theta)$, so $p_\theta(x) = \theta^x(1-\theta)^{1-x} = (1-\theta)e^{x\log\left(\frac{\theta}{1-\theta}\right)}$ is in one-parameter exponential family with $c(\theta) = 1-\theta, T(x) = x, Q(\theta) = \log\left(\frac{\theta}{1-\theta}\right), h(x) = 1$, with the support does not depend on θ.

Similarly, a random variable is in **multi-parameter exponential family** if

$$p_\theta(x) = c(\theta) \cdot e^{\sum_{j=1}^k Q_j(\theta)T_j(x)} \cdot h(x) \tag{202}$$

with the support independent of θ.

Example: $X \sim N(\mu,\sigma^2)$, so $p_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x}$ is in the multi-parameter exponential family with $c(\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}}, T_1(x) = x^2, T_2(x) = x, Q_1(\mu,\sigma) = -\frac{1}{2\sigma^2}, Q_2(\mu,\sigma) = \frac{\mu}{\sigma^2}$.

Sufficient Statistic of Exponential Family

Now let $X_1, ..., X_n$ *i.i.d.* from multi-parameter exponential family, then for $x \in \mathbb{R}^n$,

$$p_\theta(x) = c^n(\theta) \cdot e^{\sum_{j=1}^k Q_j(\theta) \sum_{i=1}^n T_j(x_i)} \cdot \prod_{i=1}^n h(x_i) \tag{203}$$

consider $\forall j = 1, 2, ..., k, T_j(x) = \sum_{i=1}^n T_j(x_i)$, then we see that

$$p_\theta(x) = c^n(\theta) \cdot e^{\sum_{j=1}^k Q_j(\theta)T_j(x)} \cdot \prod_{i=1}^n h(x_i) \tag{204}$$

by factorization theorem, $T(X) = (T_1(X), ..., T_k(X))$ is the **sufficient statistic**.

Example: $X_1, ..., X_n \sim B(1,\theta)$, so $T_1(x) = \sum_{i=1}^n T_1(x_i) = \sum_{i=1}^n x_i$, and $T(X) = \sum_{i=1}^n X_i$ is just the sufficient statistic, consistent with what we have derived.

Example: $X_1, ..., X_n \sim N(\mu,\sigma^2)$, so $T_1(x) = \sum_{i=1}^n T_1(x_i) = \sum_{i=1}^n x_i^2$, and $T_2(X) = \sum_{i=1}^n T_2(x_i) = \sum_{i=1}^n x_i$, so $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is the sufficient statistic.

Completeness of Exponential Family

Note that $T(X) = (T_1(X), ..., T_k(X))$, the sufficient statistic computed for exponential family, is also **complete provided** $\{(Q_1(\theta), ..., Q_k(\theta) : \theta \in \Theta)\}$ **contains an open set in** $\mathbb{R}^k$. So under this condition, it must also be

minimal sufficient statistic.

Theorem 11. ***(Completeness of Exponential Family Sufficient Statistic)*** *If* $\{(Q_1(\theta), ..., Q_k(\theta) : \theta \in \Theta)\}$ *contains an open set in* $\mathbb{R}^k$ ***(open set condition, OSC)****, then* $T(X) = (T_1(X), ..., T_k(X))$ *is complete.*

Example: $X_1, ..., X_n \sim B(1, \theta)$, check OSC that $Q(\theta) = \log \frac{\theta}{1-\theta}$ and $\left\{\log \frac{\theta}{1-\theta} : \theta \in (0, 1)\right\}$ contains an open set in $\mathbb{R}$, so $T(X) = \sum_{i=1}^n X_i$ is complete and minimal sufficient.

Example: $X_1, ..., X_n \sim N(\mu, \sigma^2)$, check OSC that $\left\{\left(-\frac{1}{2\sigma^2}, \frac{\mu}{\sigma^2}\right) : \mu \in \mathbb{R}, \sigma \in \mathbb{R}_{++}\right\}$ contains an open set in $\mathbb{R}^2$, so $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ is complete and minimal sufficient.

Remark. *To see an example of the violation of OSC, consider* $X_1, ..., X_n \sim N(\theta, \theta^2)$ *so* $T(X) = (\sum_{i=1}^n X_i^2, \sum_{i=1}^n X_i)$ *is minimal sufficient. However, note that* $\left\{\left(-\frac{1}{2\theta^2}, \frac{1}{\theta}\right) : \theta \in \mathbb{R}\right\}$ *is the graph of a curve in* $\mathbb{R}^2$ *so it cannot contain any open set in* $\mathbb{R}^2$ *so OSC fails. That is the easy consequence of the observation that such statistic is not complete as shown above.*

Ancillary Statistic

A statistic T is **ancillary** if its distribution does not depend on the parameter θ.

Example: $X_i \sim N(\theta, 1), T(X) = X_1 - X_2 \sim N(0, 2)$ is ancillary, $T(X) = X_n - \overline{X}$ is ancillary, and $T(X) = \sum_{i=1}^n a_i X_i \sim N(\sum_{i=1}^n a_i\theta, \sum_{i=1}^n a_i^2)$ is ancillary as long as $\sum_{i=1}^n a_i = 0$.

Theorem 12. ***(Sufficiency and Ancillarity)*** *If T is sufficient statistic and $V(X_1, ..., X_n)$ is independent of T, then V is ancillary.*

Proof. Since T sufficient, $X|_T$ is independent of θ, so $V(X)|_T$ is independent of θ. Since $V(X)$ is independent of T,

$$\forall B \in \mathscr{B}_{\mathbb{R}}, \mathbb{P}\left(V(X) \in B|T\right) = \mathbb{P}\left(V(X) \in B\right) \tag{205}$$

is independent of θ, so proved. □

Theorem 13. ***(Basu's Theorem)*** *T is bounded complete sufficient statistic, V is ancillary, then V is independent of T.*

Proof. $\forall\ g$ as bounded function, since V is ancillary, $\mathbb{E}g(V) = c$ is independent of θ.

$$\forall\theta, 0 = \mathbb{E}[g(V) - c] = \mathbb{E}[\mathbb{E}(g(V) - c|T)] \tag{206}$$

since T is bounded complete, $\mathbb{E}(g(V) - c|T)$ is a bounded function of T, conclude that $\mathbb{E}(g(V) - c|T) = 0$ *a.s.* so $\mathbb{E}[g(V)|T] = c = \mathbb{E}g(V)$ *a.s.*.

As a result, for any bounded function g, we have proved $\mathbb{E}[g(V)|T] = \mathbb{E}g(V)$ *a.s.*. This implies that V is independent of T since we can take $\forall B \in \mathscr{B}_{\mathbb{R}}, g = \mathbb{I}_B$ so $\forall B \in \mathscr{B}_{\mathbb{R}}, \mathbb{P}\left(V \in B|T\right) = \mathbb{P}\left(V \in B\right)$. □

Remark. *Basu's theorem is always applied to prove the **independence of statistics**. Let's see an example here with $X_1, X_2 \sim \mathcal{E}(\theta)$ so the density is $\theta e^{-\theta x}$ in the exponential family with $c(\theta) = \theta, T(x) = x, Q(\theta) = -\theta$. As a result, $X_1 + X_2$ is a sufficient statistic. Check OSC that $\{-\theta : \theta \in \mathbb{R}_{++}\}$ contains open set in $\mathbb{R}$ so it's also complete.*

Consider $V = \frac{X_1}{X_1+X_2}$ and use the Jacobian of the mapping $f : \mathbb{R}^2 \to \mathbb{R}^2, f(x_1, x_2) = \left(x_1, \frac{x_1}{x_1+x_2}\right)$ to see that the joint density

$$f_{X_1,V}(x_1, v) = f_{X_1,X_2}(x_1, x_2) \cdot \left|\det \frac{\partial(x_1, x_2)}{\partial(x, v)}\right| \tag{207}$$

$$= \theta^2 e^{-\theta x_1 - \theta(\frac{1}{v}-1)x_1} \cdot \frac{x_1}{v^2} \tag{208}$$

$$= \theta^2 e^{-\theta\frac{x_1}{v}} \cdot \frac{x_1}{v^2}\ (x_1 > 0, 0 < v < 1) \tag{209}$$

so the marginal density of V is given by

$$f_V(v) = \int_0^\infty f_{X_1,V}(x_1, v)\, dx_1 = 1\ (v \in (0, 1)) \tag{210}$$

so $V \sim U(0,1)$ is uniform and of course ancillary. By Basu's theorem, we have proved that X_1+X_2 ***is independent of*** *$\frac{X_1}{X_1+X_2}$* ***for exponential distributed samples****.*

Remark. *Consider* ***the location family*** *with likelihood $p_\theta(x) = f(x-\theta)$, we always have that $Y = X - \theta$ has likelihood $f(y)$, independent of θ so it's ancillary.*

As a result, $(X_1 - X_2) = (X_1 - \theta) - (X_2 - \theta) = Y_1 - Y_2$ is also ancillary and ***the sample variance*** *$S^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1} = \frac{\sum_i (Y_i - \overline{Y})^2}{n-1}$ is still* ***ancillary*** *since $X_i - \overline{X} = Y_i + \theta - (\overline{Y} + \theta) = Y_i - \overline{Y}$. Actually,* ***the difference of sample median and sample mean*** *$X_{(\frac{n}{2})} - \overline{X}$* ***is also ancillary*** *since $X_{(\frac{n}{2})} - \overline{X} = Y_{(\frac{n}{2})} + \theta - (\overline{Y} + \theta) = Y_{(\frac{n}{2})} - \overline{Y}$ whose distribution does not depend on θ. Moreover,* ***the range*** *$X_{(n)} - X_{(1)}$* ***is also ancillary*** *since $X_{(n)} - X_{(1)} = Y_{(n)} + \theta - (Y_{(1)} + \theta) = Y_{(n)} - Y_{(1)}$.* ***So for location family, any expression as a function of homogeneous linear combination is ancillary.***

As a result, for samples $X_1, ..., X_n$ coming from Gaussian with known variance $N(\theta, 1)$, sample mean $\overline{X}$ is complete sufficient. By Basu's theorem, sample mean is independent of sample variance, sample mean is independent of the difference between sample median and sample mean, sample mean is independent of the range.

For samples $X_1, ..., X_n$ coming from the distribution $p_\theta(x) = e^{-(x-\theta)}$ $(x > \theta)$,

$$p_\theta(x) = e^{n\theta - \sum_i x_i} \mathbb{I}_{x_{(1)} > \theta} \tag{211}$$

by factorization theorem, $T(X) = X_{(1)}$ is sufficient. Moreover, if for any Borel function g, $\forall \theta > 0, \mathbb{E}g(X_{(1)}) = 0$, then T has density $f_T(t) = ne^{(\theta - t)n}$ $(t > \theta)$ and

$$\forall \theta > 0, \int_\theta^\infty g(t) n e^{(\theta - t)n}\, dt = 0 \tag{212}$$

implies that $\forall \theta > 0, \int_\theta^\infty g(t) e^{-nt}\, dt = 0$ so $g = 0$ a.e. and $g(T) = 0$ a.s., so T is complete. By Basu's theorem, the minimum sample is independent of the sample variance, the minimum sample is independent of the difference of sample median and sample mean, the minimum sample is independent of the range.

Remark. *Similarly, consider the* ***scale-parameter family*** *where $p_\theta(x) = \frac{1}{\theta} f\left(\frac{x}{\theta}\right)$ so $Y = \frac{X}{\theta}$ has likelihood $f(y)$ is ancillary.*

We can verify that the ***quotient*** *$\frac{X_1}{X_2} = \frac{Y_1 \theta}{Y_2 \theta} = \frac{Y_1}{Y_2}$ is ancillary, the* ***sample percentage*** *$\frac{X_1}{X_1 + X_2} = \frac{Y_1 \theta}{Y_1 \theta + Y_2 \theta} = \frac{Y_1}{Y_1 + Y_2}$ is ancillary, and the* ***sample percentage in square*** *$\frac{X_1^2}{\sum_i X_i^2} = \frac{Y_1^2 \theta^2}{\sum_i Y_i^2 \theta^2} = \frac{Y_1^2}{\sum_i Y_i^2}$ is ancillary.* ***So for scale-parameter family, any expression that's homogeneous with the same power on numerator and denominator is ancillary.***

Statistical Decision Theory: Overview

Basic Setting

A **non-randomized decision rule** $\delta(\cdot)$ is a function from $\mathscr{X} \to D$ where $\mathscr{X}$ is the space of all possible values of the sample and D is the space of available decisions, different under different statistical settings. For example, under point estimation context, $D = \Theta$. Under confidence interval estimation, $D = \{C(x)\}$ where $C(x) \subset \Theta$ is the confidence set of θ on observing sample realization x. Under hypothesis testing, $D = \{d_0, d_1\}$ where d_0 denotes the action to accept H_0 and d_1 denotes the action to reject H_0. A decision rule is actually specifying what action $\delta(x)$ one shall take when observing the sample realization $X = x$.

A decision rule is judged by **loss function** $L : \Theta \times D \to \mathbb{R}_+$ where $L(\theta, d)$ measures the distance between the action d and the unknown true parameter θ. For example, under point estimation $L(\theta, d) = ||\theta - d||^2$ (squared error loss, SEL). Under confidence interval estimation, $L(\theta, C(x))$ has something to do with $\mathbb{I}_{\theta \notin C(x)}$.

The **risk function** $R(\theta, d)$ measures the average of loss for decision rule δ when the sample X is specified as random variables with likelihood p_θ

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(X)) \tag{213}$$

where $\mathbb{E}_\theta$ means that we are fixing θ such that the likelihood p_θ is fixed, providing a distribution for the sample X. It's easy to see that

$$R(\theta, \delta) = \int_{\mathscr{X}} L(\theta, \delta(x)) \cdot p_\theta(x)\, dx \tag{214}$$

which is the frequentist approach to take average not only on observed samples but on all possible samples in the space $\mathscr{X}$.

In order to figure out which decision rule δ_1, δ_2 is better, we have to compare the risks $R(\theta, \delta_1), R(\theta, \delta_2)$ to decide. A decision rule δ is **inadmissible** if

$$\exists \delta_0, s.t. \begin{cases} \forall \theta \in \Theta, R(\theta, \delta_0) \le R(\theta, \delta) \\ \exists \theta_0 \in \Theta, R(\theta_0, \delta_0) < R(\theta_0, \delta) \end{cases} \tag{215}$$

Remark. *Inadmissibility is a concept easy to understand. A decision rule is inadmissible if it will never be adopted, i.e. there exists another decision rule δ_0 which is strictly better. To characterize it using mathematical language, we say that for any parameter $\theta \in \Theta$, the risk of δ is always higher than the risk of δ_0. However, this does not suffice since it may happen that $\forall \theta \in \Theta, R(\theta, \delta_0) = R(\theta, \delta)$. That's why the second condition is added such that δ_0 is strictly better than δ at at least one parameter value θ_0.*

In general, there does not exist decision rule with uniformly least risk for all θ, i.e. there does not exists δ^* such that $\forall \theta \in \Theta, \forall \delta, R(\theta, \delta^*) \le R(\theta, \delta)$. For example, with SEL, if δ^* is the uniformly least risk decision rule, it has to perform better than the constant decision rule δ_0 such that $\forall x \in \mathscr{X}, \delta_0(x) = \theta_0$ for any θ_0. Then

$\forall\theta\in\Theta, R(\theta,\delta^*)\leq R(\theta,\delta_0)=||\theta-\theta_0||^2$ so $R(\theta_0,\delta^*)=0$, $\delta^*(X)=\theta_0$ *a.s.* which obviously cannot be true if $|\Theta|>1$ (parameter space is not a single element set).

Now we see that selecting decision rule according to the value of the risk $R(\theta,\delta)$ (as a function of θ) generally does not make sense. To solve this problem, we have two ways, the first is to **find the best decision rule within the desirable subclasses**, e.g. unbiased rules, invariant rules etc. The second way it to **set up an optimality criteria** for selecting the best decision rule. For example, by considering the parameter θ itself as a random variable with prior distribution $\pi(\theta)$, we can set up **the Bayes risk for taking decision** δ

$$R_\pi(\delta)=\mathbb{E}_{\theta\sim\pi}R(\theta,\delta)=\int_\Theta R(\theta,\delta)\cdot\pi(\theta)\,d\theta \tag{216}$$

and pick the best decision rule δ^* that achieves the **Bayes risk**

$$R_\pi^*=\inf_\delta R_\pi(\delta) \tag{217}$$

another useful optimality criteria can be given by the **minimax risk**

$$\inf_\delta\sup_{\theta\in\Theta}R(\theta,\delta) \tag{218}$$

so the optimal decision rule δ^* is the one such that

$$\forall\delta,\sup_{\theta\in\Theta}R(\theta,\delta^*)\leq\sup_{\theta\in\Theta}R(\theta,\delta) \tag{219}$$

i.e., minimizes the maximum risk over all parameter $\theta\in\Theta$.

Example: Point Estimation

Consider $X_1,...,X_n\sim N(\mu,\sigma^2)$ and our objective is to estimate σ^2 with SEL (squared error loss) $L(\theta,\delta(x))=||\theta-\delta(x)||^2$, we can observe the important property that

$$\begin{aligned}R(\theta,\delta)&=\mathbb{E}_\theta||\theta-\delta(X)||^2 &(220)\\&=||\theta||^2-2\theta^T\mathbb{E}_\theta\delta(X)+\mathbb{E}_\theta||\delta(X)||^2 &(221)\\&=\mathbb{E}_\theta||\delta(X)||^2-||\mathbb{E}_\theta\delta(X)||^2+||\mathbb{E}_\theta\delta(X)-\theta||^2 &(222)\\&=Var_\theta(\delta(X))+||Bias_\theta(\delta(X))||^2 &(223)\end{aligned}$$

under SEL, risk is always the sum of variance and squared bias.

If we take $\delta(X)=bS^2$, meaning that we estimate σ^2 by a multiple of sample variance with $b\geq 0$, set $Y_i=\frac{X_i-\mu}{\sigma}\sim N(0,1)$ to notice that $S_Y^2=\frac{S_X^2}{\sigma^2},(n-1)S_Y^2\sim\chi_{n-1}^2$, so we would have $\mathbb{E}S^2=\sigma^2\mathbb{E}S_Y^2=\sigma^2\frac{n-1}{n-1}=\sigma^2$ and

$Var(S^2) = \sigma^4 Var(S_Y^2) = \sigma^4 \frac{2(n-1)}{(n-1)^2} = \frac{2\sigma^4}{n-1}$. As a result,

$$R(\sigma, \delta) = Var(bS^2) + (\mathbb{E}bS^2 - \sigma^2)^2 \tag{224}$$

$$= \frac{2b^2\sigma^4}{n-1} + (b-1)^2\sigma^4 \tag{225}$$

$$= \left(\frac{2b^2}{n-1} + (b-1)^2\right)\sigma^4 \tag{226}$$

not let's specify b by minimizing the risk w.r.t. b to get $b^* = \frac{n-1}{n+1}$ so $\delta(X) = \frac{n-1}{n+1}S^2$ would be the best estimator among all decision rules in the family with the form bS^2 no matter what value σ takes (it's the only admissible decision rule in this family). This is an example where by **restricting the decision rule that has a special form**, we would get the best decision rule with uniformly least risk for all parameter values of σ.

Example: Hypothesis Testing

The decision theory framework for hypothesis testing is built as

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_0^c \end{cases} \tag{227}$$

where Θ_0, Θ_0^c is a partition of the whole parameter space Θ. The set of all decisions is $D = \{d_0, d_1\}$ where d_0 means accepting H_0 and d_1 means rejecting H_0. Generally, **0-1 loss** is taken, which means that the loss is always 1 when a wrong decision is made and the loss is always 0 when a correct decision is made (to generalize, we can assign different penalty when type I or type II mistake is made)

$$L(\theta, d_0) = \mathbb{I}_{\theta \notin \Theta_0}, L(\theta, d_1) = \mathbb{I}_{\theta \in \Theta_0} \tag{228}$$

notice that when $\theta \in \Theta_0$ but decision d_1 is taken, it's a Type I error and when $\theta \notin \Theta_0$ but decision d_0 is taken, it's a Type II error.

Calculate the risk function to find

$$R(\theta, \delta) = L(\theta, d_0) \cdot \mathbb{P}\left(\delta(X) = d_0\right) + L(\theta, d_1) \cdot \mathbb{P}\left(\delta(X) = d_1\right) \tag{229}$$

$$= \begin{cases} \mathbb{P}_\theta\left(\delta(X) = d_1\right) = \beta(\theta) & \theta \in \Theta_0 \\ \mathbb{P}_\theta\left(\delta(X) = d_0\right) = 1 - \beta(\theta) & \theta \notin \Theta_0 \end{cases} \tag{230}$$

$$= L(\theta, d_0) \cdot [1 - \beta(\theta)] + L(\theta, d_1) \cdot \beta(\theta) \tag{231}$$

where $\beta(\theta)$ is the probability of rejecting H_0 for fixed parameter θ and is called the **power function** of the test.

Example: Interval Estimation

The loss function in interval estimation problems is typically formed as

$$L(\theta, C(x)) = b\lambda(C(x)) - \mathbb{I}_{\theta \in C(x)} \tag{232}$$

where $C(x) \subset \Theta$ is the confidence region put up based on sample realization $X = x$ and λ is the Lebesgue measure on $\mathbb{R}^n$ (consider $\Theta \subset \mathbb{R}^n$). The point is to set up the **trade-off between the size of the confidence set and the coverage probability**. In other words, we want to get as large coverage probability of θ as possible but with its size as small as possible. The risk function is formed as

$$R(\theta, \delta) = b \cdot \mathbb{E}_\theta \lambda(C(X)) - \mathbb{P}_\theta \left(\theta \in C(X)\right) \tag{233}$$

Remark. *Note that the formulation of interval estimation does not always follow decision theory since the 'size' of the confidence region can sometimes be confusing. Consider $\Theta \subset \mathbb{R}$ and any countable set has zero Lebesgue measure, so for any confidence region $C(x)$, $\forall N$ countable, $C(x) \cup N$ is always better than $C(x)$ since the size of the confidence region, measured by Lebesgue measure, does not change; while the coverage probability may be higher.*

For example, consider $X \sim N(\mu, \sigma^2)$ with σ^2 known and now we want to form interval estimation for μ with confidence region of the form $C(x) = [x - c\sigma, x + c\sigma]$. So it's clear that

$$\begin{aligned} R(\mu, \delta) &= 2bc\sigma - \mathbb{P}_\mu \left(|X - \mu| \leq c\sigma\right) && (234) \\ &= 2bc\sigma - 2\Phi(c) + 1 && (235) \end{aligned}$$

taking derivative w.r.t. c to find

$$c^* = \varphi^{-1}(b\sigma) > 0 \tag{236}$$

where φ is the standard Gaussian density. As a result, if $b\sigma > \frac{1}{\sqrt{2\pi}}$, then $c^* = 0$ so $C(x) = \{x\}$, the point estimation is always the best. If $b\sigma \leq \frac{1}{\sqrt{2\pi}}$, $c^* = \varphi^{-1}(b\sigma) > 0$ is the positive inverse and the confidence region is actually an interval.

As a result, by restricting a specific form of $C(x)$, for given b, σ we have proved that the decision $\delta(X) = C(X) = [X - c^*\sigma, X + c^*\sigma]$ has the minimum risk over all possible decisions and that all other decisions are inadmissible.

Statistical Decision Theory: The Bayesian Setting

The motivation of the Bayesian setting, as described above, is that generally risk function is a function in the parameter θ and generally there does not exists the best decision rule with uniformly least risk over the whole parameter space. As a result, if we don't want to restrict ourselves to a subclass of decision rules, we have to define some optimality criterion. One of them, is to view θ as a random variable with prior distribution $\pi(\theta)$ such that **the average risk for decision δ under prior π** is defined as

$$\Lambda(\pi, \delta) = \mathbb{E}_{\theta\sim\pi} R(\theta, \delta) = \int_\Theta R(\theta, \delta) \cdot \pi(\theta)\, d\theta \tag{237}$$

is a real number not depending on the value of θ but on its prior distribution. As a result,

$$\delta_\pi = \arg\min_\delta \Lambda(\pi, \delta) \tag{238}$$

is the best decision rule under prior π, called **the Bayes rule w.r.t.** π. However, one might notice that we are not specifying the space of all possible decision rule δ to optimize the average risk. Here we are actually finding the optimal δ among all possible randomized decision rules, which is a class of more general decision rules than the non-randomized ones.

Let's introduce the **randomized decision rule** δ^* to be a mapping from $\mathscr{X}$ to the space of all probability distribution on the decision space D. So for any given sample realization $X = x$, $\delta^*(x) = \delta^*_x$ and δ^*_x is a probability measure on D. Notationally, $\delta^*_x(d_0) = 1$ is equivalent to saying on observing sample realization $X = x$ the decision is d_0 almost surely. So actually $\delta(x) = d_0$ and it's a non-randomized decision rule.

Remark. *In order to understand the difference between non-randomized and randomized decision rules, let's consider a specific example of the setting. One might also notice that the setting of decision theory is actually almost the same as that of reinforcement learning.*

$\mathscr{X}$ is the space of all possible values taken by the sample, i.e. the state space, and it contains two elements, one is that "a person is tired" and the other is that "a person is not tired". Naturally, X takes values in $\mathscr{X}$ so X is a random variable describing whether a person is tired or not (so we can think of such Xnto follow the Bernoulli distribution since it only has two possible values to take in $\mathscr{X}$). Now its distribution $X \sim B(1, \theta)$ with θ as a parameter, let's say, is "the amount of homework one has", $\theta \in (0, 1)$. The amount of homework is unknown to us but when the amount of homework θ is given, then the distribution of X is completely clear, so our task is to estimate the amount of homework θ using the observations X about whether one is tired or not.

Now we provide two decisions for the person, i.e. the space of all available decisions (actions) D has two elements, one to be "sleep early" and the other to be "sleep late". A non-randomized decision $\delta : \mathscr{X} \to D$ is a mapping from the value of the sample to the space of available decisions. For example, the decision rule that "If I am tired, I sleep early. If I am not tired, I sleep late" is just a non-randomized decision. When X takes value "a person is tired", the person makes the decision to sleep early, when X takes value "a person is not tired", the person makes the decision to sleep late.

In contrast, a randomized decision δ^* is a mapping from $\mathscr{X}$ to $\mathscr{P}(D)$, where $\mathscr{P}(D)$ is the space whose elements

are all possible probability measures on D. Simply speaking, δ^* maps $x \in \mathscr{X}$ to $\delta_x^* \in \mathscr{P}(D)$, where δ_x^* is a probability measure on D and it's called the randomized decision rule on observing the sample realization $X = x$. To raise an example, consider the same problem setting, but now the decision rule is that "If I am tired, I have 0.9 probability of sleeping early and 0.1 probability of sleeping late. If I am not tired, I have 0.2 probability of sleeping early and 0.8 probability of sleeping late." is a randomized decision rule. The reason is that δ^* maps the state "I am tired" to the probability measure $\mathbb{Q}$ on D that has 0.9 probability mass on "sleep early" and 0.1 probability mass on "sleep late".

If one is familiar with reinforcement learning (RL) settings, one might find that such randomized decision rule is just the "policy" we consider in RL, which tells us what distribution on the action space to take seeing the realization of the state we are facing. It's also easy for one to understand that ***all non-randomized decision rules can be formed as randomized decision rules*** *since non-randomized decision rules are just the δ^* such that $\forall x \in \mathscr{X}, \delta_x^*$ always puts probability mass 1 on one of the decisions in D.*

So how shall we form **the risk of the randomized decision rule**? Of course we still want to adopt the previous interpretations to form it as the expectation of loss, but the expectation should be w.r.t. both the randomness in the sample X and the randomness in the decision rule δ_x^* since it is itself a probability measure

$$R(\theta, \delta^*) = \mathbb{E}_{X \sim p_\theta} \mathbb{E}_{T \sim \delta_X^*} L(\theta, T) = \int_{\mathscr{X}} \left(\int_D L(\theta, t) \delta_x^*(t)\, dt \right) p_\theta(x)\, dx \tag{239}$$

since $\delta_x^*(t)$ is now a likelihood in t on the space D.

Remark. *In the formulation of the risk of the randomized decision rule, we first take expectation w.r.t. $T \sim \delta_X^*$ and then take expectation w.r.t. $X \sim p_\theta$ because probability measure δ_X^* is actually a* ***random measure****, i.e. $\delta_X^*|_{X=x}$ is a probability measure but δ_X^* actually has the randomness coming from sample X. That's why we first act as if we know the sample realization $X = x$ and then take expectation w.r.t. X to remove all randomness. In other words, the true definition is formed as*

$$R(\theta, \delta^*) = \mathbb{E}_{X \sim p_\theta} \mathbb{E}_{T \sim \delta_X^*} [L(\theta, T)|X] \tag{240}$$

so $\mathbb{E}_{T \sim \delta_X^*}[L(\theta, T)|X] = f(X)$ is a measurable function of X such that $\mathbb{E}_{T \sim \delta_x^*}[L(\theta, T)|X = x] = f(x)$ and the outer expectation is taken w.r.t. X following likelihood p_θ.

Combining two definitions to see that

$$\Lambda(\pi, \delta^*) = \int_\Theta \left[\int_{\mathscr{X}} \left(\int_D L(\theta, t) \cdot \delta_x^*(t)\, dt \right) p_\theta(x)\, dx \right] \cdot \pi(\theta)\, d\theta \tag{241}$$

$$= \int_\Theta \int_{\mathscr{X}} \int_D L(\theta, t) \cdot \delta_x^*(t) \cdot m(x) \cdot \pi(\theta|x)\, dt\, dx\, d\theta \tag{242}$$

$$= \int_{\mathscr{X}} m(x) \left[\int_D \left(\int_\Theta L(\theta, t) \cdot \pi(\theta|x)\, d\theta \right) \delta_x^*(t)\, dt \right] dx \tag{243}$$

by applying Fubini (integrand is non-negative) and the Bayes theorem that $p_\theta(x) \cdot \pi(\theta) = \pi(\theta|x) \cdot m(x)$.

Set $l(d, x) = \int_\Theta L(\theta, d) \cdot \pi(\theta|x)\, d\theta = \mathbb{E}_{\theta \sim \pi(\cdot|x)} L(\theta, d)$ as the **expected posterior loss (EPL) on observing**

sample realization $X = x$ and decision $d \in D$. We find that in order to minimize $\Lambda(\pi, \delta^*)$, we just need to minimize

$$\int_D l(t,x) \cdot \delta_x^*(t)\, dt \tag{244}$$

w.r.t. δ^* for $\forall x \in \mathscr{X}$. Now we observed that this expression has the structure of expectation that

$$\int_D l(t,x) \cdot \delta_x^*(t)\, dt = \mathbb{E}_{T\sim\delta_x^*} l(T,x) \tag{245}$$

in order to minimize this expectation, the most natural way is to choose

$$\forall x \in \mathscr{X}, d_0(x) = \arg\min_{d\in D} l(d,x) \tag{246}$$

such that $l(d,x)$ is minimized at $d = d_0(x)$ and then put all probability mass on $d_0(x)$ such that

$$\delta_x^*(t) = \mathbb{I}_{t=d_0(x)} \tag{247}$$

Remark. *The complete proof of why the approach above minimizes $\int_D l(t,x) \cdot \delta_x^*(t)\, dt$ is given as follows:*

$$\begin{aligned} \forall x \in \mathscr{X}, \int_D l(t,x) \cdot \delta_x^*(t)\, dt &= \mathbb{E}_{T\sim\delta_x^*} l(T,x) && (248)\\ &\geq \mathbb{E}_{T\sim\delta_x^*} l(d_0(x),x) && (249)\\ &= l(d_0(x),x) && (250) \end{aligned}$$

and the approach above achieves this lower bound. Note that $d_0(x)$ always exists if the loss function L is chosen as some convex function in d, which is often the case.

So we get the Bayes rule with $\forall x \in \mathscr{X}, \delta_x^*(d_0(x)) = 1$, and it's very interesting to see that **although we are minimizing $\Lambda(\pi, \delta^*)$ w.r.t. δ^* among all randomized decision rules, the optimal one is actually deterministic and is a function of the observed sample x!** This shows the spirit of Bayesian statistics since such decision is made based only on the observed sample realization x instead of all possible sample realizations in the space $\mathscr{X}$ (the frequentist approach).

Remark. *To conclude, finding the Bayes rule only has two steps.* ***The first step is to calculate the expected posterior loss (EPL)***

$$l(d,x) = \int_\Theta L(\theta,d) \cdot \pi(\theta|x)\, d\theta = \mathbb{E}_{Y\sim\pi(\cdot|x)} L(Y,d) \tag{251}$$

and ***the second step is to do the optimization for fixed sample realization***

$$\forall x \in \mathscr{X}, d_0(x) = \arg\min_{d} l(d,x) \tag{252}$$

so $\delta_\pi(x) = d_0(x)$ ***is the Bayes rule*** *with the least average risk under prior π on observing sample realization x.*

When the loss function has special forms, the Bayes rule can be solved out easily and has very simple forms. Let's consider $\Theta \subset \mathbb{R}^n$ with ***squared error loss (SEL)***

$$L(\theta, d) = ||\theta - d||^2 \tag{253}$$

so $l(d, x) = \mathbb{E}_{Y\sim\pi(\cdot|x)}||Y - d||^2$ and $\delta_\pi(x) = d_0(x) = \mathbb{E}_{Y\sim\pi(\cdot|x)}Y$ (since $\mathbb{E}Z = \arg\min_c \mathbb{E}||Z - c||^2$) so ***Bayes rule is just the posterior mean****!*

On the other hand, consider $\Theta \subset \mathbb{R}^n$ with ***absolute error loss (AEL)***

$$L(\theta, d) = ||\theta - d||_1 \tag{254}$$

so $l(d, x) = \mathbb{E}_{Y\sim\pi(\cdot|x)}||Y - d||_1$ and $\delta_\pi(x) = d_0(x) = Median_{\pi(\cdot|x)}$ (since $Median_Z = \arg\min_c \mathbb{E}||Z - c||_1$) so ***Bayes rule is just the posterior median****!*

Remark. *It's also an interesting question to think about for the uniqueness of such Bayes estimator δ_π under the condition that D is countable. For $\forall x \in \mathscr{X}$, denote $D_0(x) \subset D$ as the subset of decisions such that $\forall d_0(x) \in D_0(x), d_0(x) = \arg\min_{d\in D} l(d, x)$, i.e. $D_0(x)$ is the collection of all decisions that achieves argmin for sample realization $X = x$.*

It's easy to prove that ***if*** *$\forall x \in \mathscr{X}, \delta^*(x) = \delta^*_x$* ***is any probability distribution that puts all probability masses in*** *$D_0(x)$****, then*** *δ^** ***is always the Bayes estimator****. This is quite intuitive since if all decisions are as good as each other for the realized sample, then the way to distribute point mass among those decisions does not matter.*

Minimax Decision Rule

In the following context, we always refer to "decision rule" meaning "randomized decision rule". δ_0 is defined as the **minimax decision rule** if $\forall \delta, \sup_{\theta\in\Theta} R(\theta,\delta_0) \leq \sup_{\theta\in\Theta} R(\theta,\delta)$. Note that Θ is the parameter space so it has nothing to do with sample realization x.

Theorem 14. ***(Connection between Bayes Rule and Minimax Rule)*** *If δ_π is the Bayes rule under some prior π and its risk function is constant for $\forall\theta\in\Theta$, then it is minimax (such rule is called **equalized Bayes rule**).*

Proof. By definition, $\forall\delta, \Lambda(\pi,\delta_\pi) \leq \Lambda(\pi,\delta)$. Now since $R(\theta,\delta_\pi)$ is constant for $\forall\theta\in\Theta$,

$$\forall\delta, \sup_{\theta\in\Theta} R(\theta,\delta_\pi) = R(\theta,\delta_\pi) = \Lambda(\pi,\delta_\pi) \leq \Lambda(\pi,\delta) = \mathbb{E}_{\theta\sim\pi} R(\theta,\delta) \leq \sup_{\theta} R(\theta,\delta) \tag{255}$$

where the second equality comes from the fact that

$$\Lambda(\pi,\delta_\pi) = \mathbb{E}_{\theta\sim\pi} R(\theta,\delta_\pi) = \int_\Theta R(\theta,\delta_\pi)\cdot\pi(\theta)\,d\theta = R(\theta,\delta_\pi)\cdot\int_\Theta \pi(\theta)\,d\theta = R(\theta,\delta_\pi) \tag{256}$$

since $R(\theta,\delta_\pi)$ actually does not change when θ changes. □

Example: $X_1,...,X_n \sim B(1,\theta)$ Bernoulli, with prior $\theta\sim Beta(\alpha,\beta)$, so the posterior is given by

$$\pi(\theta|x) \propto \theta^S(1-\theta)^{n-S}\cdot\theta^{\alpha-1}(1-\theta)^{\beta-1} = \theta^{S+\alpha-1}(1-\theta)^{n-1+\beta-S}, \theta\in(0,1) \tag{257}$$

where $S=\sum_{i=1}^n X_i \sim B(n,\theta)$ denotes the sum of sample realizations. This means that $\theta|_X \sim Beta(S+\alpha, n+\beta-S)$. As a result, if SEL is taken as loss, the Bayes rule is just the posterior mean

$$\delta_\pi = \frac{S+\alpha}{n+\beta+\alpha} \tag{258}$$

however, if we try to calculate the risk of such Bayes estimator

$$\begin{aligned}
R(\theta,\delta_\pi) &= \mathbb{E}_{X\sim p_\theta}(\theta-\delta_\pi(X))^2 && (259)\\
&= \mathbb{E}_{S\sim B(n,\theta)}\left(\theta - \frac{S+\alpha}{n+\alpha+\beta}\right)^2 && (260)\\
&= \sum_{k=0}^n \binom{n}{k}\theta^k(1-\theta)^{n-k}\left(\theta-\frac{k+\alpha}{n+\alpha+\beta}\right)^2 && (261)\\
&= \theta^2 - \frac{2\theta}{n+\alpha+\beta}(n\theta+\alpha) + \frac{1}{(n+\alpha+\beta)^2}[n\theta(1-\theta)+n^2\theta^2+2\alpha n\theta+\alpha^2] && (262)\\
&= \frac{[(\alpha+\beta)^2-n]\theta^2+[n-2\alpha(\alpha+\beta)]\theta+\alpha^2}{(n+\alpha+\beta)^2} && (263)
\end{aligned}$$

so it's also minimax rule if

$$\begin{cases} (\alpha+\beta)^2 = n \\ 2\alpha(\alpha+\beta) = n \end{cases} \tag{264}$$

since zeroing out the coefficients of θ^2, θ ensures that the risk function is constant in θ and by the theorem above, such Bayes rule must be minimax rule. Solve out to see

$$\alpha = \beta = \frac{\sqrt{n}}{2} \tag{265}$$

so $\delta_0(X) = \frac{S+\frac{\sqrt{n}}{2}}{n+\sqrt{n}}$ is the minimax rule with minimax risk $R(\theta, \delta_0) = \frac{\alpha^2}{(n+\alpha+\beta)^2} = \frac{1}{4(\sqrt{n}+1)^2}$. This is an example how we can get minimax rule from Bayes rule by using this theorem.

Remark. *Notice the following simple facts helps reduce calculations.*

$$\sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} = 1 \tag{266}$$

$$\sum_{k=0}^{n} k \binom{n}{k} \theta^k (1-\theta)^{n-k} = \mathbb{E}S = n\theta \tag{267}$$

$$\sum_{k=0}^{n} k^2 \binom{n}{k} \theta^k (1-\theta)^{n-k} = \mathbb{E}S^2 = n\theta(1-\theta) + (n\theta)^2 \tag{268}$$

Now let's consider $\delta_1(X) = \overline{X} = \frac{S}{n}$, the most frequently used estimator in estimating Bernoulli parameter. Let's compute its risk also under SEL

$$R(\theta, \delta_1) = \mathbb{E}_{S\sim B(n,\theta)}(\theta - \overline{X})^2 \tag{269}$$

$$= \sum_{k=0}^{n} \binom{n}{k} \theta^k (1-\theta)^{n-k} \left(\theta - \frac{k}{n}\right)^2 \tag{270}$$

$$= \theta^2 - \frac{2\theta}{n} n\theta + \frac{1}{n^2}[n\theta(1-\theta) + n^2\theta^2] \tag{271}$$

$$= \frac{\theta(1-\theta)}{n} \tag{272}$$

so we will see that $\sup_{\theta\in\Theta} R(\theta, \delta_1) = \frac{1}{4n}$, with sup taken at $\theta = \frac{1}{2}$. On the other hand, $\sup_{\theta\in\Theta} R(\theta, \delta_0) = \frac{1}{4(\sqrt{n}+1)^2} \leq \frac{1}{4n}$, which is consistent with the definition of minimax rule.

However, by comparing those two decision rules in a more careful way, we see that $R(\theta, \delta_0) < R(\theta, \delta_1)$ if and only if $\theta \in \left(\frac{1}{2} - d, \frac{1}{2} + d\right)$ with $d = \frac{\sqrt{2\sqrt{n}+1}}{2(\sqrt{n}+1)} \to 0$ $(n \to \infty)$ so the advantage interval of δ_0 against δ_1 shrinks to an

empty set when the sample size is large enough. We can also see that

$$\frac{\sup_{\theta\in\Theta} R(\theta,\delta_1)}{\sup_{\theta\in\Theta} R(\theta,\delta_0)} = \frac{\frac{\theta(1-\theta)}{n}}{\frac{1}{4(\sqrt{n}+1)^2}} \tag{273}$$

$$\to 4\theta(1-\theta) \le 1 \ (n\to\infty) \tag{274}$$

so the limit is exactly 1 when $\theta = \frac{1}{2}$.

Remark. *This example shows us that* ***minimax rule is not always the best rule to take in practice****, it's just hedging against the maximum possible risk. As we can see, in the asymptotic scheme, $\delta_1(X)$ behaves better than $\delta_0(X)$ for $\forall\theta\in\Theta$.*

Theorem 15. ***(A minor extension for the theorem above)*** *Let π_k be a sequence of prior with corresponding Bayes rule δ_{π_k} and let $\Lambda(\pi_k,\delta_{\pi_k})\to\Lambda$ $(k\to\infty)$. Let δ_0 be a rule not necessarily Bayes but with constant risk function equal to Λ, then δ_0 is minimax.*

Proof.

$$\forall\delta,\forall k,\sup_\theta R(\theta,\delta) \ge \mathbb{E}_{\theta\sim\pi_k} R(\theta,\delta) = \Lambda(\pi_k,\delta) \ge \Lambda(\pi_k,\delta_{\pi_k}) \tag{275}$$

set $k\to\infty$ to see

$$\forall\delta,\sup_\theta R(\theta,\delta) \ge \Lambda = \sup_\theta R(\theta,\delta_0) \tag{276}$$

□

Remark. *This extension works if one have a sequence of prior and Bayes rule with the average risk under prior converging to a constant risk function. However, one may immediately find that we can still extend this theorem a little bit by replacing the limit with* ***upper limit***

$$\limsup_{k\to\infty} \Lambda(\pi_k,\delta_{\pi_k}) = \Lambda \tag{277}$$

and that ***the upper limit dominates the risk function of δ_0 for all possible θ***

$$\forall\theta\in\Theta, R(\theta,\delta_0)\le\Lambda \tag{278}$$

and the same proof still holds.

Example: $X_1,...,X_n\sim N(\theta,\sigma^2)$ with σ known. Consider a sequence of priors $\pi_k = N(0,\tau_k^2)$ with $\tau_k^2\to\infty$ and τ_k known. Under SEL, we know that the Bayes rule δ_{π_k} is just the posterior mean,

$$\pi_k(\theta|x) \propto e^{-\frac{\theta^2}{2\tau_k^2}}\cdot e^{-\frac{\sum_{i=1}^n (x_i-\theta)^2}{2\sigma^2}} \propto e^{-\left(\frac{1}{2\tau_k^2}+\frac{n}{2\sigma^2}\right)\left(\theta-\frac{n\bar{x}}{\frac{\sigma^2}{\tau_k^2}+n}\right)^2} \tag{279}$$

so $\delta_{\pi_k}(X) = \frac{n\overline{X}}{\frac{\sigma^2}{\tau_k^2}+n}$ and

$$\Lambda(\pi, \delta_{\pi_k}) = Var_{\theta\sim\pi_k(\cdot|x)}(\theta) = \frac{1}{\frac{1}{\tau_k^2} + \frac{n}{\sigma^2}} \to \frac{\sigma^2}{n} \ (k \to \infty) \tag{280}$$

note that for $\delta_0(X) = \overline{X}, \mathbb{E}\delta_0(X) = \theta$, so under SEL

$$R(\theta, \delta_0) = \mathbb{E}_{X\sim p_\theta} L(\theta, \delta_0(X)) = Var(\overline{X}) = \frac{\sigma^2}{n} \tag{281}$$

which is constant in θ, so according to the theorem above, it's minimax. So we have proved that **for Gaussian location model, the sample mean is always a minimax rule**.

Applications of Decision Theory in Bayesian Statistics

Credible Interval Estimation

In the Bayesian setting, if one gets the posterior likelihood $\pi(\theta|x)$ and wants to do interval estimation, one wants to find l, u such that $\mathbb{P}(l < \theta < u|x) = \int_l^u \pi(\theta|x)\, d\theta = 1 - \alpha$, called a **credible interval of confidence level** $1 - \alpha$. However, there are infinitely many choices of l, u such that this property holds, and those selections of l, u are typically not as good as each other. For example, if the posterior distribution is $N(0, 1)$ and now $\alpha = \frac{1}{2}$, one won't consider $l = -\infty, u = 0$ or $l = 0, u = +\infty$ to be the optimal choice of the credible interval but tends to select a symmetric interval w.r.t. 0, i.e. $l = -u$ since the likelihood of standard Gaussian is highest at 0, meaning that a standard Gaussian random variable has the highest probability of taking values in a neighborhood of 0 compared with other real numbers.

Naturally, the condition $(l_k, u_k) = \{\theta : \pi(\theta|x) \geq k\}$ is added such that the posterior likelihood at both endpoints is larger than k and we find l_k, u_k in this set for fixed k such that it forms a **highest posterior density (HPD) credible interval**. For HPD credible interval, the likelihood at each endpoint is the same so probability mass is distributed in the place where the posterior likelihood is the highest in priority. In some sense, HPD credible interval is the best interval estimation under the Bayesian setting we can make since it tries to cover the range where the parameter is the most likely to appear. If the posterior likelihood is unimodal and symmetric around some value c, then the HPD credible interval will also be symmetric around c. If the posterior likelihood is multimodal, then the HPD credible interval may be the union of several disjoint intervals.

Example: $X_1, ..., X_n \sim N(\theta, \sigma^2)$ *i.i.d.* with $\theta \sim N(\mu, \tau^2)$ and σ^2 known. Then

$$\pi(\theta|x) \propto e^{-\frac{\sum_i (x_i - \theta)^2}{2\sigma^2}} \cdot e^{-\frac{(\theta-\mu)^2}{2\tau^2}} \propto e^{-(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2})\left[\theta - \frac{\frac{\mu}{\tau^2} + \frac{\sum_i x_i}{\sigma^2}}{2(\frac{n}{2\sigma^2} + \frac{1}{2\tau^2})}\right]^2} \tag{282}$$

so the posterior distribution is $N(\mu_0, \tau_0^2)$ with $\mu_0 = \frac{\frac{\mu}{\tau^2} + \frac{\sum_i x_i}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \tau_0^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$. So 95% HPD credible interval is $(\mu_0 - 1.96\tau_0, \mu_0 + 1.96\tau_0)$.

Classification

Now we have $\Theta = \{\theta_1, ..., \theta_k\}$ only k possible options of the parameter θ and $D = \{d_1, ..., d_k\}$ only k possible decisions to take with d_i as the decision to take θ_i as the estimated value of the parameter. Consider 0-1 loss function $L(\theta_i, d_j) = 1 - \delta_{i,j}$ and the prior distribution of θ is given by the probability masses $\pi(\theta_i)$ $(i = 1, 2, ..., k)$ adding up to 1 so by Bayes formula

$$\pi(\theta_i|x) = \frac{p_{\theta_i}(x)\pi(\theta_i)}{\sum_{j=1}^k p_{\theta_j}(x)\pi(\theta_j)} \tag{283}$$

let's try to compute the Bayes rule for this classification problem (k parameter values are considered k different categories the sample might come from).

The expected posterior loss is

$$l(d_j, x) = \mathbb{E}_{\theta\sim\pi(\cdot|x)} L(\theta_i, d_j) \tag{284}$$

$$= \sum_{p=1}^{k} \pi(\theta_p|x) \cdot L(\theta_p, d_j) \tag{285}$$

$$= 1 - \pi(\theta_j|x) \tag{286}$$

and consider minimizing the expected posterior loss

$$d_0(x) = \arg\min_{d\in D} l(d, x) \tag{287}$$

$$= \arg\max_{d\in D} \{\pi(\theta_1|x), ..., \pi(\theta_k|x)\} \tag{288}$$

so $d_0(x)$ is equal to d_j where the correspondent $\pi(\theta_j|x) = \max\{\pi(\theta_1|x), ..., \pi(\theta_k|x)\}$. This tells us that **the Bayes rule δ_π is to pick the decision to take θ_j that maximizes the posterior likelihood**. This is called **max a posteriori (MAP) estimator** and it's the Bayes rule for the classification problem.

Hypothesis Testing

Now let's consider the simple v.s. simple hypothesis testing problem $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ with $D = \{d_0, d_1\}$, a special case for the classification problem when $k = 2$. Now we follow the Bayes rule δ_π which is just the MAP rule, we shall take decision d_0 iff $\pi(\theta_0|x) > \pi(\theta_1|x)$, i.e. $\pi(\theta_0) \cdot p_{\theta_0}(x) > \pi(\theta_1) \cdot p_{\theta_1}(x)$, i.e. $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < \frac{\pi(\theta_0)}{\pi(\theta_1)} \overset{def}{=} k(\pi)$ by comparing the likelihood ratio with the prior ratio so it's actually a **likelihood ratio test**

$$\forall x \in \mathscr{X}, \delta_\pi(x) = \begin{cases} d_0 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < k(\pi) \\ d_1 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \geq k(\pi) \end{cases} \tag{289}$$

Remark. *One can use more general loss function to specify different penalty for type I and type II error. For example, set $L(\theta_1, d_0) = a, L(\theta_0, d_1) = b$ so the expected posterior loss is*

$$l(d, x) = \mathbb{E}_{\theta\sim\pi(\cdot|x) L(\theta,d)} \tag{290}$$

$$= \pi(\theta_0|x) \cdot L(\theta_0, d) + \pi(\theta_1|x) \cdot L(\theta_1, d) \tag{291}$$

so

$$l(d_0, x) = a \cdot \pi(\theta_1|x), l(d_1, x) = b \cdot \pi(\theta_0|x) \tag{292}$$

and the Bayes rule (MAP rule) δ_π now is that

$$\forall x \in \mathscr{X}, \delta_\pi(x) = \begin{cases} d_0 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} < \frac{b}{a}k(\pi) \\ d_1 & \frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \geq \frac{b}{a}k(\pi) \end{cases} \tag{293}$$

Now we already have **the Bayes test (MAP test)** and we think of building up a minimax test. According to the theorem we have proved, if a Bayes test (w.r.t. some prior) has constant risk (not depend on θ), then it must be minimax. Let's denote δ_k as the decision rule that we reject H_0 iff $\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} \geq k$ (the likelihood ratio test) that we have proved to be the Bayes test under some prior distribution π such that $k = \frac{\pi(\theta_0)}{\pi(\theta_1)}$.

Let's first compute the risk function at θ_0, θ_1

$$\begin{aligned} R(\theta_0, \delta_k) &= \mathbb{E}_{X \sim p_{\theta_0}} L(\theta_0, \delta_k(X)) && (294) \\ &= \mathbb{P}_{\theta_0}\left(\delta_k(X) = d_1\right) && (295) \\ &\overset{def}{=} \alpha && (296) \end{aligned}$$

is the probability of rejecting H_0 while H_0 is true, so it's the probability of Type I error. Similarly,

$$\begin{aligned} R(\theta_1, \delta_k) &= \mathbb{E}_{X \sim p_{\theta_1}} L(\theta_1, \delta_k(X)) && (297) \\ &= \mathbb{P}_{\theta_1}\left(\delta_k(X) = d_0\right) && (298) \\ &\overset{def}{=} \beta && (299) \end{aligned}$$

is the probability of rejecting H_1 while H_1 is true, so it's the probability of Type II error. So we see that **the risk is just the probability of Type I and Type II error**.

By setting $\alpha = \beta$, one can solve out the value of k (or the prior π that decides such k) that provides a constant risk function so the Bayes test likelihood ratio test δ_k must also be a minimax test.

Let's show an example. $X_1, ..., X_n \sim N(\theta, \sigma^2)$ with σ^2 known, $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ where $\theta_1 > \theta_0$ so it's a simple v.s. simple hypothesis testing. Now $\pi(\theta_0) = \pi_0, \pi(\theta_1) = \pi_1$ are given. We first figure out the Bayes test under such prior distribution. Let's calculate likelihood ratio

$$\frac{p_{\theta_1}(x)}{p_{\theta_0}(x)} = e^{\frac{\theta_1 - \theta_0}{\sigma^2}\sum_{i=1}^n x_i} \cdot e^{\frac{n(\theta_0^2 - \theta_1^2)}{2\sigma^2}} \tag{300}$$

so the rejection region of H_0 is

$$e^{\frac{\theta_1 - \theta_0}{\sigma^2}\sum_{i=1}^n x_i} \cdot e^{\frac{n(\theta_0^2 - \theta_1^2)}{2\sigma^2}} \geq \frac{\pi_0}{\pi_1} \tag{301}$$

which is when $\overline{x} \geq \frac{\sigma^2}{n(\theta_1 - \theta_0)}\left[\log\frac{\pi_0}{\pi_1} - \frac{n(\theta_0^2 - \theta_1^2)}{2\sigma^2}\right]$. So **the Bayes test under the given prior is the likelihood**

ratio test below

$$\forall x \in \mathscr{X}, \delta_\pi(x) = \begin{cases} d_0 & \overline{x} < \frac{\sigma^2}{n(\theta_1-\theta_0)}\left[\log\frac{\pi_0}{\pi_1} - \frac{n(\theta_0^2-\theta_1^2)}{2\sigma^2}\right] \\ d_1 & \overline{x} \geq \frac{\sigma^2}{n(\theta_1-\theta_0)}\left[\log\frac{\pi_0}{\pi_1} - \frac{n(\theta_0^2-\theta_1^2)}{2\sigma^2}\right] \end{cases} \tag{302}$$

now to figure out the minimax test, we know that each likelihood ratio test δ_k is a Bayes test under some prior

$$\forall x \in \mathscr{X}, \delta_k(x) = \begin{cases} d_0 & \overline{x} < k \\ d_1 & \overline{x} \geq k \end{cases} \tag{303}$$

so we have to find the k such that the Type I error and Type II error probability matches, i.e.

$$\mathbb{P}_{\theta_0}\left(\overline{X} \geq k\right) = \mathbb{P}_{\theta_1}\left(\overline{X} < k\right) \tag{304}$$

$$1 - \Phi\left(\frac{k-\theta_0}{\frac{\sigma}{\sqrt{n}}}\right) = \Phi\left(\frac{k-\theta_1}{\frac{\sigma}{\sqrt{n}}}\right) \tag{305}$$

$$-\frac{k-\theta_0}{\frac{\sigma}{\sqrt{n}}} = \frac{k-\theta_1}{\frac{\sigma}{\sqrt{n}}} \tag{306}$$

so $k = \frac{\theta_0+\theta_1}{2}$. **The minimax test is given by the following likelihood ratio test**

$$\forall x \in \mathscr{X}, \delta_k(x) = \begin{cases} d_0 & \overline{x} < \frac{\theta_0+\theta_1}{2} \\ d_1 & \overline{x} \geq \frac{\theta_0+\theta_1}{2} \end{cases} \tag{307}$$

and does not depend on the given prior π. It's easy to see that when $\pi_0 = \pi_1 = \frac{1}{2}$, those two tests are the same.

Now let's look at **composite v.s. composite** hypothesis testing $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$ and Θ_0, Θ_1 are disjoint subsets of parameter space Θ. By the same argument presented above, the expected posterior loss is

$$l(d_0, x) = \mathbb{E}_{\theta\sim\pi(\cdot|x)} L(\theta, d_0) = \pi(\theta \in \Theta_1|x) \cdot L(\theta_1, d_0) = \mathbb{P}\left(\theta \in \Theta_1|x\right) \tag{308}$$

$$l(d_1, x) = \mathbb{P}\left(\theta \in \Theta_0|x\right) \tag{309}$$

so one just needs to compute posterior probability $\mathbb{P}\left(\theta \in \Theta_0|x\right), \mathbb{P}\left(\theta \in \Theta_1|x\right)$ and reject H_0 iff $\mathbb{P}\left(\theta \in \Theta_1|x\right) > \mathbb{P}\left(\theta \in \Theta_0|x\right)$, it's still an MAP estimator.

Remark. *If Θ_0, Θ_1 forms a partition of the parameter space, $\mathbb{P}\left(\theta \in \Theta_0|x\right) + \mathbb{P}\left(\theta \in \Theta_1|x\right) = 1$ so we can reject H_0 when $\mathbb{P}\left(\theta \in \Theta_1|x\right) > \frac{1}{2}$.*

There's an alternative for the composite v.s. composite hypothesis testing: define posterior odds ratio $\frac{\mathbb{P}(\theta\in\Theta_1|x)}{\mathbb{P}(\theta\in\Theta_0|x)}$ and prior odds ratio $\frac{\mathbb{P}(\theta\in\Theta_1)}{\mathbb{P}(\theta\in\Theta_1)}$ then the **Bayes factor** is defined as posterior odds ratio over prior odds ratio $\frac{\frac{\mathbb{P}(\theta\in\Theta_1|x)}{\mathbb{P}(\theta\in\Theta_0|x)}}{\frac{\mathbb{P}(\theta\in\Theta_1)}{\mathbb{P}(\theta\in\Theta_1)}}$. We reject H_0 if Bayes factor is too large, which means that the observation of the sample realizations is strongly affecting the prior so that the posterior is against H_0.

Example: $X_1, ..., X_n \sim N(\theta, \sigma^2)$ with σ^2 known and the prior $\pi(\theta) = N(\mu, \tau^2)$. We have hypothesis testing $H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$ with the posterior distribution as $N(\mu_0, \tau_0^2)$ with $\mu_0 = \frac{\frac{\mu}{\tau^2} + \frac{\sum_i x_i}{\sigma^2}}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}, \tau_0^2 = \frac{1}{\frac{n}{\sigma^2} + \frac{1}{\tau^2}}$. According to the Bayes test, since $\Theta_0 \cup \Theta_1 = \Theta$, we shall reject H_0 iff $\mathbb{P}(\theta > \theta_0 | x) > \frac{1}{2}$, i.e. $\theta_0 < \mu_0$, so the Bayes test is given by

$$\forall x \in \mathscr{X}, \delta_\pi(x) = \begin{cases} d_0 & \overline{x} \leq \frac{\theta_0 \sigma^2 \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) - \frac{\mu}{\tau^2}}{n} \\ d_1 & \overline{x} > \frac{\theta_0 \sigma^2 \left(\frac{n}{\sigma^2} + \frac{1}{\tau^2}\right) - \frac{\mu}{\tau^2}}{n} \end{cases} \tag{310}$$

Remark. *If we test simple v.s. composite hypothesis $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ and the prior is taken as a continuous distribution, then $\mathbb{P}(\theta = \theta_0) = 0$, so we may form the prior as a mixture that $\pi(\theta) = \frac{1}{3}\mathbb{I}_{\theta=\theta_0} + \frac{2}{3}\pi_0(\theta)$ for some continuous distribution π_0.*

Remark. ***In simple v.s. simple hypothesis testing, if for decision rule δ, risk $R(\theta_0, \delta) = R(\theta_1, \delta)$, then the test is always minimax regardless of whether the test is Bayes.***

Sometimes for discrete population distribution, one needs to use randomized decision rule to build up a minimax test (since discrete distribution has point mass). The example is $X_1, ..., X_5 \sim B(1, \theta)$ with hypothesis testing $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$ where $\theta_0 = \frac{1}{2}, \theta_1 = \frac{3}{4}$ and we want to build a minimax test. Then we shall consider the decision rule $\delta_{k,p}$ that rejects H_0 if $\sum_i x_i > k$ and rejects H_1 if $\sum_i x_i < k$ and rejects H_0 with probability p if $\sum_i x_i = k$.

Estimation Theory

Under general setting, we have $X_1, ..., X_n$ as *i.i.d.* samples from population p_θ and $\theta \in \Theta$ is unknown so we want to estimate such θ.

Method of Moments

Match sample and population moments to form estimates of θ, i.e. $\mathbb{E}_{X\sim p_\theta} X^k = \frac{1}{n}\sum_{i=1}^n X_i^k$ $(k = 1, 2, ...)$. Set up as many equations as needed to determine all parameters and $\hat{\theta}_{MM}$ is solved out based on those equations as a function of $X_1, ..., X_n$.

The **generalized method of moments (GMM)** applied for $\Theta \subset \mathbb{R}^p$ considers $m \geq p$ functions $g_j(x, \theta)$ $(j = 1, 2, ..., m)$ with $\mathbb{E}g_j(X, \theta) = 0$. Define

$$\hat{g_j}(X, \theta) = \frac{1}{n}\sum_{i=1}^n g_j(X_i, \theta) \tag{311}$$

as sample mean of the transformed random variables, with $\hat{g}(X, \theta)$ as the vector whose j-th component is $\hat{g_j}(X, \theta)$. So the estimator is

$$\hat{\theta} = \arg\min_\theta [\hat{g}(X, \theta)]^T W \hat{g}(X, \theta) \tag{312}$$

for some specified weight matrix $W \in \mathbb{R}^{m\times m}$ (when the weight matrix is carefully chosen, GMM estimator is consistent and asymptotically normal).

Remark. *In simple method of moments, we are using* $m = 1, g_1(X, \theta) = X - \mathbb{E}X_1$ *so* $\mathbb{E}g_1(X_1, \theta) = 0$ *and* $\hat{g_1}(X, \theta) = \frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X_1$ *so* $\hat{\theta} = \arg\min_\theta W(\overline{X} - \mathbb{E}X_1)^2$ *is the estimator that matches the population first moment with the sample first moment. Through this example, we may seen that typically* g_j *is taken as the function that* ***exhibits the difference of a certain quantity between the population and the sample****.*

Example: $X_1, ..., X_n$ are from Cauchy distribution $C(\mu, \sigma^2)$ with density

$$f(x; \mu, \sigma) = \frac{1}{\pi\sigma[1 + (\frac{x-\mu}{\sigma})^2]} \tag{313}$$

although Cauchy distribution does not have moments so we cannot apply the method of moments, it has quantiles so we can apply GMM. Let's first compute the population quantiles to find

$$Q_2 = \mu \tag{314}$$

since the density is symmetric w.r.t. μ. To compute the third quantile Q_3, compute

$$\int_{Q_3}^\infty \frac{1}{\pi\sigma[1 + (\frac{x-\mu}{\sigma})^2]}\, dx = \frac{1}{4} \tag{315}$$

change variables $u = \frac{x-\mu}{\sigma}$ to get

$$\int_{\frac{Q_3-\mu}{\sigma}}^{\infty} \frac{1}{\pi(1+u^2)}\, du = \frac{1}{4} \tag{316}$$

so $\frac{Q_3-\mu}{\sigma} = 1, Q_3 = \mu + \sigma$ and similarly $Q_1 = \mu - \sigma$. Now let's take g_j as the difference between population and sample quantiles so

$$m = 3, g_j(X, \mu, \sigma) = q_j(X) - Q_j \ (j = 1, 2, 3) \tag{317}$$

where $q_j(X)$ is the j-th sample quantile of the sample $X_1, ..., X_n$. Now $\mathbb{E}g_j(X, \mu, \sigma) = 0$ so by taking $W = I_3$ as the trivial weight matrix, we see that the GMM estimator is given by

$$(\hat{\mu}_{\text{GMM}}, \hat{\sigma}_{\text{GMM}}) = \arg\min_{\mu,\sigma} \sum_{j=1}^{3} [q_j(X) - Q_j]^2 \tag{318}$$

Minimum Chi-Square Method for Grouped data

If we have k disjoint intervals $I_1, ..., I_k$ (these intervals can be single point sets) and the observed frequency of the sample $X_1, ..., X_n$ taking values in I_i is f_i (so $f_1 + ... + f_k = n$) and the expected frequency from the population is $np_i(\theta)$ (so $p_1(\theta) + ... + p_k(\theta) = 1$), then consider

$$\chi^2 = \sum_{i=1}^{n} \frac{(f_i - np_i(\theta))^2}{np_i(\theta)} \tag{319}$$

the same statistic used in Pearson's chi-square test and is known to follow $\chi^2(k-1)$ distribution asymptotically. As a result, we can find θ that minimizes this χ^2 statistic so that it's the hardest to reject H_0 that the population is from a multinomial distribution with probability masses $p_1(\theta), ..., p_k(\theta)$. This is called **minimum χ^2 method** for estimating parameter θ.

Example: $X_1, ..., X_n \sim B(1, \theta)$ and let $X = \sum_{i=1}^{n} X_i$ so we observe X of them are 1 and $n - X$ of them are 0. Consider $\chi^2 = \frac{(X-n\theta)^2}{X} + \frac{(n-X-n(1-\theta))^2}{n-X}$ so $\hat{\theta} = \arg\min_\theta \chi^2 = \overline{X}$.

Maximum Likelihood Estimator (MLE)

Here we omit the calculations of MLE but just discuss the properties of MLE denoted $\hat{\theta}$.

Theorem 16. ***(MLE and Sufficiency)*** *If there exists any sufficient statistic T, MLE must be a function of T.*

Proof. By factorization theorem, if T is sufficient statistic, the joint likelihood $p_\theta(x) = g_\theta(T(x)) \cdot h(x)$ so $\log p_\theta(x) = \log g_\theta(T(x)) + \log h(x)$ and take derivative w.r.t. θ to see $\frac{\partial}{\partial\theta} \log p_\theta(x) = \frac{\partial}{\partial\theta} \log g_\theta(T(x))$ so maximizing $p_\theta(x)$ is equivalent to maximizing $g_\theta(T(x))$ and we see that the dependence of $\hat{\theta}(x)$ on x can only be through $T(x)$, $\hat{\theta}$ has to be a function of T. □

Theorem 17. ***(Functional Invariance)*** *For any Borel function g, $g(\hat{\theta})$ is the MLE of $g(\theta)$.*

Proof. Let $\eta = g(\theta)$ and $A_\eta = \{\theta : g(\theta) = \eta\}$ be the orbit of $g(\theta)$ taking value η, $\{A_\eta\}_{\eta\in\mathbb{R}}$ is a partition of Θ and we denote $\sup_{\theta\in A_\eta} p_\theta(x) = M(\eta)$ as the induced likelihood of η, then $\sup_{\theta\in\Theta} p_\theta(x) = \sup_{\eta\in\mathbb{R}}\sup_{\theta\in A_\eta} p_\theta(x) = \sup_{\eta\in\mathbb{R}} M(\eta)$.

Now denote $g(\hat{\theta}) = \eta^*, \hat{\theta} \in A_{\eta^*}$ so

$$p_{\hat{\theta}}(x) \le \sup_{\theta\in A_{\eta^*}} p_\theta(x) = M(\eta^*) \le \sup_{\eta\in\mathbb{R}} M(\eta) = \sup_{\theta\in\Theta} p_\theta(x) = p_{\hat{\theta}}(x) \tag{320}$$

and we see that $\sup_{\theta\in A_{\eta^*}} p_\theta(x) = \sup_{\theta\in\Theta} p_\theta(x)$, this tells us that the near optimal θ lies in $\overline{A_{\eta^*}}$ which is the orbit of $g(\theta)$ taking value η^*. As a result, MLE of $g(\theta)$ is $\eta^* = g(\hat{\theta})$. □

Note that **MLE may not exist**. Consider $X_1, ..., X_n$ from population with a mixture distribution $p_{\mu,\sigma^2}(x) = (1-\varepsilon)N(\mu, 1) + \varepsilon N(\mu, \sigma^2)$ where $\varepsilon \in (0, 1)$ is a fixed number (the mixture density is the weighted average of two Gaussian densities). Compute the joint likelihood

$$p_{\mu,\sigma^2}(x) = \prod_{i=1}^n \left[(1-\varepsilon)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} + \varepsilon\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-\mu)^2}\right] \tag{321}$$

now consider taking $\hat{\mu} = x_1$ so the joint likelihood becomes

$$\left[(1-\varepsilon)\frac{1}{\sqrt{2\pi}} + \frac{1}{\sqrt{2\pi}\sigma}\right]\prod_{i=2}^n \left[(1-\varepsilon)\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-x_1)^2} + \varepsilon\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x_i-x_1)^2}\right] \tag{322}$$

now by setting $\sigma \to 0$, we see that such joint likelihood explodes. As a result, we have shown that $\sup_{\mu\in\mathbb{R},\sigma>0} p_{\mu,\sigma^2}(x) = +\infty$ so the MLE does not exist.

MLE is generally not unbiased.

MLE may not be unique. Consider population distribution $U(\theta, \theta+1)$ so the joint likelihood is

$$p_\theta(x) = \mathbb{I}_{x_{(1)}\ge\theta, x_{(n)}\le\theta+1} \tag{323}$$

and it's obvious that $\forall\hat{\theta} \in [X_{(n)} - 1, X_{(1)}]$ is an MLE (any weighted average of $X_{(n)} - 1, X_{(1)}$ works).

MLE may need numerical solver. Consider population distribution as Cauchy $C(\theta, 1)$ so

$$p_\theta(x) = \prod_{i=1}^n \frac{1}{\pi[1 + (x_i - \theta)^2]} \tag{324}$$

$$= \pi^{-n}\prod_{i=1}^n \frac{1}{1 + (x_i - \theta)^2} \tag{325}$$

which can only be maximized w.r.t. θ numerically.

Asymptotic Normality of MLE, Fisher Information

Let's first provide the definition of **Fisher information (FI)** as

$$I(\theta) = \mathbb{E}_{X\sim p_\theta}\left(\frac{\partial}{\partial\theta}\log p_\theta(X)\right)^2 \tag{326}$$

note that here the variable x in the joint likelihood is replaced with sample X as a random variable.

For example, $X \sim B(1,\theta)$ then $\log p_\theta(x) = x\log\theta + (1-x)\log(1-\theta)$ and $\frac{\partial}{\partial\theta}\log p_\theta(X) = \frac{X}{\theta} - \frac{1-X}{1-\theta} = \frac{X-\theta}{\theta(1-\theta)}$ so $I(\theta) = \frac{\mathbb{E}(X-\theta)^2}{\theta^2(1-\theta)^2} = \frac{1}{\theta(1-\theta)}$.

Remark. *Now we know why FI represents the quantity of information. When $\theta = 0, 1$, $I(\theta) = \infty$ so there is infinitely much information since then X is a deterministic number with no randomness. When $\theta = \frac{1}{2}$, $I(\theta) = 4$ reaches its minimum since the uncertainty contained in X is the maximal, there is equal chance for X to be 0 or 1. Actually FI is closely related to concepts like entropy and information divergence.*

Theorem 18. ***(Property of FI)*** *Under certain regularity condition,*

$$\mathbb{E}_{X\sim p_\theta}\frac{\partial\log p_\theta(X)}{\partial\theta} = 0, I(\theta) = -\mathbb{E}_{X\sim p_\theta}\frac{\partial^2\log p_\theta(X)}{\partial^2\theta}, I_n(\theta) = nI(\theta) \tag{327}$$

where $I_n(\theta)$ is the FI on observing n i.i.d. samples $X_1, ..., X_n$.

Proof.

$$\begin{aligned}
\mathbb{E}_{X\sim p_\theta}\frac{\partial\log p_\theta(X)}{\partial\theta} &= \int_{\mathbb{R}}\frac{\partial\log p_\theta(x)}{\partial\theta}p_\theta(x)\,dx &(328)\\
&= \int_{\mathbb{R}}\frac{\partial}{\partial\theta}p_\theta(x)\,dx &(329)\\
&= \frac{\partial}{\partial\theta}\int_{\mathbb{R}}p_\theta(x)\,dx &(330)\\
&= \frac{\partial}{\partial\theta}1 = 0 &(331)
\end{aligned}$$

here we are interchanging the partial derivative w.r.t. θ and the integral w.r.t. x. By dominated convergence theorem, this can be done if there exists g such that $\exists\delta > 0, \forall 0 < |h| < \delta, \left|\frac{p_{\theta+h}(x)-p_\theta(x)}{h}\right| \le g(\theta, x)$ and $\int_{\mathbb{R}} g(\theta,x)\,dx < \infty$. Simply speaking, there exists g such that $\exists\delta > 0, \forall\theta' \in (\theta-\delta, \theta+\delta), \left|\frac{\partial p_\theta(x)}{\partial\theta}\Big|_{\theta=\theta'}\right| \le g(\theta, x)$ and $\int_{\mathbb{R}} g(\theta,x)\,dx < \infty$.

$$-\mathbb{E}_{X\sim p_\theta}\frac{\partial^2 \log p_\theta(X)}{\partial^2\theta} = -\int_{\mathbb{R}} \frac{\partial^2 \log p_\theta(x)}{\partial^2\theta} p_\theta(x)\,dx \tag{332}$$

$$= -\int_{\mathbb{R}} \frac{\partial}{\partial\theta}\left(\frac{\frac{\partial}{\partial\theta}p_\theta(x)}{p_\theta(x)}\right) p_\theta(x)\,dx \tag{333}$$

$$= -\int_{\mathbb{R}} \frac{\frac{\partial^2}{\partial\theta^2}p_\theta(x)\cdot p_\theta(x) - \left(\frac{\partial}{\partial\theta}p_\theta(x)\right)^2}{[p_\theta(x)]^2} p_\theta(x)\,dx \tag{334}$$

$$= -\int_{\mathbb{R}} \frac{\partial^2}{\partial\theta^2}p_\theta(x)\,dx + \int_{\mathbb{R}}\left(\frac{\frac{\partial}{\partial\theta}p_\theta(x)}{p_\theta(x)}\right)^2 p_\theta(x)\,dx \tag{335}$$

$$= -\frac{\partial^2}{\partial\theta^2}\int_{\mathbb{R}} p_\theta(x)\,dx + \mathbb{E}_{X\sim p_\theta}\left(\frac{\frac{\partial}{\partial\theta}p_\theta(X)}{p_\theta(X)}\right)^2 \tag{336}$$

$$= \mathbb{E}_{X\sim p_\theta}\left(\frac{\partial \log p_\theta(X)}{\partial\theta}\right)^2 = I(\theta) \tag{337}$$

by changing the derivative with integral once more.

For the last property,

$$I_n(\theta) = -\mathbb{E}_{X\sim p_\theta}\frac{\partial^2 \log p_\theta(X)}{\partial^2\theta} \tag{338}$$

$$= -\mathbb{E}_{X\sim p_\theta}\frac{\partial^2 \sum_{i=1}^n \log p_\theta(X_i)}{\partial^2\theta} \tag{339}$$

$$= \sum_{i=1}^n -\mathbb{E}_{X\sim p_\theta}\frac{\partial^2 \log p_\theta(X_i)}{\partial^2\theta} \tag{340}$$

$$= nI(\theta) \tag{341}$$

□

Remark. *The partial derivative of log joint likelihood w.r.t. θ is always called the score function $S(x;\theta) = \frac{\partial \log p_\theta(x)}{\partial\theta}$ and under certain regularity conditions we know that $\mathbb{E}_{X\sim p_\theta}S(X;\theta) = 0$. FI can be seen as the variance of score function $S(X;\theta)$.*

We are able to prove the asymptotic normality of MLE with the asymptotic variance given by FI.

Theorem 19. ***(Asymptotic Normality of MLE)*** *Under certain regularity condition, MLE build upon n i.i.d. samples $\hat{\theta}_n$ is asymptotically normal around the true value of θ which is denoted θ_0, i.e.*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right) \ (n\to\infty) \tag{342}$$

Proof. Let $l(\theta) = p_\theta(X)$ denote the log joint likelihood as a function of parameter θ with X random. By the definition of MLE, we know $l'(\hat{\theta}_n) = 0$ and the first-order Taylor expansion of l' at θ_0 gives (regularity condition ensures that

l is smooth enough)

$$0 = l'(\hat{\theta}_n) = l'(\theta_0) + l''(\theta_0)(\hat{\theta}_n - \theta_0) \tag{343}$$

solve out $\hat{\theta}_n - \theta_0$ to see

$$\hat{\theta}_n - \theta_0 = -\frac{l'(\theta_0)}{l''(\theta_0)} \tag{344}$$

and the quantity of our interest is

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\theta_0)} \tag{345}$$

Notice the first term on the RHS to see that

$$\frac{l'(\theta_0)}{\sqrt{n}} = \frac{\frac{\partial}{\partial \theta} \log p_\theta(X)\Big|_{\theta=\theta_0}}{\sqrt{n}} \tag{346}$$

$$= \frac{\sum_{i=1}^n \frac{\partial}{\partial \theta} \log p_\theta(X_i)\Big|_{\theta=\theta_0}}{\sqrt{n}} \tag{347}$$

so the numerator is the sum of n *i.i.d.* random variables with mean 0 and variance $I(\theta_0)$ (by the def of FI and that score function has 0 mean). By CLT, $\frac{l'(\theta_0)}{\sqrt{n}} \xrightarrow{d} N(0, I(\theta_0))$ $(n \to \infty)$.

For the denominator, notice that

$$-\frac{l''(\theta_0)}{n} = -\frac{\sum_{i=1}^n \frac{\partial^2}{\partial \theta^2} \log p_\theta(X_i)\Big|_{\theta=\theta_0}}{n} \tag{348}$$

the numerator is still the sum of n *i.i.d.* random variables with mean $-I(\theta_0)$ (by the other representation of FI). By WLLN, $-\frac{l''(\theta_0)}{n} \xrightarrow{p} I(\theta_0)$ $(n \to \infty)$.

Since the denominator has constant limit in probability, by Slutsky's theorem,

$$-\frac{\frac{1}{\sqrt{n}} l'(\theta_0)}{\frac{1}{n} l''(\theta_0)} \xrightarrow{d} \frac{N(0, I(\theta_0))}{I(\theta_0)} = N\left(0, \frac{1}{I(\theta_0)}\right) \ (n \to \infty) \tag{349}$$

proves the theorem. □

Remark. *By combining the functional invariance property of MLE and Delta method, one would be able to get the asymptotic distribution of any function of θ denoted $g(\theta)$. Let $\hat{\theta}_n$ be the MLE of θ based on n i.i.d. samples, then $g(\hat{\theta}_n)$ must be the MLE of $g(\theta)$ and $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{1}{I(\theta_0)}\right)$ $(n \to \infty)$. Apply Delta method to see that $g(\hat{\theta}_n)$ is*

also asymptotically normal with

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} N\left(0, \frac{[g'(\theta_0)]^2}{I(\theta_0)}\right) \ (n \to \infty) \tag{350}$$

note that all those results still holds when θ is a vector instead of scalar.

For example, consider $X_1, ..., X_n \sim B(1, \theta)$ so MLE $\hat{\theta}_n = \overline{X}_n$ and the asymptotic normality gives

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \theta_0(1-\theta_0)\right) \ (n \to \infty) \tag{351}$$

now if $g(\theta) = \theta(1-\theta)$, then $g'(\theta) = 1 - 2\theta$ so we have

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0)) \xrightarrow{d} N\left(0, (1-2\theta_0)^2\theta_0(1-\theta_0)\right) \ (n \to \infty) \tag{352}$$

Remark. *The problem for such asymptotic distribution is that the true value of θ which is θ_0 is always unknown so one can never practically compute $I(\theta_0)$ when doing parameter estimation. A natural idea is to replace $I(\theta_0)$ with $I(\hat{\theta}_n)$ as an estimate for the FI to practically use this result.*

Actually, if ***the MLE $\hat{\theta}_n$ is a consistent estimator for*** *θ, then*

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{I(\hat{\theta}_n)}}} = \frac{\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{I(\theta_0)}}}}{\frac{\sqrt{\frac{1}{I(\hat{\theta}_n)}}}{\sqrt{\frac{1}{I(\theta_0)}}}} \tag{353}$$

with the numerator

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{I(\theta_0)}}} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{354}$$

by asymptotic normality of MLE and the denominator

$$\frac{\sqrt{\frac{1}{I(\hat{\theta}_n)}}}{\sqrt{\frac{1}{I(\theta_0)}}} \xrightarrow{p} 1 \ (n \to \infty) \tag{355}$$

by the consistency of the estimator that $\hat{\theta}_n \xrightarrow{p} \theta_0 \ (n \to \infty)$. Since the denominator has constant limit in probability, by Slutsky's theorem, we have proved that

$$\frac{\sqrt{n}(\hat{\theta}_n - \theta_0)}{\sqrt{\frac{1}{I(\hat{\theta}_n)}}} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{356}$$

and this means that we can exactly ***replace the unknown asymptotic variance $\frac{1}{I(\theta_0)}$ with the known quantity***

$\frac{1}{I(\hat{\theta}_n)}$ ***without losing the asymptotic normality*** *under consistency condition.*

This result leads to the possibility of building asymptotic confidence interval for MLE. As a continuation of our Bernoulli example above, if we want to build up a 95% CI for $g(\theta)$, we can use the conclusion that

$$\frac{\sqrt{n}(g(\hat{\theta}_n) - g(\theta_0))}{|1 - 2\hat{\theta}_n|\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}} \xrightarrow{d} N(0, 1) \ \ (n \to \infty) \tag{357}$$

to get the 95% CI for $g(\theta_0) = \theta_0(1 - \theta_0)$ as

$$\left[g(\hat{\theta}_n) - 1.96\frac{|1 - 2\hat{\theta}_n|\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}{\sqrt{n}}, g(\hat{\theta}_n) + 1.96\frac{|1 - 2\hat{\theta}_n|\sqrt{\hat{\theta}_n(1 - \hat{\theta}_n)}}{\sqrt{n}} \right] \tag{358}$$

of course, such **asymptotic CI for MLE works under consistency and large sample size condition**.