Random Samples and Statistics

Defn: Random Sample

If $X_1, \dots, X_N$ are mutually independent rvs

"sample size"

X is a r.v.

all w/ marginal distribution f

marginal pmf/pdf

then we say $X_1, \dots, X_N$ are a random sample from f.

Denoted: $X_n \overset{iid}{\sim} f$

independent identically distributed

Notation:

$\underset{\sim}{X} = (X_1, \dots, X_N)$ mv rv (random vector)

$\underset{\sim}{x} = (x_1, \dots, x_N) \in \mathbb{R}^n$ (a real vector)

Note:

joint distribution

$$f(\underset{\sim}{x}) = f(x_1, x_2, \dots, x_N)$$

note: all marginals are same

$$= f(x_1) f(x_2) \cdots f(x_N) \quad \text{[by independence]}$$

marginal dist

$$= f(x_1)f(x_2)\cdots f(x_N)$$ [by independence]

marginal dist

$$= \prod_{n=1}^{N} f(x_n)$$

[similar notation to $\sum_{i=1}^{N} x_i = x_1 + x_2 + \cdots$]

Ex. If $X_n \overset{iid}{\sim} Exp(\lambda)$

What's the joint?

$$f(\underset{\sim}{x}) = \prod_{n=1}^{N} f(x_n)$$

$$= \prod_{n=1}^{N} \left(\lambda e^{-\lambda x_n} \mathbb{1}(x_n > 0)\right)$$

$$= \lambda^N e^{-\lambda \sum_{n=1}^{N} x_n} \cdot \mathbb{1}(x_1 > 0)\mathbb{1}(x_2 > 0)\cdots\mathbb{1}(x_N > 0)$$

If $Y \sim Exp(\lambda)$

$$\boxed{f(y) = \lambda e^{-\lambda y} \text{ for } y > 0}$$

$$= \lambda e^{-\lambda y} \underbrace{\mathbb{1}(y > 0)}_{\text{indicator function}}$$

$$\mathbb{1}(\text{statement}) = \begin{cases} 0 & \text{statement} = \text{False} \\ 1 & \text{statement} = \text{True} \end{cases}$$

Often see

$$f(y) = \text{~~~} \mathbb{1}(y \in \text{Support})$$

$$= \begin{cases} \text{~~~} & y \in \text{Support} \\ 0 & \text{else} \end{cases}$$

Defn: Statistic

Given (a random sample) $\boxed{X_n \overset{iid}{\sim} f}$ a

Statistic is a function $T: \mathbb{R}^N \to \mathbb{R}^d$.

__Statistic__ is a function $T : \mathbb{R}^{N} \to \mathbb{R}^{d}$.

sample size

Typically $d \ll N$, often $d = 1$.

input N numbers $\xrightarrow{T}$ summary in lower dimension.

__Ex.__ __Arithmetic mean__ $(d=1)$

$$T(\underline{X}) = \frac{1}{N}\sum_{n=1}^{N} X_n = \overline{X}_N$$

__Sample Variance__

$$S^2_{N-1} = \frac{1}{N-1}\sum_{n=1}^{N}\left(X_n - \overline{X}_N\right)^2$$

__Sample SD:__

$$S_{N-1} = \sqrt{S^2_{N-1}}$$

__Minimum:__ $X_{(1)}$

__Maximum:__ $X_{(N)}$

__Range:__ $X_{(N)} - X_{(1)}$.

Defn: Sampling Distribution

A sampling distribution of a statistic T is the distribution of $T(\underset{\sim}{X})$.

Theorem:

If $X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\underbrace{\overline{X_N} \sim N(\mu, \sigma^2/N)}_{\text{Sampling dist.}}$.

Theorem: Sums of RVs.

Let $g: \mathbb{R} \to \mathbb{R}$ and $X_n \overset{iid}{\sim} f$

$$\boxed{(1) \quad E\left[\sum_{n=1}^{N} g(X_n)\right] = N \underbrace{E[g(X_1)]}_{\text{expectation of any of them}}}$$

could use $X_2, X_3, \dots$

recall: $E[A+B] = E[A] + E[B]$

then

$E[g(X_1)] = E[g(X_2)] = \cdots$

$$E\left[\sum_{n=1}^{N} g(X_n)\right] = \sum_{n=1}^{N} E[g(X_n)]$$

$$= N\, E[g(X_1)].$$

$$\boxed{(2) \quad Var\left(\sum_{n=1}^{N} g(X_n)\right) = N\, Var(g(X_1))}$$

proof? Similar to above using fact that $Var(A+B) = Var(A) + \dots$

$Var(B)$

if $A \perp\!\!\!\perp B$.

(works if uncorrelated)

<u>Theorem:</u> If $X_n \overset{iid}{\sim} f$ and

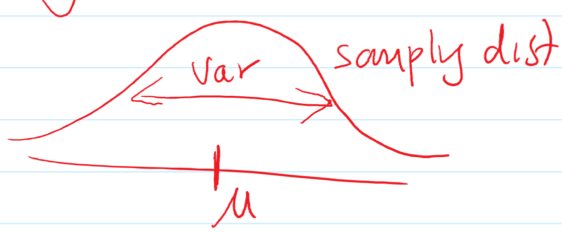
$E[X_n] = \mu$ and $Var(X_n) = \sigma^2$.

Then

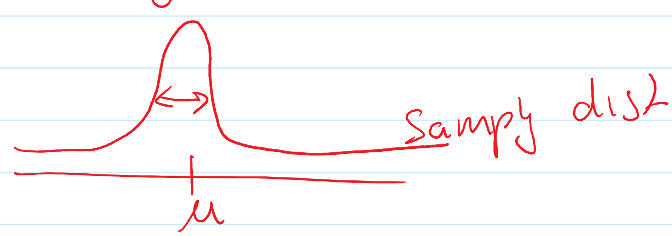
1. $E[\overline{X}_N] = \mu$ (avg. value of $\overline{X}_N$ is μ)
2. $Var(\overline{X}_N) = \sigma^2/N$ (more data is better)
3. $E[S_{N-1}^2] = \sigma^2$ (on average correctly estimates σ^2)

---

low N



higher N



<u>proof</u>. 1. $E[\overline{X}_N] = \mu$

$$E\left[\frac{1}{N}\sum_{n=1}^{N} X_n\right] = \frac{1}{N}\sum_{n=1}^{N} E[X_n] = \frac{1}{N}\sum_{n=1}^{N}\mu = \frac{1}{N}N\mu = \mu$$

(2) $\mathrm{Var}(\bar{X}_N) = \mathrm{Var}\left(\frac{1}{N}\sum_{n=1}^{N} X_n\right) = \frac{1}{N^2}\mathrm{Var}\left(\sum_{n=1}^{N} X_n\right)$

$= $ by prev. theorem

$= \frac{1}{N^2} N \,\mathrm{Var}(X_1)$

$= \frac{1}{N}\sigma^2$

(3) $\underline{E[S_{N-1}^2]}$

$=$

$\frac{1}{N-1} E\left[\sum_{n=1}^{N} X_n^2 - N\bar{X}_N^2\right]$

$= \frac{1}{N-1} E\left[\sum_{n=1}^{N} X_n^2\right] - \frac{N}{N-1} E[\bar{X}_N^2]$

$= \frac{1}{N-1}\sum_{n=1}^{N} E[X_n^2] - \frac{N}{N-1} E[\bar{X}_N^2]$

Fact! $\sum_{n=1}^{N}(X_n - \bar{X}_N)^2 = \sum_{n=1}^{N} X_n^2 - N\bar{X}_N^2$

(recall: $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$)

Fact: $E[Y^2] = \mathrm{Var}(Y) + E[Y]^2$

recall $E[X_n] = \mu \quad E[\bar{X}_N] = \mu$

$\mathrm{Var}(X_n) = \sigma^2 \quad \mathrm{Var}(\bar{X}_N) = \sigma^2/N$

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

$=$ algebra ...

$= \sigma^2.$

Saw: $\mathbb{E}[\bar{X}_N] = \mu$ $\quad$ $\bar{X}_N$ is unbiased for μ

$\mathbb{E}[S_{N-1}^2] = \sigma^2$ $\quad$ S_{N-1}^2 is unbiased for σ^2

Defn: Unbiased

We say a statistic T is unbiased for a quantity θ if $\mathbb{E}[T(\underline{X})] = \theta$.

Theorem: MGFs of $\bar{X}_N$.

Let $X_n \overset{iid}{\sim} f$ then

$$\boxed{M_{\bar{X}_N}(t) = \left(M_{X_1}(t/N)\right)^N}$$

proof.

$$M_{\bar{X}}(t) = \mathbb{E}\left[e^{t\bar{X}}\right] = \mathbb{E}\left[e^{t\frac{1}{N}\sum_{n=1}^{N} X_n}\right]$$

$$= \mathbb{E}\left[\prod_{n=1}^{N} e^{t\frac{1}{N}X_n}\right] \quad \left(\text{property of exp } e^{a+b} = e^a e^b\right)$$

$$= \prod_{n=1}^{N} \mathbb{E}\left[e^{t/N\, X_n}\right] \quad (\text{by independence})$$

$\mathbb{E}[AB] = \mathbb{E}[A]\,\mathbb{E}[B]$

$$= \prod_{n=1}^{N} M_{X_n}(t/N)$$

$$= \underbrace{M_{X_1}(t/N)\, M_{X_1}(t/N) \cdots M_{X_1}(t/N)}_{N}$$

$$= \left(M_{X_1}(t/N)\right)^N$$

Theorem

Let $X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ then $\overline{X}_N \sim N(\mu, \sigma^2/N)$.

proof

$$M_{\overline{X}}(t) = \left(M_{X_1}(t/N)\right)^N$$

For $X_1 \sim N(\mu, \sigma^2)$

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

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

algebra: $\boxed{(e^a)^b = e^{ab}}$

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

Notice that is the MGF of a $N(\mu, \sigma^2/N)$

So $\overline{X} \sim N(\mu, \frac{\sigma^2}{N})$.