Recall :- $X_1, X_2, \ldots, X_n$ be i.i.d. random variables

Emperical distribution is the discrete distribution with p.m.f

$$f_n(t) = \frac{1}{n} \#\{i : X_i = t\}$$

Descriptive statistics { - Tools of probability

- Make no additional assumptions

Emperical distribution - Random Quantity ≡ change with each sample

as n - gets larger - "captures information" about the underlying distribution.

Given by underlying distribution

if we compute Probabilities of events that we are interested in from Emperical distribution then these will approach the true Probabilities of the events

Focus of this week :- How to make this idea rigorous ?

Empirical Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "empirical distribution" based on these is the discrete distribution with probability mass function given by

$$f(t) = \frac{1}{n}\#\{X_i = t\}.$$

Ex:- Y is a r.v. with p.m.f $f_n(\cdot)$

i.e $\mathbb{P}(Y=t) = f_n(t)$

$$E[Y] = \sum_{t \in T} t\, \mathbb{P}(Y=t) = \overline{X}$$

(see below for definition)

Sample Mean

let X be a random variable
$E[X] = \mu$ $\quad$ $Var[X] = \sigma^2$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables. The "sample mean" of these is

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

linearity of expectation

- $E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$

$$= \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \frac{n\, E[X]}{n} = \mu$$

How well does $\bar{X}$ estimate μ?

$$\boxed{E[\bar{X}] = \mu} \quad (1)$$

$$\cdot \ \mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right]$$

Properties of variance $= \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}[X_i]$ $\left[\because \ \mathrm{Cov}[X_i, X_j] = 0, \ i \neq j \right]$ independence

$$= \frac{1}{n^2} \, n \cdot \mathrm{Var}[X]$$

$$= \boxed{\frac{1}{n} \sigma^2}$$

$$\therefore \ SD[\bar{X}] = \boxed{\frac{\sigma}{\sqrt{n}}}$$ — S.D reduction as n gets larger

"Range of $\bar{X}$" $=$ "$[-SD[\bar{X}] + \mu, \ SD[\bar{X}] + \mu]$"

$$= \underbrace{\left(\mu - \frac{\sigma}{\sqrt{n}}, \ \mu + \frac{\sigma}{\sqrt{n}}\right)}_{\text{interval gets smaller as } n \text{ gets larger}}$$

n- gets large $\bar{X}$ — concentrates around μ.

(1) - $E[\bar{X}] = \mu$ $\equiv$ unbiased estimate

(2) - $SD[\bar{X}] \to 0$ as $n \to \infty$ $\equiv$ Consistent estimate.

Question :- let A be an event of interest

$$\mathbb{P}(X \in A) = ?$$

Let $X_1, X_2, \ldots, X_n$ be i.i.d. X

let Y be a random variable with p.m.f

the Emperical distribution $\equiv f_n(\cdot)$

- $\mathbb{P}(X \in A) \approxeq \mathbb{P}(Y \in A) = \sum_{t \in A} f_n(t)$.

approximation

$$\mathbb{P}(Y \in A) = \frac{1}{n} \# \{i : X_i \in A\}$$

$f_n(\cdot)$ gave mass $\frac{1}{n}$ to each x_i

Proportion of $X_i \in A$

– Sample proportion

Question:- Is $\mathbb{P}(Y \in A)$ a good estimator for $p = \mathbb{P}(X \in A)$?

Reasoning :-

$$Z_i = \begin{cases} 1 & \text{if } X_i \in A \\ 0 & \text{if } X_i \notin A \end{cases}$$

- $\{Z_i\}_{i \geq 1}$ are independent and $\mathbb{P}(Z_i = 1) = p$.

$$\Rightarrow \quad W = \sum_{i=1}^{n} Z_i \sim \text{Binomial}(n, p)$$

Note $\mathbb{P}(Y \in A) = \frac{W}{n}$.

Sample Proportion

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables with the same distribution as a random variable X, and suppose that we are interested in the value $p = P(X \in A)$ for an event A. Let

$$\hat{p} = \frac{\#\{X_i \in A\}}{n}.$$

Then, $E(\hat{p}) = P(X \in A)$ and $Var(\hat{p}) \to 0$ as $n \to \infty$.

– "Sample proportion" $\hat{p}$

n-large $\cong$ close

"True Proportion" $p = P(X \in A)$

Probability of A from the empirical distribution unbiased

As: $\hat{p} = \mathbb{P}(\text{YEA})$ and $E[\hat{p}] = E[\frac{W}{n}] = \frac{1}{n} np = p$

$W \sim \text{Binomial}(n, p)$

$$Var[\hat{p}] = Var[\frac{W}{n}] = \frac{1}{n^2} Var[W]$$

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

$Var[\hat{p}] \longrightarrow 0$ as $n \to \infty$

Consistent

Relative frequency in n trials $\xrightarrow[\text{close}]{n - \text{large}}$ Probability

Let $X_1, X_2, \ldots X_n$ be i.i.d X. $E[X] = \mu$
and $Var[X] = \sigma^2$

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n)$$

Already observe $E[\bar{X}] = \mu$

$Var[\bar{X}] = \frac{\sigma^2}{n}$.

Quantity / Event of interest :- $\{|\bar{X} - \mu| > \varepsilon\}$ for some $\varepsilon > 0$

Markov's Inequality

$$\mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{1}{\varepsilon^2}\,\mathbb{E}|\bar{X}-\mu|^2$$

$$= \frac{\mathrm{Var}[\bar{X}]}{\varepsilon^2}$$

$$= \frac{\sigma^2}{n\varepsilon^2}$$

$$\therefore \quad 0 \leq \mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

$$\boxed{\lim_{n\to\infty} \frac{\sigma^2}{\varepsilon^2}\frac{1}{n} = 0}$$

Squeeze Theorem

$\forall \varepsilon > 0$

$$\therefore \quad \lim_{n\to\infty} \mathbb{P}(|\bar{X}-\mu| \geq \varepsilon) = 0$$

$$\Updownarrow$$

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(\bar{X} \notin (\mu-\varepsilon, \mu+\varepsilon)) = 0$$

$$\Updownarrow$$

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(\bar{X} \in (\mu-\varepsilon, \mu+\varepsilon)) = 1$$

Weak Law of Large Numbers

$\boxed{\text{Proof is above}}$

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and finite variance σ^2. Then for any $\epsilon > 0$

$$\lim_{n\to\infty} P(\mid \bar{X}_n - \mu \mid > \epsilon) = 0, \tag{1}$$

$$\lim_{n\to\infty} \mathbb{P}(\bar{X}_n \notin (\mu-\varepsilon, \mu+\varepsilon) = 0$$

$$\lim_{n\to\infty} \mathbb{P}(\bar{X}_n \in (\mu-\varepsilon, \mu+\varepsilon) = 1$$

Strong Law of Large Numbers

- "stronger" than the weak law, ie SLLN $\Downarrow$ WLLN $\not\Uparrow$

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables. Assume that X_1 has finite mean μ and $E \mid X_1 \mid < \infty$

$$A = \left\{ \lim_{n \to \infty} \frac{X_1 + X_2 + \ldots + X_n}{n} = \mu \right\},$$

then

$$P(A) = 1.$$

- Proof is much harder to do.

A - an event of interest

Question: $\mathbb{P}(X \in A) = p = ?$

Take $X_1, X_2, \dots, X_n$ i.i.d X

Sample

$$Z_i = \begin{cases} 1 & X_i \in A \\ 0 & X_i \notin A \end{cases}$$

$\{Z_i\}_{i=1}^{n}$ i.i.d. Bernoulli(p)

$E[Z_i] = p$

$Var[Z_i] = p(1-p)$

<u>WLLN</u>:

$$\bar{Z} = \frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\#\{i : X_i \in A\} \equiv \hat{p}$$

$\forall \varepsilon > 0$, $\mathbb{P}(|\bar{Z} - p| > \varepsilon) \longrightarrow 0$ as $n \to \infty$

$\forall \varepsilon > 0$, $\mathbb{P}(|\hat{p} - p| > \varepsilon) \longrightarrow 0$ as $n \to \infty$

<u>SLLN</u>

$$\mathbb{P}\left(\lim_{n\to\infty} \hat{p} = p\right) = 1$$

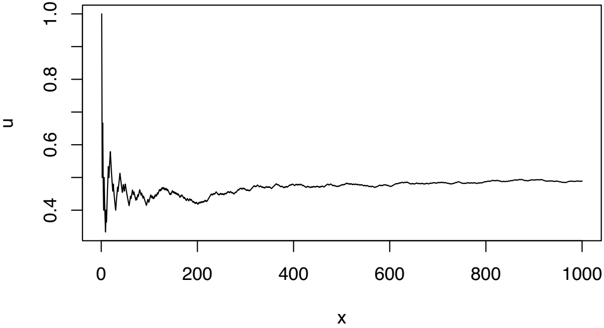
$$\mathbb{P}\left(\lim_{n\to\infty} \text{"relative frequency"} = \text{"True Probability"}\right) = 1$$

Law of Large Numbers

```
> runningmean = function (x,N){
+ y = sample(x,N, replace=TRUE)
+ c = cumsum(y)
+ n = 1:N
+ c/n
+ }
> u = runningmean(c(0,1), 1000)
```

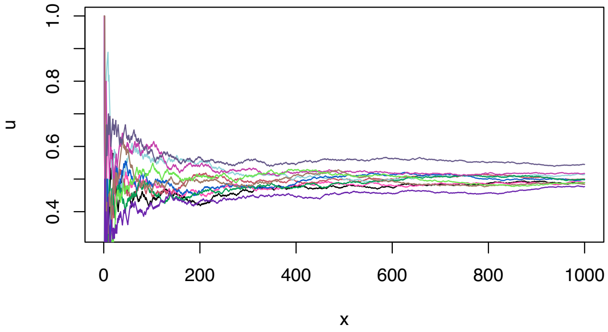
Law of Large Numbers

```
> x=1:1000; plot(u~x, type="l");
>
```



Law of Large Numbers

```
> x=1:1000; plot(u~x, type="l");
> replicate(10, lines(runningmean(c(0,1), 1000)~x, type="l",  col=rgb(runif(3),runif(3),runif(3))))
```



Law of Large Numbers