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 $p_\theta(x), \theta \in \Omega$ where θ is the parameter and Ω 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 $\mathscr{P} = \{p_\theta, \theta \in \Omega\}$ 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

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

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.

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

$$\sqrt{n}\left(W_n - \frac{1}{\theta^2}\right) \xrightarrow{d} N\left(0, \frac{1}{\theta^8}\right) \ (n \to \infty) \tag{176}$$