INTRODUCTION TO DATA ANALYTICS

Class #9

Sampling Distributions

Dr. Sreeja S R

Assistant Professor

Indian Institute of Information Technology

IIIT Sri City

IN THIS PRESENTATION…

- Basic concept of sampling distribution
- Usage of sampling distributions
- Issue with sampling distributions
- Central limit theorem
- Application of Central limit theorem
- Major sampling distributions
  - **χ^2 distribution**
  - **t-distribution**
  - **F distribution**

Introduction

As a task of statistical inference, we usually follow the following steps:

- **Data collection**
  - Collect a sample from the population.
- **Statistics**
  - Compute a statistics from the sample.
- **Statistical inference**
  - From the statistics we made various statements concerning the values of population parameters can be inferred.
    - For example, population mean from the sample mean, etc.

Basic terminologies

Some basic terminology which are closely associated to the above-mentioned tasks are reproduced below.

- **Population:** A population consists of the totality of the observation, with which we are concerned.
- **Sample:** A sample is a subset of a population.
- **Random variable:** A random variable is a function that associates a real number with each element in the sample.
- **Statistics:** Any function of the random variable constituting random sample is called a statistics.
- **Statistical inference:** It is an analysis basically concerned with generalization and prediction.

Basic terminologies

Probability distribution: A function that shows the probabilities of the outcomes of an event or experiment.

Normal (Gaussian) distribution: A probability distribution that looks like a bell. Two terms that describe a normal distribution are **mean** and **standard deviation**. Mean is the average value that has the highest probability to be observed. **Standard deviation** is a measure of how spread out the values are. As standard deviation increases, the normal distribution curve gets wider.

Statistical Inference

There are two facts, which are key to statistical inference.

1. Population parameters are fixed number whose values are usually unknown.
2. Sample statistics are known values for any given sample, but vary from sample to sample, even taken from the same population.

- In fact, it is unlikely for any two samples drawn independently, producing identical values of sample statistics.
- In other words, the variability of sample statistics is always present and must be accounted for in any inferential procedure.
- This variability is called sampling variation.

Note:

A sample statistics is random variable and like any other random variable, a sample statistics has a probability distribution.

Sampling Distribution

More precisely, sampling distributions are probability distributions and used to describe the variability of sample statistics.

Definition 7.1: Sampling distribution

The sampling distribution of a statistics is the probability distribution of that statistics.

- The probability distribution of sample mean (hereafter, will be denoted as $\bar{X}$) is called the sampling distribution of the mean (also, referred to as the distribution of sample mean).
- Like $\bar{X}$, we call sampling distribution of variance (denoted as S^2).
- Using the values of $\bar{X}$ and S^2 for different random samples of a population, we are to make inference on the parameters μ and σ^2 (of the population).

Sampling Distribution

Example 7.1:

Consider five identical balls numbered and weighing as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls.

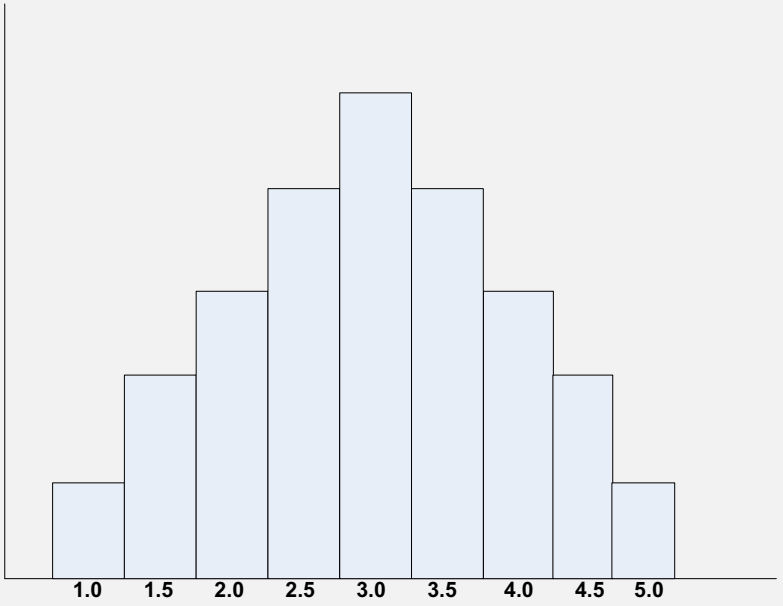
Following table lists all possible samples and their mean.

Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)	Sample (X)	Mean ($\bar{X}$)
[1,1]	1.0	[2,4]	3.0	[4,2]	3.0
[1,2]	1.5	[2,5]	3.5	[4,3]	3.5
[1,3]	2.0	[3,1]	2.0	[4,4]	4.0
[1,4]	2.5	[3,2]	2.5	[4,5]	4.5
[1,5]	3.0	[3,3]	3.0	[5,1]	3.0
[2,1]	1.5	[3,4]	3.5	[5,2]	3.5
[2,2]	2.0	[3,5]	4.0	[5,3]	4.0
[2,3]	2.5	[4,1]	2.5	[5,4]	4.5
				[5,5]	5.0

Sampling Distribution

Sampling distribution of means

$\bar{X}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0
$f(\bar{X})$	$\frac{1}{25}$	$\frac{2}{25}$	$\frac{3}{25}$	$\frac{4}{25}$	$\frac{5}{25}$	$\frac{4}{25}$	$\frac{3}{25}$	$\frac{2}{25}$	$\frac{1}{25}$



Issues with Sampling Distribution

1. In practical situation, for a large population, it is infeasible to have all possible samples and hence probability distribution of sample statistics.

2. The sampling distribution of a statistics depends on

   - the size of the population
   - the size of the samples and
   - the method of choosing the samples.

Theorem on Sampling Distribution

Famous theorem in Statistics

Theorem 7.1: **Sampling distribution of mean and variance**

The sampling distribution of a random sample of size *n* drawn from a population with mean μ and variance σ^2 will have mean $\bar{X} = \mu$ and variance $S^2 = \frac{\sigma^2}{n}$

Example 7.2: Consider the following small population consisting of N=6 patients who recently underwent total hip replacement. Three months after surgery they rated their pain-free function on a scale of 0 to 100 (0=severely limited and painful functioning to 100=completely pain free functioning). The data are shown below and ordered from smallest to largest.

Pain-Free Function Ratings in a Small Population of N=6 Patients:

25, 50, 80, 85, 90, 100

Example 7.2: 25, 50, 80, 85, 90, 100 For the population, $\mu = 71.4$ $\sigma = 28.40$

Suppose we did not have the population data and instead we were estimating the mean functioning score in the population based on a sample of n=4. The table below shows all possible samples of size n=4 from the population of N=6. The rightmost column shows the sample mean based on the 4 observations contained in that sample.

Sample	Observations in the Sample (n=4)				Mean
1	25	50	80	85	60.0
2	25	50	80	90	61.3
3	25	50	80	100	63.6
4	25	50	85	90	62.5
5	25	50	85	100	65.0
6	25	59	90	100	66.3
7	25	80	85	90	70.0
8	25	80	85	100	72.5
9	25	80	90	100	73.8
10	25	85	90	100	75.0
11	50	80	85	90	76.3
12	50	80	85	100	78.8
13	50	80	90	100	80.0
14	50	85	90	100	81.3
15	80	85	90	100	88.8

From the table,

$$\bar{X} = 71.4$$

Central Limit Theorem

The Theorem 7.1 is an amazing result and in fact, also verified that if we sampling from a population with unknown distribution, the sampling distribution of $\bar{X}$ will still be approximately normal with mean μ and variance $\frac{\sigma^2}{n}$ provided that the sample size is large.

This further, can be established with the famous "central limit theorem", which is stated below.

Theorem 7.3: Central Limit Theorem

If $\bar{X}$ is the mean of a random sample of size n taken from a population having the mean μ and the finite variance σ^2, then

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

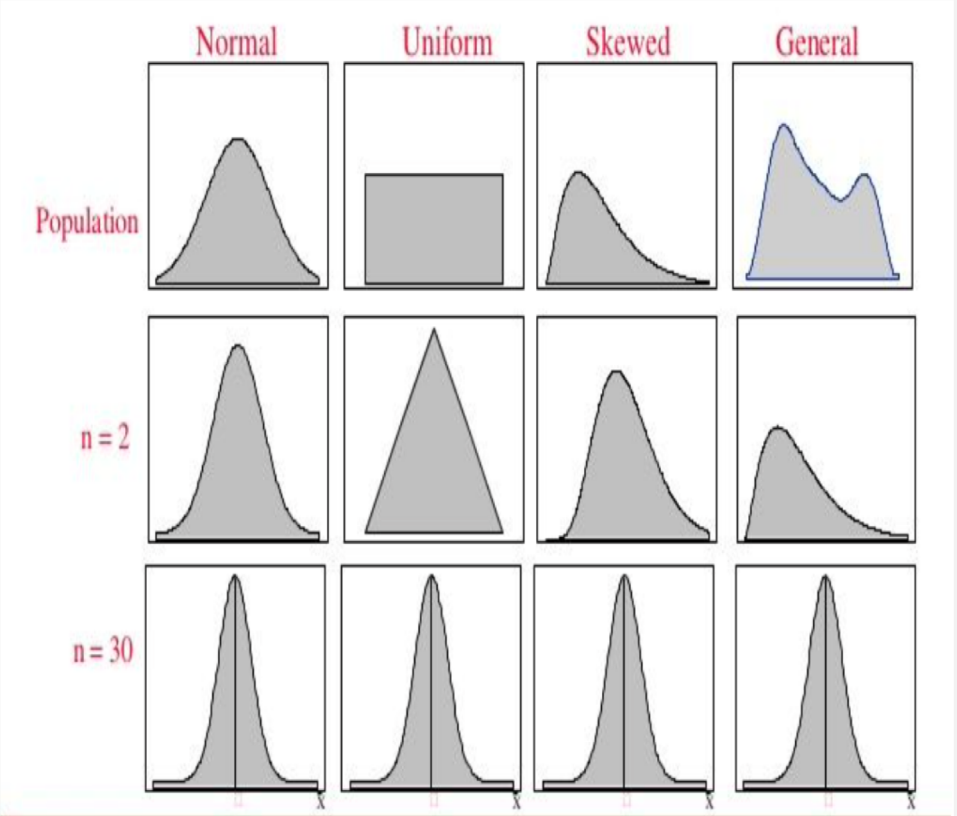
is a random variable whose distribution function approaches that of the standard normal distribution as $n \to \infty$

Central Limit Theorem

CLT states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger – no matter what the shape of the population distribution. This fact holds especially true for sample sizes over 30.

Why is it so important to have a normal distribution?

Normal distribution is described in terms of mean and standard deviation which can easily be calculated. And, if we know the mean and standard deviation of a normal distribution, we can compute pretty much everything about it.



Example for central limit theorem:

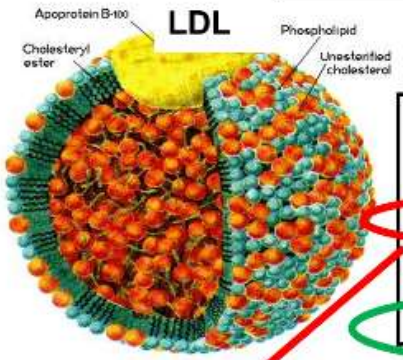
Different classes of these lipid transport carriers can be separated (fractionated)based on their density and where they layer out when spun in a centrifuge. High density lipoprotein cholesterol (HDL) is sometimes referred to as the "good cholesterol," because higher concentrations of HDL in blood are associated with a lower risk of coronary heart disease. In contrast, high concentrations of low density lipoprotein cholesterol (LDL) are associated with an increased risk of coronary heart disease. The illustration on the right outlines how total cholesterol levels are classified in terms of risk, and how the levels of LDL and HDL fractions provide additional information regarding risk.

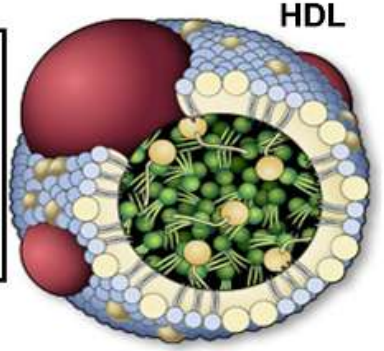
Total Cholesterol (mg/deciliter)

- Desirable: <200
- Borderline High: 200-239
- High: 240+

Total Fractionated into Lipoprotein Carriers

- Chylomicrons
- VLDL (very low density)
- LDL (low density)
- IDL (intermediate density)
- HDL (high density)





LDL (mg/deciliter)	
<100	Optimal
100-129	Above optimal
130-159	Borderline high
160-189	High
>190	Very high

HDL Concentration (mg/deciliter)	
<40	Low (High risk)
40-50	Average risk
60+	High (Low risk)

Example for central limit theorem:

Data from the Framingham Heart Study found that subjects over age 50 had a mean HDL of 54 and a standard deviation of 17. Suppose a physician has 40 patients over age 50 and wants to determine the probability that the mean HDL cholesterol for this sample of 40 men is 60 mg/dl or more (i.e., low risk).

- Probability questions about a sample mean can be addressed with the Central Limit Theorem, as long as the sample size is sufficiently large.
- In this case n=40, so the sample mean is likely to be approximately normally distributed, so we can compute the probability of HDL>60 by using the standard normal distribution table.
- The population mean is 54, but the question is what is the probability that the sample mean will be >60?

Solution:

$\overline{X}$ **= 60,** μ **= 54,** σ **= 17,** n **= 40.**

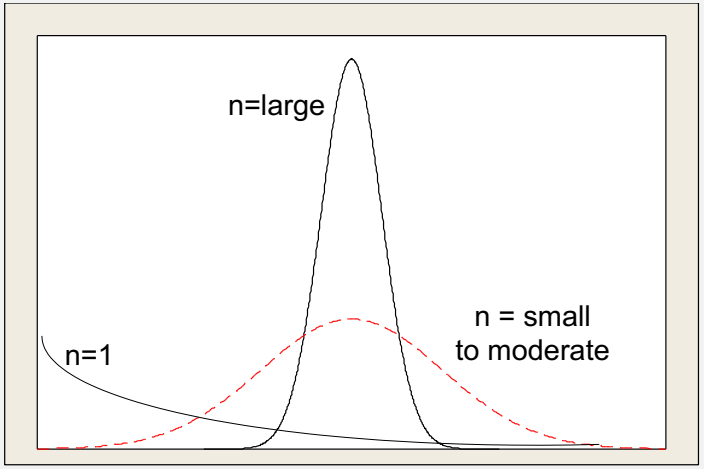
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{60 - 54}{17/\sqrt{40}} = 2.22$$

P(Z > 2.22) = 1 - 0.9868 = 0.0132.

Therefore, the probability that the mean HDL in these 40 patients will exceed 60 is 1.32%.

Applicability of Central Limit Theorem

- The normal approximation of $\overline{X}$ will generally be good if $n \geq 30$
- The sample size $n = 30$ is a guideline for the central limit theorem.
- The normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



One very important application of the **Central Limit Theorem** is the determination of reasonable values of the population mean μ and variance σ^2.

STANDARD SAMPLING DISTRIBUTIONS

- Apart from the normal distribution to describe sampling distribution, there are some other quite different sampling, which are extensively referred in the study of statistical inference.

  - t: Describes the sampling distribution of the mean when σ is unknown
  - χ^2: Describes the distribution of variance.
  - F: Describes the distribution of the ratio of two variables.

The t Distribution

1. To know the sampling distribution of mean we make use of Central Limit Theorem with $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$

2. Central Limit Theorem require the known value of σ a priori.

3. However, in many situation, σ is certainly no more reasonable than the knowledge of the population mean μ.

4. In such situation, only measure of the standard deviation available may be the sample standard deviation S.

5. It is natural then to substitute S for σ. The problem is that the resulting statistics is not normally distributed!

6. The t distribution is to alleviate this problem. This distribution is called $student's\ t$ or simply $t-distribution$.

The t Distribution

Definition 7.4: t −distribution

If $\bar{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, then

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with the parameter $v = n - 1$

Example for t-distribution:

A manufacturer of fuses claims that with a 20% overload, the fuses will blow in 12.40 minute on the average. To test this claim, a sample of 20 of the fuses was subjected to a 20% overload, and the time it took them to blow had a mean of 10.63 minutes and a std. dev. of 2.48 minutes. If it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim?

Solution:

$\overline{X}$ **= 10.63,** μ **= 12.40,** S **= 2.48,** n **= 20.**

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{10.63 - 12.40}{2.48/\sqrt{20}} = -3.19$$

is a random variable having the t distribution with the parameter $v = n-1 = 19$ degrees of freedom. From the t-distribution table, for t = -3.19 and v =19, we see that a t value of 2.861 already has only 0.5% probability (the probability is 0.005). Since the probability is very small, we conclude that the data refute the manufacturer's claim. In all likelihood, the mean blowing time of his fuses with a 20% overload is less than 12.40 minutes.

THE χ^2 DISTRIBUTION

- A common use of the χ^2 distribution is to describe the distribution of the sample variance.
- It is concerned with the sampling distribution of the sample variance for random samples from normal populations.

Definition 7.5: χ^2 −distribution

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then

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

is a random variable having the chi square distribution with the parameter $v = n - 1$

The F Distribution

- The F distribution finds enormous applications in comparing sample variances.

Definition 7.5: *F* distribution

If ${S_1}^2$ and ${S_2}^2$ are the variances of independent random samples of size n_1 and n_2, respectively, taken from two normal populations having the same variance, then

$$F = \frac{{S_1}^2}{{S_2}^2}$$

is a random variable having the F distribution with the parameter $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$

Therefore, if we assume that we have sample of size n_1 from a population with variance ${\sigma_1}^2$ and an independent sample of size n_2 from another population with variance ${\sigma_2}^2$, then the statistics

$$F = \frac{{S_1}^2/{\sigma_1}^2}{{S_2}^2/{\sigma_2}^2}$$

Representation of χ^2 random variable

Definition 7.6: χ^2 random variable

Let $Z_1, Z_2 \ldots \ldots \ldots \ldots Z_n$ be independent standard normal random variables.

$$\chi_v^2 = \sum_{i=1}^{v} Z_i^2$$

has a *chi square* distribution with v degrees of freedom.

Representation of t random variable

Definition 7.7: t random variable

Let the standard normal Z and χ^2 with v degrees of freedom be independent.

$$t = \frac{standard\ normal}{\sqrt{\frac{chi\ square}{degress\ of\ freedom}}} = \frac{Z}{\sqrt{\frac{\chi^2}{v}}}$$

has a t distribution with v degrees of freedom.

Representation of *F* random variable

Definition 7.8: *F* random variable

Let the chi square variables χ_1^2, with v_1 degrees of freedom, and χ_2^2, with v_2 degrees of freedom, be independent.

$$\mathsf{F}(v_1, v_2) = \frac{\chi_1^2 / v_1}{\chi_2^2 / v_2}$$

has a *F* distribution with (v_1, v_2) degrees of freedom.

REFERENCE

- The detail material related to this lecture can be found in

Probability and Statistics for Enginneers and Scientists (8th Ed.) by Ronald E. Walpole, Sharon L. Myers, Keying Ye (Pearson), 2013.

Any question?

QUESTIONS OF THE DAY…

1. What are the degrees of freedom in the following cases.

   Case 1: A single number.

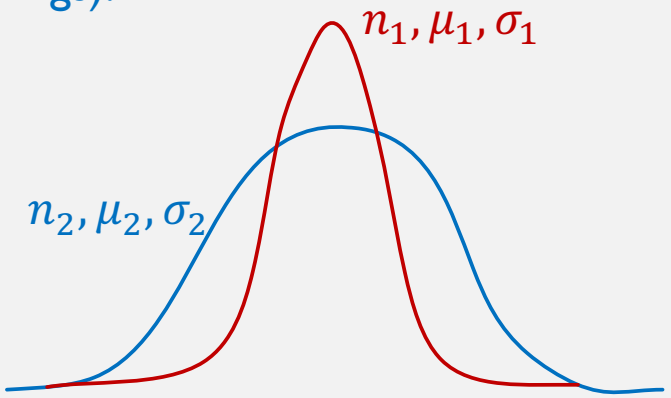
   Case 2: A list of n numbers.

   Case 3: a table of data with m rows and n columns.

   Case 4: a data cube with dimension $m \times n \times p$.

QUESTIONS OF THE DAY…

2. In the following, two normal sampling distributions are shown with parameters n, μ and σ (all symbols bear their usual meanings).



What are the relations among the parameters in the two?

QUESTIONS OF THE DAY…

3. Suppose, $\bar{X}$ and S denote the sample mean and standard deviation of a sample. Assume that population follows normal distribution with population mean μ and standard deviation σ. Write down the expression of z and t values with degree of freedom n.