innovate achieve lead

BITS Pilani
Pilani Campus

Course No: MATH F113

Probability and Statistics

BITS Pilani
Pilani Campus

Descriptive Statistics

Sumanta Pasari

sumanta.pasari@pilani.bits-pilani.ac.in

Recall: What is Statistics?

- A discipline of science that pertains to the **collection**, **analysis**, **interpretation**, and **presentation** of data
- Science of ***learning from data***
- Mathematical Statistics: application of Mathematics to Statistics
- In 18th century, “Statistics” designated to a systematic collection of demographic and economic data by states.
- Main concern is to collect, analyze, and present data in the context of **uncertainty** and **decision making**

Information Extraction from Data

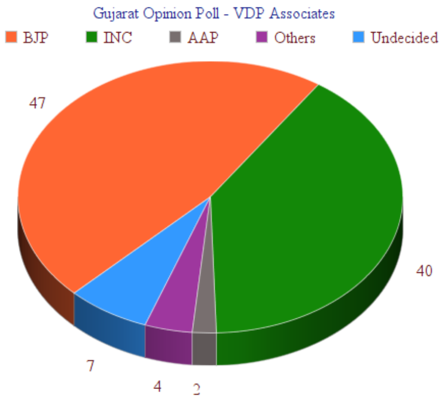
Example 6.1: Rural Business of LIC

Rural Business of LIC	Rural Business		% of Rural Business to Total Business	
	Polices(in Lakhs)	Sum Assured (in Crore)	Polices	Sum Assured
1970	4.61	251.764	33.00	24.54
1975	5.72	464.27	31.85	26.37
1980	5.91	603.77	28.20	22.09
1985	9.52	1569.62	35.26	29.20
1990	30.48	8086.35	41.23	34.33
1995	49.02	21571.00	45.10	39.10
1996	52.57	21263.59	47.70	41.00
1997	60.33	24278.73	49.20	42.80
1998	68.40	27550.69	51.40	43.30
1999	81.23	35372.94	54.70	47.00

Example 6.2: Sales Statistics

Year	Sales (in million rupees)
Mar 1990	2167.2
Mar 1991	2273
Mar 1993	2120.3
Mar 1994	2615.9
Mar 1995	3490.6
Mar 1996	5173.3
Mar 1997	6246.2
Mar 1998	7502.1
Mar 1999	12503.6
Mar 2000	7328.2
Mar 2001	6337.5
Mar 2002	5428.7
Mar 2004	11116.3
Mar 2005	6962.7
Mar 2007	3653.8

Example 6.3: Election Forecast



Example 6.4: Road Accidents

State/UT	2008	2009	2010	2011 (P)
Top* 5 States: Share in total number of road accidents (in %)				
Share of 5 States	55.4	55.3	55.5	54.8
Maharashtra	15.6	14.8	14.3	13.8
Tamil Nadu	12.5	12.5	13.0	13.2
Madhya Pradesh	9.0	9.7	10.0	9.9
Karnataka	9.5	9.3	9.3	9.0
Andhra Pradesh	8.8	9.0	8.9	8.9

Example 6.5: Student Information

	Age	Gender	Height (cm)	Weight (kg)	Nose length (mm)
1	20	M	165	62	40
2	21	F	152	56	44
3	20	F	158	62	42
4	21	F	**191**	54	40
5	19	M	167	65	41
6	20	F	159	60	43
7	18	M	190	101	47
8	19	M	182	95	46
9	20	M	170	81	44
10	21	M	172	74	41
11	22	F	170	55	42
12	19	M	178	75	45
13	20	F	157	55	40
14	21	M	169	70	48
15	20	M	164	63	45
16	19	M	174	67	44
17	18	F	154	56	**72**
18	21	F	156	58	47
19	19	M	171	59	42
20	19	F	151	**102**	45

Observe and Answer:

1. *Any outliers present in this raw data?*
2. Any relation between height and weight of students?
3. Do the female students have more nose-length?
4. What is the average height of male students?
5. *Do the male students come from a rich family?*
6. *Are the female students more intelligent than male students?*

Example 6.6: Sales of Passenger Vehicles (Processed Data)

Passenger Vehicles	Oct '15	Oct '14
Maruti Suzuki	1,21,063	97,069
Hyundai Motor India	47,015	38,010
Mahindra & Mahindra	24,060	20,255
Honda Cars India	20,166	13,242
Tata Motors	12,798	11,511
Toyota Kirloskar Motors	12,403	12,556
Volkswagen India	3,255	4,663
TOTAL	2,40,760	1,97,306

Let’s Watch…

https://www.youtube.com/watch?v=wlGCwoU264Q

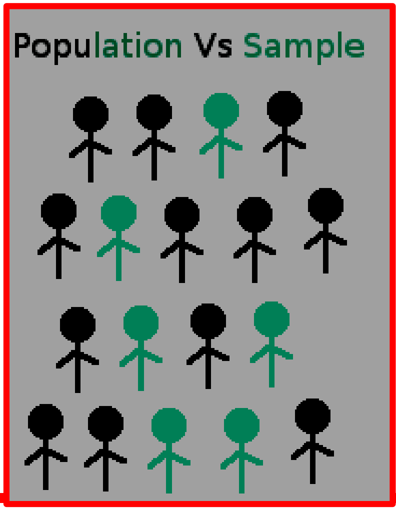
Descriptive and Inferential Statistics

1. **Descriptive Statistics** consists of the collection, organization, summarization, and presentation of data.
   - process of describing data and trying to reach a conclusion
   - data and charts observed in general newspapers or articles

2. **Inferential Statistics** provides a scientific procedure to make inferences about a population based on sample.
   - generalizing from samples to populations, performing estimations and hypothesis testing, determining relationships among variables, and making predictions
   - ages of students of a class are 25, 24, 28, 29, 30, 25, 26, 25, 28, 25, and 25 years. Are the students (overall) follow a normal distribution?

Population and Sample

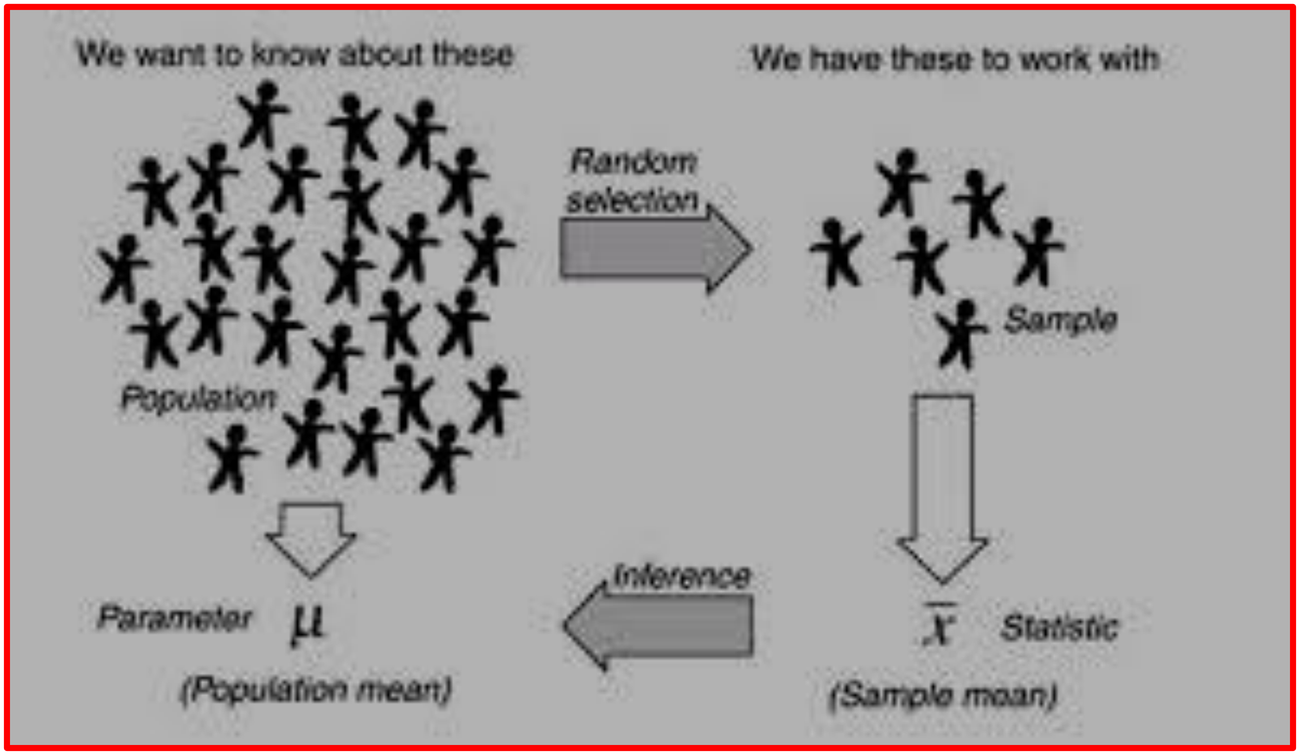
- **Population** is a collection of all distinct individuals or objects or items under study; Size of population: ***N***
- **Sample** is a portion (representative portion) of a population; Size of a sample: ***n***
- **Sampling**: How to choose a sample? What are the desired criteria?
- To represent the population well, a sample should be **randomly** collected and adequately **large**.



Ex: Election forecast, lifetime of tube light,
efficiency of a drug, campus placements

Parameter and Statistic

- **Parameter** is a descriptive measure of some characteristics of the population. Parameter is usually unknown.
- **Statistic** is a descriptive measure obtained from a sample; It is a function of observations in a random sample.



Outliers

- What is an outlier?
- Why does it occur?
- Is it really an outlier?
- How to detect?
- What to do then?

(Out of present syllabus)

Example

- If the population has gamma distribution, the shape and scale are described by α and β.

- For a binomial distribution consisting of n trials, the shape is determined by p.

Example

- However, often the values of ***parameters*** that specify the exact form of a distribution are ***unknown***.

- You must rely on the ***sample*** to make inference about these parameters.

Random Sample: Finite Population

- Suppose there are 2500 managers in a company. We would like to develop a profile of managers, with (a) their mean annual salary, (b) proportion of managers who completed company's management training program. …..suppose the population mean salary is known as ***μ = $51,800*** and ***population s.d. σ = $4000,*** and 1500 managers have already completed the training program, i.e., ***population proportion p =0.6***

- Suppose there are 2000 oak trees in a managed small forest. We need to estimate tree diameter at breast height.

- How to choose *n* samples in each case? **Random number table?**

Random Number Table

10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
09893	20505	14225	68514	46427	56788	96297	78822	54382	14598

Random sample

- Recall previous example: once 30 managers are randomly selected, we calculate **sample statistic (?)**.

Salary	Training?	Salary	Training?	Salary	Training?
49094.3	Yes	45922.6	Yes	45120.9	Yes
53263.9	Yes	57268.4	No	51753.0	Yes
49643.5	Yes	55688.8	Yes	54391.8	No
49894.9	Yes	51564.7	No	50164.2	No
47621.6	No	56188.2	No	52973.6	No
55924.0	Yes	51766.0	Yes	50241.3	No
49092.3	Yes	52541.3	No	52793.6	No
51404.4	Yes	44980.0	Yes	50979.4	Yes
50957.7	Yes	51932.6	Yes	55860.9	Yes
55109.7	Yes	52973.0	Yes	57309.1	No

Random Sample: Infinite Population

- Consider the population of 8 million school students in a certain state. Suppose, we are interested to determine $\boldsymbol{\mu}$, the unknown population mean distance from students' schools to their hometowns. Suppose, the sample mean from 100 students show a distance of 1.2 km.

- Suppose a tyre manufacturer is interested to test whether the new design provides increasing mileage. To estimate the mean useful life, the manufacturer choose a sample of 120 tires. Test result shows a sample mean of 36000 miles.

Other Examples

- Suppose you would like to select a random sample of 50 customers in a restaurant to complete a feedback survey – how to go ahead?
- In order to improve the facilities of an amusement park (e.g., Nicco Park, Kolkata), the manager wants to collect a sample of 50 persons.
- A quality control manager is concerned about the proper machine-filling of breakfast boxes, filled with 100 gm of cereals.
- **Discuss, in each of the above examples, how to avoid selection bias.**
- Infinite populations: customers entering a retail store, repeated experimental trials in a laboratory, number of trees in a forest, telephone calls arriving at a technical support centre
- **For practical purposes, if $n/N>0.05$, we consider it as finite population, otherwise if $n/N\leq 0.05$, one may consider infinite population.**

Random Sample

- The random variables X_i constitute **a random sample of size *n*** if and only if,

  1) Random variables X_i are independent, and
  2) Random variables X_i are identically distributed, that is, each X_i has same distribution having pdf $f(x)$, mean μ, and variance σ^2. **We say that X_i are i.i.d.**

- **Examples?**

Remarks

(1) Here the sample of size *n* can be thought as ordered and with repetition allowed, i.e., an *n*-tuple $(x_1, \ldots, x_n)$ represents a sample.

(2) Let r.v. *X* with density $f_X(x)$ denote the population. The ***random sample of size n from (infinite) population X*** can be thought as the *n*-dimensional random variable $(X_1, \ldots, X_n)$ whose joint density is

$$f_{X_1 \ldots X_n}(x_1, \ldots, x_n) = f_X(x_1) \ldots f_X(x_n)$$

Remarks

3. If we are choosing n values of X randomly and independently with replacement and in order, whether population is finite or infinite, we can interpret as sample of size n from infinite population.
4. For infinite populations or large populations compared to sample (say 20 times larger than sample size), even if each choice is done randomly and replacement is not allowed, we more or less have these assumptions satisfied.

Remarks

(5) The random sample and its characteristics will be denoted by capital letters and a particular sample and its characteristics by corresponding small letters. This is in tune with the practice of denoting random variables by capital letters and its values by corresponding small letters. Thus a particular sample is a value $(x_1, \ldots, x_n)$ of the n-dimensional r.v. $(X_1, \ldots, X_n)$, or equivalently, a point in n-dimensional space. The values $x_1, \ldots, x_n$ are also called observed values of X.

Sec 6.2 (graphical representation and analysis of data) not in syllabus.

Sec. 6.3: Statistic

- A **statistic** of a random sample *(X_1, ..., X_n)* from population *X* is a function *H (X_1, ..., X_n)* of the n-dim r.v. *(X_1, ..., X_n)*. For example,

(i) Sample mean, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

(ii) Sample variance,

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

Other Statistic

(iii) Sample minimum $X^{(1)} = \underset{i}{Min}\{X_i\}$

(iv) Sample maximum $X^{(n)} = \underset{i}{Max}\{X_i\}$

(v) Sample range $= X^{(n)} - X^{(1)} = \underset{i}{Max}\{X_i\} - \underset{i}{Min}\{X_i\}$

(vi) Sample median $= \begin{cases} X^{\left(\frac{n+1}{2}\right)};\ n \text{ odd} \\ \dfrac{X^{\left(\frac{n}{2}\right)} + X^{\left(\frac{n+1}{2}\right)}}{2};\ n \text{ even} \end{cases}$

Order Statistic: $X^{(1)}, X^{(2)}, \cdots, X^{(n)}$ is called an ordered statistic.

Sample Statistic

- Recall previous example: once 30 managers are randomly selected, we calculate **sample statistic (?)**.

Salary	Training?	Salary	Training?	Salary	Training?
49094.3	Yes	45922.6	Yes	45120.9	Yes
53263.9	Yes	57268.4	No	51753.0	Yes
49643.5	Yes	55688.8	Yes	54391.8	No
49894.9	Yes	51564.7	No	50164.2	No
47621.6	No	56188.2	No	52973.6	No
55924.0	Yes	51766.0	Yes	50241.3	No
49092.3	Yes	52541.3	No	52793.6	No
51404.4	Yes	44980.0	Yes	50979.4	Yes
50957.7	Yes	51932.6	Yes	55860.9	Yes
55109.7	Yes	52973.0	Yes	57309.1	No

From sample,

$$\bar{x} = \$51,814$$

$$s = \$3,348$$

$$\bar{p} = \frac{19}{30} = 0.63$$

Actually,

$$\mu = \$51800$$

$$\sigma = \$4000$$

$$p = 0.60$$

Sampling Distributions

- Suppose we collect another random sample of size 30, and we found sample mean = $ 52670, and sample proportion = 0.70
- Let us repeat this "**random experiment**" 500 times – random variables come into picture.

Sample Number	Sample mean	Sample proportion
001	51814	0.63
002	52670	0.70
003	51780	0.67
004	51588	0.53
…	…	…
…	…	…
500	51752	0.50

Each $X_1, X_2, \cdots, X_n$ is a random variable (and, X_i are i.i.d), with mean μ and standard deviation σ.

So, how does $\bar{X}$ behave?

Can we get a histogram of $\bar{X}$?

What is the distribution of $\bar{X}$?

What are the mean and s.d. of $\bar{X}$?

Some Properties

innovate achieve lead

If $X_1, X_2, \cdots, X_n$ is a random sample (that is, X_i are i.i.d), with mean μ and standard deviation σ, then

(i) $E\left(X_1 + X_2 + \cdots + X_n\right) = \sum_{i=1}^{n} E\left(X_i\right) = n\mu$ (true, without independent)

(ii) $\mathrm{Var}\left(X_1 + X_2 + \cdots + X_n\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i\right) = n\sigma^2$ $\left(\text{as, } \mathrm{cov}\left(X_i, X_j\right) = 0, i \neq j\right)$

(iii) $E\left(X_1 X_2 \cdots X_n\right) = \left[E\left(X_i\right)\right]^n = \mu^n$

(iv) Joint pdf $f_{X_1 X_2 \cdots X_n}\left(x_1, x_2, \cdots x_n\right) = f_{X_1}\left(x_1\right) f_{X_2}\left(x_2\right) \cdots f_{X_n}\left(x_n\right)$

(v) Joint cdf $F_{X_1 X_2 \cdots X_n}\left(x_1, x_2, \cdots x_n\right) = F_{X_1}\left(x_1\right) F_{X_2}\left(x_2\right) \cdots F_{X_n}\left(x_n\right)$

So, how does $\bar{X}$ behave? What is the distribution of $\bar{X}$?

What are the mean and standard deviation of $\bar{X}$?

Mean and Variance of Sample Mean

Ex.6.1. If $X_1, X_2, \cdots, X_n$ is a random sample, each X_i having mean μ and standard deviation σ, then

$$\text{(i) } \mu_{\bar{X}} = E(\bar{X}) = \mu \qquad \text{(ii) } \sigma_{\bar{X}}^2 = \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Proof. (i) $E(\bar{X}) = E\left(\dfrac{X_1 + X_2 + \cdots + X_n}{n}\right) = \dfrac{1}{n}\sum_{i=1}^{n} E(X_i) = \dfrac{n\mu}{n} = \mu$

$$\text{(ii) } \mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \quad \left(\text{as, } \mathrm{Var}(aX) = a^2\mathrm{Var}(X)\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) \quad \left(\text{as, } X_1, X_2, \cdots, X_n \text{ are independent}\right) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Example: Page No 218

Ex. 25 (Approximation of σ via range) :

If *X* is normal, $P[-2\sigma < X-\mu < 2\sigma] = 0.95$ (close to 1)

So we can assume 4σ to be the whole range, or

Approximately σ = (estimated range)/4

If *X* is not normal, by Chebyshev's inequality :

$$P[-3\sigma < X-\mu < 3\sigma] \geq 0.89,$$

So we can approximately take

σ = (estimated range)/6.

Problem Solving

Ex.6.2. Let $X_1, X_2, \cdots, X_{25}$ be a random sample from the distribution of X having mean 10 and variance 50. Find the mean and standard deviation of (i) $a\bar{X}+5$ (ii) $7\bar{X}+5a$, a is a scalar.

Sol.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{25}\sum_{i=1}^{25} X_i,$$

$$E\left(\bar{X}\right) = \mu = 10$$

$$\mathrm{Var}\left(\bar{X}\right) = \frac{\sigma^2}{n} = \frac{50}{25} = 2$$

$$\text{(i)}\ \mu_{a\bar{X}+5} = E\left(a\bar{X}+5\right) = aE\left(\bar{X}\right)+5 = 10a+5$$

$$\sigma^2_{a\bar{X}+5} = \mathrm{Var}\left(a\bar{X}+5\right) = a^2\mathrm{Var}\left(\bar{X}\right) = 2a^2 \Rightarrow \sigma_{a\bar{X}+5} = |a|\sqrt{2}$$

Problem Solving

HW.6.1. Let the mean and variance of a sample mean are 5 and 2, respectively. If the random sample $X_1, X_2, \cdots, X_n$ comes from a distribution of X having variance 10, find the mean of X_i and the sample size *n*.

HW.6.2. Let the mean and standard deviation of a sample mean are 10 and 2, respectively. If the random sample $X_1, X_2, \cdots, X_n$ comes from a distribution of X having variance 60, find the sample size *n*. $\left(\textbf{Sol.}\ n = 15?\right)$

Problem Solving

innovate achieve lead

HW.6.3. Let $X_1, X_2, \cdots, X_n$ be a random sample from the distribution of X having mean 10 and variance 50. Find the minimum number of sample size n, such that variance of $3\bar{X}$ becomes less than 5.

HW.6.4. Let $X_1, X_2, \cdots, X_n$ be a random sample from the distribution of X having mean 10 and variance 50. Find the minimum number of sample size n, such that the standard deviation of $8\bar{X}$ becomes less than 10.

Problem Solving

HW.6.5. A random sample of size 5 provides the following observations on X (height of students, in cm) and Y (weight of students, in kg).

X_{obs}: 185 175 165 170 156 Y_{obs}: 69 67 61 64 56

(a) Calculate sample means and sample standard deviations of height and weight. Can we compare the standard deviations of height and weight?

(b) Is there a (linear) relationship between observed height and weight data? (wait, sample covariance/correlation will be discussed later!)

HW.6.6. Analyze the datset in Example 6.5. Find observed values of sample mean, sample variance, sample maximum, sample minimum, sample median, sample range for height, weight, and nose lengths.