3.1 Sampling Distributions

datamingtools

Contents:

Sampling Distributions

1. Populations and Samples
2. The Sampling Distribution of the mean (σ known)
3. The Sampling Distribution of the mean (σ unknown)
4. The Sampling Distribution of the Variance

Populations and Samples

Population

Finite Population

Infinite Population

datamingtools

Populations and Samples

Population: A set or collection of all the objects, actual or conceptual and mainly the set of numbers, measurements or observations which are under investigation.

Finite Population : All students in a College

Infinite Population : Total water in the sea or all the sand particle in sea shore.

Populations are often described by the distributions of their values, and it is common practice to refer to a population in terms of its distribution.

Finite Populations

- Finite populations are described by the actual distribution of its values and infinite populations are described by corresponding probability distribution or probability density.
- **"Population $f(x)$"** means a population is described by a frequency distribution, a probability distribution or a density $f(x)$.

Infinite Population

If a population is infinite it is impossible to observe all its values, and even if it is finite it may be impractical or uneconomical to observe it in its entirety. Thus it is necessary to use a **sample**.

Sample: A part of population collected for investigation which needed to be representative of population and to be large enough to contain all information about population.

datamingtools

Random Sample (finite population):

- A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from a finite population of size N, if its values are chosen so that each subset of n of the N elements of the population has the same probability of being selected.

datamingtools

Random Sample (infinite Population):

A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from the infinite population $f(x)$ if:

1. Each X_i is a random variable whose distribution is given by $f(x)$
2. These n random variables are independent.

We consider two types of random sample: those drawn *with replacement* and those drawn *without replacement*.

datamingtools

Sampling with replacement:

- In sampling with replacement, each object chosen is returned to the population before the next object is drawn. We define a random sample of size *n* drawn with replacement, *as* an ordered *n*-tuple of objects from the population, repetitions allowed.

The Space of random samples drawn with replacement:

- If samples of size n are drawn with replacement from a population of size N, then there are N^n such samples. In any survey involving sample of size n, each of these should have same probability of being chosen. This is equivalent to making a collection of all N^n samples a probability space in which each sample has probability of being chosen $1/N^n$.
- Hence in above example There are $3^2 = 9$ random samples of size 2 and each of the 9 random sample has probability 1/9 of being chosen.

datamingtools

Sampling without replacement:

In sampling without replacement, an object chosen is not returned to the population before the next object is drawn. We define random sample of size n, drawn without replacement as an unordered subset of n objects from the population.

The Space of random samples drawn without replacement:

- If sample of size n are drawn without replacement from a population of size N, then there are $\binom{N}{n}$ such samples. The collection of all random samples drawn without replacement can be made into a probability space in which each sample has same chance of being selected.

datamingtools

Mean and Variance

If $X_1, X_2, \ldots, X_n$ constitute a random sample, then

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

is called the **sample mean** and

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

is called the **sample variance.**

Sampling distribution:

- The probability distribution of a random variable defined on a space of random samples is called a *sampling distribution*.

datamingtools

The Sampling Distribution of the Mean (σ Known)

Suppose that a random sample of n observations has been taken from some population and $\bar{x}$ has been computed, say, to estimate the mean of the population. If we take a second sample of size n from this population we get some different value for $\bar{x}$. Similarly if we take several more samples and calculate $\bar{x}$, probably no two of the $\bar{x}$'s would be alike. The difference among such $\bar{x}$'s are generally attributed to chance and this raises important question concerning their distribution, specially concerning the extent of their chance of fluctuations.

Let $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}}$ be mean and variance for sampling *distributi on* of the mean $\overline{X}$.

The Sampling Distribution of the Mean (σ Known)

Formula for $\mu_{\bar{X}}$ and $\sigma^2_{\bar{X}}$:

Theorem 1: If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\bar{X}$ is a random variable whose distribution has the mean μ.

For samples from infinite populations the variance of this distribution is $\frac{\sigma^2}{n}$.

For samples from a finite population of size N the variance is $\frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}$.

That is $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \begin{cases} \frac{\sigma^2}{n} & \text{(for infinite Population)} \\ \frac{\sigma^2}{n}\frac{N-n}{N-1} & \text{(for finite Population)} \end{cases}$

The Sampling Distribution of the Mean (σ Known)

Proof of $\mu_{\bar{X}} = \mu$ for infinite population for the continuous

case: From the definition we have

$$\mu_{\bar{X}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \bar{x}\, f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\ldots dx_n$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \sum_{i=1}^{n} \frac{x_i}{n} f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\ldots dx_n$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} x_i\, f(x_1, x_2, \ldots, x_n)\,dx_1\,dx_2\ldots dx_n$$

The Sampling Distribution of the Mean (σ Known

Where $f(x_1, x_2, \ldots, x_n)$ is the joint density function of the random variables which constitute the random sample. From the assumption of random sample (for infinite population) each X_i is a random variable whose density function is given by $f(x)$ and these n random variables are independent, we can write

$$f(x_1, x_2, \ldots, x_n) = f(x_1)\, f(x_2) \ldots\ldots f(x_n)$$

and we have

$$\mu_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \ldots\ldots \int_{-\infty}^{\infty} x_i f(x_1) f(x_2) \ldots\ldots f(x_n) dx_1 dx_2 \ldots\ldots dx_n$$

$$= \frac{1}{n}\sum_{i=1}^{n} \int_{-\infty}^{\infty} f(x_1) dx_1 \ldots \int_{-\infty}^{\infty} x_i f(x_i) dx_i \ldots\ldots \int_{-\infty}^{\infty} f(x_n) dx_n$$

The Sampling Distribution of the Mean (σ Known)

Since each integral except the one with the integrand $x_i f(x_i)$ equals 1 and the one with the integrand $x_i f(x_i)$ equals to μ, so we will have

$$\mu_{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \mu = \frac{1}{n} n\mu = \mu .$$

Note: for the discrete case the proof follows the same steps, with integral sign replaced by $\sum$'s.

For the proof of $\sigma_{\bar{X}}^2 = \sigma^2 / n$ we require the following result

Result: If X is a continuous random variable and $Y = X - \mu_X$, then $\mu_Y = 0$ and hence $\sigma_Y^2 = \sigma_X^2$.

Proof: $\mu_Y = E(Y) = E(X - \mu_X) = E(X) - \mu_X = 0$ and

$\sigma_Y^2 = E[(Y - \mu_Y)^2] = E[((X - \mu_X) - 0)^2] = E[(X - \mu_X)^2] = \sigma_X^2$.

The Sampling Distribution of the Mean (σ Known)

- Regardless of the form of the population distribution, the distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n whenever n is large.
- In practice, **the normal distribution provides an excellent approximation to the sampling distribution of the mean for n as small as 25 or 30**, with hardly any restrictions on the shape of the population.
- **If the random samples come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample**.

datamingtools

The Sampling Distribution of the mean (σ unknown)

- Application of the theory of previous section requires knowledge of the population standard deviation σ.
- If n is large, this does not pose any problems even when σ is unknown, as it is reasonable in that case to use for it the sample standard deviation s.
- However, when it comes to random variable whose values are given by

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

very little is known about its exact sampling distribution for small values of n unless we make the assumption that the sample comes from a normal population.

The Sampling Distribution of the mean (σ unknown)

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

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

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

is a random variable having the t distribution with the parameter $\nu = n - 1$.

This theorem is more general than Theorem 6.2 in the sense that it does not require knowledge of σ; on the other hand, it is less general than Theorem 6.2 in the sense that it requires the assumption of a normal population.

The Sampling Distribution of the mean (σ unknown)

- The *t* distribution was introduced by William S.Gosset in 1908, who published his scientific paper under the pen name "Student," since his company did not permit publication by employees. That's why *t* distribution is also known as the **Student-*t* distribution**, or **Student's *t* distribution**.
- The shape of ***t* distribution** is similar to that of a normal distribution i.e. both are bell-shaped and symmetric about the mean.
- Like the standard normal distribution, the *t* distribution has the mean 0, but its variance depends on the parameter ν, called the number of **degrees of freedom**.

The Sampling Distribution of the mean (σ unknown)

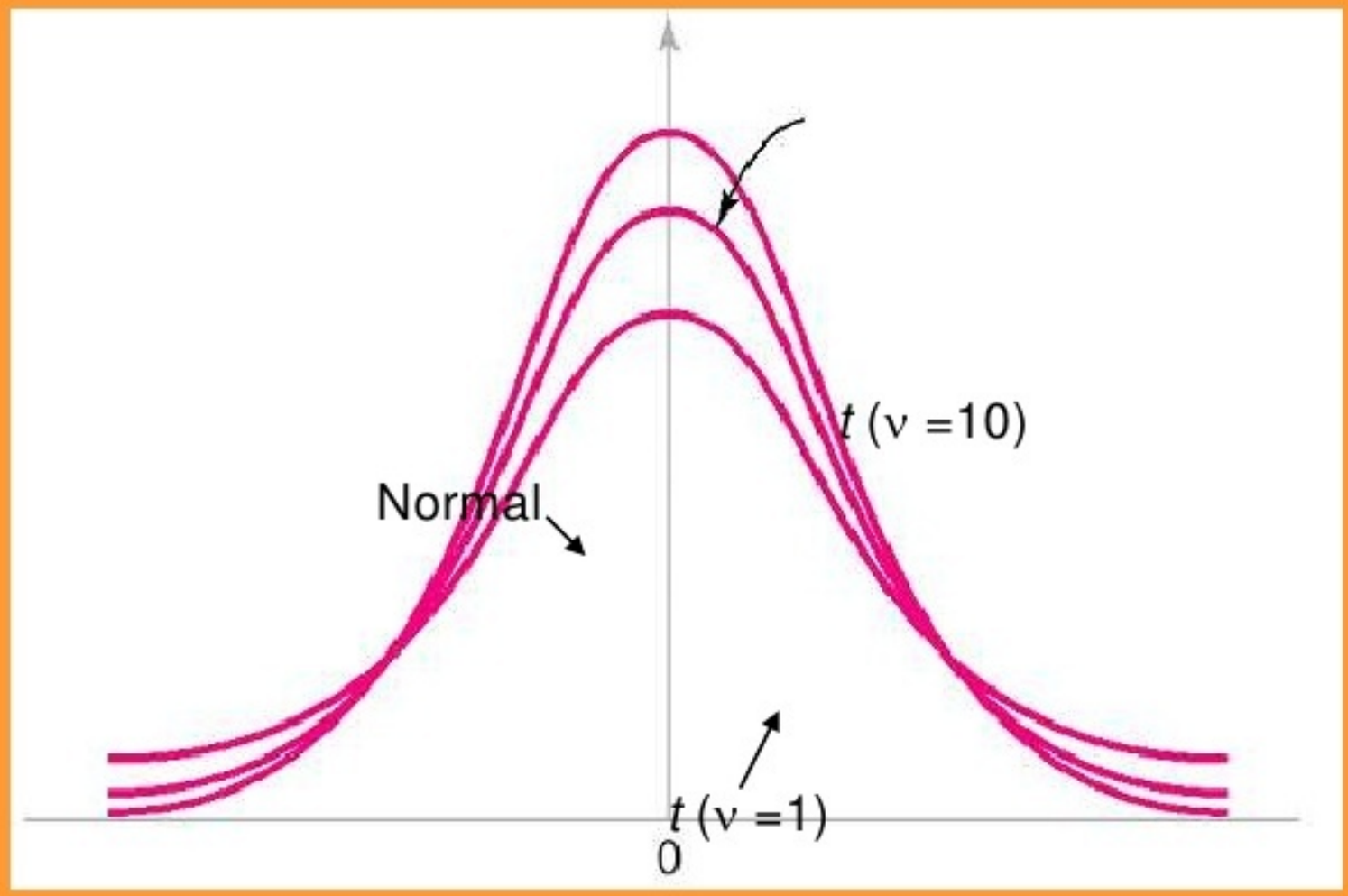

Figure: *t* distribution and standard normal distributions

datamingtools

The Sampling Distribution of the mean (σ unknown)

- When $\nu \to \infty$, the *t* distribution approaches the standard normal distribution i.e. when $\nu \to \infty$, $t_\alpha \to z_\alpha$.
- The standard normal distribution provides a good approximation to the *t* distribution for samples of size 30 or more.

datamingtools