Probability :- is the study of models for (random) experiments when the model is known.

Statistics :- when the model is not (fully) known, then one tries to infer about the unknown aspects based on the observed outcomes of the experiment.

Suppose we sample from a large population. and record a numerical fact. We have seen several examples of data in R, [done our own dice experiment], Study in current HW. etc.

Typically we will get

$$X_1, X_2, \ldots, X_n$$

independent & identically distributed observations.

We assume these are "with" replacement. [Caution!] $n \ll N$.

So far we understand the underlying distribution by:-

- Summary :- mean, median, quantiles
- Plots : Histogram, boxplot, ..
- Distribution :- Q-Q plots.

Here are some keys:-

- Assumption :- Equal probability at each observed point

Empirical distribution :- $S = \{X_1, .. X_n\}$ ← Sample may include repeat observations

is based on discrete distribution on S

with p.m.f.

$$f(t) = \sum_{t \in S} \frac{1}{n} \#\{i \mid X_i = t\}$$

- Random
- Every sample produces a different p.m.f.
- We do not make any additional assumption about the underlying distribution.

Inference based on Empirical distribution is called Descriptive Statistics.

As n grows we will get closer to a better understanding of the true distribution

Note:- Suppose Y was a r.v. with p.m.f f.

$$E(Y) = \sum_{t \in S} t\, f(t)$$

$$= \sum_{t \in S} t \, \frac{\#\{i : X_i = t\}}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{[induction]}$$

Sample variance: $X_1, X_2, \ldots, X_n$ are i.i.d observations (random variables) then we define sample mean to be $\frac{X_1 + X_2 + \ldots + X_n}{n} := \overline{X}$ (Suppress n)

Theorem 1 :- Suppose $X_1, \ldots, X_n$ are i.i.d r.v.

$E X_1 = \mu$ $\quad var(X_1) = \sigma^2$ Then

$$E\,\overline{X} = \mu \quad \text{and} \quad var(\overline{X}) = \frac{\sigma^2}{n}$$

$$\text{or } SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Proof :- $E[\overline{X}] = E\left[\frac{\sum_{i=1}^{n} X_i}{n}\right]$

(induction & linearity of Expectation) $= \frac{1}{n} \sum_{i=1}^{n} E[X_i]$

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

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

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

- $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\,\mathrm{Cov}(X,Y)$
- induction
- $\mathrm{Cov}(X,Y) = 0$ if X & Y are independent

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

$= \frac{1}{n^2}\,\sigma^2 n = \frac{\sigma^2}{n}$ □

On an average, the quantity $\bar{X}$ describes the unknown μ. [In statistics $\bar{X}$ is called an unbiased estimator of μ]

Further, $\mathrm{Var}(\bar{X}) \to 0$ as $n \to \infty$

$\Rightarrow$ $\bar{X}$ is closer to μ as n gets larger

[In statistics, $\bar{X}$ is called a consistent Estimator of μ]

If it is possible to sample from an unknown distribution then averaging a large sample should produce a value close to the expected value of the distribution.

Sample variance :- $X_1, X_2, \ldots X_n$ be i.i.d r.v. Then sample variance is defined as

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

- normalisation in $n-1$ & not n because we want an unbiased estimator of true variance

Theorem :- $X_1, X_2 \ldots X_n$ be i.i.d r.v such that $\mathbb{E} X_1 = \mu$ and $\mathrm{Var}(X_1) = \sigma^2$. Then

$$\mathbb{E} S^2 = \sigma^2$$

Proof :

$$\mathbb{E} \bar{X}^2 = \mathrm{Var}(\bar{X}) + (\mathbb{E}\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2$$

$1 \le i \le n$, $\mathbb{E} X_i^2 = \sigma^2 + \mu^2$

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

(linearity of Expectation)

$$= \sum_{i=1}^{n} \left[E X_i^2 + E \bar{X}^2 - 2 E X_i \bar{X} \right]$$

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

$$= (n+1)\sigma^2 + 2n\mu^2 - 2(E\bar{X}^2)\,n$$

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

$$= (n-1)\sigma^2 \quad \Rightarrow \quad \boxed{E S^2 = \sigma^2}$$. D

$$\tilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n} (X_i - \bar{X})^2 \quad \Rightarrow \quad \tilde{S}^2 = \frac{n-1}{n} S^2$$

$$\therefore \quad E\tilde{S}^2 = \frac{n-1}{n}\sigma^2$$

[not unbiased but consistent]

Can check :-

$\text{var}(S^2) \to 0$

$\text{var}(\tilde{S}^2) \to 0$

Sample Proportions :-

Suppose our interest is in the event $(X \in A)$

& $p = P(X \in A)$.

One can use the random sample $X_1, \ldots X_n$ [i.i.d observations from X] to define

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

$$\hat{p}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i \qquad [\text{Sample proportion}]$$

$$= \frac{\#\{X_i \in A\}}{n}$$

Note that $E[\hat{p}_n] = E(Z_1) = P(X_1 \in A) = p$

$$Var(\hat{p}_n) = \frac{Var(Z_1)}{n} = \frac{p(1-p)}{n}$$

Empirical cumulative distribution of $X_1, \ldots, X_n$ ..

$$F_n(t) = \frac{\#\{X_i \le t\}}{n} \qquad t \in \mathbb{R}.$$

(think as df) $= P(Y \le t)$

Seen in Probability :-

$X_1, X_2, \ldots X_n$ are i.i.d $E X_1 = \mu$,

($E|X_1| < \infty$ Enough) then

SLLN :- $P\left(\frac{X_1 + \ldots + X_n}{n} \to \mu \text{ as } n \to \infty\right) = 1$

One of the key questions of interest is : uncertainty of how close $\bar{X}_n$ is to μ? or $\hat{p}_n$ into p?

Sampling from a given distribution

- we can use the `sample` function.
- takes a sample of the specified size (specified by `size`) from the elements of x using either with or without replacement (specified by `replace`).
- The optional `prob` argument can be used to give a vector of weights for obtaining the elements of the vector being sampled.

```
> x = c(1,2,3,4,5,6)
> probx= c(1/6,1/6,1/6,1/6,1/6,1/6)
> Rolls=sample(x, size=1800,  replace=T, prob=probx)
```

Uniform(1,2,3,4,5,6)

```
> table(Rolls)

Rolls
  1   2   3   4   5   6
300 329 285 273 306 307

> hist(Rolls,breaks = seq(1,6, by=0.25))
```

Functions in R: Variance of Uniform

- Let us try to compute the variance of x

```
> x
[1] 1 2 3 4 5 6
> ourvariance = function(x) {
+  sum((x -ourmean(x))^2)/length(x)
+ }
```

- Note that this differs from sample variance in the normalisation.

Uniform(1,2,3,4,5,6)

```
> var(Rolls)
[1] 2.980381
> ourvariance(x)
[1] 2.916667
```

- `ourvariance` gives the variance of the uniform random variable.

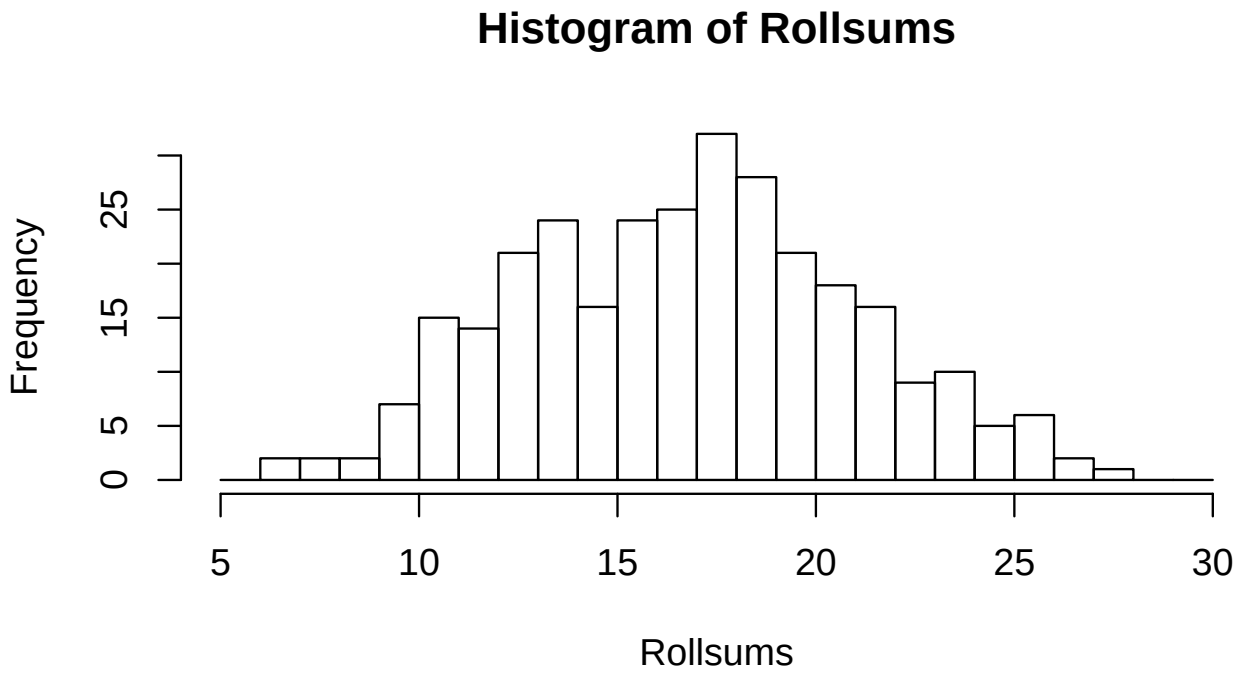
Sums of Rolls

Suppose we wish to simulate in R the experiment that we did in class of Rolling a die and noting down its sum. We can use the `sample` , `matrix` and `apply`.

```
> x = c(1,2,3,4,5,6)
> probx= c(1/6,1/6,1/6,1/6,1/6,1/6)
> Rolls=sample(x, size=1500,  replace=T, prob=probx)
> Rollm=matrix(Rolls, 5)
> # above creates a matrix 5 columns and 30 Rows
> Rollsums = apply(Rollm, 2, sum)
```

Sums of Rolls

```
> table(Rollsums)

Rollsums
 7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
 2  2  2  7 15 14 21 24 16 24 25 32 28 21 18 16  9 10  5  6  2  1

> hist(Rollsums,breaks = seq(5,30, by=1))
```



Class experiment: Sums of Rolls

This was the histogram that we got when we did the experiment of rolling a die 5times and noting down its sum.

Sampling distribution

Suppose we want to verify the below result via simulations:

Let $X_1, X_2, \ldots, X_n$ be an i.i.d. sample of random variables whose distribution has finite expected value μ and finite variance σ^2. Let $\bar{X}$ represent the sample mean. Then

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

Sampling distribution

```
> x = c(1,2,3,4,5,6)
> probx= c(1/6,1/6,1/6,1/6,1/6,1/6)
```

Let us generate 3 sets of data:

500,5000,150000 samples from x and probx.

```
> Rolls=sample(x,size=500,replace=T,prob=probx)
> Rolls5000=sample(x,size=5000,replace=T,prob=probx)
> Rolls150000=sample(x,size=150000,replace=T,prob=probx)
```

Sampling distribution

We split them up into sets of $5, 50, 5000$ rolls.

```
> Rollm=matrix(Rolls, 5)
> Rollm5000=matrix(Rolls5000, 50)
> Rollm150000=matrix(Rolls150000, 5000)
```

Thus each gives us sets of 100,100,30 trials respectively fo $5, 50, 5000$

Sampling distribution

Let us compute the the mean of each row which are of size $5, 50, 5000$

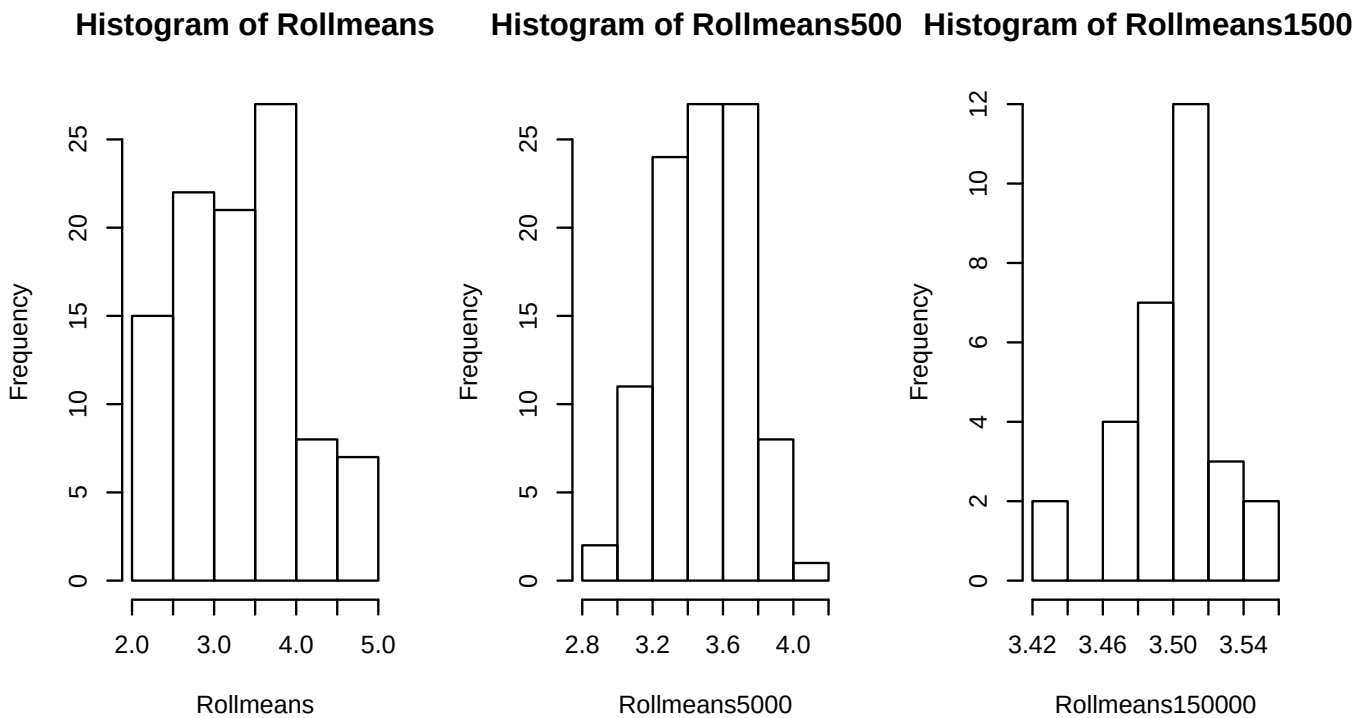
```
> Rollmeans = apply(Rollm, 2, mean)
> Rollmeans5000 = apply(Rollm5000, 2, mean)
> Rollmeans150000 = apply(Rollm150000, 2, mean)
```

Mean of Rolls

```
> table(Rollmeans)

Rollmeans
  2 2.2 2.4 2.6 2.8   3 3.2 3.4 3.6 3.8   4 4.2 4.4 4.8   5
  3   3   9  10   7   5  10  11  11   8   8   5   3   5   2

> table(Rollmeans5000)

Rollmeans5000
2.88    3 3.02 3.04 3.06 3.08  3.1 3.12 3.14 3.16 3.18  3.2 3.22 3.24 3.26 3.28
   1    1    1    1    1    1    1    1    1    1    2    3    2    1    2
3.32 3.34 3.36 3.38  3.4 3.42 3.44 3.46 3.48  3.5 3.52 3.54 3.56 3.58  3.6 3.62
   3    1    6    3    3    2    3    5    2    2    2    4    3    2    2    5
3.64 3.66 3.68  3.7 3.72 3.74 3.76 3.78  3.8 3.82 3.84 3.86 3.88  3.9 3.92 3.96
   1    6    4    4    2    1    2    1    1    2    1    1    1    1    1    1
4.02
   1

> table(Rollmeans150000)

Rollmeans150000
3.4326 3.4334 3.4644  3.474  3.477 3.4774 3.4826 3.4828 3.4844 3.4868 3.4902
     1      1      1      1      1      1      1      1      1      1      1
3.4912 3.4938 3.5006 3.5016 3.5028  3.503  3.505  3.507  3.508 3.5084 3.5102
     1      1      1      1      1      1      1      1      1      1      2
 3.511 3.5172 3.5268 3.5316 3.5334 3.5482  3.551
     1      1      1      1      1      1      1
```

Centered around 3.5

```
> par(mfrow=c(1,3))
> hist(Rollmeans)
> hist(Rollmeans5000)
> hist(Rollmeans150000)
```



Variance Reduction

Observe that there is real variance reduction in the sample means.

```
> ourvariance(x) # Variance of Uniform (1,2,3,4,5,6)

[1] 2.916667

> var(Rollmeans) # S^2, 100 Trials, mean of 5 Rolls

[1] 0.5527919

> var(Rollmeans5000)# S^2,100 Trials, mean of 50 Rolls

[1] 0.056544

> var(Rollmeans150000)# S^2, 100 Trials, mean of 5000 Rolls

[1] 0.0007484258
```