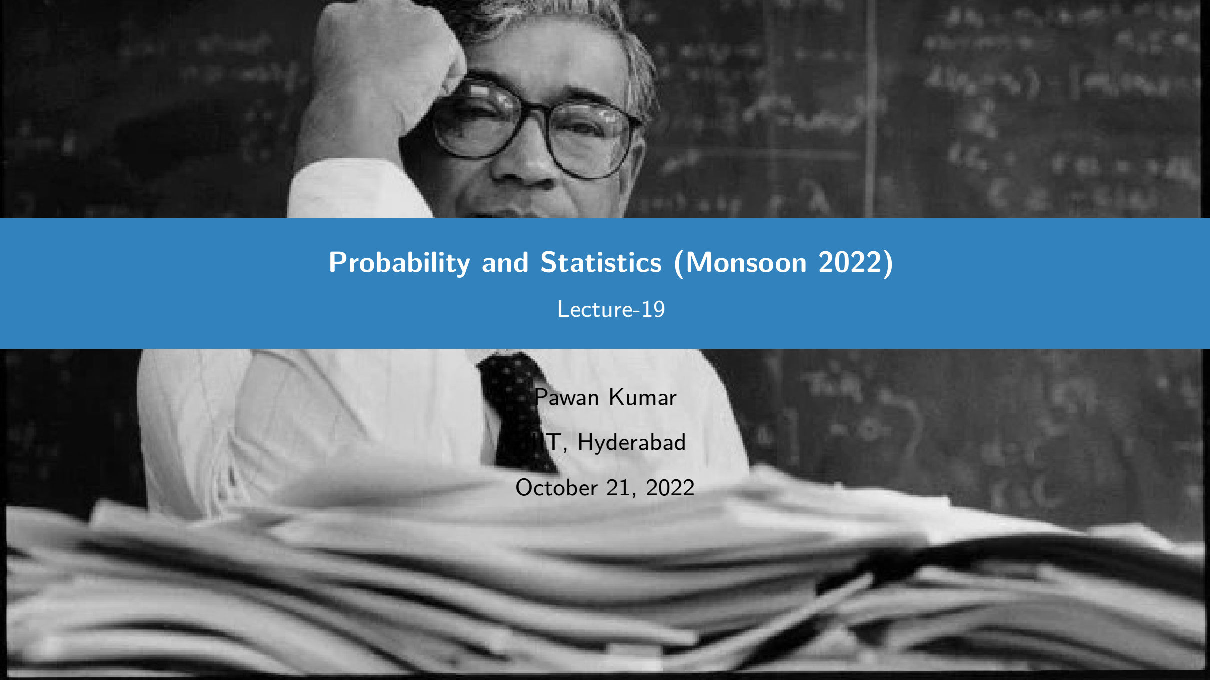


Probability and Statistics (Monsoon 2022)

Lecture-19

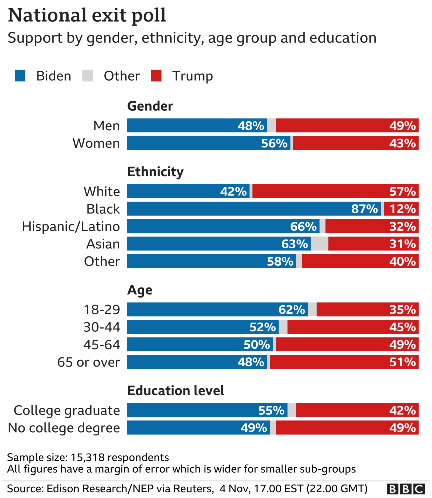
Pawan Kumar

IIT, Hyderabad

October 21, 2022

Outline

Motivation for Statistical Inference...



- On the left, US exit poll results
- Poll on Trump Vs Biden
- Sample size of 15,318
- Error margin shown in grey
- Draw conclusions from the sample data
- Will inference fail? How much it can fail?
- How confident we are of this?

Motivation for Statistical Inference...

POLL OF ALL POLLS

	NDA	MAHAGATHBANDHAN	LJP	OTHERS
JAN KI BAAT	104	128	6	5
C-VOTER	116	120	1	6
MY AXIS	80	150	4	9
TV9 BHARATVARSH	115	120	4	4
POLL OF POLLS	104	129	4	6

BIHAR ASSEMBLY ELECTIONS
RESULTS 2020
TOTAL SEATS 243

NDA	125	MGB	110	OTH	8
BJP	74	RJD	75	LJP	1
JD(U)	43	CONG	19	AIMIM	5
HAM	4	CPI-ML	11	BSP	1
VIP	4	CPM	3	OTHERS	1
		CPI	2		

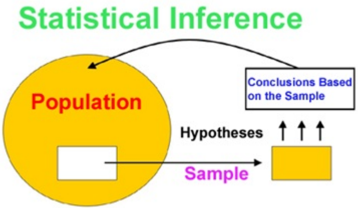
- On the left, poll of polls showing clear majority for MAHAGATHBANDHAN
- After election, NDA has full majority
- How do we estimate such errors?

Definition of Statistical Inference...

Definition of Statistical Inference

Statistical inference is a collection of methods that deal with drawing conclusions from data that are prone to random variation.

- knowledge of probability is used
- we need to work with real data
- distribution of the data may not be known



- Two types: Frequentist and Bayesian

Statistical Inference Problem

To determine an unknown quantity, get some data, and then estimate the required quantity using this data.

Frequentist or Classical Inference...

Recall: A statistical inference problem is to estimate an unknown quantity

Frequentist Inference

Here the unknown quantity is assumed to be fixed quantity and not random. So, the unknown quantity θ is to be estimated by the observed data.

- Let θ be the percentage of people who will vote for a given candidate
    - $\hat{\Theta} = \dfrac{Y}{n}$, Y is the number of people among randomly chosen ones who will vote for candidate
    - Although, θ is non random, we estimate it via $\hat{\Theta}$, a random variable
    - Here $\hat{\Theta}$ is random variable, because it depends on random sample

Bayesian Inference...

What is Bayesian Inference?

Here the unknown quantity Θ is assumed to be a random variable. Furthermore, we assume to have some initial guess about the distribution of Θ. After we observe the data, we can update the distribution of Θ using Bayes rule.

- Consider communication systems in which information is transmitted in the form of bits
- In each transmission, the transmitter sends a 1 with probability p, and sends a 0 with probability $1-p$
- Hence, if Θ is the transmitted bit, then $\Theta \sim \text{Bernoulli}(p)$
- Let us assume that at receiver end we get the output X
- The problem now is to estimate Θ from the noisy output X
- We use the prior knowledge that $\Theta \sim \text{Bernoulli}(p)$

What is Random Sampling? Motivation with an example...

1. Choose a random sample of size $n: X_1, \ldots, X_n$ with replacement
2. We chose a person uniformly at random from the population and let X_1 be the height of that person. Here, every person in the population has the same chance of being chosen
3. To determine the value of X_2, again we choose a person uniformly (and independently from the first person) at random and let X_2 be the height of that person. Again, every person in the population has the same chance of being chosen

- In general, X_i is the height of the ith person that is chosen uniformly and independently from the population
- why do we do the sampling with replacement?
  - if the population is large, then the probability of choosing one person twice is extremely low
  - big advantage of sampling with replacement is that X_i's will be independent
  - that is, working with independently and identically distributed makes analysis simpler

Definition of Random Sample...

Definition of Random sample

The collection of random variables $X_1, X_2, X_3, ..., X_n$ is said to be a random sample of size n if they are independent and identically distributed (i.i.d.), i.e.,

1. $X_1, X_2, X_3, ..., X_n$ are independent random variables, and
2. they have the same distribution, i.e,

$$F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x), \quad \text{for all } x \in \mathbb{R}$$

Point Estimator and Sample Mean...

Definition of Sample Mean

Let $X_1, X_2, \ldots, X_n$ be random sample. That is, here $X_1, X_2, \ldots, X_n$ are i.i.d. That is, following holds true for i.i.d. random variables

1. The X_i's are independent (since they are i.i.d.)
2. $F_{X_1}(x) = F_{X_2}(x) = \cdots = F_{X_n}(x) = F_X(x)$ (the CDFs are same)
3. $E[X_i] = E[X] = \mu < \infty$
4. $0 < \mathrm{Var}(X_i) = \mathrm{Var}(X) = \sigma^2 < \infty$

Then the sample mean is defined as follows

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

Recall: Properties of Sample Mean...

Properties of sample mean, $\bar{X}$

1. $E[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \dfrac{\sigma^2}{n}$
2. Weak law of large numbers (WLLN)

$$\lim_{n\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$$

3. Central limit theorem: The random variable

$$Z_n = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\sigma}$$

converges in distribution to the standard normal random variable

$$\lim_{n\to\infty} P(Z_n \leq x) = \Phi(x), \quad \text{for all } x \in \mathbb{R},$$

where $\Phi(x)$ is standard normal CDF.

Order Statistics and its PDF and CDF

Let $X_1, X_2, \ldots, X_n$ be random sample from a continuous distribution with CDF $F_X(x)$. If we order the random variables from smallest to largest i.e., $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$ with

$$X_{(1)} = \min(X_1, X_2, \cdots, X_n) \quad \text{and} \quad X_{(n)} = \max(X_1, X_2, \ldots, X_n),$$

then $X_{(i)}'$s is called order statistics. The CDF and PDF of $X_{(i)}$ are given by

$$f_{X_{(i)}} = \frac{n!}{(i-1)!(n-i)!} f_X(x) [F_X(x)]^{i-1} [1 - F_X(x)]^{n-i}$$

$$F_{X_{(i)}} = \sum_{k=i}^{n} \binom{n}{k} [F_X(x)]^k [1 - F_X(x)]^{n-k}$$

Also, the joint PDF of $X_{(1)}, X_{(2)}, \cdots, X_{(n)}$ is given by

$$f_{X_{(1)},\ldots,X_{(n)}}(x_1, \ldots, x_n) = \begin{cases} n!\, f_X(x_1, f_X(x_2) \cdots f_X(x_n)) & \text{for } x_1 \le x_2 \le \cdots \le x_n \\ 0 & \text{otherwise} \end{cases}$$

Example of Order Statistics...

Example (Order Statistics)

Let $X_1, X_2, \ldots, X_4$ be a random variable from the Uniform(0,1) distribution, and let $X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}$ be the order statistics of $X_1, X_2, \ldots, X_4$.
Find the PDFs of $X_{(1)}, X_{(2)}$, and $X_{(4)}$.

Answer to previous problem...

Point Estimator, Biased and Unbiased Estimators...

Definitions: point estimator, bias and unbiased estimators

1. Let θ be an unknown parameter to be estimated. For example, $\theta = E[X]$
2. Let $X_1, X_2, \ldots, X_n$ be a random sample using which we want to estimate θ. Here X_i's have same distribution
3. To estimate θ we define point estimator $\hat{\Theta}$ as follow

$$\hat{\Theta} = h(X_1, X_2, \ldots, X_n)$$

4. There can be many possible point estimators, which one to choose?
   - For example if $\theta = E[X]$, then $\hat{\Theta} = h(X_1, \ldots, X_n) = \dfrac{X_1 + \cdots + X_n}{n}$
5. **Bias:** The bias of a point estimator $\hat{\Theta}$ is defined as

$$B(\hat{\Theta}) = E[\hat{\Theta}] - \theta$$

   - If bias is close to 0, then $\hat{\Theta}$ is closer to θ
   - We say that $\hat{\Theta}$ is an unbiased estimator for a parameter θ if

$$B(\hat{\Theta}) = 0, \quad \text{for all possible values of } \theta$$

Example

Example

Example Let $X_1, \ldots, X_n$ be a random sample. Show that the sample mean

$$\hat{\Theta} = \overline{X} = \frac{X_1 + X_2 + ... + X_n}{n} \tag{1}$$

is an unbiased estimator of $\theta = EX_i$.

Unbiased Estimator is not Necessarily a Good Estimator...

Fact

Show that unbiased estimator is not necessarily a good estimator.

Solution:

In the above example, if we choose $\hat{\Theta}_1 = X_1$, then $\hat{\Theta}_1$ is also an unbiased estimator of θ.

- However, we suspect that $\hat{\Theta}_1$ is probably not as good as the sample mean $\bar{X}$.
- We need other measures to ensure that an estimator is a good estimator.

Mean Squared Error...

Mean squared error

The mean squared error (MSE) of a point estimator $\hat{\Theta}$ denoted by $\text{MSE}(\hat{\Theta})$ is defined as

$$\text{MSE}(\hat{\Theta}) = E[(\hat{\Theta} - \theta)^2]$$

Example (Application of MSE)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean $E[X_i] = \theta$, and variance $\text{Var}(X_i) = \sigma^2$. For the following two estimators for θ

1. $\hat{\Theta}_1 = X_1$
2. $\hat{\Theta}_2 = \bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$

Find $\text{MSE}(\hat{\Theta}_1)$ and $\text{MSE}(\hat{\Theta}_2)$ and show that for $n > 1$

$$\text{MSE}(\hat{\Theta}_1) > \text{MSE}(\hat{\Theta}_2).$$

Answer to previous problem...

Answer to previous problem...

Relationship of MSE, Variance, and Bias...

Property

If $\hat{\Theta}$ is a point estimator for θ,

$$\mathrm{MSE}(\hat{\Theta}) = \mathrm{Var}(\hat{\Theta}) + B(\hat{\Theta})^2$$

Solution:

$$\begin{aligned} MSE(\hat{\Theta}) &= E\left[(\hat{\Theta} - \theta)^2\right] && (2) \\ &= \mathrm{Var}(\hat{\Theta} - \theta) + \left(E[\hat{\Theta} - \theta]\right)^2 && (3) \\ &= \mathrm{Var}(\hat{\Theta}) + B(\hat{\Theta})^2. && (4) \end{aligned}$$

Consistent Estimator...

Definition of Consistent Estimator

Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots, \hat{\Theta}_n, \ldots,$ be a sequence of point estimators of θ. We say that $\hat{\Theta}_n$ is a consistent estimator of θ, if

$$\lim_{n\to\infty} P(|\hat{\Theta}_n - \theta| \geq \epsilon) = 0, \quad \text{for all } \epsilon > 0$$

Theorem

Let $\hat{\Theta}_1, \hat{\Theta}_2, \ldots,$ be a sequence of point estimators of θ. If

$$\lim_{n\to\infty} \mathrm{MSE}(\hat{\Theta}_n) = 0$$

then $\hat{\Theta}_n$ is a consistent estimator of θ

Answer to previous problem...

Example

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample with mean $E[X_i] = \theta$ and variance $\mathrm{Var}(X_i) = \sigma^2$. Show that $\hat{\Theta}_n = \bar{X}$ is a consistent estimator of θ.

Definition of Sample Variance and Sample Standard Deviation...

Sample Variance and Sample Standard Deviation

Let $X_1, X_2, \ldots, X_n$ be a random variable with mean $E[X_i] = \mu < \infty$, and variance $0 < \mathrm{Var}(X_i) < \sigma^2 < \infty$. The sample variance of this random sample is defined as

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

We can check that sample variance is an unbiased estimator of σ^2. The sample standard deviation is defined as

$$S = \sqrt{S^2}$$

and it is usually used as an estimator for σ. Also, S is an unbiased estimator of σ.

Example

Example

Example Let $X_1, \ldots, X_n$ be a random variable with mean $E[X_i] = \mu$ and variance $Var(X_i) = \sigma^2$. Suppose that we use

$$\overline{S}^2 = \frac{1}{n}\sum_{k=1}^{n}(X_k - \overline{X})^2 = \frac{1}{n}\left(\sum_{k=1}^{n} X_k^2 - n\overline{X}^2\right) \tag{5}$$

to estimate σ^2. Find the bias of this estimator

$$B(\overline{S}^2) = E[\overline{S}^2] - \sigma^2. \tag{6}$$

Example (Sample Mean, Sample Variance, Sample Standard Deviation)

Let T be the time that is needed for a specific task in a factory to be completed. In order to estimate the mean and variance of T, we observe a random sample $T_1, T_2, \cdots, T_6$. Thus, T_i's are i.i.d. and have the same distribution as T. We obtain the following values (in minutes):

$$18, 21, 17, 16, 24, 20.$$

Find the values of the sample mean, the sample variance, and the sample standard deviation for the observed sample.

Answer to previous problem...

Example (Prediction Using Data)

I have a bag that contains 3 balls. Each ball is either red or blue, but I have no information in addition to this. Thus, the number of blue balls, call it θ, might be 0, 1, 2, or 3. I am allowed to choose 4 balls at random from the bag with replacement. We define the random variables X_1, X_2, X_3, and X_4 as follows

$$X_i = \begin{cases} 1 & \text{if the } ith \text{ chosen ball is blue} \\ 0 & \text{if the } ith \text{ chosen ball is red} \end{cases}$$

We observe here that X_i's are i.i.d. and $X_i \sim \text{Bernoulli}\left(\frac{\theta}{3}\right)$. After the experiment, we observe the values for X_i's

$$x_1 = 1,\ x_2 = 0,\ x_3 = 1,\ x_4 = 1.$$

From above, we have 3 blue balls and 1 red ball. Answer the following

1. Find the probability of the observed sample $(x_1, x_2, x_3, x_4) = (1, 0, 1, 1)$ for each possible θ
2. Find the value of θ that maximizes the probability of the observed sample

Answer to previous problem...

Answer to previous problem...