Sampling and Descriptive statistics

Probability :- study of models for random experiments when the model is fully unknown.

Statistics :- model is not fully known and one tries to infer unknown aspects of the model based on outcomes of an experiment.

Assume :- Large population N Q:- Height distribution?

Random Experiment
- Sample n people in the population
- Record $X_1, X_2, \dots, X_n$:- their heights.

Assume :- $X_1, X_2, \dots, X_n$ i.i.d. X

Sample with replacement

$\approx$

$N \ggg$ Sample without replacement $\longleftarrow$ Ex in Hw6

Empirical Distribution — Descriptive statistics is inferences based on Empirical distribution.

- can study Empirical distribution using tools of Probability
  - Do not make any assumptions about the underlying 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

$$f(t) = \frac{1}{n}|\{X_i = t\}|. \equiv \frac{1}{n}|\{i : X_i = t,\ 1 \leq i \leq n\}|$$

Remarks:-

- Empirical distribution is a random quantity

as $n \to \infty$, intuitively we expect the

- Empirical distribution to approach the true / underlying distribution.

need to make rigorous.

Sample Mean $\equiv$ $\bar{X}$ is an unbiased estimate of μ, i.e $E[\bar{X}] = \mu$

$\bar{X}$ is a consistent estimate of μ, ie $Var(\bar{X}) \to 0$ as $n \to \infty$.

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

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

Result: $X_1, X_2, \ldots, X_n$ are i.i.d X. $E[X] = \mu$ $SD[X] = \sigma$ then

$$E[\bar{X}] = \mu \quad \text{and} \quad SD[\bar{X}] = \frac{\sigma}{\sqrt{n}}.$$

<u>Proof:-</u>

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \cdots + X_n}{n}\right] = \frac{1}{n}\sum_{i=1}^{n} E[X_i]$$

linearity of Expectation

$X_1 \ldots X_n$ are i.i.d X & $E[X] = \mu$

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

(unbiased)

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{X_1+X_2+\dots+X_n}{n}\right)$$

$$= \frac{1}{n^2}\mathrm{Var}(X_1+X_2+\dots+X_n)$$

Variance reduction

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i)$$

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

$$\therefore \quad SD(\bar{X}) = \sqrt{\mathrm{Var}(\bar{X})} = \frac{\sigma}{\sqrt{n}}$$

Remark:

- $\mathrm{Var}(\bar{X}) \longrightarrow 0$ as $n \to \infty$. $\equiv$ $\bar{X}$ concentrates around μ.
  - Consistency
- "Effective" Range of $\bar{X}$:- $\left(\mu - 3\frac{\sigma}{\sqrt{n}},\ \mu + 3\frac{\sigma}{\sqrt{n}}\right)$

Sample Variance

- normalisation by $n-1$ instead of n is artificial.

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

$$S^2 = \frac{(X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2}{n-1}.$$

Result: S^2 is an unbiased estimator of σ^2, i.e.

$$E[S^2] = \sigma^2.$$

Proof :- Exercise $\square$

Remark :- One can show that

$$\mathrm{Var}(S^2) \longrightarrow 0 \quad \text{as} \quad n \to \infty.$$

• Sample mean and variance – key summary statistics from sample $X_1, \dots, X_n$ i.i.d X

• Question of interest :- A – event of interest

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

Answer :- $$\hat{p}_n = \frac{|\{i : X_i \in A, 1 \le i \le n\}|}{n}$$

we will say $\hat{p}_n$ is an estimate for p. $\square$

Question 2 :- How good of an estimate is $\hat{p}_n$ for p ?

claim :-

• unbiased estimate of p

• consistent

– $\hat{p}_n$

• "effective" range of $\hat{p}_n$: $\left(-3\sqrt{\frac{p(1-p)}{n}} + p,\ p + 3\sqrt{\frac{p(1-p)}{n}}\right)$

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

$$\hat{p}_n = \frac{\#\{X_i \in A\}}{n}. = \frac{|\{i : X_i \in A, 1 \le i \le n\}|}{n}$$

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

<u>Proof</u> :-

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

<u>Easy to see</u> :-

- $\mathbb{P}(Z_i = 1) = p \quad \forall i \ge 1$
- $\{Z_i\}_{i \ge 1}$ are also independent

- $Z_i \sim \text{Bernoulli}(p) \quad 1 \le i \le n$
- $\{Z_i\}_{i \ge 1}$ are independent
- $\sum_{i=1}^{n} Z_i \sim \text{Binomial}(n, p)$
- $$\sum_{i=1}^{n} Z_i = |\{i : X_i \in A, 1 \le i \le n\}|$$

$$\text{and} \quad \hat{p}_n = \frac{\sum_{i=1}^{n} Z_i}{n}$$

Now $$E[\hat{p}_n] = E\left[\frac{\sum_{i=1}^{n} Z_i}{n}\right]$$

Linearity of Expectation $$= \frac{1}{n} E\left(\sum_{i=1}^{n} Z_i\right)$$

mean of Binomial $$= \frac{1}{n} np$$

$$= p$$

$$\text{Var}(\hat{p}_n) = \text{Var}\left[\frac{\sum_{i=1}^{n} Z_i}{n}\right]$$

$\text{Var}(aU) = a^2 \text{Var}(U)$ $$= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^{n} Z_i\right)$$

Variance of Binomial $$= \frac{np(1-p)}{n}$$

$$= \frac{p(1-p)}{n} \longrightarrow 0 \quad \text{as} \quad n \to \infty \qquad \square$$

Weak Law of Large Numbers

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 $\epsilon > 0$

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

Proof:-

$$\mathbb{P}(|\bar{X}_n - \mu| > \varepsilon)$$

Tschebyschev in equality

$$\leq \frac{E|\bar{X}_n - \mu|^2}{\varepsilon^2}$$

shown

$E[\bar{X}_n] = \mu$

$var[\bar{X}_n] = \frac{\sigma^2}{n}$

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

$$\therefore \quad 0 \leq \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \implies (1) \qquad \square$$

Summarize

- (WLLN) : $\bar{X}_n \equiv$ close to μ as $n \to \infty$

ie $\forall\, \varepsilon > 0 : \quad \mathbb{P}(|\bar{X}_n - \mu| > \varepsilon) \longrightarrow 0$ as $n \to \infty$

- $p = \mathbb{P}(X \in A)$ $\quad \hat{p}_n =$ relative frequency of A

$\hat{p}_n \approx$ close to $\approx p$

[unbiased and consistent]

$\hat{p}_n \in$ effective range

$$\left(-3\frac{\sqrt{p(1-p)}}{\sqrt{n}} + p \; , \; p + 3\sqrt{\frac{p(1-p)}{n}} \right)$$

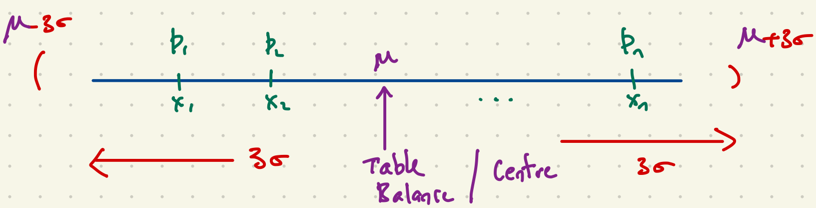
- Relative frequency $\xrightarrow[\text{close}]{n \to \infty}$ Probability

---

Effective Range of X

X - Discrete random variable and

$\mu = E[X] \qquad \sigma = SD[X]$



" $\mathbb{P}(X \in (\mu - 3\sigma, \mu + 3\sigma)) \approx 1$ "

Strong Law of Large Numbers

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

$$P(A) = 1.$$

$$\mathbb{P}\left(\lim_{n\to\infty} \overline{X}_n = \mu \right) = 1$$

$y = (y_1, \ldots, y_N)$

Sampling N points with replacement from x.

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

$c = (y_1, y_1+y_2, y_1+y_2+y_3, \ldots, \sum_{i=1}^{n} y_i)$

$n = (1, 2, \ldots N)$

$(\overline{X}_1, \overline{X}_2, \overline{X}_3, \ldots, \overline{X}_N)$

with Probability 1

$\overline{X}_n \longrightarrow \frac{1}{2}$ (SLLN)

as $n \to \infty$

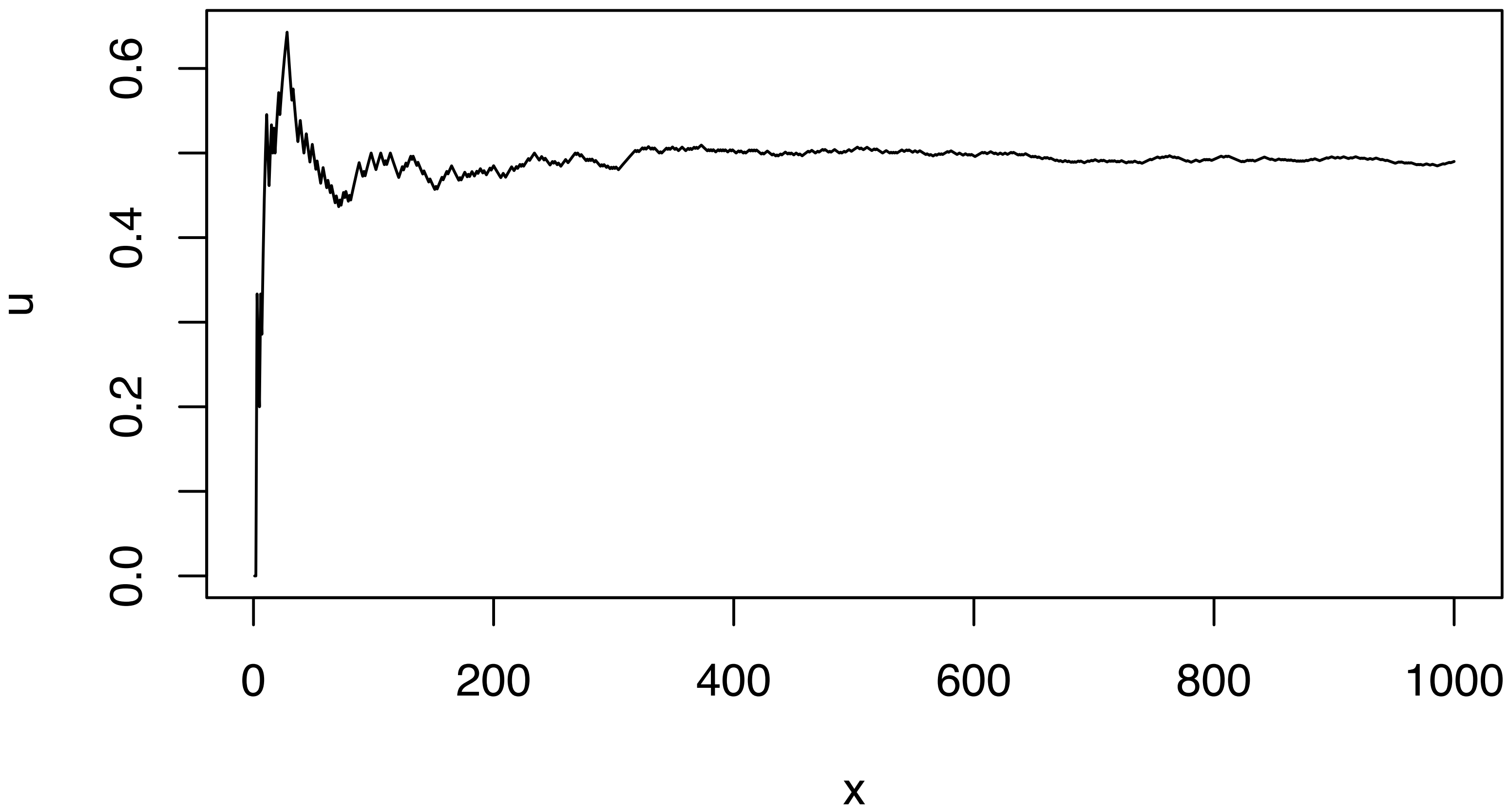
$(\overline{X}_1, \overline{X}_2, \ldots, \overline{X}_{1000})$

$X_1, X_2, \ldots X_{1000}$ are i.i.d Bernoulli $(1/2)$

Law of Large Numbers

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

Base - R code



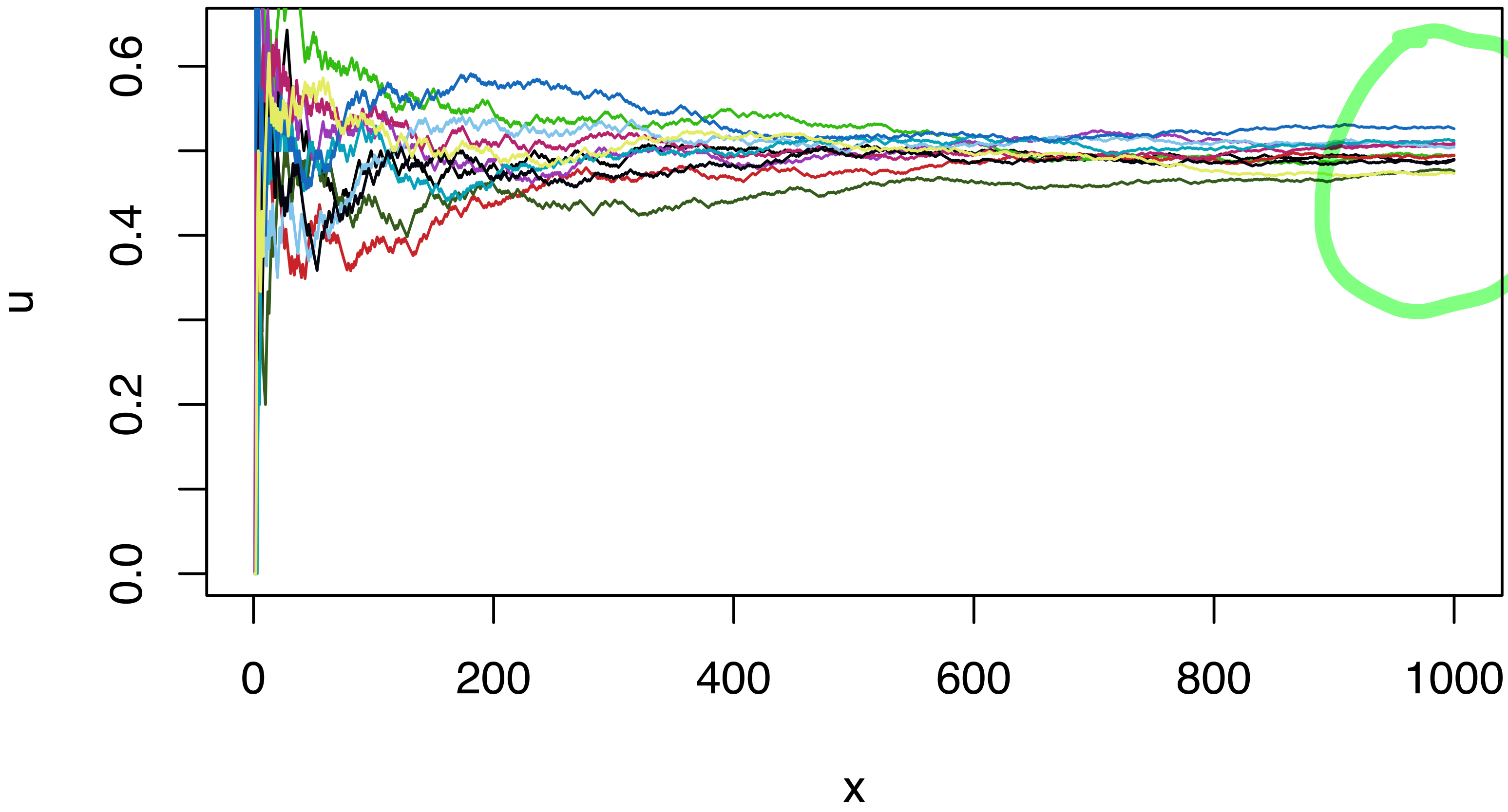
Law of Large Numbers

```
> 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))))
```

$E[\overline{X}_n] = \frac{1}{2}$

$Var[\overline{X}_n] = \frac{1}{4n}$



Observe variance reduction of $\overline{X}_n$

Law of Large Numbers — "Proof by Simulation"