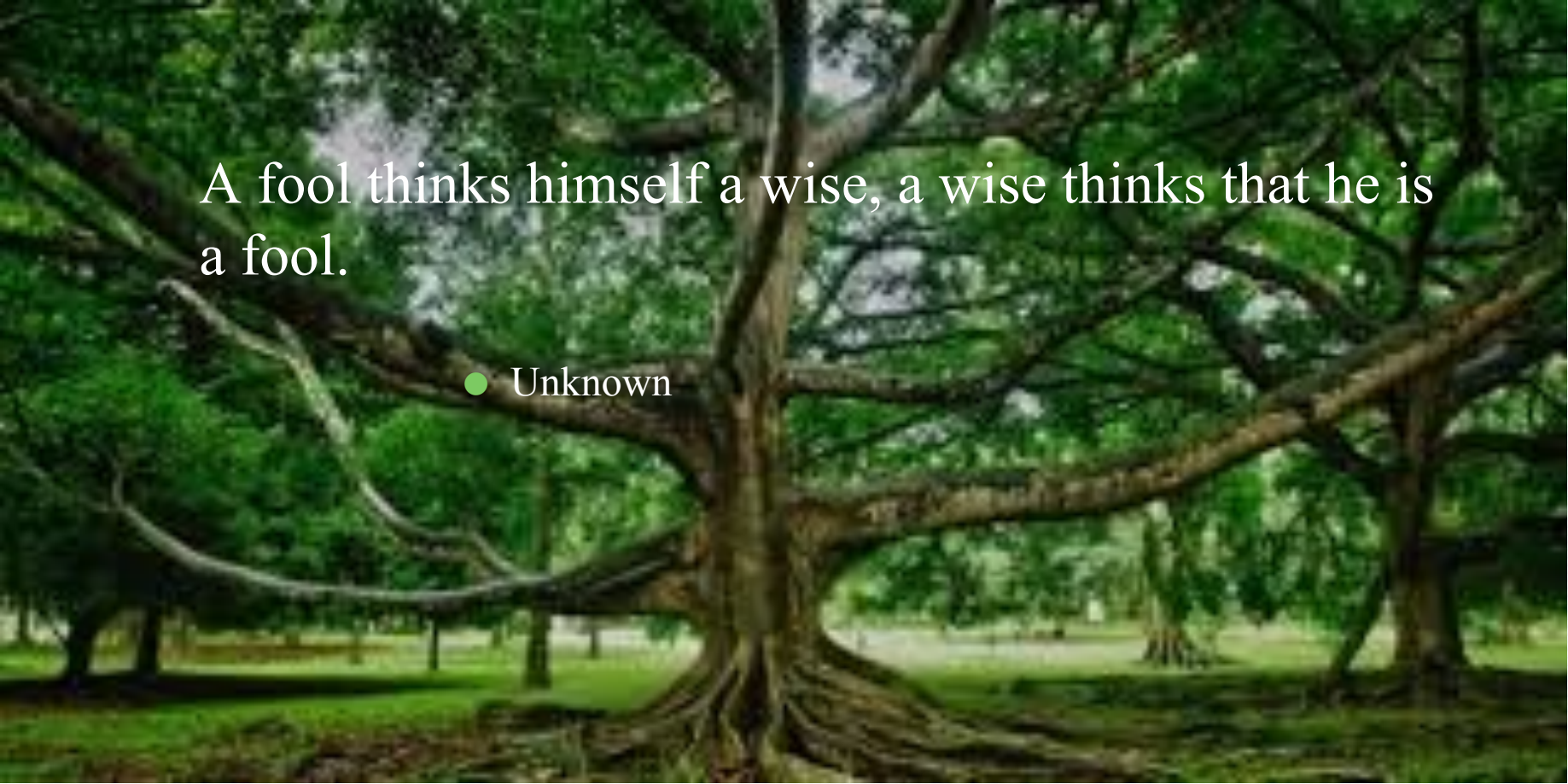


Sampling Distributions

Quote of the day..



A fool thinks himself a wise, a wise thinks that he is a fool.

● Unknown

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.
 - 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.

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.

Why probability distribution for random variable is not applicable to sample statistics?

Sampling Distribution

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

Definition 5.1: 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 5.1:

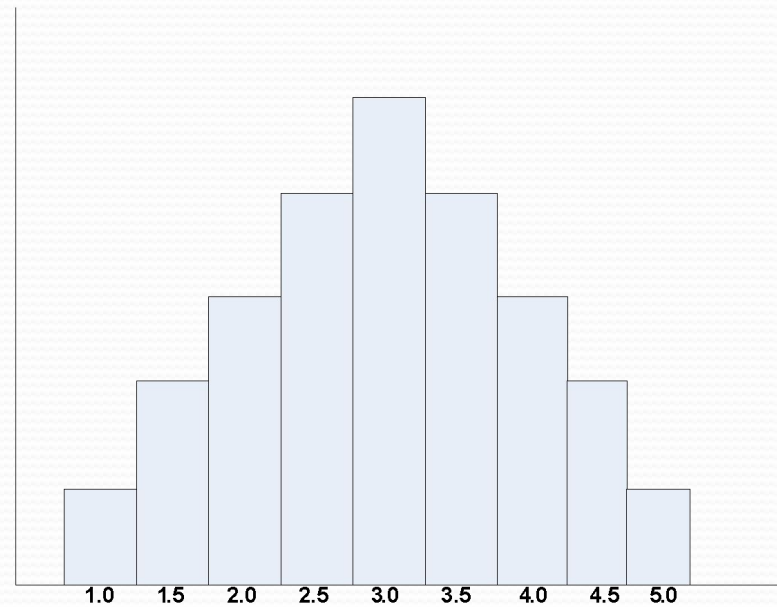
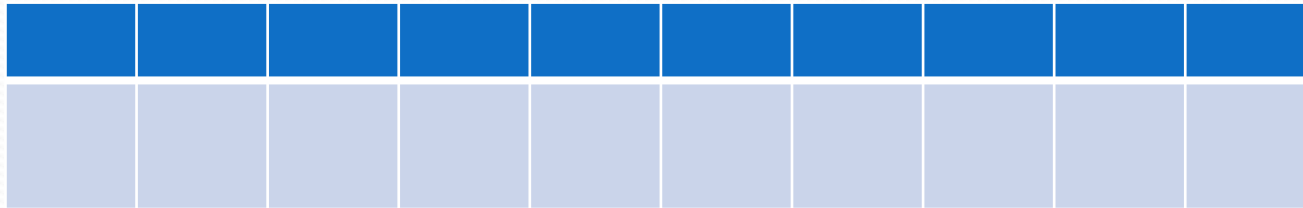
Consider five identical balls numbered and weighting as 1, 2, 3, 4 and 5. Consider an experiment consisting of drawing two balls, replacing the first before drawing the second, and then computing the mean of the values of the two balls.

Following table lists all possible samples and their mean.

[1,1]		[2,4]		[4,2]	

Sampling Distribution

Sampling distribution of means



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 5.1: 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 5.2: With reference to data in Example 5.1

For the population, $\mu = \frac{1+2+3+4+5}{5} = 3$

$$\sigma^2 = \frac{(25-1)}{12} = 2$$

Applying the theorem, we have $\bar{X} = 3$ and $S^2 = 1$

Hence, the theorem is verified!

Central Limit Theorem

- The Theorem 5.1 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.

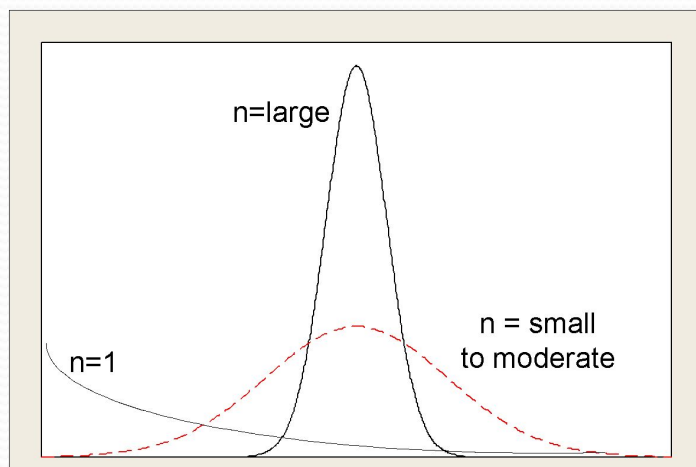
Theorem 5.3: Central Limit Theorem

If random samples each of size n are taken from any distribution with mean μ and variance σ^2 , the sample mean \bar{X} will have a distribution approximately normal with mean μ and variance $\frac{\sigma^2}{n}$.

The approximation becomes better as n increases.

Applicability of Central Limit Theorem

- The normal approximation of \bar{X} will generally be good if $n \geq 30$
- The sample size $n = 30$ is, hence, a guideline for the central limit theorem.
- The normality on the distribution of \bar{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 .

For standard normal distribution, we have the z-transformation

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

Extension

Theorem 5.2: Reproductive property of normal distribution

If X_1, X_2, \dots, X_n are independent random variables, having **normal distribution** with mean $\mu_1, \mu_2, \dots, \mu_n$ and variance $\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2$ then the random variable $\bar{X} = a_1X_1 + a_2X_2 + \dots + a_nX_n$ has uniform distribution with mean, $\mu_{\bar{X}} = a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n$
variance $\sigma_{\bar{X}}^2 = a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2$

Note:

If all samples X_1, X_2, \dots, X_n are uniformly distributed then $a_1 = a_2 = \dots = \frac{1}{n}$

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.
 - χ^2 : Describes the distribution of variance.
 - t : Describes the distribution of normally distributed random variable standardized by an estimate of the standard deviation.
 - F : Describes the distribution of the ratio of two variables.

The χ^2 Distribution

- ▲ common use of the χ^2 distribution is to describe the distribution of the sample variance. In order to arrive into a deduction for χ^2 distribution for a sample variance, we rely on the following theorems, whose proof can be available in any book on Statistics.

Theorem 5.4: Linear combination of random variable

If X_1, X_2, \dots, X_n are mutually independent random variables that have, respectively Chi-squared distribution with v_1, v_2, \dots, v_n degrees of freedom, then the random variable.

$$Y = X_1 + X_2 + \dots + X_n$$

has a Chi squared distribution with $v_1 + v_2 + \dots + v_n$ degrees of freedom.

The χ^2 Distribution

An important corollary of the Theorem 5.4 is stated below.

Corollary 5.1: Reference Theorem 5.4

If x_1, x_2, \dots, x_n are independent random variables having identical normal distribution with mean μ and variance σ^2 , then the random variable

$$Y = \sum_{i=1}^n \left(\frac{x_i - \mu}{\sigma} \right)^2$$

has a Chi squared distribution with $n-1$ degrees of freedom

The χ^2 Distribution

Note:

- Each of the n independent random variable $\left(\frac{x_i - \mu}{\sigma}\right)^2, i = 1, 2, 3, \dots, n$ has Chi-squared distribution with 1 degree of freedom.

Now we can derive χ^2 - distribution for sample variance.

We can write

$$\begin{aligned}\sum_{i=1}^n (x_i - \mu)^2 &= \sum_{i=1}^n [(x_i - \bar{x}) + (\bar{x} - \mu)]^2 \\ &= \sum_{i=1}^n (x_i - \bar{x})^2 + n \cdot (\bar{x} - \mu)^2\end{aligned}$$

or	$\frac{1}{\sigma^2} \sum (x_i - \mu)^2$	=	$\frac{(n-1)S^2}{\sigma^2}$	+	$\frac{(\bar{x} - \mu)^2}{\sigma^2/n}$
Chi-square distribution with n-degree			Chi-square distribution with (n-1) degree of freedom		Chi-square distribution with 1 degree of freedom [= Z^2]

The χ^2 Distribution

Definition 5.2: χ^2 -distribution for Sampling Variance

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

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

Has a chi-squared distribution with $\nu = n - 1$ degrees of freedom

- This way χ^2 -distribution is used to describe the sampling distribution of S^2 .

The χ^2 Distribution

Definition 5.3: χ^2 -distribution for sampling variance

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

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\sigma} \right)^2$$

has a Chi-squared distribution with $\nu = n - 1$ degrees of freedom.

This way, χ^2 -distribution is used to describe the sampling distribution of S^2

The t Distribution

● 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. This 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

● The t Distribution

Definition 5.4: t –distribution

The t –distribution with ν degrees of freedom actually takes the form

$$t(\nu) = \frac{Z}{\sqrt{\frac{\chi^2(\nu)}{\nu}}}$$

where Z is a standard normal random variable, and $\chi^2(\nu)$ is χ^2 random variable with ν degrees of freedom.

The t Distribution

Corollary: Let X_1, X_2, \dots, X_n be independent random variables that are all normal with mean μ and standard deviation σ .

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

Using this definition, we can develop the sampling distribution of the sample mean when the population variance, σ^2 is unknown.

That is,

$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has the standard normal distribution.

$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ has the χ^2 distribution with $(n-1)$ degrees of freedom.

$$\text{Thus, } T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} \quad \text{or}$$

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

This is the t – *distribution* with $(n-1)$ degrees of freedom.

The F Distribution

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

Definition 5.5: F distribution

The statistics F is defined to be the ratio of two independent Chi-Squared random variables, each divided by its number of degrees of freedom. Hence,

$$F(v_1, v_2) = \frac{\chi^2(v_1)/v_1}{\chi^2(v_2)/v_2}$$

Corollary: Recall that $\chi^2 = \frac{(n-1)S^2}{\sigma^2}$ is the Chi-squared distribution with $(n - 1)$ degrees of freedom.

Therefore, if we assume that we have sample of size n_1 from a population with variance σ_1^2 and an independent sample of size n_2 from another population with variance σ_2^2 , then the statistics

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