Chapter 5. Properties of a Random Sample

5.1 Basic concepts of random samples

Defn: The r.v.'s $X_1, \dots, X_n$ are called a random sample of size n from the pop^n $f(x)$ if $X_1, \dots, X_n \overset{iid}{\sim} f(x)$.

- The joint pdf of $X_1, \dots, X_n$ is given by

$$f_{X_1,\dots,X_n}(x_1, \dots, x_n) = f(x_1) f(x_2) \cdots f(x_n) = \prod_{i=1}^{n} f(x_i)$$

(e.g.) $X_1, \dots, X_n$: a random sample from $Exp(\beta)$ population

$$f(x_1, \dots, x_n \mid \beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-x_i/\beta} = \frac{1}{\beta^n} e^{-\sum_{i}^{n} x_i/\beta}$$

//

- $\underbrace{\text{Sampling from an infinite pop}^n}_{\text{iid !}}$ vs. $\underbrace{\text{Sampling from a finite pop}^n}$.

① with replacement ; iid !

② w/o replacement ? iid ✗

- Sampling w/o replacement is also called simple random sampling.

(e.g.) $\{1, 2, \cdots, 1000\}$: pop^n.

$\{X_1, \cdots, X_{10}\}$ is drawn w/o replacement.

$P(\text{all ten sample values} > 200) = ?$

sol. · If drawn with replacement, then

$P(X_1 > 200, \cdots, X_{10} > 200) = P(X_1 > 200) \cdots P(X_{10} > 200) = \left(\frac{800}{1000}\right)^{10} = 0.1074$.

· Let $Y = \#$ of X_i's > 200. Then $P(Y=10)$ is the answer.

$P(Y=10) = \binom{800}{10}\binom{200}{0} \Big/ \binom{1000}{10} = 0.1062$ //

5.2 Sums of random variables from a random sample.

$X_1, \cdots, X_n$: a random sample from $f(x)$.

$Y = T(X_1, \cdots, X_n)$: a statistic, $\sim \boxed{?}$ ← sampling distribution

Theorem : $X_1, \cdots, X_n \overset{iid}{\sim} \cdot(\mu, \sigma^2)$. Then,

(a) $E\bar{X} = \mu$

(b) $Var\,\bar{X} = \sigma^2/n$

(c) $ES^2 = \sigma^2$

pf: (a), (b) : trivial.

(c) : $ES^2 = E\,\frac{1}{n-1}\sum_1^n (X_i - \bar{X})^2 = E\,\frac{1}{n-1}\left\{\sum_1^n X_i^2 - n(\bar{X})^2\right\} = \frac{1}{n-1}\left\{nEX_1^2 - nE(\bar{X})^2\right\}$

$= \frac{1}{n-1}\left\{n(\mu^2+\sigma^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right\} = \sigma^2$ //

Theorem : $X_1, \cdots, X_n$, a random sample. Then,

$$M_{\bar{X}}(t) = \{ M_X(t/n) \}^n.$$

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

Then $M_{\bar{X}}(t) = \{ \exp(\mu \cdot \frac{t}{n} + \frac{1}{2}\sigma^2 (\frac{t}{n})^2 \}^n = \exp\left(\mu t + \frac{1}{2} \cdot (\frac{\sigma^2}{n}) t^2\right) : N(\mu, \frac{\sigma^2}{n})$ //

Theorem : $X \sim f_X$, $Y \sim f_Y$: indep.

Then $Z \sim f_Z$ with $f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z-x)\, dx$

(e.g.) $Z_1, Z_2, \cdots, Z_n \overset{iid}{\sim}$ Cauchy $(0, 1)$

Then $\sum_1^n Z_i \sim$ Cauchy $(0, n)$ and $\bar{Z} \sim$ Cauchy $(0, 1)$.

Theorem : $X_1, \cdots, X_n \overset{iid}{\sim} f(x|\theta)$ where $f(x|\theta) = h(x)\, c(\theta) \exp\left(\sum_{i=1}^{k} w_i(\theta)\, t_i(x)\right)$.

$T_i(X_1, \cdots, X_n) = \sum_{j=1}^{n} t_i(X_j)$, $i = 1, 2, \cdots, k$.

If $\{(w_1(\theta), \cdots, w_k(\theta)),\ \theta \in \Theta\}$ contains an open set in $\mathbb{R}^k$, then

$$f_{T_1, \cdots, T_k}(u_1, \cdots, u_k \mid \theta) = H(u_1, \cdots, u_k)\, \{c(\theta)\}^n \exp\left(\sum_{i=1}^{k} w_i(\theta)\, u_i\right).$$

Remark : The open set condition eliminates curved exp'l families.

(e.g.) $X_1, \cdots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$.

$$f(x|p) = p^x (1-p)^{1-x} = \underbrace{(1-p)}_{c(p)} \cdot \exp\left(\underbrace{\log \frac{p}{1-p}}_{w_1(p)} \cdot \underbrace{x}_{t_1(x)}\right)$$

$$T_1(X_1, \cdots, X_n) = \sum_{j=1}^{n} X_j \sim f_{T_1}(u_1|p) = H(u_1)\, (1-p)^n \exp\left(\log \frac{p}{1-p} \cdot u_1\right) \propto p^{u_1} (1-p)^{n-u_1}$$

//

5.3 Sampling from the normal distribution

Theorem : $X_1, \cdots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

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

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

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

Pf. (a) $S^2 = \frac{1}{n-1}\left[(X_1 - \bar{X})^2 + \sum_{i=2}^{n} (X_i - \bar{X})^2 \right]$

WLOG, assume $\mu=0$ and $\sigma=1$.

$= \frac{1}{n-1}\left[\left\{ \sum_{i=2}^{n} (X_i - \bar{X}) \right\}^2 + \sum_{i=2}^{n} (X_i - \bar{X})^2 \right] = \text{ft}^n \text{ of } (X_2 - \bar{X}, \cdots, X_n - \bar{X})$.

$$\begin{cases} Y_1 = \bar{X} \\ Y_2 = X_2 - \bar{X} \\ \vdots \\ Y_n = X_n - \bar{X} \end{cases} \Rightarrow f_{Y_1,\cdots,Y_n}(y_1,\cdots,y_n) \propto e^{-ny_1^2/2} \cdot e^{-\frac{1}{2}\left(\sum_2^n y_i^2 + \left(\sum_2^n y_i\right)^2\right)}$$

Lemma : (a) $Z \sim N(0,1) \Rightarrow Z^2 \sim \chi^2_1$

(b) $X_1, \cdots, X_n \overset{\text{indep.}}{\sim} \chi^2_{p_i} \Rightarrow X_1 + \cdots + X_n \sim \chi^2_{p_1+\cdots+p_n}$.

pf of the theorem (cont'd) :

$\frac{1}{n-1}\sum_{1}^{n-1} X_i$

(c) $(n-1)S_n^2 = (n-2)S_{n-1}^2 + \frac{n-1}{n}(X_n - \overline{X}_{n-1})^2$

Consider $n=2$: $S_2^2 = (X_2 - X_1)^2/2 = \{\underbrace{(X_2 - X_1)/\sqrt{2}}_{N(0,1)}\}^2 \sim \chi^2_1$

Assume that for $n=k$, $(k-1)S_k^2 \sim \chi^2_{k-1}$.

$$kS_{k+1}^2 = \underbrace{\underbrace{(k-1)S_k^2}_{\sim \chi^2_{k-1}} + \underbrace{\frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2}}$$

$X_{k+1} - \overline{X}_k \sim N(0, \frac{k+1}{k})$

$\therefore \frac{k}{k+1}(X_{k+1} - \overline{X}_k)^2 \sim \chi^2_1$

need to prove independence: $X_{k+1} - \overline{X}_k \perp S_k^2$ by (a).

//

- t-distⁿ & F-distⁿ : HW (page 222 ~ 225)

5.4 Order statistics

$X_1, \cdots, X_n \overset{iid}{\sim} f(x)$

$X_{(1)}, \cdots, X_{(n)}$: the order statistics

- $X_1, \cdots, X_n \overset{iid}{\sim} f_X(x_i) = p_i$, $x_1 < x_2 < \cdots$ (discrete)

$P(X_{(j)} \le x_i) = \sum_{k=j}^{n} \binom{n}{k} P_i^k (1-P_i)^{n-k}$, $P_i = p_1 + p_2 + \cdots + p_i$.

$P(X_{(j)} = x_i) = \sum_{k=j}^{n} \binom{n}{k} \left[P_i^k (1-P_i)^{n-k} - P_{i-1}^k (1-P_{i-1})^{n-k} \right]$.

- $X_1, \dots, X_n \overset{iid}{\sim} f_X, F_X$ (continuous)

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

Pf. $F_{X_{(j)}}(x) = P(X_{(j)} \le x) = \sum_{k=j}^{n} \binom{n}{k} \{F_X(x)\}^{j} \{1 - F_X(x)\}^{n-j}$

$f_{X_{(j)}}(x) = \frac{d}{dx} F_{X_{(j)}}(x) = \cdots$ //

- $1 \le i < j \le n$:

$$f_{X_{(i)}, X_{(j)}}(u, v) = \frac{n!}{(i-1)!\,(j-1-i)!\,(n-j)!} f_X(u) f_X(v) \{F_X(u)\}^{i-1} \{F_X(v) - F_X(u)\}^{j-1-i} \{1 - F_X(v)\}^{n-j}$$

for $-\infty < u < v < \infty$.

5.5 Convergence concepts.

$X_1, X_2, \cdots$: a seq. of random variables

- $X_n \xrightarrow{p} X$ if for every $\varepsilon > 0$ $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0$

  [<] [1]

  (convergence in prob.)

Theorem: $X_1, X_2, \cdots \overset{iid}{\sim} \cdot(\mu, \sigma^2)$

$\bar{X}_n = \frac{1}{n}\sum_1^n X_i \xrightarrow{p} \mu$. (WLLN)

pf $P(|\bar{X}_n - \mu| \geq \varepsilon) = P((\bar{X}_n - \mu)^2 \geq \varepsilon^2) \leq \frac{E(\bar{X}_n-\mu)^2}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \longrightarrow 0$ as $n\to\infty$. //

Theorem: $X_n \xrightarrow{p} X$ then $h(X_n) \xrightarrow{p} h(X)$ when h is a conti. func.

pf Given $\varepsilon > 0$, $\exists \delta > 0$ s.t. $|h(x_n) - h(x)| < \varepsilon$ whenever $|x_n - x| < \delta$.

$P(|h(X_n) - h(X)| > \varepsilon) = 1 - P(|h(X_n) - h(X)| \leq \varepsilon) \geq 1 - P(|X_n - X| < \delta)$

$\longrightarrow 0$. //

(e.g.) $X_1, X_2, \cdots \overset{iid}{\sim} N(\mu, \sigma^2)$. Then,

$S_n \xrightarrow{p} \sigma$.

sol. $P(|S_n^2 - \sigma^2| \geq \varepsilon) \leq \dfrac{Var\, S_n^2}{\varepsilon^2} = \dfrac{2\sigma^4}{(n-1)\,\varepsilon^2} \longrightarrow 0$ as $n \to \infty$. //

$h(x) = \sqrt{x}$: conti. in x.

$\therefore \; S_n \xrightarrow{p} \sigma$.

- $X_n \xrightarrow{a.s.} X$ if for every $\varepsilon > 0$ $P\left[\lim_{n\to\infty}\{|X_n - X| < \varepsilon\}\right] = 1$.

(almost sure convergence)

Theorem: $X_1, X_2, \cdots \overset{iid}{\sim} \bullet(\mu, \sigma^2)$

$\overline{X}_n \xrightarrow{a.s.} \mu$ (SLLN)

- $X_n \xrightarrow{d} X$ if $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$ at all $x \in \mathrm{Conti}(F_X)$.

(convergence in distribution)

(e.g.) $X_1, X_2, \cdots \overset{iid}{\sim} \mathrm{Uniform}(0,1)$

$X_{(n)} = \max_{1\le i\le n} X_i \xrightarrow{p} 1$

$\because P(|X_{(n)} - 1| \geq \varepsilon) = P(X_{(n)} \leq 1-\varepsilon) = (1-\varepsilon)^n \longrightarrow 0.$

If we take $\varepsilon = t/n$, then for all $t>0$ we have

$$P(X_{(n)} \leq 1 - t/n) = P\left(n(1 - X_{(n)}) \geq t\right) \longrightarrow e^{-t} \quad \text{as } n\to\infty.$$

That is, $P(n(1-X_{(n)}) \leq t) \longrightarrow 1 - e^{-t}$ as $n\to\infty$

$\therefore\ n(1-X_{(n)}) \xrightarrow{d} \mathrm{Exp}(1).$ //

Theorem : $X_n \xrightarrow{p} X$ implies $X_n \xrightarrow{d} X$.

$X_n \xrightarrow{p} \mu \iff X_n \xrightarrow{d} \mu$.

Theorem (Central limit theorem, CLT) :

$X_1, X_2, \cdots \overset{iid}{\sim} \cdot(\mu, \sigma^2)$ whose mgfn exists in a nbhd. of 0.

Then, $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0, 1)$.

pf. $Y_i = (X_i - \mu)/\sigma$, $i = 1, 2, \cdots$

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i .$$

$$M_{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma}}(t) = M_{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Y_i}(t) = M_{\sum_{i=1}^{n} Y_i}(t/\sqrt{n}) = \left\{M_{Y_1}(t/\sqrt{n})\right\}^n = \left(1 + \frac{1}{2}\cdot\frac{t^2}{n} + o(n^{-1})\right)^n$$

$$M_{Y_1}(t/\sqrt{n}) = 1 + \frac{(t/\sqrt{n})^2}{2} + o(n^{-1}) \qquad \longrightarrow e^{\frac{1}{2}t^2}$$

//.

Theorem (Slutsky's theorem):

$X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} a$

$\Longrightarrow$ $Y_n X_n \xrightarrow{d} aX$, $X_n + Y_n \xrightarrow{d} X + a$.

(e.g.) $\dfrac{\sqrt{n}(\bar{X}_n - \mu)}{S_n} \xrightarrow{d} N(0,1)$.

pf. (HW).

Theorem (Delta method):

$\sqrt{n}(Y_n - \theta) \rightarrow N(0, \sigma^2)$

g : a func. s.t. $g'(\theta) = 0$

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

pf. $g(Y_n) \simeq g(\theta) + g'(\theta)(Y_n - \theta)$. By Slutsky's theorem ... //

(e.g.) $\sqrt{n}\,(1/\bar{X} - 1/\mu) \xrightarrow{d} N\left(0, \left(\frac{1}{\mu}\right)^4 \sigma^2\right)$

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

Theorem (Second-order Delta method):

$\sqrt{n}\,(Y_n - \theta) \xrightarrow{d} N(0, \sigma^2)$

g: a function s.t. $g'(\theta) = 0$ and $g''(\theta) \neq 0$.

Then, $n(g(Y_n) - g(\theta)) \xrightarrow{d} \sigma^2 \cdot \frac{g''(\theta)}{2} \cdot \chi_1^2$

5.6 Generating a random sample

" To generate a r.s. $X_1, \cdots, X_n$ from a given $f(x|\theta)$."

(1) Direct methods

: when there is a closed formula $g(u)$ s.t. $Y = g(U)$, $U \sim Unif(0,1)$

- Prob. integral transform

To generate a r.v. $Y \sim F_Y$: $F_Y^{-1}(U) \sim F_Y$ when $U \sim Unif(0,1)$.

(e.g.) $F_Y = Exp(\lambda)$.

$F_Y^{-1}(u) = -\lambda \log(1-u)$, $U \sim Unif(0,1)$.

- Box - Müller

$U_1, U_2 \overset{iid}{\sim} Unif(0,1)$. $R = \sqrt{-2\log U_2}$, $\theta = 2\pi U_1$.

Then, $X = R\cos\theta$ and $Y = R\sin\theta$ are iid $N(0,1)$.

- $B(n, p)$

$$U \sim \text{Unif}(0,1)$$

$$Y = \begin{cases} 0 & \text{if} \quad 0 < U \le F_Y(1) \\ 1 & \text{if} \quad F_Y(1) < U \le F_Y(2) \\ 2 & \text{if} \quad F_Y(2) < U \le F_Y(3) \\ \vdots & \qquad \vdots \\ n & \text{if} \quad F_Y(n) < U \le 1. \end{cases}$$

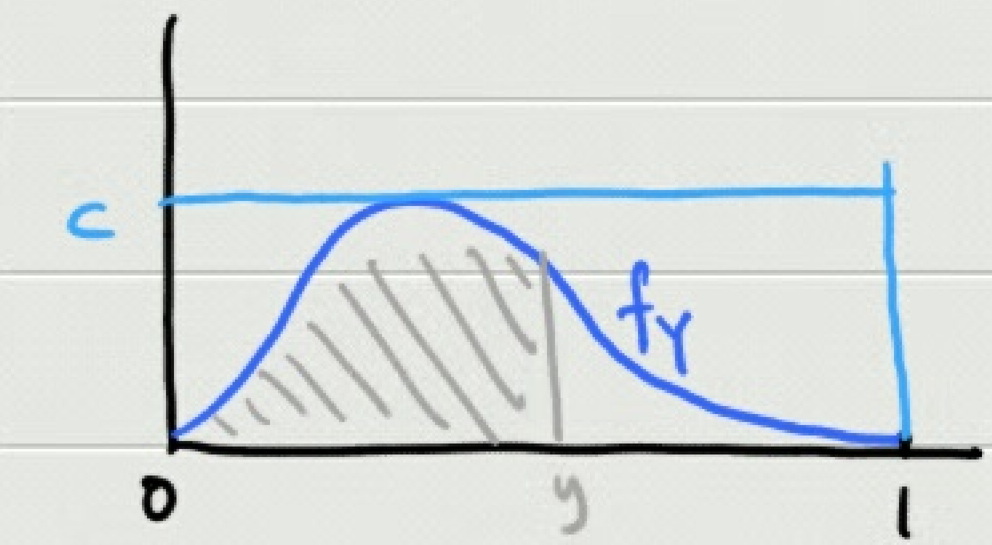
or

$$U_1, U_2, \cdots, U_n \overset{iid}{\sim} \text{Unif}(0,1)$$

$$Y = \sum_{i=1}^{n} I(U_i \le p).$$

(2) Indirect methods : Accept–reject algorithm

• A motivating example : to generate $Y \sim f_Y$.



$U, V \sim \text{iid Unif}(0,1)$

$$\underbrace{P(V \le y,\ U \le \tfrac{1}{c} f_Y(V))}_{\text{prob. of shaded area}} = \int_0^y \int_0^{f_Y(v)/c} du\, dv$$

$$= \frac{1}{c} \int_0^y f_Y(v)\, dv = \frac{1}{c} P(Y \le y).$$

If we set $y=1$, then $P(U \le \frac{1}{c} f_Y(V)) = \frac{1}{c}$.

$$\therefore F_Y(y) = P(Y \le y) = P(V \le y,\ U \le \tfrac{1}{c} f_Y(V)) \,/\, P(U \le \tfrac{1}{c} f_Y(V)) = P(V \le y \mid U \le \tfrac{1}{c} f_Y(V)).$$

//

a. Generate $(U, V) \overset{\text{iid}}{\sim} \text{Unif}(0,1)$

b. If $U < \frac{1}{c} f_Y(V)$, set $Y = V$; o.w. return to step (a).

Yes, this algorithm generates Beta(a,b) r.v. as long as $c \ge \max_y f_Y(y)$. Now, $c = ?$
Any efficient choice?

Theorem: $Y \sim f_Y$, $V \sim f_V$, $\operatorname{supp} f_Y = \operatorname{supp} f_V$

V can be easily generated, and $f_V \approx f_Y$.

$M = \sup_y f_Y(y) / f_V(y) < \infty$

To generate $Y \sim f_Y$:

a. Generate $U \sim \text{Unif}(0,1)$, $V \sim f_V$, indep.

b. If $U < \frac{1}{M} f_Y(V) / f_V(V)$, set $Y = V$; o.w., return to step (a).

Pf:
$$P\left(V \le y \,\middle|\, U < \frac{1}{M} f_Y(V)/f_V(V)\right) = \frac{\int_{-\infty}^{y} \int_0^{\frac{1}{M} f_Y(v)/f_V(v)} du \, f_V(v)\, dv}{\int_{-\infty}^{\infty} \int_0^{\frac{1}{M} f_Y(v)/f_V(v)} du \, f_V(v)\, dv} = \int_{-\infty}^{y} f_Y(v)\, dv$$

//

Remark: For $M < \infty$, f_V is required to have heavier tails than f_Y.

the candidate density

the target density

(e.g.) $Y \sim \text{Beta}(2.7,\ 6.3)$

a. Generate $U \sim \text{Unif}(0,1)$ and $V \sim \text{Beta}(2,6)$

b. If $U < \frac{1}{M} \cdot \frac{f_Y(V)}{f_V(V)}$, set $Y = V$; o.w., return to step (a).

* There are cases where the target density has heavy tails, and it is difficult to get candidate densities such that $M < \infty$.

Then, Accept-reject algorithm will no longer apply $\cdots\Rightarrow$ MCMC!