# Chapter 5. Properties of Random Variable

**Notation:** Big $\mathcal{O}$

We have $g(n) = \mathcal{O}(f(n))$ if $\frac{g(n)}{f(n)} = \mathcal{O}(1)$.

$\Rightarrow \frac{g(n)}{f(n)} = c$ as $n \to \infty$.

$\Rightarrow f(n)$ & $g(n)$ grow or decay at same rate

~~Example~~: Big $\mathcal{O}$

$f(n) = an^2 + bn + c \quad a \neq 0 \quad \Rightarrow f(n) = \mathcal{O}(n^2)$

**Notation:** Small o

$f(n) = o(g(n))$ iff for every $c > 0$. $\exists$ $N$ s.t. if $n > N$.

$|f(n)| < c \cdot g(n)$

i.e. $\frac{f(n)}{g(n)} \to 0$ as $n \to \infty$.

Note: these notations often applicable to statistics

$\{X_n\} \sim$ R.V.s.

then $X_n = o_p(a_n)$ if $\frac{X_n}{a_n} \xrightarrow{P} 0$ in probability as $n \to \infty$

~~Example~~: Statistical Application

$\{X_n\} = o_p(\frac{1}{n}) \Rightarrow \frac{X_n}{\frac{1}{n}} = n \cdot X_n \xrightarrow{P} 0$ as $n \to \infty$.

**Notation:** Tight Sequence

$\{X_n\} = \mathcal{O}_p(a_n)$ if $\frac{X_n}{a_n}$ stochastically bounded.

i.e. Given $\varepsilon > 0$. $\exists$ $N$ s.t. $P(|\frac{X_n}{a_n}| > N) < \varepsilon$ for $n > N$.

**Theorem:** Central Limit Theorem. C & B. 5.5.14.

Let $X_1, X_2, \ldots$ be iid R.V.s. whose MGF exist in a neighborhood of 0.

($M_{X_i}(t)$ exist for $|t| < h$ for some positive $h$) Let $E[X_i] = \mu$ and

$Var(X_i) = \sigma^2 > 0$ (Both $\mu$ & $\sigma^2 < \infty$. Since MGF exists).

Define $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ Then $\sqrt{n}\left(\frac{\bar{X}_n - \mu}{\sigma}\right) \xrightarrow{d} N(0,1)$

Note: How big does n have to be to get a good approximation?

ANS: 30 or more observations are needed usually.

But more skewed distribution needs more observations.

take bootstrap sample to see if a statistics is Normally distributed.

Proof: Based on Taylor's Series. (Wood's book).

Let $y_i = \frac{\bar{X}-\mu}{\sigma}$ $\quad Z_n = \frac{1}{\sqrt{n}} \cdot \sum_{i=1}^{n} y_i = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$

By theorem 2.3.15.

$$MGF_{Z_n} = MGF_{\frac{\sqrt{n}(\bar{x}-\mu)}{\sigma}}(t) = MGF_{\frac{\sum Y_i}{\sqrt{n}}}(t) \overset{\text{Theorem 2.3.15}}{=} MGF_{\sum Y_i}\left(\frac{t}{\sqrt{n}}\right)$$

$$\overset{\text{Theorem 4.6.7}}{=} \left(MGF_{Y_i}\left(\frac{t}{\sqrt{n}}\right)\right)^n$$

Now we expand $MGF_{Y_i}\left(\frac{t}{\sqrt{n}}\right)$ in a Taylor Series.

$$M_Y\left(\frac{t}{\sqrt{n}}\right) = M_Y(0) + M_Y'(0)\left(\frac{t}{\sqrt{n}}\right) + \frac{1}{2!} M_Y''(0) \cdot \frac{t^2}{n} + \frac{1}{3!} M_Y'''(0) \left(\frac{t}{\sqrt{n}}\right)^3 + \cdots$$

$$= 1 + \underbrace{0}_{\text{mean}} + \underbrace{1}_{\text{Variance}} \cdot \frac{t^2}{2n} + o\left(\frac{1}{n}\right)$$

$$M_{Z_n}(t) = \left[1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n$$

$$\lim_{n\to\infty} \left[1 + \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n = e^{\frac{t^2}{2}} \Rightarrow MGF \text{ of } N(0,1)$$

$$\Rightarrow Z_n \xrightarrow{d} N(0,1)$$

However in practice, we need more general form of CLT.

Theorem: Lindeberg's Central Limit Theorem.

$\{X_{nj}, j = 1, 2, \ldots k_n\}$ is independent R.V.s with $\sigma_n^2 = Var\left(\sum_{j=1}^{k_n} X_{nj}\right) < \infty$

for $n = 1, 2, 3 \ldots$ as $k_n \to \infty$.

if $\sum_{j=1}^{k_n} E\left[(X_{nj} - E(X_{nj}))^2 \cdot I\left(|X_{nj} - \sum_{j=1}^{k_n} X_{nj}| > \varepsilon \sigma_n\right)\right] = o(\sigma_n^2)$

Then $\frac{1}{\sigma_n} \cdot \sum_{j=1}^{k_n} (X_{nj} - E(X_{nj})) \xrightarrow{d} N(0,1)$

* For the CLT above, the iid case is a special case.

* Typically, $k_n = n$ in practice.

However Lindeberg condition is often hard to verify in practice.

Theorem: Lyapunov condition for CLT. (Alternative for Lindeberg's CLT).

$\{X_1, \dots X_n\}$. is independent R.V.s. Each with finite expected value $\mu_i$ & finite variance $\sigma_i^2$

Define $S_n^2 = \sum_{i=1}^{n} \sigma_i^2$, if for some $\delta > 0$. Lyapunov condition

$\lim_{n\to\infty} \frac{1}{S_n^{2+\delta}} E\left[|X_i - \mu|^{2+\delta}\right] = 0$. is satisfied, Then. $\frac{1}{S_n}\sum_{i=1}^{n}(X_i - \mu_i) \xrightarrow{d} N(0,1)$

In practice it's easiest to check Lyapunov condition for $\delta = 1$.

Lyapunov condition $\underset{\not\Leftarrow}{\Longrightarrow}$ Lindeberg condition

Example: Lyapunov condition for Bernuli distribution

$\{X_i\} \sim$ Bernulli $(p_i)$. Then we have $E|X_i - p_i|^{2+\delta} \le E|X_i - p_i|^2 = p_i(1-p_i)$. Variance of $X_i$

Thus. $\frac{1}{\sigma_n^{2+\delta}} \sum_{j=1}^{n} E|X_j - p_j|^{2+\delta} \le \frac{1}{\sigma_n^{2+\delta}} \sum_{j=1}^{n} p_j(1-p_j)$

$= \frac{1}{\sigma_n^{\delta}} \to 0$. if $\sum_{j=1}^{n} p_j(1-p_j) \to \infty$.

$\Rightarrow$ Lyapunov condition Holds.

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} X_i$ Asympotically normal by Lindeberg CLT.

Note: Two central limit theorem can be used when R.V.s are iid or independent.

What about when R.V.s are not independent?

Theorem: Central limit theorem for not independent R.V.s.

Finite population problem

Let say. We are sampling without replacement from a finite population size N.

Then $\{X_i\}$ is not independent But various of CLT still Holds for $\bar{X}$ when $N \to \infty$ & $n \to \infty$. (large population & Large sample size)

See paper Hajek 1960

In practice, We need $\frac{1}{\min(n, N-n)} \cdot \frac{m_N}{V_N} \to 0$ as $N \to \infty$.

$m_N = \max(X_{Ni} - \bar{X}_N)^2$

$V_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_{Ni} - \bar{X}_N)^2$

**Theorem: Multivariate Central Limit Theorem**

$\underset{\sim}{X_1}, \underset{\sim}{X_2}, \dots, \underset{\sim}{X_n}$ iid. $X_i$ are $p$-variate random vectors.

$$\underset{\sim}{X_i} = \begin{pmatrix} X_{i1} \\ \vdots \\ X_{ip} \end{pmatrix} \; \& \; E(\underset{\sim}{X_i}) = \underset{\sim}{\mu_i} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_p \end{pmatrix} \; \& \; \operatorname{cov}(\underset{\sim}{X_i}) = \Psi_{p\times p} = E\left[(\underset{\sim}{X}-\underset{\sim}{\mu})(\underset{\sim}{X}-\underset{\sim}{\mu})'\right]$$

$$\sqrt{n}\,(\bar{\underset{\sim}{X}} - \bar{\underset{\sim}{\mu}}) \xrightarrow{d} N(\underset{\sim}{0}, \Psi).$$

Proof: proof of this is similar to the 1 dimension case.

Set $y_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\underset{\sim}{X_i} - \underset{\sim}{\mu_i})$ & look at MGF of $\underset{\sim}{t}'\underset{\sim}{X_n}$.

$\frac{1}{\sqrt{n}} \underset{\sim}{t}'(\underset{\sim}{X_i} - \underset{\sim}{\mu_i})$ are "iid" 1-dimensional R.V.s.

Apply 1-dimension CLT to prove.

**Theorem: Cramer Wold Device**

$\underset{\sim}{X_n}$ iid $p$-dimensional $n = 1, 2, \dots$

$\underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{X}$ iff $\underset{\sim}{\alpha}'\underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{\alpha}'\underset{\sim}{X}$ for all $\underset{\sim}{\alpha} = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathbb{R}^p$:

Proof:

Necessary:

If $\underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{X} \implies$ for any $\underset{\sim}{\alpha} \in \mathbb{R}^p$

MGF of $\underset{\sim}{\alpha}'\underset{\sim}{X_n}$ is $E\left[e^{t\underset{\sim}{\alpha}'\underset{\sim}{X}}\right] = M_{\underset{\sim}{X_n}}(t\underset{\sim}{\alpha}) \to M_{\underset{\sim}{X}}(t\underset{\sim}{\alpha}) \implies \underset{\sim}{\alpha}'\underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{\alpha}'\underset{\sim}{X}$.

Sufficient:

Assume $\underset{\sim}{\alpha}'\underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{\alpha}'\underset{\sim}{X}$ $\forall\, \underset{\sim}{\alpha}' \in \mathbb{R}^p$

Then $M_{X_n}(t) = E\left[e^{\underset{\sim}{t}'\underset{\sim}{X_n}}\right] = M_{\underset{\sim}{t}'\underset{\sim}{X_n}}(1) = M_{\underset{\sim}{t}'\underset{\sim}{X}}(1) = M_{\underset{\sim}{X}}(\underset{\sim}{t}) \implies \underset{\sim}{X_n} \xrightarrow{d} \underset{\sim}{X}$.

**Theorem: Slusky theorem**

Suppose $X_n \xrightarrow{d} X$ & $Y_n \xrightarrow{P} a$. $a$ is a constant.

(i) $X_n + Y_n \xrightarrow{d} X + a$

(ii) $X_n Y_n \xrightarrow{d} aX$

proof: Part (i)

a.

$$F_{X_n+Y_n}(t) = P(X_n+Y_n \le t)$$

$$= P(X_n+Y_n \le t, |Y_n-a| < \varepsilon) + P(X_n+Y_n \le t, |Y_n-a| > \varepsilon)$$

$$\le P(X_n+Y_n \le t, a-\varepsilon < Y_n < a+\varepsilon) + P(|Y_n-a| > \varepsilon)$$

$$\le P(X_n \le t-a+\varepsilon) + P(|Y_n-a| > \varepsilon) \qquad \rightarrow 0 \because Y_n \xrightarrow{P} a$$

$$\limsup_{n\to\infty} F_{X_n+Y_n}(t) \le \limsup_{n\to\infty} P(X_n \le t-a+\varepsilon) + \boxed{\limsup_{n\to\infty} P(|Y_n-a| > \varepsilon)}$$

$$\limsup_{n\to\infty} F_{X_n+Y_n}(t) \le F(X_n \le t-a+\varepsilon) = F_X(t-a+\varepsilon)$$

$$\limsup_{n\to\infty} F_{X_n+Y_n}(t) \le F_X(t-a+\varepsilon)$$

b.

$$1 - F_{X_n+Y_n}(t) = P(X_n+Y_n \ge t)$$

$$= P(X_n+Y_n \ge t, |Y_n-a| < \varepsilon) + P(X_n+Y_n \ge t, |Y_n-a| > \varepsilon)$$

$$\le P(X_n+Y_n \ge t, a-\varepsilon < Y_n < a+\varepsilon) + P(|Y_n-a| > \varepsilon)$$

$$\le P(X_n \ge t-a-\varepsilon) + P(|Y_n-a| > \varepsilon) \qquad \rightarrow 0 \because Y_n \xrightarrow{P} a$$

$$\limsup_{n\to\infty} (1-F_{X_n+Y_n}(t)) \le \limsup_{n\to\infty} P(X_n \ge t-a-\varepsilon) + \boxed{\limsup_{n\to\infty} P(|Y_n-a| > \varepsilon)}$$

$$\liminf_{n\to\infty} F_{X_n+Y_n}(t) \le 1 - F(X < t-a-\varepsilon) = 1 - F_X(t-a-\varepsilon)$$

$$\liminf_{n\to\infty} F_{X_n+Y_n}(t) \ge F_X(t-a-\varepsilon)$$

Thus we have follow sequence

$$F_X(t-a-\varepsilon) \le \liminf_{n\to\infty} F_{X_n+Y_n}(t) \le \limsup_{n\to\infty} F_{X_n+Y_n}(t) \le F_X(t-a+\varepsilon)$$

$$F_{X-a}(t-\varepsilon) \le \liminf_{n\to\infty} F_{X_n+Y_n}(t) \le \limsup_{n\to\infty} F_{X_n+Y_n}(t) \le F_{X-a}(t+\varepsilon)$$

Since $\varepsilon$ is an arbitrary number & $\varepsilon > 0$.

We have $\lim_{n\to\infty} F_{X_n+Y_n}(t) = F_{X+a}(t) \implies X_n + Y_n \xrightarrow{d} X + a$.

**Theorem: Delta Method** C&B 5.5.24.

Let $Y_n$ be a sequence of R.V.s that satisfied $\sqrt{n}(Y_n-\theta) \xrightarrow{d} N(0, \sigma^2)$.

For a given function $g$ and a specific value of $\theta$, suppose that $g'(\theta)$ exists and $g'(\theta) \ne 0$.

Then We Have:

$\sqrt{n}\,(g(Y_n)-g(\theta)) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2)$.

Proof: Also uses Taylar Expansion & slusky Theoerm.

$g(Y_n) = g(\theta) + g'(\theta)(Y_n-\theta) +$ Reminder.

$g(Y_n) - g(\theta) = g'(\theta)(Y_n-\theta) +$ Reminder.

$\sqrt{n}\,(g(Y_n)-g(\theta)) = \sqrt{n}\,g'(\theta)(Y_n-\theta) + \sqrt{n}$ (Reminder).

$P(|Y_n-\theta|<\varepsilon) = P(|\sqrt{n}(Y_n-\theta)| < \sqrt{n}\,\varepsilon)$

Therefore $\lim_{n\to\infty} P(|Y_n-\theta|<\varepsilon) = \lim_{n\to\infty} P(|\sqrt{n}(Y_n-\theta)| < \sqrt{n}\,\varepsilon)$

$= \lim_{n\to\infty} P(|Z|<\infty) = 1$. Where $Z \sim N(0,\sigma^2)$.

Thus We have $Y_n \xrightarrow{P} \theta$.

By Slusky theoerm (a). $g'(\theta)\sqrt{n}(Y_n-\theta) \longrightarrow g'(\theta)\cdot Z$. Where $Z \sim N(0,\sigma^2)$.

Therefore $\sqrt{n}[g(Y_n)-g(\theta)] = g'(\theta)\sqrt{n}(Y_n-\theta) \xrightarrow{d} N(0, \sigma^2[g'(\theta)]^2)$.

Example: Delta Method.

$\{X_i\}$ ~ Bernulli $(p)$ iid R.V.s.

$\sqrt{n}\dfrac{\bar{X}-p}{\sqrt{p(1-p)}} \longrightarrow N(0,1)$. Look at $g(p) = p(1-p)$. $g'(p) = 1-2p$

$\sqrt{n}\,(\hat{p}(1-\hat{p}) - p(1-p)) \longrightarrow N(0, p(1-p)(1-2p)^2)$.

Theoerm: Multivariate Delta Method.

Suppose $\underset{\sim}{X_n} = \begin{pmatrix} X_{n1} \\ \vdots \\ X_{np} \end{pmatrix}$ is a $p$-dimension vector

Where $\sqrt{n}(\underset{\sim}{X_n} - \underset{\sim}{\mu}) \xrightarrow{d} N(\underset{\sim}{0}, \Psi)$

& $g: \mathbb{R}^p \to \mathbb{R}^1$ with continous partial derivatives in a Neighborhood of $\underset{\sim}{\mu}$.

Then $\sqrt{n}\,(g(\underset{\sim}{X_n}) - g(\underset{\sim}{\mu})) \xrightarrow{d} N(0, (\nabla g(\underset{\sim}{\mu}))'\,\Psi\,\nabla g(\underset{\sim}{\mu})))$

provided that $(\nabla g(\underset{\sim}{\mu}))'\,\Psi\,\nabla g(\underset{\sim}{\mu})) \neq 0$.

Note: $\nabla g(\underset{\sim}{\mu}) = \dfrac{\partial g}{\partial x} = \begin{pmatrix} \partial g/\partial x_1 \\ \vdots \\ \partial g/\partial x_p \end{pmatrix}$ gradient of $g$ at $\underset{\sim}{x} \in \mathbb{R}^p$

Example. Non-Linear Regression (least square) Application of Multivariate Delta Method

$y_i = \frac{\alpha_1}{\beta_1}(1-e^{-\beta_1 t}) + \frac{\alpha_2}{\beta_2}(1-e^{-\beta_2 t}) + \varepsilon_i$. Compartmental Model.

$\underset{\sim}{\theta} = \begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix}$ ~ non-negative parameters ( Also $\sigma^2 = Var(\varepsilon_i)$ )

We can use least square estimate $\theta$. ("nls" package)

$\hat{\underset{\sim}{\theta}} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\alpha}_2 \\ \hat{\beta}_2 \end{pmatrix}$ Needs to use iteration algorithm to find $\hat{\underset{\sim}{\theta}}$.

Concentration of drug in blood accumulate.

As $t \to \infty$. $t$ = time. Concentration $\longrightarrow \frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}$.

$$g(\theta) = g\begin{pmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{pmatrix} = \frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}.$$

$$\frac{\partial g(\theta)}{\partial \theta} = \begin{pmatrix} \partial g(\theta)/\partial \alpha_1 \\ \partial g(\theta)/\partial \beta_1 \\ \partial g(\theta)/\partial \alpha_2 \\ \partial g(\theta)/\partial \beta_2 \end{pmatrix} = \begin{pmatrix} 1/\beta_1 \\ -\alpha_1/\beta_1^2 \\ 1/\beta_2 \\ -\alpha_2/\beta_2^2 \end{pmatrix}$$

$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0, \Psi_{4\times4})$.

By Delta Method.

$$\sqrt{n}\left(\left(\frac{\hat{\alpha}_1}{\hat{\beta}_1} + \frac{\hat{\alpha}_2}{\hat{\beta}_2}\right) - \left(\frac{\alpha_1}{\beta_1} + \frac{\alpha_2}{\beta_2}\right)\right) \xrightarrow{d} N\left(\underset{\sim}{0}, \left(\frac{1}{\beta_1}, -\frac{\alpha_1}{\beta_1^2}, \frac{1}{\beta_2}, -\frac{\alpha_2}{\beta_2^2}\right) \underset{\sim}{\Psi} \begin{pmatrix} 1/\beta_1 \\ -\alpha_1/\beta_1^2 \\ 1/\beta_2 \\ -\alpha_2/\beta_2^2 \end{pmatrix}\right)$$

Example: Approximate mean & Variance Using Taylor expansion.

$\{X\}$ ~ R.V. with mean $\mu$ & variance $\sigma^2$.

let $h(x)$ be a smooth function. (use transformation to get distribution)

Taylor Expansion $h(x) \simeq h(\mu) + h'(\mu)(x-\mu)$.

$\Longrightarrow (h(x)-h(\mu))^2 \approx (h'(\mu))^2(x-\mu)^2$

$\Longrightarrow E[(h(x)-h(\mu))^2] \approx (h'(\mu))^2 \cdot E[(x-\mu)^2]$

$\Longrightarrow Var(h(x)) \approx (h'(\mu))^2 \cdot Var(x)$.

Also. By the Taylor expansion $h(x) \approx h(\mu) + h'(\mu)(x-\mu) + \frac{1}{2} h''(\mu)(x-\mu)^2$

$$\Rightarrow E[h(x)] \approx h(\mu) + h'(\mu) \cdot E(x-\mu) + \frac{1}{2} h''(\mu) E[(x-\mu)^2].$$

$$\Rightarrow E[h(x)] \approx h(\mu) + h'(\mu) \cdot 0 + \frac{1}{2} h''(\mu) \cdot \sigma_x^2$$

$$\Rightarrow E[h(x)] \approx h(\mu) + \frac{h''(\mu)}{2} \sigma_x^2$$

**Theorem: Quantiles of R.V.s also follows CLT.** (proof using Lyapunov CLT.)

Let F be a CDF that is continously differentiable in a neighborhood of $\xi_p$. When $F(\xi_p) = p$. $\xi_p$ = Sample quantiles Assume $F'(\xi_p) = f(\xi_p) > 0$.

We have $\sqrt{n}(\xi_p - \hat{\xi}_p) \xrightarrow{d} N\left(0, \frac{p(1-p)}{f^2(\xi_p)}\right)$ as $n \to \infty$.

**Theorem: finite population Central Limit theorem (Simple Random Sample).**

$\{Y\} = \{y_1, y_2 \cdots y_N\} \sim$ R.V.s. N: population size, n: Sample size.

Each of samples have equal probability of selection

let $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ Approximate to Normal distribution as $n, N \to \infty$.

let $\mathcal{L}$ = one of the $C_N^n$ samples.

$E(\bar{y})$ = Average Sample mean over all $C_N^n$ samples.

$$= \frac{1}{C_N^n} \sum_{\mathcal{L}} \bar{y_s} = \frac{1}{C_N^n} \sum_{\mathcal{L}} \frac{1}{n} (y_{s1} + y_{s2} + \cdots + y_{sn}).$$

$$= \frac{(N-n)!(n-1)!}{N!} \sum_{i=1}^{N} y_i \{\# \text{ of samples } y_s \text{ that containing } y_i\}.$$

$$= \frac{(N-n)!(n-1)!}{N!} \sum_{i=1}^{N} y_i \cdot C_{N-1}^{n-1}$$

$$= \frac{(N-n)!(n-1)!}{N!} \times \frac{(N-1)!}{(N-n)!(n-1)!} \sum_{i=1}^{N} y_i.$$

$$= \frac{1}{N} \sum_{i=1}^{N} y_i = \hat{\mu}_y \Rightarrow \hat{\mu}_y \text{ is unbiased for } \mu_y \text{ in SRS}.$$

**Theorem: finite Population correction factor.**

$$\text{let } z_i = \begin{cases} 1 & \text{if unit } i \text{ is selected.} \\ 0 & \text{otherwise,} \end{cases}$$

$$P(Z=1) = \frac{C_1^1 C_{N-1}^{n-1}}{C_N^n} = \frac{n}{N} = E(Z)$$

$$P(Z_i, Z_j = 1) = \frac{C_2^2 C_{N-2}^{n-2}}{C_N^n} = \frac{n(n-1)}{N(N-1)} = E(Z_i, Z_j)$$

$$\left.\begin{aligned} \mathrm{cov}(Z_i, Z_j) &= E(Z_i, Z_j) - E(Z_i)E(Z_j) \\ &= \frac{n(n-1)}{N(N-1)} - \frac{n}{N}\cdot\frac{n}{N} \neq 0 \end{aligned}\right\} \Rightarrow Z_i, Z_j \text{ are not } \text{independent}$$

Example: Delta Method on Survival Analysis.

Simulating survival time in a Cox model with covariate $x$, with cumulative hazard function $H_0(t) = \lambda t$. (Generating survival time to simulate Cox proportional hazard models. Bender, Statistics in Medicine).

$T = H_0^{-1}\left[-\log(u)\, e^{-\beta x}\right]$ ~ survival time

$u \sim$ uniform $(0, 1)$. $\quad -\frac{1}{\lambda}\log(u)\cdot e^{-\beta x}$

~~Goal~~: Find $\beta$ ~~that~~ give a specific percentile.

i.e. $p = p^{th}$ percentile & we want $p = P(T \le 12)$ ~ probability of death in year.

$P\left(-\frac{1}{\lambda}\log(u) e^{-\beta x} < 12\right) = p$. Solve for $\beta$.

$P\left(-\frac{1}{12\lambda}\log(u) < e^{\beta x}\right) = p$ $\quad$ let $y = -\frac{1}{12\lambda}\log(u) \sim \exp(12\lambda)$.

$W = e^{\beta x} \sim$ log-normal.

$x \sim N(\mu, \sigma^2)$.

$P(T < 12) = P(Y < W)$.

$$= \int_0^{+\infty}\int_0^{w} 12\lambda e^{-12\lambda y}\, \frac{e^{-\frac{\left(\frac{1}{\beta}\log(w) - \mu\right)^2}{2\sigma^2}}}{\sqrt{2\pi}\,\sigma\beta w}\, dy\, dw$$

$$= 1 - E\left[e^{-12\lambda W}\right] \qquad \beta = \sigma \quad \mu = 0$$

We can estimate $E\left[e^{-12\lambda W}\right]$ by Delta Method with $g(w) = e^{-12\lambda w}$.

$W \sim$ lognormal $(0, \beta)$ $\quad E(W) = e^{\frac{\beta^2}{2}}$ $\quad \mathrm{Var}(W) = \left(e^{\beta^2} - 1\right)e^{\beta^2}$

By Mean & Variance Approximation from Delta Method

$$E\left[e^{-12\lambda W}\right] = \exp\left\{-12\lambda\cdot e^{\frac{\beta^2}{2}}\right\} + \frac{1}{2}(12\lambda)^2 \exp\left\{-12\lambda\cdot e^{\frac{\beta^2}{2}}\right\}\cdot\left(e^{\beta^2} - 1\right)e^{\beta^2}$$

# Chapter 6. Principle of Data Reduction.

Principles of Data Reduction:

1. Sufficient principle: use a statistics that contain more information about $\theta$.

2. Likelihood principle: A function of parameters determined by data $\Rightarrow$ what value of $\theta$ makes the observed data most likely.

3. Equivalent principle: Are the data reduction method processing those important features.

## Definition: Sufficient Statistics

A statistics $T(x)$ is sufficient statistics for $\theta$ if conditional distribution of sample $x$ given the value of $T(x)$ does not depends on $\theta$.

Example: Sufficient Statistics.

$\{X_i\}$ denotes number of accident in NYC on a particular day

Assume independent with poission distribution.

$X_i \sim \text{Poission}(\lambda)$.

$$f(\underset{\sim}{x}|\lambda) = f(x_1 \dots x_n|\lambda) \quad \text{Joint PDF}$$
$$= \frac{e^{-\lambda}\lambda^{x_1}}{x_1!} \frac{e^{-\lambda}\lambda^{x_2}}{x_2!} \cdots \frac{e^{-\lambda}\lambda^{x_n}}{x_n!}$$
$$= \frac{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

$T(x) = \sum_{i=1}^n x_i$ is the sufficient statistics.

proof. $T(x) = \sum_{i=1}^n x_i$ is a sufficient statistics by definition of sufficient.

$$P(X_1 = x_1, X_2 = x_2 \dots X_n = x_n \mid T(x) = t)$$
$$= \frac{P(X_1 = x_1 \dots X_n = x_n, T(x) = t)}{P(T(x) = t)} = \frac{P(X_1 = x_1 \dots X_n = x_n, \sum_{i=1}^n X_i)}{P(\sum_{i=1}^n x_i)}$$
$$= \frac{P(X_1 = x_1 \dots X_n = x_n)}{P(\sum_{i=1}^n x_i)} \quad \text{Since } \{X_1 = x_1 \dots X_n = x_n\} \subseteq \{T(\underset{\sim}{x}) = t(\underset{\sim}{x})\}$$
$$= \frac{e^{-\lambda}\lambda^{x_1}, \dots, e^{-\lambda}\lambda^{x_n}}{e^{-n\lambda}\lambda^{\sum_{i=1}^n x_i}} = \sim \text{ independent of } \lambda.$$

Theorem 6.2.2.

If $f(\underset{\sim}{x}|\theta)$ is a joint PDF of $\underset{\sim}{X}$ and $q(t|\theta)$ is PDF or PMF of $T(x)$

Then $T(\underset{\sim}{X})$ is a sufficient statistic for $\theta$ if for every $x$ in the sample space the ratio of $\dfrac{f(\underset{\sim}{x}|\theta)}{q(T(\underset{\sim}{x})|\theta)}$ is constant as a function of $\theta$.

proof:

$P_\theta(X_1=x_1 \dots X_n=x_n) = P_\theta(X_1=x_1 \dots X_n=x_n, T(\underset{\sim}{X})=t(\underset{\sim}{x}))$. $\{\underset{\sim}{x}\} \subseteq \{T(x)\}$.

$P_\theta(\underset{\sim}{X}=\underset{\sim}{x}) = P_\theta(\underset{\sim}{X}=\underset{\sim}{x} \mid T(\underset{\sim}{X})=t(\underset{\sim}{x})) \cdot P_\theta(T(\underset{\sim}{X})=t(\underset{\sim}{x}))$.

$$\frac{P_\theta(\underset{\sim}{X}=\underset{\sim}{x})}{P_\theta(T(\underset{\sim}{X})=t(\underset{\sim}{x}))} = \underbrace{P_\theta(\underset{\sim}{X}=\underset{\sim}{x} \mid T(\underset{\sim}{X})=t(\underset{\sim}{x}))}_{\text{independ of } \theta \text{ if } T(\underset{\sim}{X}) \text{ is a sufficient statistics.}}$$

Theorem : Factorization Theorem (&B 6.6.2.

Let $f(\underset{\sim}{x}|\theta)$ denote the joint PDF or PMF of $\underset{\sim}{X}$. A statistics $T(\underset{\sim}{X})$ is a sufficient statistics for $\theta$ iif there exists function $g(t|x)$ and $h(x)$ s.t. for all sample points $x$ and all parameter points $\theta$,

$f(\underset{\sim}{x}|\theta) = g(T(x)|\theta)\, h(x)$.

proof : discrete case prove both necessary & sufficient.

Suppose $T(x)$ is a sufficient statistics. Choose $g(t|\theta) = P_\theta(T(x)=t)$ and $h(x) = P(\underset{\sim}{X}=\underset{\sim}{x} \mid T(\underset{\sim}{X})=t(\underset{\sim}{x}))$. Because $T(x)$ is sufficient, the conditional probability defining $h(x)$ does not depends on $\theta$. Thus the choice of $h(x)$ & $g(t|\theta)$ is legitimate, and for this choice, we have

$$\begin{aligned}
f(\underset{\sim}{x}|\theta) &= P_\theta(\underset{\sim}{X}=\underset{\sim}{x}) \\
&= P_\theta(\underset{\sim}{X}=\underset{\sim}{x}, T(\underset{\sim}{X})=t(\underset{\sim}{x})) \\
&= P_\theta(T(\underset{\sim}{X})=t(\underset{\sim}{x})) \cdot P_\theta(\underset{\sim}{X}=\underset{\sim}{x} \mid T(\underset{\sim}{X})=t(\underset{\sim}{x})) \\
&= g(T(\underset{\sim}{x})|\theta)\, h(\underset{\sim}{x})
\end{aligned}$$

The above exhibited the factorization theorem. We also see from the last two lines above that.

$P_\theta(T(\underset{\sim}{X})=T(\underset{\sim}{x}))=g(T(\underset{\sim}{x})|\theta)$. so, $g(T(\underset{\sim}{x})|\theta)$ is PDF or PMF of $T(x)$.

Now, we assume factorization theorem exists. let $q(t|\theta)$ be the PMF of $T(x)$. Show that $T(x)$ is sufficient, we examine the ratio $\dfrac{f(\underset{\sim}{x}|\theta)}{q(T(\underset{\sim}{x})|\theta)}$

Define $A_{T(x)}=\{\underset{\sim}{y}: T(\underset{\sim}{y})=T(x)\}$. Then,

$$\frac{f(\underset{\sim}{x}|\theta)}{q(T(\underset{\sim}{x})|\theta)}=\frac{g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{x})}{q(T(\underset{\sim}{x})|\theta)}$$ Since factorization theorem exists.

$$=\frac{g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{x})}{\sum_{A_{T(x)}} f(\underset{\sim}{y}|\theta)}$$

$\left(T(\underset{\sim}{x})=\sum x_i \quad \begin{matrix} x_1=1 & x_1=2 \\ x_2=2 & x_2=1 \end{matrix}\right)$.

$P(T(x))= f(x_1=1, x_2=2)+f(x_1=2, x_2=1)$.

$$=\frac{g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{x})}{\sum_{A_{T(x)}} g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{y})}$$

$$=\frac{g(T(\underset{\sim}{x})|\theta)h(\underset{\sim}{x})}{g(T(\underset{\sim}{x})|\theta)\sum_{A_{T(x)}} h(\underset{\sim}{y})}$$ Since $T$ is constant on $A_{T(x)}$

$$=\frac{h(\underset{\sim}{x})}{\sum_{A_{T(x)}} h(\underset{\sim}{y})}$$ $\sim$ not depends on $\theta$.

Since the ratio does not depends on $\theta$. by theorem 6.2.2.

$T(x)$ is a sufficient statistics for $\theta$.

Various version of this proved by Fisher & Savage.

See "Testing statistical Hypothesis by F. Lehman

Also see Billingsly P471 for general framework.

Example: Poission Re-visit.

$$f_\lambda(x_1 \ldots x_n)=\frac{e^{-n\lambda}\cdot\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i}=h(\underset{\sim}{x})\,e^{-n\lambda}\cdot\lambda^{\sum_{i=1}^n x_i}$$

$\Rightarrow$ $\sum_{i=1}^n x_i$ is a sufficient statistics.

$\updownarrow$ one-to-one function.

$\frac{1}{n}\sum_{i=1}^n x_i=\bar{x}$ is also a sufficient statistics.

This form holds for any distribution in general.

Example. Uniform distribution. German Tank Problem

Use serial number on Tank to estimate total number of tank. $\theta$

Set. $\{x_i\} \sim$ uniform. distribution

$f(x|\theta) = \frac{1}{\theta}$ for $x = 1, 2, \cdots \theta$.

Select. n. serial. number without. replacement We will have $C_n^\theta$ possible

Samples. each with. PDF $\frac{1}{C_n^\theta}$

$$f(\underset{\sim}{x}|\theta) = \frac{1}{C_n^\theta} \cdot I(1 \le x_1 \le \theta,\ 1 \le x_2 \le \theta \cdots 1 \le x_n \le \theta).$$

$$= (C_n^\theta)^{-1} \cdot I(\max(x_i) \le \theta).$$

$$= (C_n^\theta)^{-1} \cdot I(\underbrace{x_{(n)}}_{n^{th} \text{ order statistic}})$$

let. $h(\underset{\sim}{x}) = 1$. by. factorization theoerm.

$X_{(n)}$ is a sufficient. Statistic for $\theta$.

However, $E[X_{(n)}] < \theta$. so. it's a biased. sufficient. statistic

So. it need. to be. rescaled. to get. the. unbiased. estimator.

**Note: Sufficient. Statistics. with. extreme. value Statistics.**

Typically, in examples where support. of. distribution. depends on some unknown.

parameters (e.g. uniform $(0, \theta)$) The. sufficient. statistics. will involve some.

extreme order statistics.

Example: Sufficient. Statistics. for. Normal. distribution.

$\{x_i\} \sim N(\mu, \sigma^2)$. Both. $\mu$, & $\sigma^2$. are. unknown.

$$f(\underset{\sim}{x}|\mu, \sigma^2) = (2\pi)^{-\frac{n}{2}} (\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \frac{\sum_{i=1}^n (x_i - \mu)^2}{\sigma^2}\right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x} + \bar{x} - \mu)^2\right\}.$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2 + 2 \sum_{i=1}^n \overset{=0}{\underline{\underline{(x_i - \bar{x})}}} (\bar{x} - \mu)\right\}.$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2\right\}. \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2} (n-1) S^2 + n(\bar{x} - \mu)^2\right\}.$$

By. factorization. Theoerm. $T(x) = (\bar{x}, s^2)$ are. sufficient. statistics for $(\mu, \sigma^2)$.

~~Example~~: Sufficient Statistics for Multi-variate Normal distribution.

$$\mathbf{x}_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ip} \end{pmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Psi}).$$

$$f(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \mid \mu, \Psi) = (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \Psi^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr\left( \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' \Psi^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} tr\left( \boldsymbol{\Psi}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})' (\mathbf{x}_i - \boldsymbol{\mu}) \right) \right\}.$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr\left( n \boldsymbol{\Psi}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \boldsymbol{\mu})' (\mathbf{x}_i - \boldsymbol{\mu}) \right) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr\left( n \boldsymbol{\Psi}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu})' (\mathbf{x}_i - \bar{\mathbf{x}} + \bar{\mathbf{x}} - \boldsymbol{\mu}) \right) \right\}.$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr\left( n \boldsymbol{\Psi}^{-1} \sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}})' (\mathbf{x}_i - \bar{\mathbf{x}}) + n(\bar{\mathbf{x}} - \boldsymbol{\mu})' (\bar{\mathbf{x}} - \boldsymbol{\mu}) + 2 \sum_{i=1}^{n} \underbrace{(\mathbf{x}_i - \bar{\mathbf{x}})'}_{=0} (\bar{\mathbf{x}} - \boldsymbol{\mu}) \right) \right\}$$

$$= (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Psi}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} tr\left( n \boldsymbol{\Psi}^{-1} \left( \mathbf{S}^2 + n(\bar{\mathbf{x}} - \boldsymbol{\mu})' (\bar{\mathbf{x}} - \boldsymbol{\mu}) \right) \right) \right\}.$$

By factorization theorem, $(\bar{\mathbf{x}}, \mathbf{S}^2)$ are sufficient statistics for $(\boldsymbol{\mu}, \boldsymbol{\Psi})$.