Statistical Thinking for Data Science

Sumit Kumar Yadav

Department of Management Studies

May 15, 2022

Learning Outcomes Session - 1

- Descriptive Statistics
- Concept of Probability
- Random Variable, Discrete and Continuous
- Expected Value, Variance, Correlation

Definitions of Statistics

- Art of learning from data – Sheldon M. Ross, Introduction to Probability and Statistics for Engineers and Statistics
- Statistics is a branch of mathematics working with data collection, organization, analysis, interpretation and presentation. - Wikipedia
- Statistics may be regarded as (i) the study of populations, (ii) as the study of variation, (iii) as the study of methods of the reduction of data. – Fisher, 1925

Descriptive Statistics

- Describing the data, Summarize the data, etc
- Using numbers, pictures etc. → Box Plot
- https://www.ted.com/talks/hans_rosling_the_best_stats_you_ve_ever_seen
- https://www.ted.com/talks/hans_rosling_asia_s_rise_how_and_when

Summary Statistics

$X_1, X_2, \ldots, X_n$

$\mu = \frac{X_1 + X_2 + \ldots + X_n}{n}$

S-1	S-2
100	80
150	90
200	1000

Average

- Measures of Central Tendency(mean, median, mode)
- Measures of Dispersion (Range, Variance)
- Chebyschev Inequality

(Max value - Min value)

D1	D2		D3
8	(-10)	0	9
9	(-5)	5	9
10	0	10	9
11	5	15	9
12	10	20	14

Summary Statistics(multiple data-sets)

Co-variance and Correlation

Visually describing the data

Scatter Plot, Histogram

Need for visually describing the data

Anscombe's Quartet Counter Example

Box-Plot (in Python Example)

A Few More terminologies

- Cross-sectional Data
- Time Series Data
- Panel Data

- Qualitative Data
  1. Nominal
  2. Ordinal
- Quantitative Data
  1. Interval
  2. Ratio

Probability Concepts

A - Australia
B - Bangladesh

1. A wins — 0.7
2. B wins — 0.25
3. It is a draw — 0.05

{ Head, Tail }
{ 1, 2, 3, 4, 5, 6 }

- What is probability??
- Concept of Experiment, Sample Space, Events
- A number associated with each Sample Point $P(E_i)$
- Less than 1
- Sum of all probabilities = 1
- $P(A_1 \cup A_2) <= P(A_1) + P(A_2)$
- Intersection of Events, Independent Events (Card Example)

A wins — 0.9
B wins — 0.09
Draw — 0.01

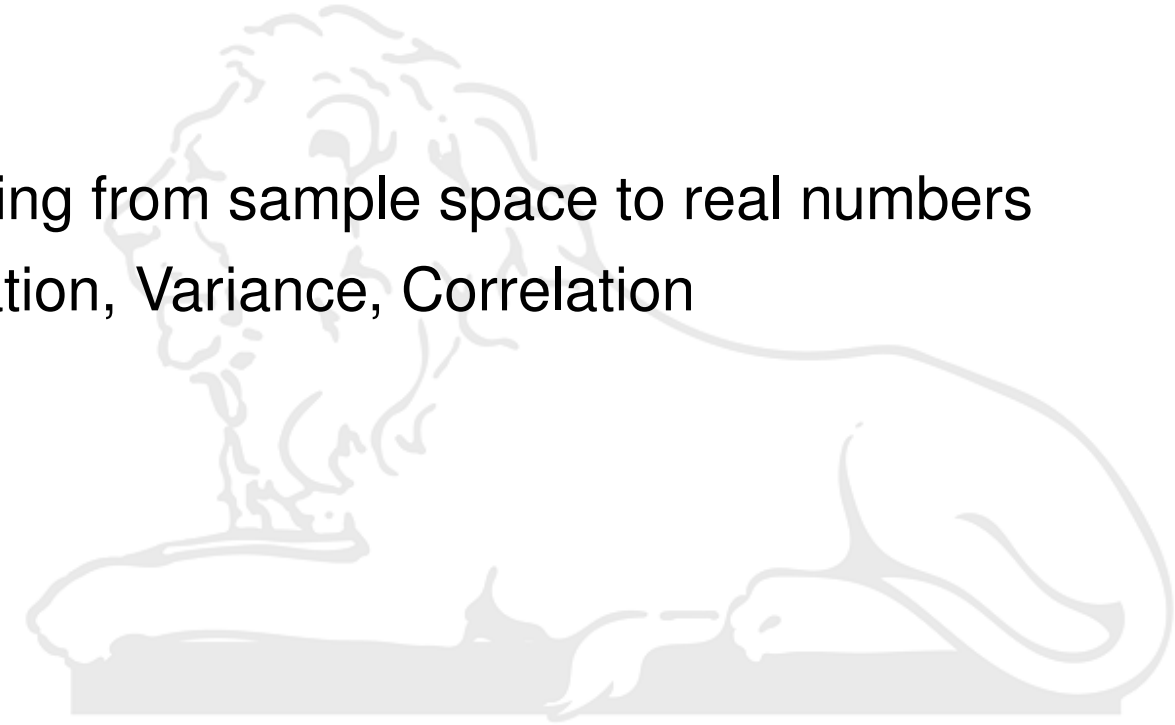
Probability Concepts

- What is probability??
- Concept of Experiment, Sample Space, Events
- A number associated with each Sample Point $P(E_i)$
- Less than 1 & ≥ 0
- Sum of all probabilities = 1
- $P(A_1 \cup A_2) <= P(A_1) + P(A_2)$
- Intersection of Events, Independent Events (Card Example)

GodBole's Problem

Random Variable

- A mapping from sample space to real numbers
- Expectation, Variance, Correlation

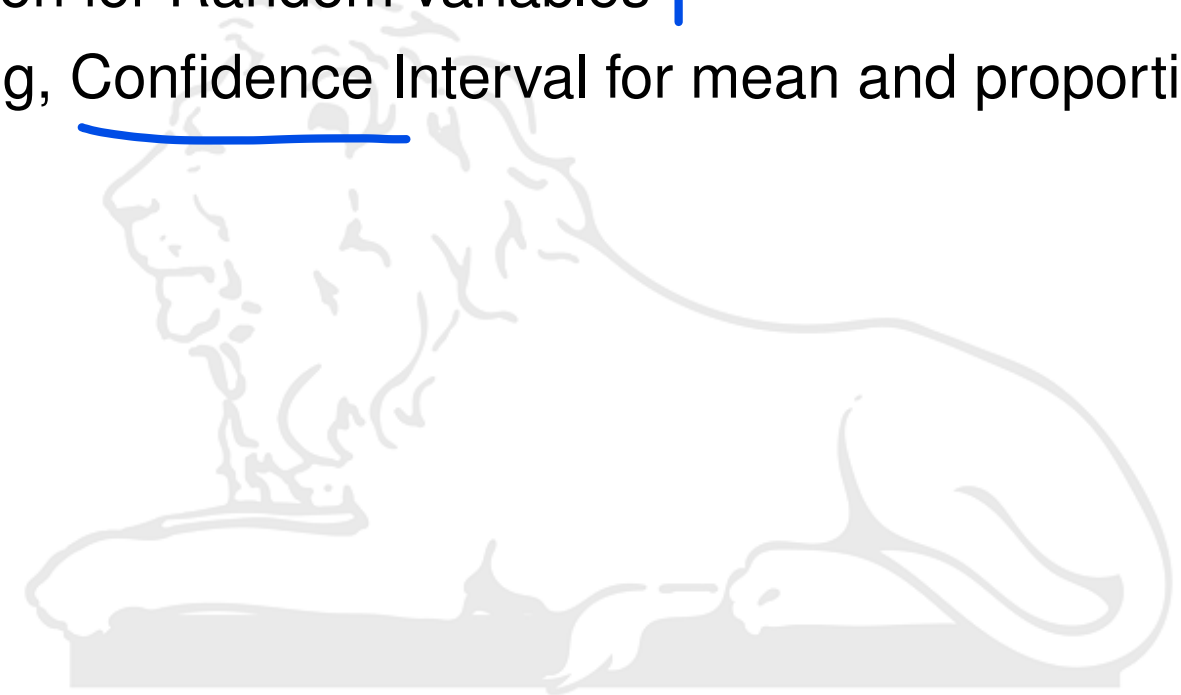


Pooled Testing Example

A LAB doing Covid testing gets 1000 samples to test everyday. However, due to the positivity rate drop in cases of Covid samples, the LAB is contemplating if it is better to mix the samples to get the result in lesser number of tests. Assuming 3% positivity rate, what is the number of samples that should be pooled together?

Learning Outcomes Session - 2

- Standard discrete and Continuous Distributions
- Binomial, Poisson, Geometric, Normal, Uniform, Exponential
- Simulation for Random variables
- Sampling, Confidence Interval for mean and proportion



Standard Distributions

- Discrete
  1. Binomial
  2. Poisson
  3. Geometric
- Continuous
  1. Normal
  2. Uniform
  3. Exponential

Simulation of Random numbers in Pytho

Basic Ideas of Sampling

Confidence Interval

Case I : 500 — 9000 → (900, 9100)

Case II : 100 — 8800 → (8200, 9400)

1. Population (Sometimes, it is not even observable and only abstract)
2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
3. Subject
4. Parameter(Constant - might be unknown)
5. Statistic (Random Variable)

Population

↓

20 lakh

(20 − 0.5)

19.5 lakh → 10,000

+

Sample

[3900]

(50,000)

AAP — (38%)

Basic Ideas of Sampling

1. Population (Sometimes, it is not even observable and only abstract)
2. Sampling Frame (if you are lucky, you might get this, not guaranteed in most practical situations)
3. Subject
4. Parameter(Constant - might be unknown)
5. Statistic (Random Variable)
6. Statistic ceases to be a random variable after it is observed

Types of Sampling

Population

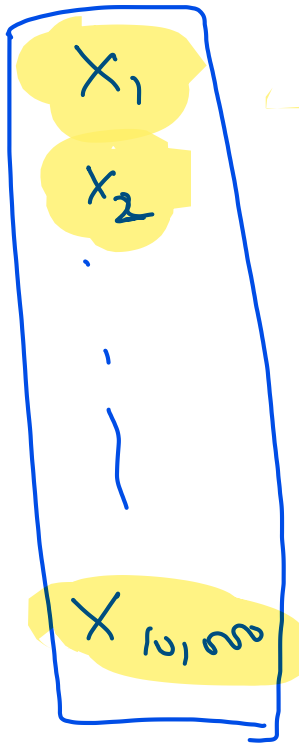
$P(X_1 = A_1) = \frac{1}{20 \text{ lakh}}$

$P(X_1 = A_2) = \frac{1}{20 \text{ lakh}}$

- Simple Random Sample With replacement
- Simple Random Sample without replacement
- Cluster Sampling
- Stratified Sampling

$[X_1 = A_{4823}, \; X_2 = A_{4823}]$

Sample



Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??

Potential Causes of Bias

- Convenience Sampling
- Volunteer Sampling
- Systematic Sampling
- Non-response Bias
- Response Bias

Does that mean we shouldn't use any of these types of sampling??
NO, one can use, but with caution. Make sure it is not leading to a systematic error

Sampling Using Excel in Presence of Sampling Frame

Generate a random number for each observation and then sort the observations

What after sampling?

- Ask, why did we do sampling? Objective is to learn about the population
- Statistical Inference - Learn about parameters from sample statistic
- Usually, the quantities of interest are mean and proportion in the population (depending on the context)
- We deal with them separately

Estimating from the statistic

$E(\bar{x}) = \mu$

$\bar{x} = \frac{x_1 + x_2}{3}$

- The rational behind estimating is expectation of statistic should be equal to the population parameter
- Biased and Unbiased Estimator
- Variance of estimate should be minimized to the extent possible

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

Estimating Population Mean from Sampl

1. Let the true values in the population be $A_1, A_2, A_3, \cdots, A_N$
2. Population mean is denoted by μ and equals $\frac{\sum_{i=1}^{N} A_i}{N}$
3. Also, population variance is denoted by σ^2 and equals $\frac{\sum_{i=1}^{N}(A_i - \mu)^2}{N}$
4. Let the sample be a SRS of size n
5. Observations are $X_1, X_2, \ldots, X_n$
6. Sample mean is denoted by $\overline{X}$ and defined as follows
   $\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$
7. $E(\overline{X}) = \mu$

$Var(X_i) = \sigma^2$

Estimating Population Variance from Sample Observations

- If the sampling scheme is WITH REPLACEMENT

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

- If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N-1}{N}\right)\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- Why is estimating population variance important?

Estimating Population Variance from Sample Observations

- If the sampling scheme is WITH REPLACEMENT

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

- If the sampling scheme is WITHOUT REPLACEMENT

$$s_{X,WOR}^2 = \left(\frac{N-1}{N}\right)\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

Why is estimating population variance important?

To get an idea about error in estimation of sample mean

Standard Error in Sample Mean

- If the sampling scheme is WITH REPLACEMENT
  $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$
- If the sampling scheme is WITHOUT REPLACEMENT
  $$\mathrm{Var}(\overline{X}) = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$
- $\frac{N-n}{N-1}$ is called the finite population correction
- Typically, can be ignored if sampling fraction $\frac{n}{N} \leq 0.05$
- Standard deviation of $\overline{X}$ is called the Standard error of the sample mean
- Do we know σ^2?? What is the remedy??

Central Limit Theorem

- Can we do better about our inference from sample mean??
- Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?

Central Limit Theorem

- Can we do better about our inference from sample mean??
- Maybe as the sample size increases, can we say something more than just expectation and variance of sample mean?
- Thank you Central Limit Theorem

Central Limit Theorem

$\mu = \frac{A_1 + A_2 + \cdots + A_N}{N}$

A_1
A_2
$\vdots$
A_N

Theorem

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

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

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

x_1
x_2
$\vdots$
x_{500}

$\overline{X} = \frac{x_1 + x_2 + \cdots + x_{500}}{500}$

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

CLT when population is normally distribu

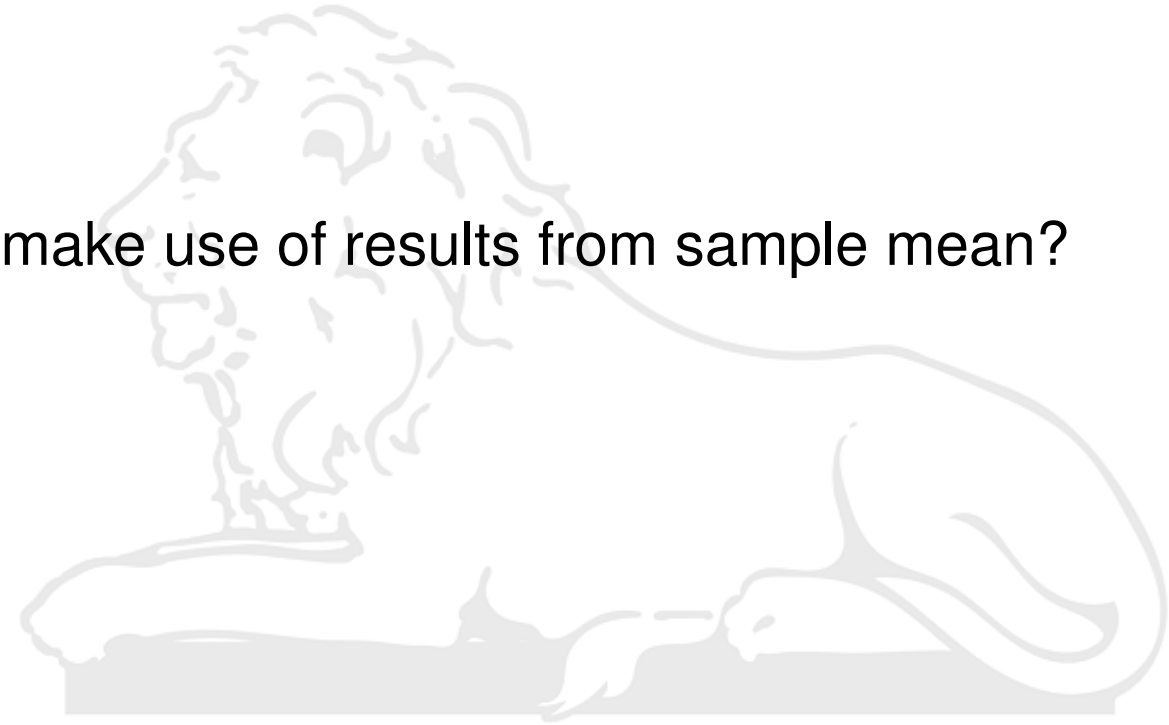
1. If underlying population is normally distributed, CLT is not required
2. If population is approximately normal, then $\overline{X}$ is approximately normal even for small n

Sample Proportion

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

Sample Proportion

- Can we make use of results from sample mean?



Sample Proportion

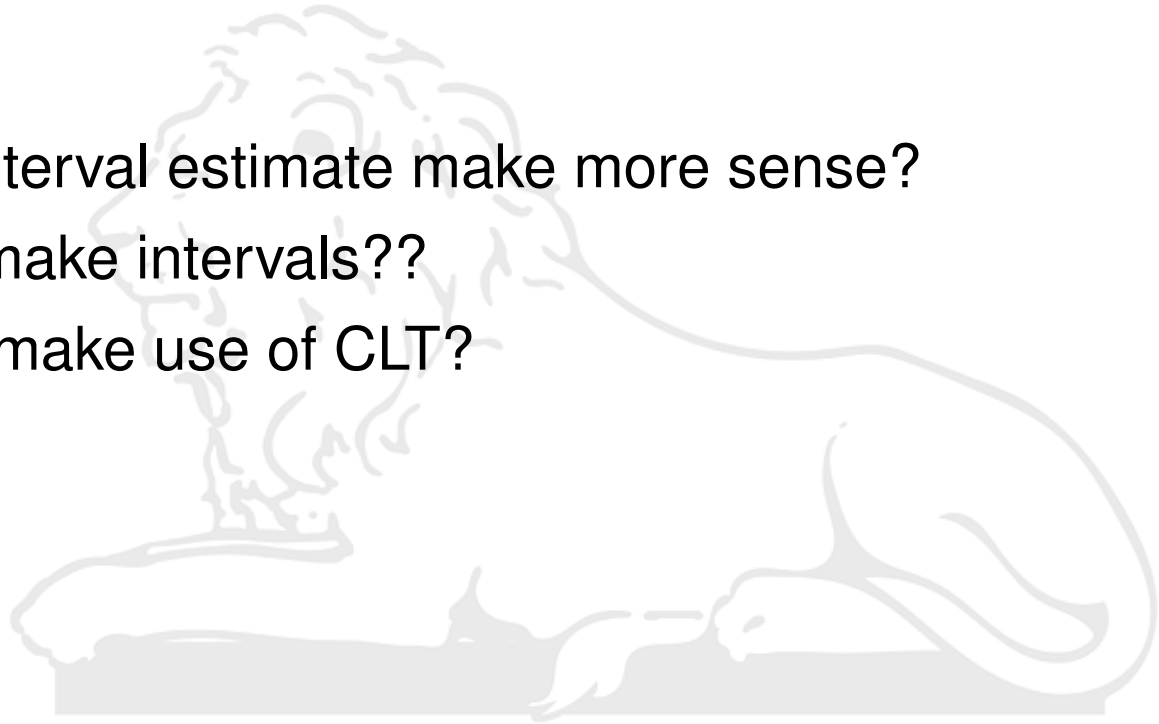
- Can we make use of results from sample mean?
- If the i^{th} respondent says YES, model it as $X_i = 1$
- If the i^{th} respondent says NO, model it as $X_i = 0$
- Denote by n_{YES} and n_{NO} are the responses in the sample of size n
- Denote by N_{YES} and N_{NO} are the actual values in the population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

Sampling Proportion

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

Confidence Interval

1. Would interval estimate make more sense?
2. How to make intervals??
3. Can we make use of CLT?



Confidence Interval

1. Confidence Interval typically computed by adding and subtracting standard error in the point estimate
2. Point estimate +/- some multiple of Standard error
3. What is confidence level?

Confidence Interval Idea

1. Confidence Interval for any parameter of the population is a random interval that contains the parameter with some probability
2. If sampling is done repeatedly and confidence intervals are constructed, a 95% confidence interval will contain the values about 95% of the times
3. Typically, significance level is used, denoted by α
4. What is confidence level?

Confidence Interval Idea

1. 95% is the confidence level of the interval generated
2. We pick a sample, construct the interval. Can we say that the probability that the interval contains the true value is 0.95??
3. Different schools of thought, most don't agree on the above made statement
4. But, everyone agrees on the fact that confidence is on the procedure used to construct the confidence interval

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.

Easier way for check unbiasedness of sample proportion

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

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Standard Normal Distribution

- is a normal distribution with mean = 0, variance = 1
- Can you convert any normal distribution to a standard normal distribution by change of origin and change of scale??
- Let $X \sim N(\mu, \sigma^2)$
- Hence, $X - \mu \sim N(0, \sigma^2)$
- Thus, $Z = \dfrac{X - \mu}{\sigma} \sim N(0, 1)$
- Typically, Z is used to denote a standard normal distribution

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $\overline{X}$ is a random variable
2. Under certain conditions, large sample size, etc. We use CLT to get better idea about $\overline{X}$
3. Using properties of normal distribution, what can be said about
4. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right) =$ 0.9544
5. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
6. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
7. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
8. Magic here, we have created an interval for μ
9. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

Confidence Interval

Back to Sample Mean $\overline{X}$

1. $P\left(\overline{X} \text{ is in between } \mu - 2\frac{\sigma}{\sqrt{n}} \text{ and } \mu + 2\frac{\sigma}{\sqrt{n}}\right)$
2. Is is not 0.9544 approximately?? Why approximately?? Because CLT is approximate result.
3. Thus, $P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \leq \overline{X} \leq \mu + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
4. Or, by rearrangement of terms,
   $P\left(\overline{X} - 2\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + 2\frac{\sigma}{\sqrt{n}}\right) = 0.9544$
5. Magic here, we have created an interval for μ
6. This is nothing but the confidence interval

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	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

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

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

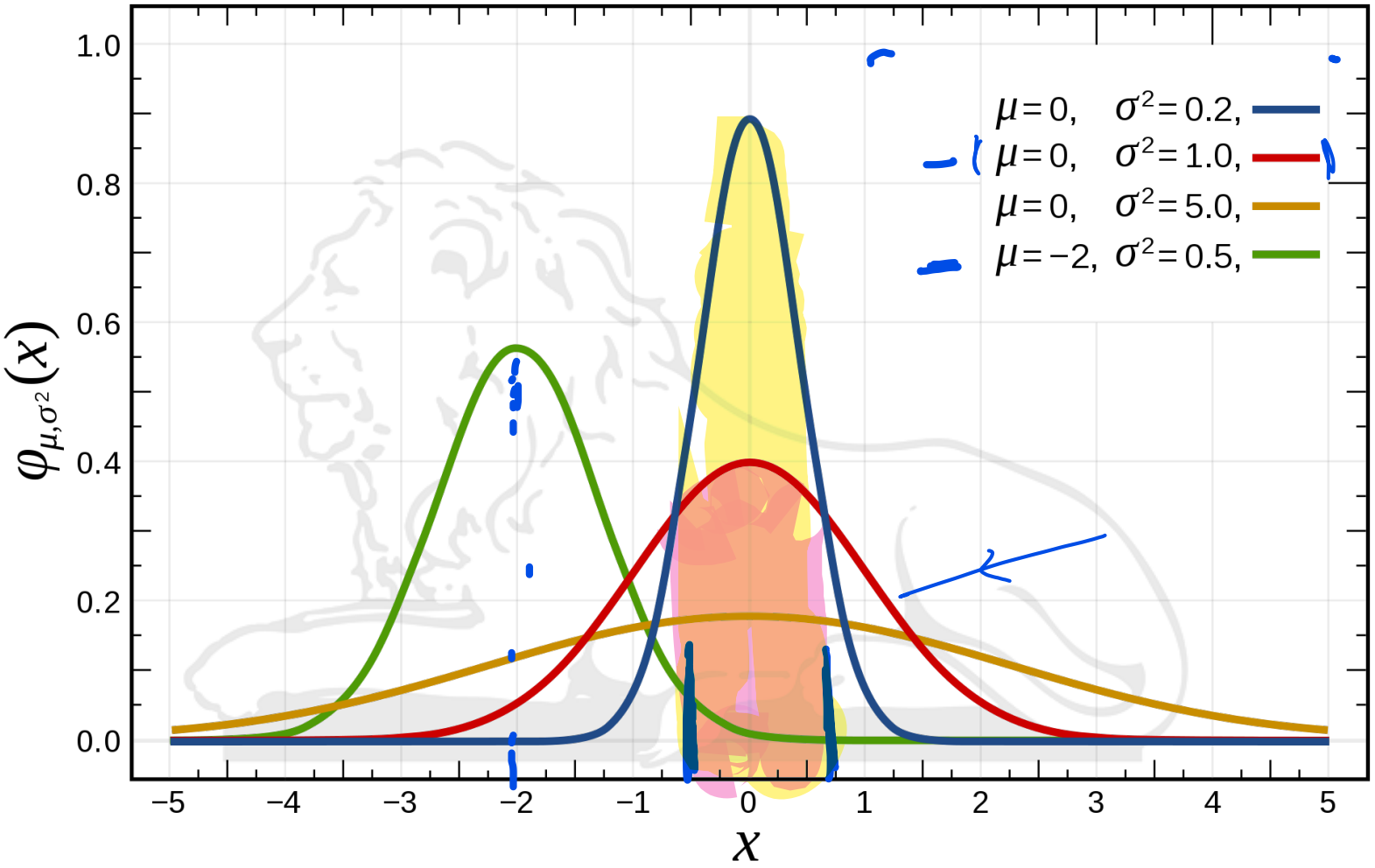
- A survey asked 500 randomly selected students, "the average time spent in physical exercise daily". Sample mean was 20 minutes, and standard deviation of the sample was 5 minutes. Construct a 95% confidence interval of the population mean of time spent in physical exercise daily.
- We want to repeat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

Basics about Random Variable

- A variable whose value depends on outcome of a random phenomenon
- A random variable is characterized by its distribution
- Sum of two or more different random variables is also a random variable, and thus will also have a distribution (we might not cover the mathematical tools required to find the distribution, but it is important to appreciate that it will have a distribution)
- Similarly, any other algebraic operation of two or more random variables also remain a random variable
- If X and Y are random variables, $X + Y, X - Y, XY, \frac{X}{Y}$ are all random variables

Standard Normal Distribution

A normal distribution with mean 0 and standard deviation as 1



Source - https://en.wikipedia.org/wiki/Normal_distribution#Standard_normal_distribution

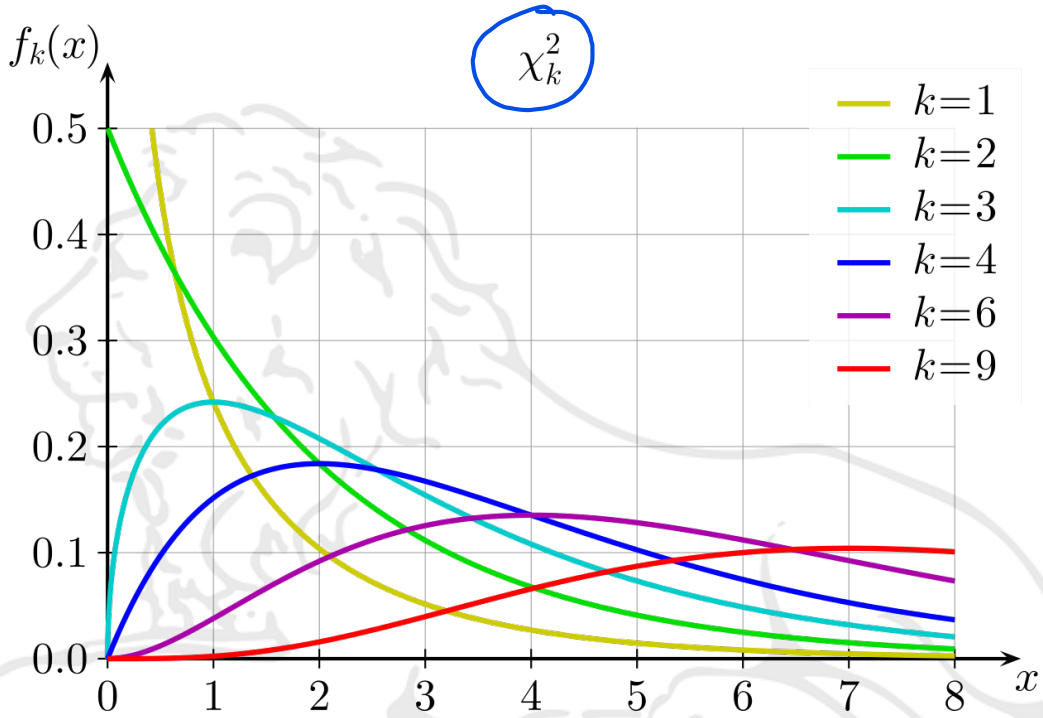
Chi-square distribution

If $Z_1, Z_2, \ldots, Z_n$ are all independent and normally distributed with mean 0 and variance 1, then $U = \sum_{i=1}^{n} Z_i^2 \sim \chi_n^2$

Alternatively, U is said to follow a χ^2 distribution with n degrees of freedom

$U = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$

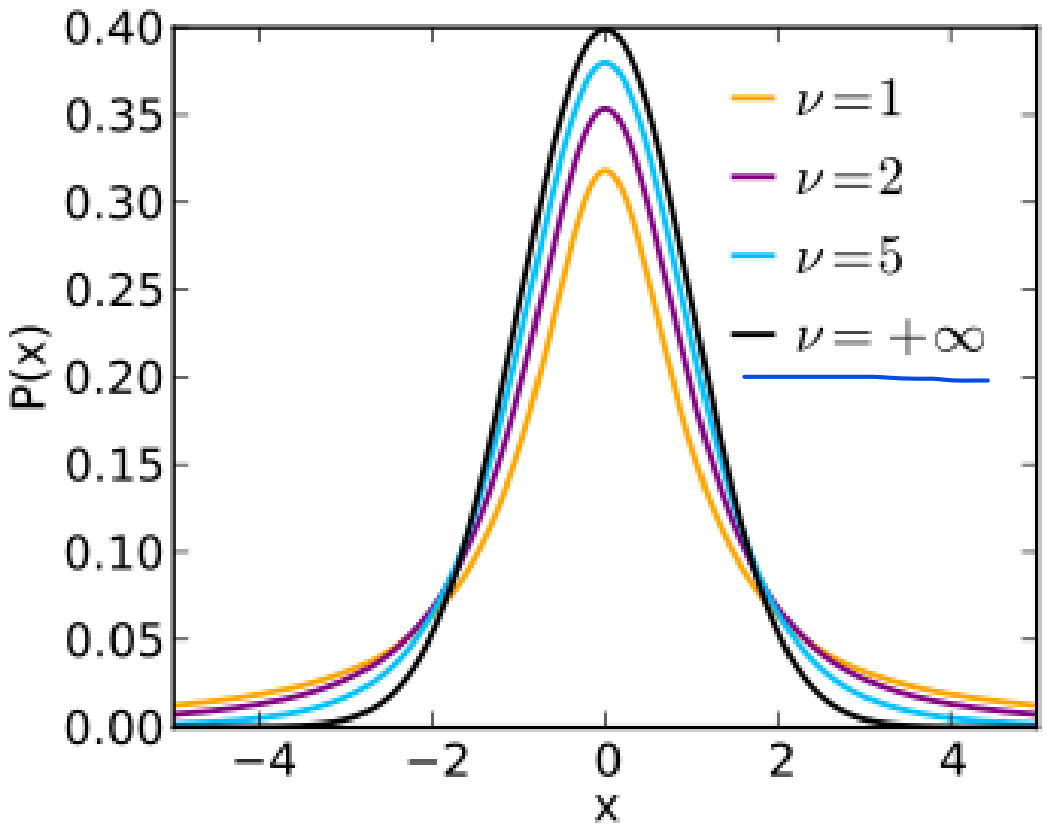
Chi-square distribution



Source - `https://en.wikipedia.org/wiki/Chi-squared_distribution`

t-distribution

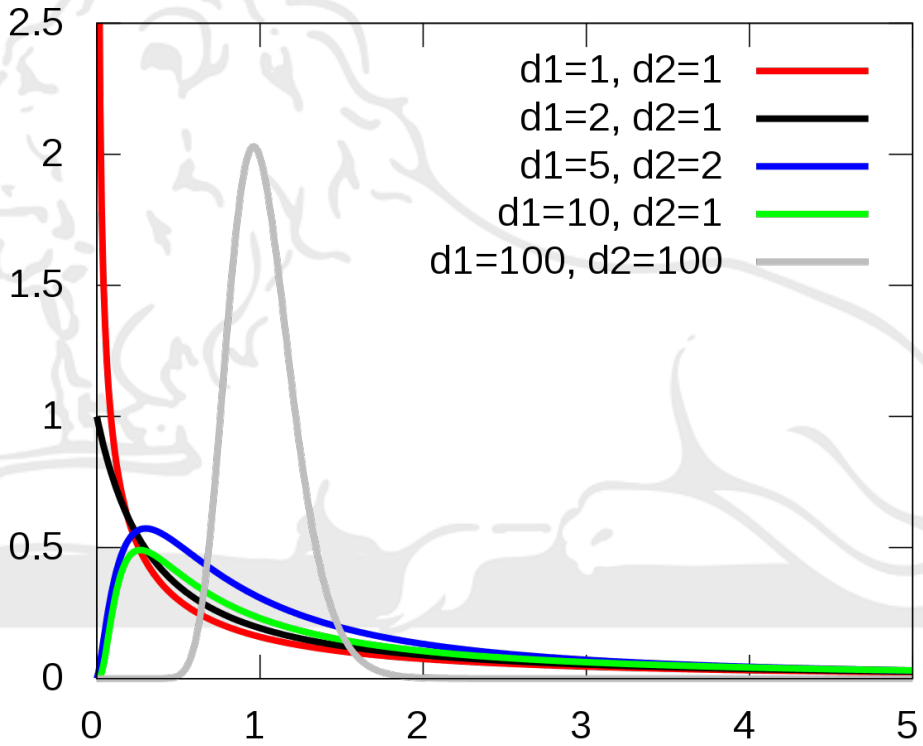
If $Z \sim N(0,1)$ and $U \sim \chi^2_n$ are independent, then the distribution of $\frac{Z}{\sqrt{\frac{U}{n}}}$ is called the t-distribution with n degrees of freedom



Source - https://en.wikipedia.org/wiki/File:Student_t_pdf.svg

F distribution

If $U \sim \chi^2_m$ and $V \sim \chi^2_n$ are independent, then the distribution of $\frac{\frac{U}{m}}{\frac{V}{n}}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



$$\frac{\left(\frac{U}{m}\right)}{\left(\frac{V}{n}\right)}$$

Source - https://en.wikipedia.org/wiki/F-distribution

Application of distributions that we just s

$X_1 \sim N(\mu, \sigma^2)$ $X_2 \sim N(\mu, \sigma^2)$.

- Does having the idea of population distribution itself a useful information?
- If yes, how do we make use of it?
- Let us concern ourselves with sample mean
- Assume you have the information that the population distribution is normal.
- How do you use this??
- Is CLT required??

$$\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Case of normal population

- Sample mean is denoted by $\overline{X}$ and defined as follows
$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$
- If the sampling scheme is WITH REPLACEMENT, sample variance equals
$$s_X^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n-1}$$
- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence

Case of normal population

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

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X} - \mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Case of normal population

- It can be shown that -
  1. $\overline{X}$ and s_X^2 are independent
  2. $\frac{(n-1)s_X^2}{\sigma^2} \sim \chi^2_{n-1}$
- Can you guess the distribution of $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}$
- If σ is unknown, we replace by s_X, but then the distribution ceases to be $N(0,1)$. So what is it then??
- Distribution of $\frac{\overline{X}-\mu}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}$
- Hence, even for small sample size, we can make confidence intervals if we know that the population is normally distributed

Moral of the story - results for 100(1-α)% C.I. for Mean

Population distribution	n	σ^2	C.I. Type	Symmetric C.I.
Any	Large	known	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Any	Large	unknown	Approximate	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}s}{\sqrt{n}}\right)$
Normal	Any	known	Exact	$\left(\overline{X} - \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}, \overline{X} + \frac{z_{\frac{\alpha}{2}}\sigma}{\sqrt{n}}\right)$
Normal	Any	unknown	Exact	$\left(\overline{X} - \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}, \overline{X} + \frac{t_{n-1,\frac{\alpha}{2}}s}{\sqrt{n}}\right)$

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

Typically, $t_{n-1,\frac{\alpha}{2}}$ is used only for small n, because for large n, $z_{\frac{\alpha}{2}}$ gives a good approximation

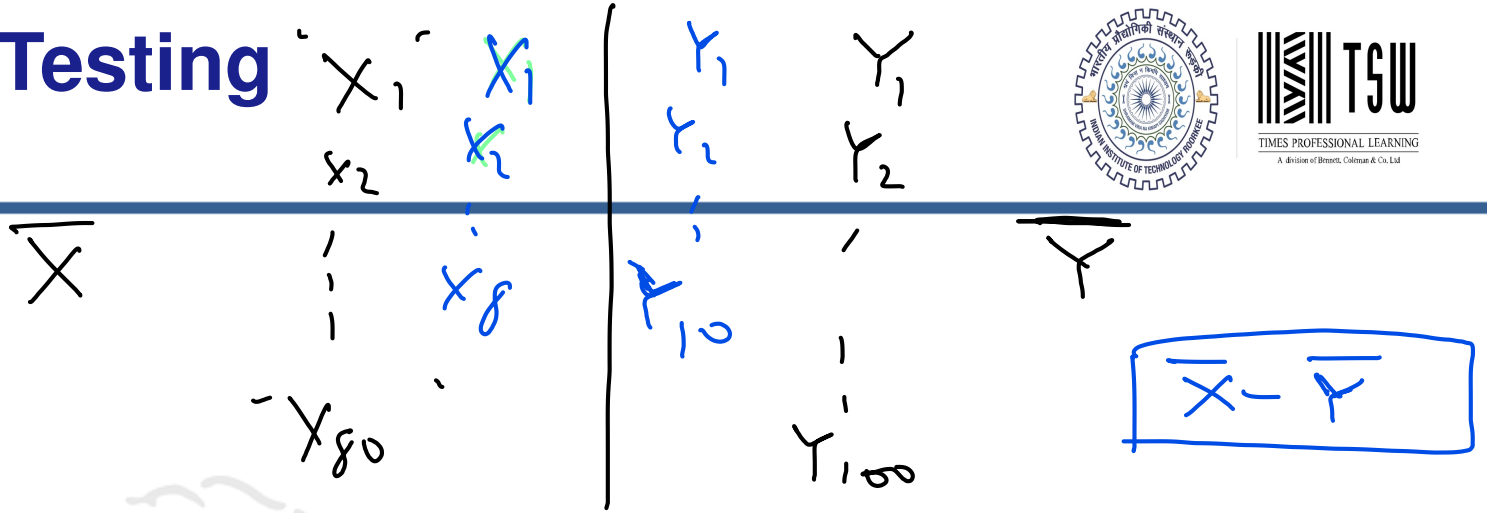
Moral of the story - results for 100(1-α)% C.I. for Proportion

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

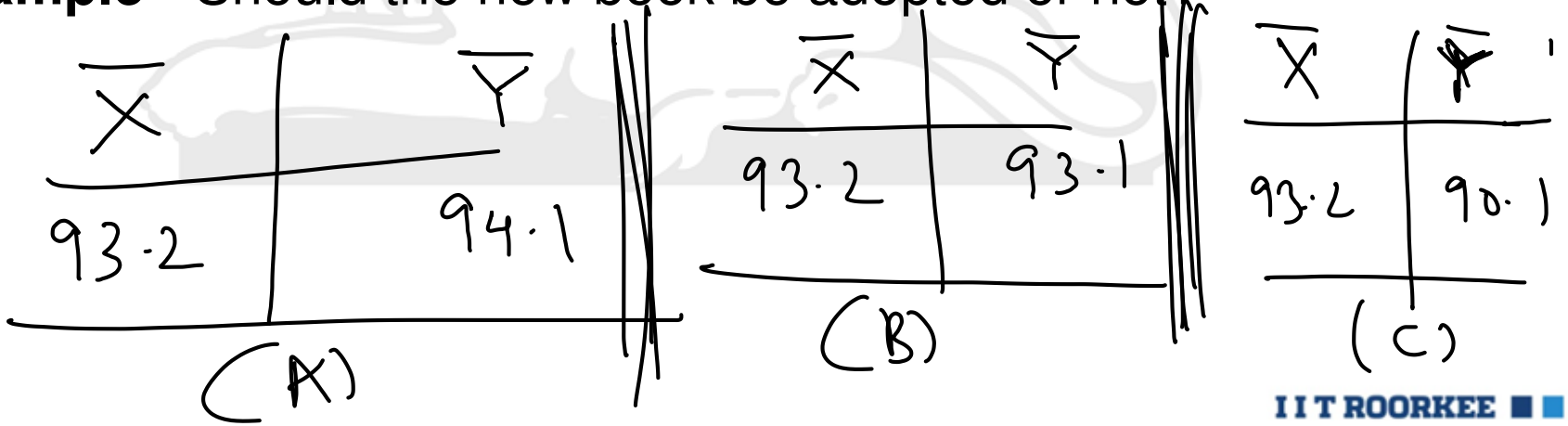
Table: C.I. for population proportion p, $\hat{p}$ is sample proportion

For case of proportion, it is advised to use these formulae only when apart from n being large, $n\hat{p} \geq 10$ and also $n(1-\hat{p}) \geq 10$

Hypothesis Testing



- Hypothesis testing is a way of making statistical decisions using observed or experimental data
- **Example** - Do students perform better if tests are given in the morning?
- **Example** - Is the coin biased or not?
- **Example** - Should the new book be adopted or not?



Key Terms in Hypothesis Testing

- Null Hypothesis, usually denoted by H_0
- Alternative Hypothesis, usually denoted by H_1 or H_a
- Hypothesis testing is performed after data is available for a sample - $X_1, X_2, \ldots, X_n$
- Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- Basis - from the value of a **statistic** T, which is calculated from the sample, called the **test statistic**
- Whenever the **null hypothesis** is **rejected**, **alternative hypothesis** is **accepted**

Key Terms in Hypothesis Testing

- Null Hypothesis, usually denoted by H_0
- Alternative Hypothesis, usually denoted by H_1 or H_a
- Hypothesis testing is performed after data is available for a sample - $X_1, X_2, \ldots, X_n$
- Decision of testing the hypothesis is whether to reject the null hypothesis or not to reject the null hypothesis
- Basis - from the value of a **statistic** T, which is calculated from the sample, called the **test statistic**
- Whenever the **null hypothesis** is **rejected**, **alternative hypothesis** is **accepted**

Errors, Level and Power

	Null Rejected	Null not rejected
Null is true	**Type 1 error**	
Null is false		**Type 2 error**

Table: Type 1 and Type 2 error

- Probability of type 1 error - level of the test or significance level of the test, denoted by α
- Probability of type 2 error usually denoted by β
- 1-β is called the power of the test

Remarks on α and β

- We control the type 1 error by our choice of level of significance α
- We then try to maximize power $1 - \beta$
- We cannot reduce the chances of both type 1 and type 2 error simultaneously

Example

BCCI was accused of using a biased coin in the toss for cricket match. A test was thus performed to check whether the given coin is biased towards heads. The coin was tossed 10 times.

1. Observed result - HHTHHHHHTHH
2. Null Hypothesis $p_H = 0.5$
3. Alternative Hypothesis is $p_H > 0.5$

It was decided that the null will be rejected if the number of heads is greater than 7?

What is the level of the test?

What is the power of the test?

$0.05 \approx {}^{10}C_8 (0.5)^8 (0.5)^2 + {}^{10}C_9 (0.5)^9 (0.5) + {}^{10}C_{10} (0.5)^{10}$

Excel commands (for 2016 version) for building confidence interval for mean

- To find $\overline{X}$, AVERAGE function in excel
- To find s, STDEV.S function in excel
- To find the value of $z_{\frac{\alpha}{2}}$: =NORM.S.INV$(1 - \frac{\alpha}{2})$
- To find the value of $t_{n-1,\frac{\alpha}{2}}$: =TINV$(\alpha, n-1)$

Hypothesis Testing

- Null Hypothesis and Alternative Hypothesis
- Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- Alternative - Everything else comes into this hypothesis
- There should be no common possibility in null and alternative hypothesis
- There should be no possibility outside of H_0 and H_a

Hypothesis Testing

- Null Hypothesis and Alternative Hypothesis
- Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- Alternative - Everything else comes into this hypothesis
- There should be no common possibility in null and alternative hypothesis
- There should be no possibility outside of H_0 and H_a

H_0:

H_a or H_1:

Hypothesis Testing

- Null Hypothesis and Alternative Hypothesis
- Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- Alternative - Everything else comes into this hypothesis
- There should be no common possibility in null and alternative hypothesis
- There should be no possibility outside of H_0 and H_a
- Philosophy of testing - Innocent until proven guilty
- What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- Null Hypothesis and Alternative Hypothesis
- Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- Alternative - Everything else comes into this hypothesis
- There should be no common possibility in null and alternative hypothesis
- There should be no possibility outside of H_0 and H_a
- Philosophy of testing - Innocent until proven guilty
- What is meant by proven?? Depends on the criteria set by the decision maker.

Hypothesis Testing

- Null Hypothesis and Alternative Hypothesis
- Null - typically the status quo or the one that people put more faith on, or the one that is easier to describe
- Alternative - Everything else comes into this hypothesis
- There should be no common possibility in null and alternative hypothesis
- There should be no possibility outside of H_0 and H_a
- Philosophy of testing - Innocent until proven guilty
- What is meant by proven?? Depends on the criteria set by the decision maker.

Building Hypothesis - Examples

$H_0: w \leq 5$

$H_a: w > 5$

- Coin is biased towards heads
- Coin is not an unbiased coin
- Gravity Fitness Gym, Roorkee claims that if you join the gym, you will lose more than 5 kgs in one month on an average.
- BJP IT team claims that with the marketing strategy that they have adopted, more than 40% people who were voters of other parties will vote for BJP in the coming elections

$H_0: p \leq 0.4$

$H_a: p > 0.4$

Building Hypothesis - Example

There is a perception that TSW-IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that TSW-IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that TSW-IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

Building Hypothesis - Example

There is a perception that TSW-IITR students are made to work much harder than students from other similar programs. A recent study states that on average a student in such a program works for 18 hours per week with an SD of 4 hours per week

- What would be your hypothesis?
- What would be your test statistic?
- How to go about testing the hypothesis?

$H_0: A_w \leq 18$

$H_a: A_w > 18$

One tail versus two tail test

In one-tailed test, we can reject null hypothesis only on one side of the hypothesized value of population parameter

In two-tailed test, we can reject null hypothesis on either side of the hypothesized value of population parameter

P-value

- p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- If the p-value is low, the null hypothesis should go
- In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- *p*-value for the textbook ratings problem? How to make a decision??

P-value

- p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- If the p-value is low, the null hypothesis should go
- In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- *p*-value for the textbook ratings problem? How to make a decision??

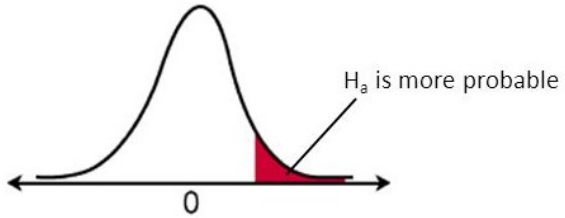
P-value

- p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- If the p-value is low, the null hypothesis should go
- In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- *p*-value for the textbook ratings problem? How to make a decision??

P-value

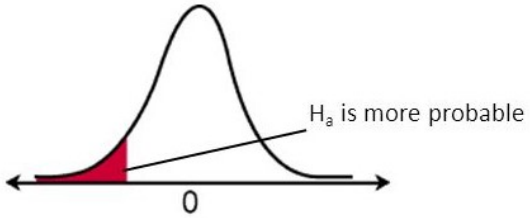
- p-value is the probability of obtaining a result as extreme as the one that was actually observed, under the assumption that null hypothesis is true
- If the p-value is low, the null hypothesis should go
- In the toss experiment, observed number of heads is 8. What is the p-value of this test statistic??
- *p*-value for the textbook ratings problem? How to make a decision??

One Tail and Two Tail Test



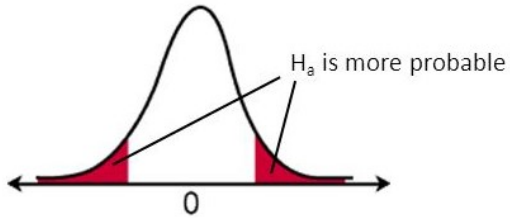
Right-tail test

$H_a: \mu > \text{value}$



Left-tail test

$H_a: \mu < \text{value}$



Two-tail test

$H_a: \mu \neq \text{value}$

Source - https://www.fromthegenesis.com/difference-between-one-tail-test-and-two-tail-test/

Hypothesis testing for population mean

- Case when σ is known
- Case when σ is unknown

Steps in Hypothesis Testing

1. Formulate the problem by clearly writing null hypothesis and alternative hypothesis
2. Decide on the level, α, for the test
3. Decide on the test statistic T. Calculate the value of the test statistic based on the sample
4. Find the p-value of the test statistic (or find the rejection criteria)
5. Give a verdict

ANOVA : Analysis of Variance