Mathematical Statistics

S : event space (~~sample space~~)

E : Random Experiment

X : Random Variable Connected with E

$F_X(x)$: distribution function <is unknown in Statistics>

Aim:

Extract information experimentally about the population distribution of X.

The experiment depends on a finite set of observed values of X (for $\underbrace{\text{finite performances of the random experiment}}$ E). $\langle x_1, x_2, \ldots, x_n \rangle$ (*)

Population and Sample

Population:

$X : S \to \mathbb{R}$.

E: Random Experiment

X: Random Variable connected with E

Infinitely many repetions of E (under uniform conditions) will give rise to an infinite sequence of observed values of X, totality of which is called the population of the random variable X connected with the random experiment E.

Ex: E: Throwing a die

Observed values: $x_1, x_2, \ldots, x_n, \ldots$

: 1, 3, 2, 6, 4, 1, 1, 2, 2, ...

Note: The d.f. $F_X(x)$ of X determines the distribution of the population.

→ Hypothetical Population
(population doesn't exists, but we can imagine)
(finite).

→ Existent Population, e.g. 1. consider mathematics marks of 1000 students in a class.
(population exists as a set)

(infinite) 2. atmospheric pressure at every point in a room.

Q: Sampling theory:
i.e. How to choose a good sample from a population?

Sample :

Let X be a random variable connected with a random experiment E.

(Probability distribution of X is usually unknown.)

- If E is performed $\underbrace{n \text{ times}}$, then we shall get n observed values $\underbrace{x_1, x_2, \ldots, x_n}$ of X (not necessarily all are distinct). → Observed sample values.

The ordered n-tuple $(x_1, x_2, \ldots, x_n)$: a sample of size n drawn from the population of X.

eg. $E \to$ throwing a die, $X \to$ value of a face

$(1, 2, 2, 6, 5, 2, 1, 6, 1)$ is sample of size 10 from the population of X

Note: Such a sample from a population will be called a "random sample" since different samples of size n are repeatedly drawn, under uniform conditions, the set of observed values of X will vary at random.

$$\left(x_1^{(1)}, x_2^{(1)}, \ldots, x_n^{(1)}\right) : \text{Sample of size } n$$

$$\left(x_1^{(2)}, x_2^{(2)}, \ldots, x_n^{(2)}\right) : \ldots$$

$$\vdots$$

$$\left(x_1^{(m)}, x_2^{(m)}, \ldots, x_n^{(m)}\right)$$

I. Distribution of Sample / Empirical Distribution

II. Sampling Distribution

Distribution of Sample/Empirical Distribution

Let $(x_1, x_2, \ldots, x_n)$ be a random sample of size n drawn from the population of X (all of the observed values may not be distinct)

Define a Fake Random Variable $\mathring{X}$ on $\{x_1, x_2, \ldots, x_n\}$ with

$$P(\mathring{X} = x_i) = \frac{1}{n}, \quad i = 1, 2, \ldots, n$$

Th The distribution function of a random variable X connected with a random experiment E is approximately equal to empirical distribution for large sample.

Pf: $F_{\overset{\circ}{X}}(x)$ = distribution function of $\overset{\circ}{X}$

$$= P(\overset{\circ}{X} \le x)$$

$$= \frac{\nu}{n}$$

where ν is the no. of observed sample values for the event $(\overset{\circ}{X} \le x)$

If we take $n \to \infty$

population distribution

$$\lim_{n\to\infty} F_{\overset{\circ}{X}}(x) = \lim_{n\to\infty} \frac{\nu}{n} \approx F_X(x)$$

i.e. as $n \to \infty$, empirical distribution $\Leftrightarrow$ population distribution

empirical distribution is a statistical image of population distribution

Sample Characteristics

$X \to \mathring{X}$

the characteristics of the distribution of the fake random variable $\mathring{X}$: $P(\mathring{X}=x_i)=\frac{1}{n}, i=1,2,\dots,n$

- Sample Mean: $\bar{x} = E(\mathring{X}) = \sum_{i=1}^{n} x_i P(\mathring{X}=x_i)$

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

- Sample Variance: $S^2 = E\{(\mathring{X}-\bar{x})^2\} = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$

- Sample Moment of Order k:

$$a_k = E\{\mathring{X}^k\} = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

- Sample Central Moment of order k:

$$m_k = E\{(\overset{\circ}{X} - \bar{x})^k\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

Study Yourself: median, lower quartile, upper quartile, semi-interquartile range, mode, coeff. of skewness, coeff. of kurtosis..

(*) ③ Mathematical Statics – Miller & Miller

Alt. Book: ② Mathematical Statistics S.K. De

① Mathematical Prob. & Stat. S. Sen A. Gupta

Th (i) $m_k = \sum_{i=0}^{k} (-1)^i \binom{k}{i} a_{k-i} \bar{x}^i$

(ii) $a_k = \sum_{j=0}^{k} \binom{k}{j} m_{k-j} \bar{x}^j$

Pf: Try yourself!

Tabular and Graphical Representations of samples

(i) Discrete Population :

$(x_1, x_2, \ldots, x_n)$: random sample of size n

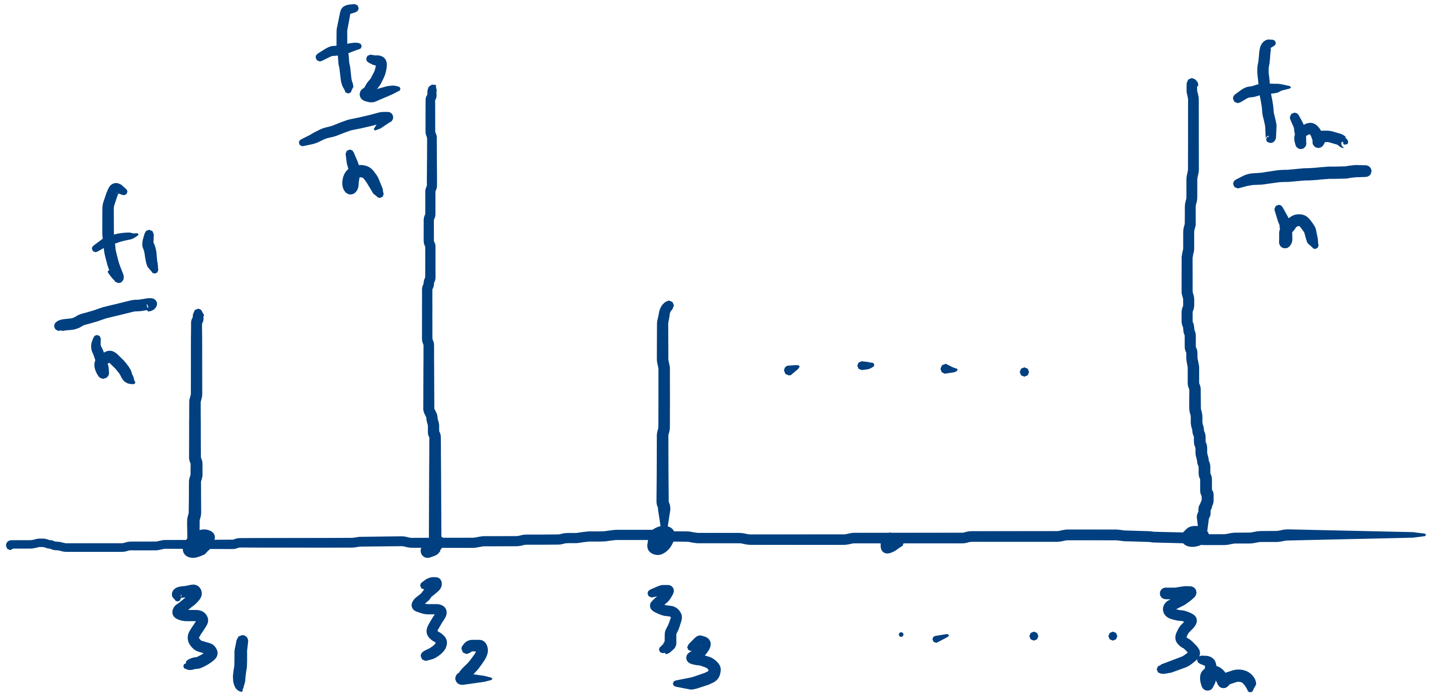
Let, $\xi_1 < \xi_2 < \ldots < \xi_m$, $(m \leq n)$ are the distinct sample values and let ξ_i occurs f_i times

ξ	f
ξ_1	f_1
ξ_2	f_2
$\vdots$	$\vdots$
ξ_m	f_m

(ξ-f table)

$$\sum_{i=1}^{m} f_i = n.$$

Frequency Diagram:



$$P(\mathring{x} = z_i) = \frac{f_i}{n}$$

for, $i = 1, 2, \ldots, m$.

(ii) Continuous Population

$(x_1, x_2, \ldots, x_n)$: random sample of size n

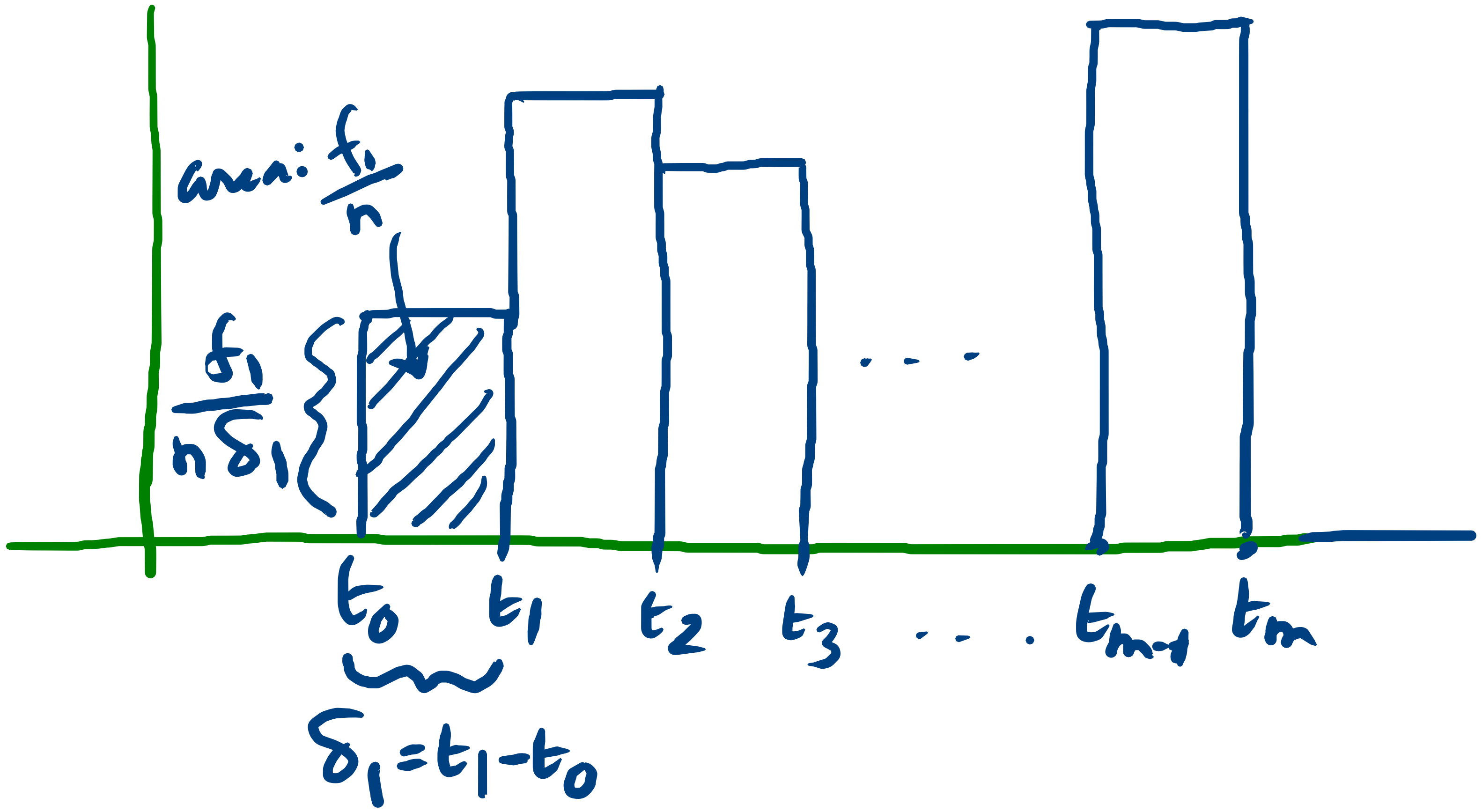
$a < \min_{1\le i\le n} x_i$, $b > \max_{1\le i\le n} x_i$

(a,b) is divided into m $(m<n)$ subintervals by partition: $a = t_0 < t_1 < \ldots < t_{m-1} < t_m = b$

(Grouped data table)

Class Interval	frequency
$t_0 - t_1$	f_1
$t_1 - t_2$	f_2
$\vdots$	$\vdots$
$t_{m-1} - t_m$	f_m

Histogram:



height of each rectangle is $\dfrac{f_i}{n\,\delta_i}$

$$\delta_i = t_i - t_{i-1}$$

$$\text{area} = \frac{f_i}{n}.$$

Sampling Distribution

Q: How good is a "sample characteristic" in approximating a population characteristic?

$$(x_1^{(1)}, x_2^{(1)}, \ldots\ldots, x_n^{(1)}) : \bar{x}^{(1)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(1)}$$

$$(x_1^{(2)}, x_2^{(2)}, \ldots\ldots, x_n^{(2)}) : \bar{x}^{(2)} = \frac{1}{n}\sum_{i=1}^{n} x_i^{(2)}.$$

To answer this we need to find the distribution of the sample mean for all possible random samples.

Statistic: Any real-valued function $f(x_1, x_2, \ldots, x_n)$ of sample values $x_1, x_2, \ldots, x_n$ (where $x_1, x_2, \ldots, x_n$ are taken as variables) is called a statistic. e.g. $\bar{x}$, s^2 etc.

Sampling Distribution: The probability distribution of the random variable $f(X_1, X_2, \ldots, X_n)$ is called the sampling distribution of the statistic $f(x_1, x_2, \ldots, x_n)$ where $X_1, X_2, \ldots, X_n$ are mutually independent random variables all having the same distribution as the population random variable X.

e.g. $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$, $S^2 = \dfrac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2$ etc.

random sample : $(x_1, x_2, \dots, x_n)$

$E \to X$ $\to$ $\underbrace{X_1 \quad X_2 \quad \dots \quad X_n}$

(i) mutually independent

(ii) all have same distribution as X.

$\beta(X_1, X_2, \dots, X_n)$ is called (i.i.d)

distribution

sampling distribution

random point in $\mathbb{R}^n$ $(X_1, X_2, \dots, X_n)$ - n-dimentional random variable