Random samples and iid variables

Definitions

- *Definition:* If $X_1, \dots, X_n$ are **independent identically distributed** (iid) with $X_i \sim f_X(x_i)$, then we call $X_1, \dots, X_n$ a **random sample** from the population $f_X(x)$.

R.V. $\Longrightarrow \mathbb{P}(Y \le y) \rightarrow$ Sampling distribution

$\mathbb{E}(Y) = \mathbb{E}(T(X_1, \dots, X_n))$

$\mathrm{Var}(Y)$

- $Y = T(X_1, \dots, X_n)$ is called a statistic.
  i.e., $T(X_1, \dots, X_n)$ is computable from the data

- The distribution of a statistic Y is sometime called the **sampling distribution** of the statistic.

- Examples

  1. sample mean: $\bar{X}_n = \sum_{i=1}^{n} X_i / n$

  2. sample variance:

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 = \frac{1}{n-1} \left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2 \right)$$

  $X_{(1)}$ 3. minimum: $\min\{X_1, \dots, X_n\}$

  $X_{(n)}$ 4. maximum: $\max\{X_1, \dots, X_n\}$

Random samples and iid variables

Distribution of $\bar{X}_n$

Let $X_1, \ldots, X_n$ be a random sample from $f_X(x)$ with $\mu = \mathrm{E}X_i$ and $\sigma^2 = \mathrm{Var}(X_i)$

i.i.d

Important Results for $\bar{X}_n$: If $X_1, \ldots, X_n$ is a sample random with $\mu = \mathrm{E}X_i$ and $\sigma^2 = \mathrm{Var}(X_i)$, then

1. $\mathrm{E}\bar{X}_n = \mu$

2. $\mathrm{Var}(\bar{X}_n) = \dfrac{\sigma^2}{n}$

$\boxed{M_{\bar{X}_n}(t) = \left[M_{X_1}(t/n)\right]^n \text{ If } X_1 \cdots X_n \text{ are i.i.d}}$

MGF approach can sometimes apply for determining the exact distribution of $\bar{X}_n$

(MGF of i.i.d X_i's)

$$M_{\bar{X}_n}(t) \underset{\text{def of MGF}}{=} \mathrm{E}e^{t\bar{X}_n} \underset{\text{def of }\bar{X}}{=} \mathrm{E}e^{n^{-1}t(X_1+\cdots+X_n)} \underset{e^{a+b}=e^a \cdot e^b}{=} \mathrm{E}\prod_{i=1}^{n} e^{n^{-1}tX_i} = \prod_{i=1}^{n} \mathrm{E}e^{n^{-1}tX_i} = [M_{X_1}(t/n)]^n$$

$$\mathbb{E}\left[e^{\frac{t}{n}X_1} e^{\frac{t}{n}X_2} \cdots e^{\frac{t}{n}X_n}\right] \quad X_i \text{ are ind.}$$

$$\mathbb{E}\left[e^{\frac{t}{n}X_1}\right] \cdots \mathbb{E}\left[e^{\frac{t}{n}X_n}\right] \quad X_i \text{ are identically dis.}$$

$$\left(\mathbb{E}\left[e^{\frac{t}{n}X_1}\right]\right)^n = \boxed{\left(M_{X_1}(t/n)\right)^n}$$

Examples

1. Suppose $X_1, \ldots, X_n$ are iid Gamma(α, β)

$$M_{\bar{X}_n}(t) = \left[M_{X_1}(t/n)\right]^n \ \frac{X_1 \sim \text{Gamma}(\alpha, \beta)}{M_X(t) = (1-\beta t)^{-\alpha}} \left[M_{X_1}(t/n)\right]^n$$

$$= \left[\left(1 - \beta \tfrac{t}{n}\right)^{-\alpha}\right]^n = \left(1 - \frac{\beta t}{n}\right)^{-n\alpha} \Rightarrow \bar{X}_n \sim \text{Gamma}\left(n\alpha, \frac{\beta}{n}\right)$$

2. Suppose $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$

$$M_{\bar{X}_n}(t) = \left[M_{X_1}(t/n)\right]^n \ \frac{X_i \sim N(\mu, \sigma^2)}{M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}} \left[e^{\frac{\mu t}{n} + \frac{\sigma^2 t^2}{2n^2}}\right]^n = e^{\mu t + \frac{\sigma^2}{n}\frac{t^2}{2}}$$

$$\Longrightarrow \boxed{\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)}$$

Random samples and iid variables

Distribution of S^2 (Sample Variance S^2)

Let $X_1, \ldots, X_n$ be a random sample from $f_X(x)$ with $\mu = \mathrm{E}X_i$ and $\sigma^2 = \mathrm{Var}(X_i)$

i.i.d

- The exact sampling distribution of sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}_n^2\right)$$

is difficult to obtain in general

- However, if $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$ then the sampling distribution of S^2 can be found (later... after scaling, the distribution is chi-square with $n-1$ degrees of freedom)

- **Result:** For random samples with $\mu = \mathrm{E}X_i$ and $\sigma^2 = \mathrm{Var}(X_i)$,

$$\mathrm{E}S^2 = \sigma^2 = \mathrm{Var}(X_i)$$

$\mathbb{E}X_i^2 = \mathrm{Var}\, X_i + (\mathbb{E}X_i)^2$

$\mathrm{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$ (*)

$\mathrm{Var}(\bar{X}_n) = \mathbb{E}[(\bar{X}_n)^2] - (\mathbb{E}(\bar{X}_n))^2$

$\frac{\sigma^2}{n} = \mathbb{E}[(\bar{X}_n)^2] - \mu^2$

$\Rightarrow \mathbb{E}[(\bar{X}_n)^2] = \mu^2 + \frac{\sigma^2}{n}$

def of S^2

Proof: $\mathbb{E}[S^2] = \mathbb{E}\left[\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right]\right]$

$= \frac{1}{n-1}\sum_{i=1}^{n}\mathbb{E}(X_i^2) - n\,\mathbb{E}[(\bar{X}_n)^2]$

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

$= \frac{1}{n-1}\left[n\sigma^2 + \cancel{n\mu^2} - \cancel{n\mu^2} - \sigma^2\right] = \frac{\sigma^2[\cancel{n-1}]}{\cancel{n-1}} = \sigma^2$

$\Rightarrow \boxed{\mathbb{E}[S^2] = \sigma^2}$ ■

Important

Random samples and iid variables

Distribution of Maximum and Minimum

$F_X(x) = \mathbb{P}(X \le x)$ where X has the same disribution of $X_1, \dots, X_n$

Let $X_1, \dots, X_n$ be a random sample (i.i.d) with common cdf $F_{X_1}(x) = P(X_1 \le x)$

Let $X_{(n)} = \max\{X_1, \dots, X_n\}$ and $X_{(1)} = \min\{X_1, \dots, X_n\}$

Important results:

1. $F_{X_{(n)}}(x) \underset{\text{def of CDF}}{=} P(X_{(n)} \le x) = [F_{X_1}(x)]^n$, for $x \in \mathbb{R}$

2. $F_{X_{(1)}}(x) \underset{\text{def of CDF}}{=} P(X_{(1)} \le x) = 1 - [1 - F_{X_1}(x)]^n$, for $x \in \mathbb{R}$

3. If the population cdf $F_{X_1}(x) = P(X_1 \le x)$ is continuous with pdf $f_{X_1}(x) = \frac{dF_{X_1}(x)}{dx}$, then $X_{(n)}$ and $X_{(1)}$ both have pdfs given by

$$f_{X_{(n)}}(x) = n f_{X_1}(x)[F_{X_1}(x)]^{n-1}, \qquad f_{X_{(1)}}(x) = n f_{X_1}(x)[1 - F_{X_1}(x)]^{n-1}$$

$$f_{X_{(n)}}(x) \overset{\text{def}}{=} \frac{d}{dx} F_{X_{(n)}}(x) = \frac{d}{dx}\left[F_{X_1}(x)\right]^n = n \left(\frac{d}{dx}F_{X_1}(x) = f_{X_1}(x)\right)\left(F_{X_1}(x)\right)^{n-1} = n f_{X_1}(x)\left[F_{X_1}(x)\right]^{n-1}$$

Proofs: (These are proofs that are useful to remember.)

$$F_{X_{(n)}}(x) \overset{\text{def of cdf}}{=} \mathbb{P}(X_{(n)} \le x) = \mathbb{P}(\max(X_1, \dots, X_n) \le x)$$

$$\overset{\text{def of } X_{(n)}}{=} \mathbb{P}(X_1 \le x, X_2 \le x, \dots, X_n \le x)$$

$$\overset{X_i\text{'s are independent}}{=} \mathbb{P}(X_1 \le x)\,\mathbb{P}(X_2 \le x) \cdots \mathbb{P}(X_n \le x)$$

$$\xrightarrow{X_i\text{'s are identically dist.}} [\mathbb{P}(X_1 \le x)]^n = [F_{X_1}(x)]^n$$

$$\boxed{F_{X_{(1)}}(x)} \overset{\text{def of CDF}}{=} \underbrace{\mathbb{P}(X_{(1)} \le x)} \overset{P(A^c) = 1 - \mathbb{P}(A)}{=} 1 - \mathbb{P}(X_{(1)} \ge x)$$

$$= 1 - \mathbb{P}(X_1 \ge x, X_2 \ge x, \dots, X_n \ge x)$$

$$= 1 - \mathbb{P}(X_1 \ge x)\,\mathbb{P}(X_2 \ge x) \cdots \mathbb{P}(X_n \ge x)$$

$$= 1 - \underbrace{[\mathbb{P}(X_1 \ge x)]}^n = 1 - [1 - \mathbb{P}(X_1 \le x)]^n$$

$$= \boxed{1 - [1 - F_{X_1}(x)]^n}$$

$$\boxed{F_{X_{(1)}}(x) = 1 - [1 - F_{X_1}(x)]^n}$$

Random samples and iid variables

Order statistics

- *Definition:* The **order statistics** for a sample $X_1, \ldots, X_n$ are the values in ascending order denoted as

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

- Primarily interested in iid $X_1, \ldots, X_n$ having a continuous distribution

- For random samples we may be interested in

  1. the distribution of a single order statistic $X_{(i)}$

  2. the distribution of two or more order statistics $(X_{(i)}, X_{(j)})$

  3. function of two or more order statistics

     e.g., range $R = X_{(n)} - X_{(1)}$

- order statistics are a type of (discontinuous) transformation of $X_1, \ldots, X_n$

Random samples and iid variables

Distribution of kth order statistic

i.i.d

Result 1: If $X_1, \dots, X_n$ are a random sample with common cdf $F_{X_1}(x)$, then the cdf of the kth order statistic (given some $k = 1, \dots, n$) is given by

$$F_{X_{(k)}}(x) = P(X_{(k)} \leq x) = P(\text{at least } k\ X_i\text{'s} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [F_{X_1}(x)]^j [1 - F_{X_1}(x)]^{n-j}$$

(out of n)

Proof: Let $Y = \#$ of X_i's which are less than or equal to x

$$\Rightarrow Y \sim \mathrm{Bin}(n, F_{X_1}(x))$$

$$\mathbb{P}(Y \geq k) \underset{Y \sim \mathrm{Bin}(n, F_{X_1}(x))}{=} \sum_{j=k}^{n} \mathbb{P}(Y = j) = \sum_{j=k}^{n} \binom{n}{j} (F_{X_1}(x))^j (1 - F_{X_1}(x))^{n-j}$$

Result 2 (pdf in continuous case): If $X_1, \dots, X_n$ are a random sample with common continuous cdf $F_{X_1}(x)$ and pdf $f_{X_1}(x)$, the pdf of the kth order statistic is

$$f_{X_{(k)}}(x) = \frac{dF_{X_{(k)}}(x)}{dx} = \frac{n!}{(k-1)!(n-k)!} f_{X_1}(x) [F_{X_1}(x)]^{k-1} [1 - F_{X_1}(x)]^{n-k}$$

- Heuristic argument for the form of the pdf $f_{X_{(k)}}(x)$:

  $k-1$ observations $\leq x$; 1 observation in $(x, x + dx)$; $n-k$ observations $> x$

- A formal proof uses derivative of cdf + algebra (see next slide)

Note: in the discrete case, the pmf of $X_{(k)}$ is obtained as

$$f_{X_{(k)}}(x) = P(X_{(k)} = x) = P(X_{(k)} \leq x) - P(X_{(k)} < x) = F_{X_{(k)}}(x) - \lim_{y \uparrow x} F_{X_{(k)}}(y)$$