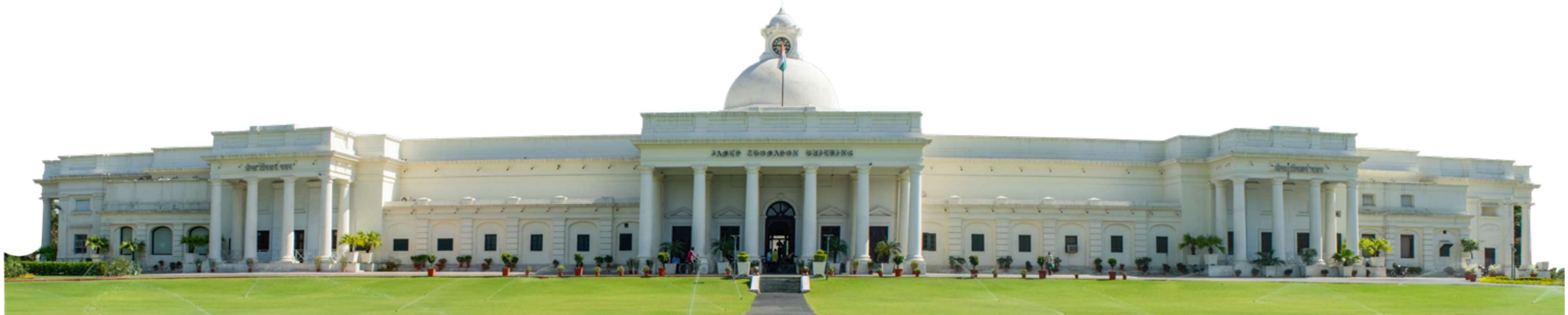
INDIAN INSTITUTE OF TECHNOLOGY ROORKEE

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 7, 2023



Recap

$X_1, X_2, \ldots, X_{50} \;\|\; \bar{X}, s^2 (\to 49)$

$\bar{X} \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right)$

Central Limit Theorem
Confidence Interval for the case of Mean

Errors in the Process of Estimation

- **Sampling Error** - Because we are only considering a subset of population, the point estimate is rarely exactly correct. Unavoidable error, but we can estimate the error and hence have some control over it
- **Non-sampling Error** - If there is bias in the observations, or sampling wasn't done properly. Can't be dealt with mathematically. Should be avoided

Central Limit Theorem

Theorem

If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean $\overline{X}$ is approximately normal with

1. *mean* = μ
2. *variance* = $\frac{\sigma^2}{n}$

What is meant by large n? Typically, $n \geq 30$

Comments about Central Limit Theorem

1. Don't use CLT when population size is not too large compared with sample size, CLT is approximate result
2. If sample size is large and population size is also much larger as compared to population, one can use CLT even when the sampling scheme is Without Replacement

Example of Confidence Interval

A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.

Sample Proportion

P_1 : Yes 1
P_2 : No 0
⋮
P_N : No 0

$$\boxed{p = \frac{N_{yes}}{N}}$$

X_1 : yes 1
X_2 : yes 1
⋮
X_n : No 0

n_{yes}
n_{No}

- Sometimes, one is interested in estimating population proportion
- What is the proportion of IBM312 students who like statistics?
- One can attempt the answer to this using sampling

$$\boxed{E(\hat{p}) = p}$$

??

$$\boxed{\hat{p} = \frac{n_{yes}}{n}}$$

Sample Proportion

$$\hat{p} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

- Can we make use of results from sample mean?

$$X_1 = \begin{cases} 1 & ; \; p \\ 0 & ; \; (1-p) \end{cases}$$

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

Sample Proportion

$$E\left(\frac{\hat{p}(1-\hat{p})}{n}\right) \neq \frac{p(1-p)}{n}$$

- Can we make use of results from sample mean?
- If the i^{th} respondent says YES, model it as $X_i = 1$
- If the i^{th} respondent says NO, model it as $X_i = 0$
- Denote by n_{YES} and n_{NO} are the responses in the sample of size n
- Denote by N_{YES} and N_{NO} are the actual values in the population of size N

$$E(\hat{p}^2) = p^2 + \frac{p(1-p)}{n}$$

Sampling Proportion

$$E\left(\frac{\hat{p}(1-\hat{p})}{n}\right) = \frac{1}{n}\left[p - p^2 - \frac{p(1-p)}{n}\right]$$

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- $E(\hat{p}) = p$. Do we need to prove this??
- $\mathrm{Var}(\hat{p}) = \dfrac{p(1-p)}{n}$. Why??
- Is p known?
- State CLT for sample proportion
- Additional conditions - $np \geq 10$ and $n(1-p) \geq 10$

$$\frac{(n-1)(p(1-p))}{n^2}$$

$$\left(\frac{\hat{p}(1-\hat{p})}{n-1}\right)$$

$$E(\underline{f(\hat{p})}) = \frac{p(1-p)}{n}$$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- What kind of random variable is n_{YES}??
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

$\hat{p}$, $\dfrac{\hat{p}(1-\hat{p})}{n-1}$

$P\left(\hat{p} - 1.96\sqrt{\boxed{\dfrac{p(1-p)}{n}}} < p\right)$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Easier way for check unbiasedness of sample proportion

- We denote the estimate by $\hat{p}$
- The population proportion is denoted by p
- $\hat{p} = \dfrac{n_{YES}}{n}$
- n_{YES} is Binomial random variable with parameters p and n
- Hence, $E(\hat{p}) = E\left(\dfrac{n_{YES}}{n}\right) = \dfrac{E(n_{YES})}{n} = p$
- Also, $Var(\hat{p}) = Var\left(\dfrac{n_{YES}}{n}\right) = \dfrac{Var(n_{YES})}{n^2} = \dfrac{p(1-p)}{n}$
- But, we don't know p
- $\mathrm{Var}(\hat{p}) = \dfrac{\hat{p}(1-\hat{p})}{n-1}$. Why??
- To provide an unbiased estimator of $\mathrm{Var}(\hat{p})$

Confidence Interval Discussions

- Can you also do similar calculations and make a confidence interval for Population proportion? (Hint - Use CLT and our remark that sample proportion can be given a similar treatment as sample mean)
- Khan Academy Video
  `https://www.youtube.com/watch?v=bGALoCckICI`
- Which is bigger - 99% confidence interval or 95% confidence interval?

Summary of results for 100(1-α)% C.I.

n	σ^2	C.I. Type	Symmetric C.I.
Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

$2 \times 1.96 \times \frac{s}{\sqrt{n}} \leq 20$

Table: C.I. for population mean μ, s is sample standard deviation

$n \geq \left(\frac{392}{20}\right)^2$

n	C.I. Type	Symmetric C.I.
Large	Approximate	$\left(\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right)$

Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

Sample Size Determination

- A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- We want to repeat this study, how many students should you survey so that the 99% confidence interval's width is no more than 2 minutes?