

Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

February 2, 2023

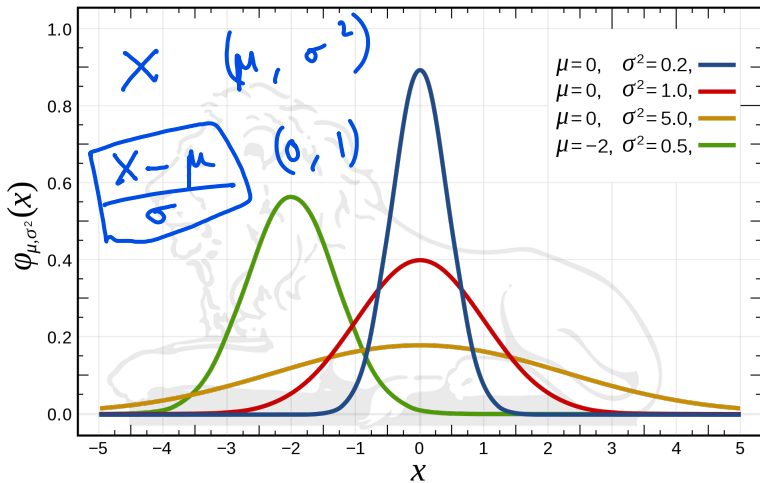


Basics about Random Variable

- ❑ 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



Chi-square distribution

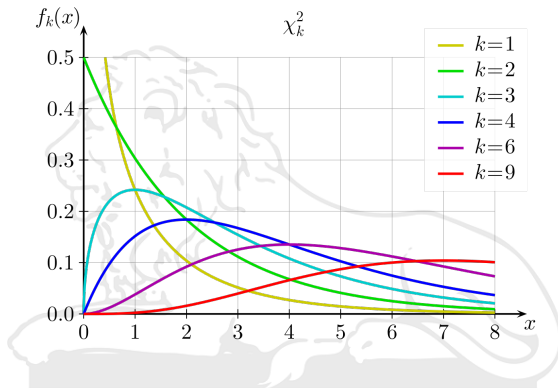
$$U_5 = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 + Z_5^2$$

If Z_1, Z_2, \dots, 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

$$E(U_5) = ?? \quad 5$$

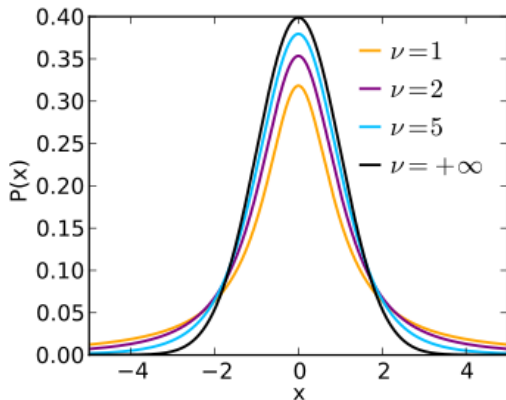
Chi-square distribution



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

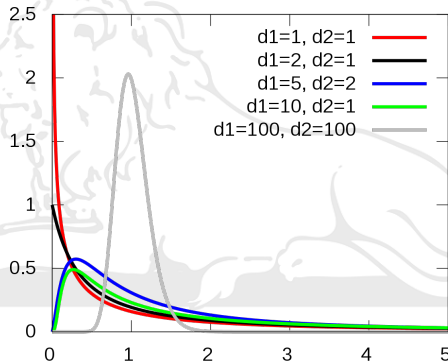
t-distribution

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



F distribution

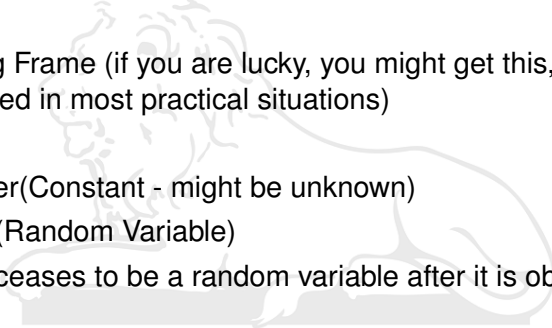
If $U \sim \chi_m^2$ and $V \sim \chi_n^2$ are independent, then the distribution of $\frac{U/m}{V/n}$ is called the F-distribution with m and n degrees of freedom and is denoted by $F_{m,n}$



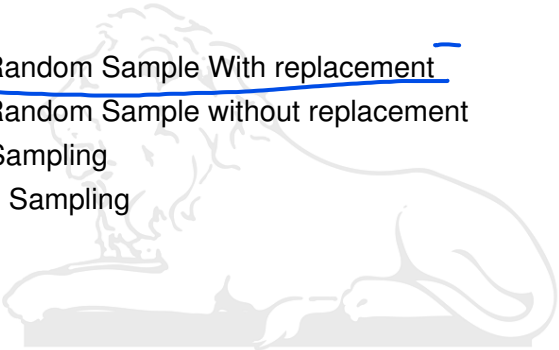
Basic Ideas of Sampling

- 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) ✓

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

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

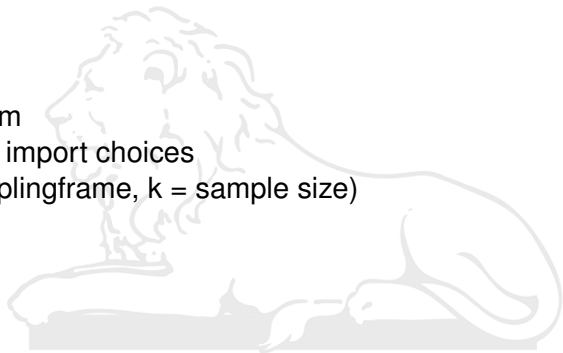
Potential Causes of Bias

- ☐ 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 Python in Presence of Sampling Frame

```
import random  
from random import choices  
choices(samplingframe, k = sample size)
```

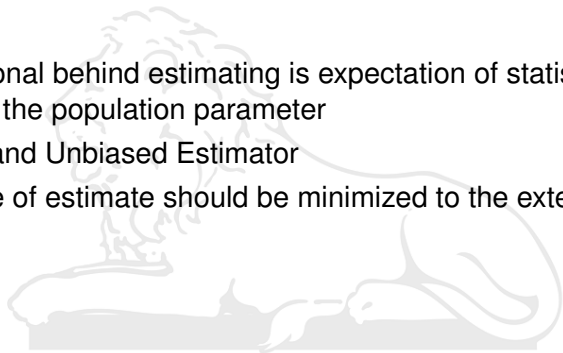


What after sampling?

- ❑ 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

- ❑ The rationale 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 Sample Mean

$$N = 500$$

$$n = 100$$

1. Let the true values in the population be $\bar{A}_1, \bar{A}_2, \bar{A}_3, \dots, \bar{A}_N$

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

$$\frac{\sum_{i=1}^N (A_i - \mu)^2}{N}$$

- with replacement -

4. Let the sample be a SRS of size n

5. Observations are X_1, X_2, \dots, X_n

6. Sample mean is denoted by \bar{X} and defined as follows

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

7. $E(\bar{X}) = \mu$

$$\begin{aligned} X_i &\rightarrow A_1 - \frac{1}{N} \\ &A_2 - \frac{1}{N} \\ &\vdots \\ &A_N - \frac{1}{N} \end{aligned}$$

$$\boxed{\text{Var}(X_i) = \sigma^2 \quad E(X_i) = \mu}$$

Estimating Population Variance from Sample Observations

- If the sampling scheme is WITH REPLACEMENT

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

- If the sampling scheme is WITHOUT REPLACEMENT

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

Why is estimating population variance important?

Estimating Population Variance from Sample Observations

- If the sampling scheme is WITH REPLACEMENT

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

- If the sampling scheme is WITHOUT REPLACEMENT

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

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

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

- If the sampling scheme is WITHOUT REPLACEMENT

$$\text{Var}(\bar{X}) = \left(\frac{N-n}{N-1} \right) \frac{\sigma^2}{n}$$

- $\frac{N-n}{N-1}$ is called the finite population correction
- Typically, can be ignored if sampling fraction $\frac{n}{N} \leq 0.05$
- Standard deviation of \bar{X} is called the Standard error of the sample mean
- Do we know σ^2 ?? What is the remedy??

Proof for $n - 1$ in the denominator

Proof of *sample variance is the unbiased estimator of population variance* (Sampling is done with replacement)

Let the true values in the population be $A_1, A_2, A_3, \dots, A_N$

Population mean is denoted by μ and equals $\frac{\sum_{i=1}^N A_i}{N}$

Also, population variance is denoted by σ^2 and equals $\frac{\sum_{i=1}^N (A_i - \mu)^2}{N}$

We will discuss the case when sampling is done WITH REPLACEMENT

Proof for $n - 1$ in the denominator

We now take a sample of size n (With replacement). Let the values that come in the sample are $X_1, X_2, X_3, \dots, X_n$

Our objective now is to use these n numbers to estimate population variance. The rational behind our approach is that the expectation of the estimate should be equal to the population variance.

We note that because the samples are taken with replacement, X_1, X_2, \dots, X_n are independent random variables.

We have seen that sample mean is an unbiased estimator of population mean. Sample mean is denoted by \bar{X} and defined as follows

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

We now claim that if we define the **sample variance** (denoted by s_X^2) as follows, it would be an unbiased estimator of population variance.

Proof for $n - 1$ in the denominator

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \text{ We need to show that } E(s_X^2) = \sigma^2$$

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

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

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

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

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

Proof for $n - 1$ in the denominator

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

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

$$\Rightarrow E(s_X^2) = \frac{\sum_{i=1}^n E(X_i^2) - nE(\bar{X}^2)}{n-1}$$

Since, X_i can take any values from A_1, A_2, \dots, A_N each with probability $\frac{1}{N}$, X_i^2 can take any values from $A_1^2, A_2^2, \dots, A_N^2$ each with probability $\frac{1}{N}$,

$$E(X_i^2) = \frac{A_1^2 + A_2^2 + \dots + A_N^2}{N}, \text{ for all } i.$$

Proof for $n - 1$ in the denominator

We can easily see the following result. As an exercise, it is recommended that you do this.

$$\sigma^2 = \frac{A_1^2 + A_2^2 + \cdots + A_N^2}{N} - \mu^2$$
$$\implies E(X_i^2) = \sigma^2 + \mu^2, \text{ for all } i.$$

$$\implies E(s_X^2) = \frac{n\sigma^2 + n\mu^2 - nE(\bar{X}^2)}{n-1} \quad (1)$$

We would now compute $E(\bar{X}^2)$ and substitute in the above equation (1) to see the result.

$$E(\bar{X}^2) = E\left(\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)^2\right)$$
$$\implies E(\bar{X}^2) =$$
$$E\left(\frac{X_1^2 + X_2^2 + \cdots + X_n^2 + 2X_1X_2 + 2X_1X_3 + \cdots + 2X_{n-1}X_n}{n^2}\right)$$

Proof for $n - 1$ in the denominator

$$\Rightarrow E(\bar{X}^2) = \frac{E(X_1^2) + E(X_2^2) + \cdots + E(X_n^2) + 2E(X_1X_2) + 2E(X_1X_3) + \cdots + 2E(X_{n-1}X_n)}{n^2}$$

$$\Rightarrow E(\bar{X}^2) = \frac{n\sigma^2 + n\mu^2 + 2E(X_1X_2) + 2E(X_1X_3) + \cdots + 2E(X_{n-1}X_n)}{n^2}$$

To deal with terms like $E(X_jX_k)$, we will use a result about expectations of product of independent random variables.

If X and Y are independent random variables, the following result holds.

$$E(XY) = E(X)E(Y)$$

Proof for $n - 1$ in the denominator

Hence, $E(X_j X_k) = E(X_j)E(X_k) = (\mu)(\mu) = \mu^2$, for all j and k (because X_j and X_k are independent random variables when the sampling scheme is with replacement)

Also, there are $\frac{n(n-1)}{2}$ such terms. Thus,

$$\Rightarrow E(\bar{X}^2) = \frac{n\sigma^2 + n\mu^2 + 2\frac{n(n-1)}{2}\mu^2}{n^2}$$

$$\Rightarrow E(\bar{X}^2) = \frac{\sigma^2 + \mu^2 + (n-1)\mu^2}{n}$$

We now substitute the value of $E(\bar{X}^2)$ in equation (1) to obtain-

$$\Rightarrow E(s_X^2) = \frac{n\sigma^2 + n\mu^2 - \sigma^2 - \mu^2 - (n-1)\mu^2}{n-1}$$

$$\Rightarrow E(s_X^2) = \frac{(n-1)\sigma^2}{n-1} = \sigma^2$$

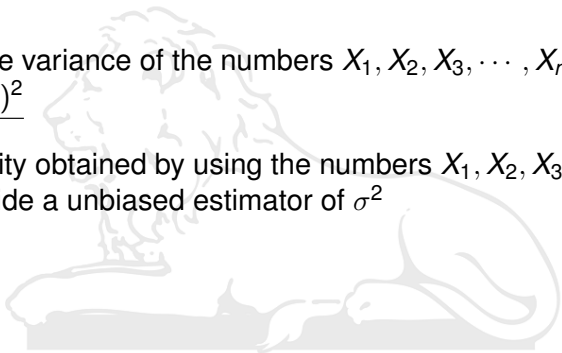
Thus, proved.

Remarks on the result

Remark - The variance of the numbers $X_1, X_2, X_3, \dots, X_n$ is equal to

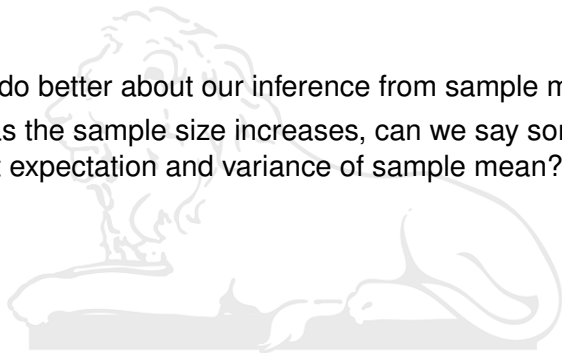
$$\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

s_X^2 is a quantity obtained by using the numbers $X_1, X_2, X_3, \dots, X_n$ in order to provide a unbiased estimator of σ^2



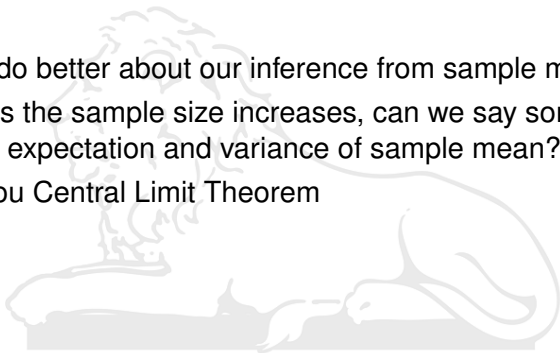
Central Limit Theorem

- ❑ 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

Theorem

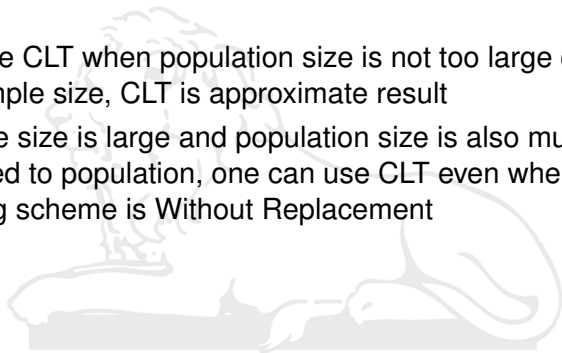
If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean \bar{X} is approximately normal with

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

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

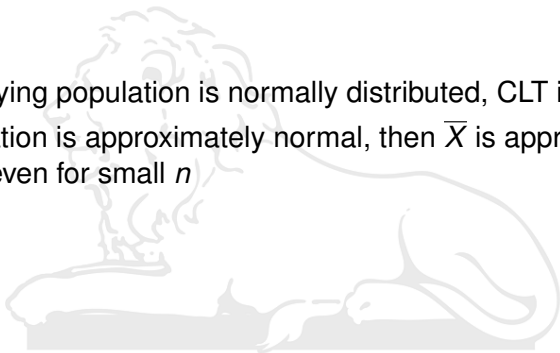
Comments about Central Limit Theorem

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



CLT when population is normally distributed

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



Confidence Interval Ideas



Sample Proportion

- ☐ Sometimes, one is interested in estimating population proportion
- ☐ What is the proportion of IBM-312 students participants who like statistics?
- ☐ One can attempt the answer to this using sampling

