Sampling & Descriptive statistics

A view. Probability is study of models for (random) experiments when the model is fully known.

Example: Toss a fair coin
Model – Bernoulli($\frac{1}{2}$)

– When model is not fully known and one tries to infer about unknown aspects of the model based on observed outcomes of experiment, this is where statistics enters the picture.

. We sample from a large population and record a numerical fact associated with each selection.

Example – Recording Height of students in Bangalore University

. View Sampling as a random experiment.

Suppose $X_1, X_2, X_3, \ldots, X_{100}$ are recordings.

— $X_1, \ldots X_{100}$ are i.i.d if sampling is done with replacement

$\times$ iid if without replacement

Typically, sampling is done with replacement

- but if sample size is relatively small to Population size, these methods are not that dramatically different.

- Questions :- From a sample size n, we need to understand the distribution. How should we use the sample to answer the objective?

Empirical distribution

- One natural quantity we create from observed data, is a discrete distribution that puts equal probability at each observed data point

Definition 1.1 $X_1, \ldots, X_n$ be i.i.d random variables. Then

The "Empirical distribution" based on these is the discrete distribution with probability mass function given by

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

Example: Suppose $X_1 = 1,\ X_2 = 2,\ X_3 = 4,\ X_4 = 4,\ X_5 = 6$

$$f(t) = \begin{cases} 1/6 & t = 1 \\ 1/6 & t = 2 \\ 2/6 & t = 4 \\ 1/6 & t = 5 \\ 0 & \text{otherwise} \end{cases}$$

We can use these discrete distributions and tools from Probability to understand the population.

- Performing inference on population based on "f" is called "descriptive statistics".

Sample Mean

Definition: let $X_1, X_2, \dots, X_n$ be i.i.d random variables. Then

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$ is called the sample mean.

Theorem 1 :- $X_1, X_2, \ldots, X_n$ i.i.d random variables such that $E[X_1] = \mu < \infty$ and $Var[X_1] = \sigma^2 < \infty$. Let $\bar{X}$ be the sample mean. Then

$$E[\bar{X}] = \mu \quad \text{and} \quad Var[\bar{X}] = \frac{\sigma^2}{n}$$

Proof.

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

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

$$= \frac{1}{n} E[X_1 + X_2 + \cdots + X_n]$$

$$= \frac{1}{n}\left(E[X_1] + E[X_2] + \cdots + E[X_n]\right)$$

$$= \frac{1}{n}(\mu + \mu + \cdots + \mu) = \frac{n\mu}{n} = \mu$$

$\left(E[aX+bY] = aE[X] + bE[Y] \right)$
if $E[X] < \infty$
$E[Y] < \infty$

$\left(X_i$ are i.i.d.
$E[X_i] = \mu \right)$

So Vo…

$$\mathrm{Var}[\bar{X}] = \mathrm{Var}\left[\frac{X_1+X_2+\cdots+X_n}{n}\right] = \frac{1}{n^2}\mathrm{Var}[X_1+\cdots+X_n]$$

$\mathrm{Var}[aX] = a^2\,\mathrm{Var}[X]$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{var}[X_i]$$

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)$

independence $\Rightarrow$ $\mathrm{Cov}[X,Y] = 0$

iid

So $\mathrm{Var}[X_i] = \sigma^2$ $\forall i \geq 1$

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

$$= \sigma^2/n$$

Sample Variance

Definition: let $X_1 \dots X_n$ be i.i.d random variables. The "sample variance" is given by

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

Theorem 2:- let $X_1, X_2, \dots X_n$ be i.i.d sample of random variables whose distribution has finite mean μ and finite variance σ^2.

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

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

Proof :-

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

$1 \le j \le n$ $\quad E[(X_j)^2] = \sigma^2 + \mu^2$

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

Observe $(X_i - \overline{X})^2 = X_i^2 - 2X_i\overline{X} + (\overline{X})^2, \quad 1 \le i \le n$

So $$\sum_{i=1}^{n} (X_i - \overline{X})^2 = \sum_{i=1}^{n} X_i^2 - 2\overline{X}\left(\sum_{i=1}^{n} X_i\right) + n(\overline{X})^2$$

$\boxed{n\overline{X} = X_1 + \dots + X_n}$

$$= \sum_{i=1}^{n} X_i^2 - 2n(\overline{X})^2 + n(\overline{X})^2$$

$$= \sum_{i=1}^{n} X_i^2 - n(\overline{X})^2$$

$$\Rightarrow E[S^2] = \frac{1}{n-1} E\left[\sum_{i=1}^{n}(X_i - \overline{X})^2\right] = \frac{1}{n-1} E\left[\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2\right] =$$

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

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

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

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

Ex: $Var[S^2] = ?$