Properties of Random Sample

Definition 5.1.1 :

The random variables $X_1, \ldots, X_n$ are called a random sample of size n from the population $f(x)$ if $X_1, X_2, \ldots, X_n$ are mutually independent random variables and the marginal pdf or pmf of each X_i is the same function $f(x)$. Alternatively, $X_1, X_2, \ldots, X_n$ are called independent and identically distributed random variables with pdf or pmf $f(x)$. This is commonly abbreviated to iid random variables.

X_1 is the 1^{st} observation made from $f(x)$

X_2 is the 2^{nd} observation made from $f(x)$

$\vdots$

X_n is the n^{th} observation made from $f(x)$

* under the random sampling model each X_i is an observation on the same variable and each X_i has a marginal distribution given by $f(x)$.

* Furthermore, the observation's are taken in such a way that the value of one observation has no effect on or relationship with any of the other observation's. ⟹ $X_1, X_2, \ldots X_n$ are mutually independent.

$$f(x_1, x_2, x_3 \ldots x_n) = \prod_{i=1}^{n} f(x_i)$$

⟹ if the population pdf or pmf is a member of a parametric family with pdf $f(x|\theta)$ then the joint pdf or pmf

$$f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta)$$

θ true parameter is unknown.

Example 5.1.2:

Let $X_1, X_2, \ldots, X_n$ be a random sample from a exponential (β)

$$f(x_1, x_2, \ldots, x_n \mid \beta) = \prod_{i=1}^{n} f(x_i \mid \beta)$$

$$= \prod_{i=1}^{n} \frac{1}{\beta} e^{-\frac{x_i}{\beta}}$$

$$= \frac{1}{\beta^n} e^{-\frac{(x_1+x_2+\cdots+x_n)}{\beta}}$$

$$\mathbb{P}(X_1>2, X_2>2, \ldots, X_n>2)$$

$$= \int_2^{\infty}\int_2^{\infty}\cdots\int_2^{\infty} \prod_{i=1}^{n} \frac{1}{\beta} e^{-\frac{x_i}{\beta}} dx_1 \ldots dx_n$$

$$= \int_2^{\infty}\int_2^{\infty}\cdots\int_2^{\infty} -e^{-\frac{x_1}{\beta}}\Big]_2^{\infty} dx_2 \ldots$$

$$= \int_2^{\infty}\int_2^{\infty}\cdots\int_2^{\infty} e^{-\frac{2}{\beta}}$$

$$= \left(e^{-\frac{2}{\beta}}\right)^{n} = e^{-\frac{2n}{\beta}}$$

This is called Sampling from infinite population.

when sampling is from finite population, The above def may not apply.

- A finite population is a finite set of number's $\{x_1, \ldots x_N\}$.

A sample $x_1, x_2, \ldots x_n$ is drawn from the finite population. Four way's of drawing this sample.

① finite sampling with replacement

② finite sampling without replacement.

(simple random sampling)

5.2 Sums of Random Variables from a Random Sample.

When a sample $X_1, X_2, .. X_n$ is drawn, some summary of the values is usually computed.

$\Rightarrow$ $T(x_1, x_2, .. , x_n)$ is a function of n random sample's, whole domain includes the sample space of the random vector $(X_1, X_2, .. X_n)$

$\Rightarrow$ The function can be real valued or vector valued

$\Rightarrow$ $Y = T(X_1, X_2, .. X_n)$ is a random vector or random variable.

$\Rightarrow$ Since $X_1, X_2, \dots, X_n \overset{iid}{\sim} X$

The distribution of Y is tractable.

Because this distribution is usually derived from the distribution of the variables in the random sample. it is called the "Sampling distribution of Y".

in this section, we will discuss some properties of sampling distribution especially for functions $T(x_1, \ldots x_n)$ defined by sum's of r.v's.

Definition 5.2.1 :

Let $X_1, X_2, \ldots X_n$ be a random Sample of size n from a population and let $T(x_1, x_2, \ldots x_n)$ be a real-valued or vector-valued function whose domain includes the sample space $(X_1, X_2, \ldots X_n)$ Then the random variable or random vector $Y = T(X_1, \ldots X_n)$ is called a Statistic. The Probability distribution of Statistic Y is called the sampling distribution of Y.

Statistic cannot be a function of parameter.

Three Statistics that are often used and provide good summaries of the sample are now defined.

Definition 5.2.2 :

The "sample mean" is the arithmetic average of the values in a random sample. It is usually denoted by.

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_i$$

Definition 5.2.3 :

The "sample variance" is the statistic defined by

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

The "Sample standard deviation" is the statistic defined by $S = \sqrt{S^2}$

The observed values of statistics are denoted with lowercase letter. $\bar{x}$, s^2, s.

Properties of sample mean & variance.

Theorem 5.2.4:- Let $x_1, x_2, \dots x_n$ be any numbers and $\bar{x} = \frac{(x_1+x_2+\cdots+x_n)}{n}$. Then

a) $$\min_a \sum_{i=1}^{n} (x_i-a)^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2$$

b) $$(n-1)s^2 = \sum_{i=1}^{n} (x_i-\bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Proof:

a) $$\sum_{i=1}^{n} (x_i-a)^2 = \sum_{i=1}^{n} (x_i-\bar{x}+\bar{x}-a)^2$$

$$= \sum_{i=1}^{n} (x_i-\bar{x})^2 + (\bar{x}-a)^2 + 2(x_i-\bar{x})(\bar{x}-a)$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n \cdot (\bar{x} - a)^2 + 2(\bar{x} - a) \sum_{i=1}^{n} (x_i - \bar{x})$$

$$= \sum_{i=1}^{n} (x_i - \bar{x})^2 + n \cdot (\bar{x} - a)^2 + 2(\bar{x} - a)\, \underbrace{n(\bar{x} - \bar{x})}_{0}$$

$$= \underbrace{\sum_{i=1}^{n} (x_i - \bar{x})^2}_{\text{Fixed}} + \underbrace{n(\bar{x} - a)^2}_{\text{for min } \bar{x} - a = 0 \Rightarrow a = \bar{x}}$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

(b)

$$\sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (\bar{x} - a)^2$$

take $a = 0$

$$\Rightarrow \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - a)^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2$$

$$\Rightarrow \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$\boxed{\Rightarrow (n-1) S^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$

Lemma 5.2.5:

Let $X_1, \ldots X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathbb{E}[g(X_1)]$ and $\mathrm{Var}(g(X_1))$ exists then.

$$\mathbb{E}\left[\sum_{i=1}^{n} g(X_i)\right] = n\,\mathbb{E}[g(X_1)]$$

$$\mathrm{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\,\mathrm{Var}(g(X_1))$$

Proof:

$$\mathbb{E}\left[\sum_{i=1}^{n} g(X_i)\right] = \sum_{i=1}^{n} \mathbb{E}[g(X_i)]$$

as X_i's are identically distributed

$$= \quad n\,\mathbb{E}\,g(x_1)$$

2 $$\mathrm{Var}\left(\sum_{i=1}^{n} g(x_i)\right)$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} g(x_i) - \mathbb{E}\left[\sum_{i=1}^{n} g(x_i)\right]\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} g(x_i) - \sum_{i=1}^{n} \mathbb{E}[g(x_i)]\right)^2\right]$$

$$= \mathbb{E}\left[\left(\sum_{i=1}^{n} \left(g(x_i) - E[g(x_i)]\right)\right)^2\right]$$

$$\Rightarrow \left[\sum_{i=1}^{n} \left(g(x_i) - \mathbb{E}[g x_i]\right)\right]^2 = n \text{ term's}, \; n^2 - n \text{ term'}$$

$$\Rightarrow \left\{g(x_i) - E\,g(x_i)\right\}^2 \longrightarrow n \text{ term's}$$

$$\left(g(x_i) - Eg(x_i)\right)\left(g(x_j) - Eg(x_j)\right) \quad i \neq j$$

$\longrightarrow n(n-1)$ terms

$$\Rightarrow \mathbb{E}\left[\left(g(x_i) - Eg(x_i)\right)^2\right] = \text{var}(g(x_i))$$

$$= \text{var}(g(x_1))$$

$$\mathbb{E}\left[\left(g(x_i) - Eg(x_i)\right)\left(g(x_j) - Eg(x_j)\right)\right]$$

$i \neq j$

$\Rightarrow$ x_i, x_j are independent.

$\Downarrow$

$$\text{Cov}\left(g(x_i), g(x_j)\right) = 0$$

$$\Rightarrow \text{var}\left(\sum_{i=1}^{n} g(x_i)\right) = n\,\text{var}(g(x_i))$$

<u>Theorem 5.2.6</u> :-

Let $X_1, X_2, \ldots X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$, then

ⓐ $\mathbb{E}[\bar{X}] = \mu$

ⓑ $Var(\bar{X}) = \frac{\sigma^2}{n}$

ⓒ $\mathbb{E}[S^2] = \sigma^2$

Proof:

ⓐ $\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right]$

$$= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] =$$

$$= \frac{1}{n}\cdot n\mu = \mu$$

ⓑ $Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$

$$= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2}\cdot n\cdot Var(X_i) = \frac{\sigma^2}{n}$$

(c) $\mathbb{E}[s^2] = \mathbb{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]$

$= \frac{1}{n-1}\mathbb{E}\left\{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2\right\}$

$= \frac{1}{n-1}\left[\mathbb{E}\left[\sum_{i=1}^{n}x_i^2\right] - n\,\mathbb{E}[\bar{x}]^2\right]$

$= \frac{1}{n-1}\left[n\mathbb{E}[x_1^2] - n\mathbb{E}[\bar{x}]^2\right]$

$= \frac{1}{n-1}\left[n(\sigma^2+\mu^2) - n(\mathrm{var}(\bar{x}) + \mathbb{E}[\bar{x}]^2)\right]$

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

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

$= \frac{(n-1)}{n-1}\sigma^2 = \sigma^2$

in theorem 5.2.6, relationship's between a statistic and a population parameter, are examples of unbiased statistics.

$\Rightarrow$ $\overline{X}$ is a unbiased estimator of μ

$\Rightarrow$ S^2 is a unbiased estimator of σ^2

$$\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

if $Y = X_1 + X_2 + \cdots + X_n$

& $Y \sim f(y)$

then $\overline{X} = \frac{Y}{n} \Rightarrow Y = n\overline{X}$

$\Rightarrow \frac{dY}{d\bar{x}} = n$

$$f_{\overline{X}}(\bar{x}) = f_Y(n\bar{x}) \cdot n$$

$$\Rightarrow f_{\overline{X}}(\bar{x}) = n \cdot f_Y(n\bar{x})$$

$$M_{\bar{X}}(t) = \mathbb{E}\left[e^{t\bar{X}}\right]$$

$$= \mathbb{E}\left[e^{t\frac{(X_1+X_2+\cdots+X_n)}{n}}\right]$$

$$= \mathbb{E}\left[e^{\frac{t}{n}Y}\right]$$

$$M_{\bar{X}}(t) = M_Y(t/n)$$

$$f_{\bar{X}}(\bar{x}) = n \cdot f_Y(n\bar{x})$$

Theorem 5.2.7 :-

Let $X_1, X_2, \ldots X_n$ be a random sample from a population with mgf $M_X(t)$. Then the mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[M_X(t/n)\right]^n$$

Proof:

$$M_{\bar{X}}(t) = \mathbb{E}\left[e^{t\bar{X}}\right]$$

$$= \mathbb{E}\left[e^{\frac{t}{n}(X_1 + X_2 + \cdots + X_n)}\right]$$

$$= \mathbb{E}\left[e^{\frac{t}{n}X_1} \cdot e^{\frac{t}{n}X_2} \cdot e^{\frac{t}{n}X_3} \cdots e^{\frac{t}{n}X_n}\right]$$

$$= \mathbb{E}\left[e^{\frac{t}{n}X_1}\right] \cdot \mathbb{E}\left[e^{\frac{t}{n}X_2}\right] \cdots \mathbb{E}\left[e^{\frac{t}{n}X_n}\right]$$

$$= M_{X_1}\left(\frac{t}{n}\right) M_{X_2}\left(\frac{t}{n}\right) \cdots M_{X_n}\left(\frac{t}{n}\right)$$

$X_1, X_2, \ldots X_n \overset{iid}{\sim} X$

$$= M_X\left(\frac{t}{n}\right) M_X\left(\frac{t}{n}\right) \cdots M_X\left(\frac{t}{n}\right)$$

$$= \left[M_X\left(\frac{t}{n}\right)\right]^n$$

Example 5.2.8:

$X_1, X_2, \dots X_n \overset{iid}{\sim} n(\mu, \sigma^2)$

The the mgf of the sample mean is

$$M_{\bar{X}}(t) = \left[M_X(t/n)\right]^n$$

$$= \exp\left(\mu \frac{t}{n} + \frac{\sigma^2 t^2}{2n^2}\right)^n$$

$$= \exp\left(\mu t + \frac{\sigma^2/n \cdot t}{2}\right)$$

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

another.

$X_1, X_2, \dots X_n \sim \text{gamma}(\alpha, \beta)$

$$M_X(t) = \left(\frac{1}{1-\beta t}\right)^\alpha$$

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

$$= \left(\frac{1}{1 - \frac{\beta t}{n}} \right)^{\alpha n}$$

$$= \left(\frac{1}{1 - \frac{\beta}{n} \cdot t} \right)^{\alpha n}$$

$$\Rightarrow \quad \bar{X} \sim \text{gamma}\left(n\alpha, \frac{\beta}{n} \right)$$

Theorem 5.29 :-

if X and Y are independent continuous r.v with pdf's $f_X(x)$ $f_Y(y)$ then a pdf of $Z = X + Y$ is

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(w) \, f_Y(z-w) \, dw$$

Proof:

$$Z = X + Y \qquad\qquad X = W$$

$$W = X \quad \Rightarrow \quad Y = Z - W$$

$$J = \begin{vmatrix} 0 & 1 \\ 1 & -1 \end{vmatrix} = -1$$

$$f_{Z,W}(z,w) = f_{X,Y}(w, z-w) \cdot 1 = f_X(w) f_Y(z-w)$$

$$\Rightarrow \quad f_Z(z) = \int_w f_{Z,W}(z,w)$$

$$= \int_{-\infty}^{\infty} f_X(w)\, f_Y(z-w)\, dw$$

Example 5.2.10

Sum of Cauchy r.v.'s

Sol^n

$$Z_1, Z_2, \ldots, Z_n \sim \text{Cauchy}(0,1)$$

$$\bar{Z} = \frac{Z_1 + Z_2 + \cdots + Z_n}{n}$$

* Let $U \sim \text{Cauchy}(0, \sigma)$

$V \sim \text{Cauchy}(0, \tau)$

$$Z = U + V \; ?$$

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{\pi\sigma} \frac{1}{1+\left(\frac{\omega}{\sigma}\right)^2} \cdot \frac{1}{\pi\gamma} \frac{1}{1+\left(\frac{z-\omega}{\gamma}\right)^2} \, d\omega$$

$$= \frac{1}{\pi\sigma \cdot \pi\gamma} \int_{-\infty}^{\infty} \frac{\sigma^2 \cdot \gamma^2}{(\sigma^2+\omega^2)(\gamma^2+(z-\omega)^2)} \, d\omega$$

$$= \frac{\sigma\gamma}{\pi^2} \int_{-\infty}^{\infty} \left[\frac{A}{\sigma^2+\omega^2} + \frac{B}{\gamma^2+(z-\omega)^2} \right] d\omega$$

$$f_Z(z) = \frac{1}{\pi(\sigma+\gamma)} \frac{1}{1+\left(\frac{z}{\sigma+\gamma}\right)^2}$$

$$Z \sim \text{Cauchy}(0, \gamma+\sigma)$$

$$\frac{Z_1+Z_2+\cdots+Z_n}{n} \sim \text{Cauchy}(0,1)$$

Suppose $X_1, X_2, \dots X_n \overset{iid}{\sim} X$

$$f_X(x) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

$\Rightarrow \exists \; Z_1, Z_2, \dots Z_n \sim Z$

where $X_i = \sigma Z_i + \mu$

Theorem 5.2.11:

Suppose $X_1, X_2, \dots X_n$ is a random sample from a pdf or pmf $f(x|\theta)$

$$f(x|\theta) = h(x) c(\theta) \exp\left[\sum_{i=1}^{k} w_i(\theta) t_i(x)\right]$$

Define statistics $T_1, T_2, \dots T_k$ by

$$T_i(X_1, X_2 \dots X_n) = \sum_{j=1}^{n} t_i(X_j)$$

$i = 1, \dots, k$

if the set $\{(w_1(\theta), w_2(\theta) \dots, w_k(\theta)), \theta \in \Theta\}$

contain's on a open set $\mathbb{R}^k$, the

the distribution of $(T_1, \ldots, T_k)$ is an exponential family of the form

$$f_T(u_1, \ldots, u_k | \theta)$$

$$= H(u_1, \ldots u_k) \left[c(\theta)\right]^n \exp\left(\sum_{i=1}^{k} w_i(\theta) u_i\right)$$