Sampling Distributions and Central Limit Theorem

Statistics

Any function of the sample values is known as a statistics. Eg: Sample mean, Sample median, Sample variance etc. are all statistics.

Sampling Distribution

A sampling distribution is a distribution of a statistic over all possible samples. That is sampling distribution is the probability distribution of the statistics.

The commonly used sampling distributions are, Normal distribution, t distribution, χ^2 distribution and F distribution.

Sampling distribution of the mean of sample chosen from a Normal population N(μ,σ)

Let $x_1, x_2, \ldots, x_n$ be a sample chosen from Normal population N(μ,σ) we requires the distribution of

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Case-I when σ known

In this case $\bar{x}$ has normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ (i.e $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$)

Case-II when σ unknown

In this case $\bar{x}$ has Students t distribution with 'n-1' degrees of freedom.

Students t distribution

The t distribution was discovered by William. S. Gosset in 1908 who wrote under the pen name 'Student'.

A continues random variable t is said to follow a student's t distribution with n degrees of freedom if its pdf is given by

$$f(t) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad -\infty < t < \infty$$

or

$$f(t) = \frac{1}{\sqrt{n}\mathrm{B}\left(\frac{1}{2}, \frac{n}{2}\right)} \left(1 + \frac{t^2}{n}\right)^{-\left(\frac{n+1}{2}\right)}, \quad -\infty < t < \infty$$

the mean of the distribution is zero and variance is $\frac{n}{n-2}$ (n>2).

Properties

- A t-distribution is like a Z distribution, except has slightly fatter tails to reflect the uncertainty added by estimating σ.
- It is unimodel distribution
- It is a symmetric distribution (i.e. $\beta_1 = 0$)
- It is mesocourtic (i.e. $\beta_2 = 3$)
- The bigger the sample size (i.e., the bigger the sample size used to estimate σ), then the closer t becomes to Z.
- If n>100, t approaches Z.

Application or Uses of t-distribution

1. To find out the confidence interval for population mean of a normal distribution when population standard deviation σ unknown
2. To test sample mean $\bar{x}$ differs significantly from the hypothetical value of the population mean μ when population standard deviation σ unknown
3. To the significance of the difference between two population means when population standard deviation σ unknown
4. To test the significance of correlation coefficients and regression coefficients

Result: If X is a standard normal variate (i.e.X~N(0,1)) and Y is a χ^2 variate with n degree of freedom, then $t = \frac{X}{\sqrt{\frac{Y}{n}}}$ has t distribution with n degree of freedom.

Chi-Square distribution

A continues random variable χ^2 is said to follow a chi-square distribution with n degrees of freedom if its pdf is given by

$$f(\chi^2) = \frac{\left(\frac{1}{2}\right)^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)} e^{-\frac{\chi^2}{2}} (\chi^2)^{\left(\frac{n}{2}-1\right)}, \ 0 \leq \chi^2 < \infty.$$

the mean of the distribution is n and variance is $2n$.

Properties

- It is positively skewed distribution
- It is unimodel distribution
- If U and V are independent chi-square variables with n_1 and n_2 degrees of freedom then Z = U+V has chi-square variables with $n_1 + n_2$ degrees of freedom.
- If X is a standard normal variate (i.e.X~N(0,1)) then $Y = X^2$ is a χ^2 variate with 1 degree of freedom.

Application or Uses of χ^2 -distribution

1. To find out the confidence interval for population variance of a normal distribution
2. To test the hypothetical value of the population variance (i.e. to test whether $\sigma^2 = \sigma_0^2$)
3. To test the goodness of fit
4. To test the independence of attribute
5. To test the homogeneity of independent estimate of the population variance
6. To test the homogeneity of independent estimate of the population correlation coefficients.

Sampling distribution of the variance of sample chosen from a Normal population N(μ,σ)

Let $x_1, x_2, \ldots, x_n$ be a sample chosen from Normal population N(μ,σ) and let $s^2 = \frac{1}{n}\sum(x_i - \bar{x})^2$. Define $u = \frac{ns^2}{\sigma^2}$ then u has chi-square variables with n -1 degrees of freedom.

Sendecor's F distribution

A continues random variable F is said to follow a F distribution with (n_1 , n_2) degrees of freedom if its pdf is given by

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{B\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \frac{F^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1}{n_2}F\right)^{\frac{n_1+n_2}{2}}},\ 0 \le F < \infty.$$

the mean of the distribution is $\frac{n_2}{n_2-2}$, n_2>2 and variance is $\frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$ n_2>4.

Properties

- It is positively skewed distribution
- It is unimodel distribution
- If U and V are independent chi-square variables with n_1 and n_2 degrees of freedom then F = $\frac{U/n_1}{V/n_2}$ has F distribution with (n_1,n_2) degrees of freedom.
- If t is a student's t variate with n degree of freedom then t^2 follows F distribution with (1,n) degrees of freedom.

Application or Uses of F -distribution

1. To test the equality of two population variances (i.e. to test whether $\sigma_1^2 = \sigma_2^2$)
2. To test the significant of an observed multiple correlation
3. To test the significant of correlation ratio
4. To test the model fit or Linearity of Regression
5. To test the equality of several means

Standard Error

The standard deviation of the sampling distribution of a statistic is called standard error (SE)

Uses of Standard Error

1. It is used for finding confidence intervals
2. It is used for testing a given hypothesis
3. It gives an idea about the unreliability of the sample. Reciprocal of SE is taken as a measure of reliability.

Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ are independently and identically distributed random variables with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$, $i = 1, 2, \ldots, n$ then the sum $S_n = X_1 + X_2 + \ldots + X_n$ is asymptotically normally distributed with mean $n\mu$ and variance $n\sigma^2$.

Assumption made about the component variables for CLT

1. The variables are independent
2. All the variable follows the same distribution (no restriction that they are normal)
3. The mean and variance exist for all the variables
4. All the variables have the same mean and same variance.

If S_n is defined as above then $Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$ follows Standard normal distribution when n become very large ($n \to \infty$).

Application:- Central limit theorem plays a vital role in testing of hypothesis and interval estimation when the sample size become very large (generally n>50).

Tchebycheff's Inequality

Let X be any random variables with mea μ and variance σ^2 exist. Then for any positive number t

$$P\{|X - \mu| \geq t\sigma\} \leq \frac{1}{t^2}$$

Or

$$P\{|X - \mu| \leq t\sigma\} \geq 1 - \frac{1}{t^2}.$$

Advantages of the Inequality

1. It gives an upper bound to the probability of a random variable deviating from its mean by more than 't' times its standard deviation.
2. It is applicable to all random variables for which mean and standard deviation exist.
3. It is valid for both discrete and continuous random variables.
4. It justifies the importance given to the standard deviation as a measure of dispersion.

Disadvantages of the Inequality

1. If 't' is chosen as a number less than 1, the upper bound obtain will be more than 1 and hence is of no use as we know that any probability is less than 1.
2. The upper bound given in most cases is much larger than the true probability.
3. This will give a useful result only when the distribution of the variable is not known.

Bernoulli's law of large number

Consider a random experiment with only two possible outcome 'success' and 'failure'. Let the experiment be repeated n times and x be the number of successes among them. Let p be the probability of success in a single trail. Then Bernoulli's law of large number states that for any small positive number $\in$,

$$P\left\{\left|\frac{x}{n} - p\right| \leq \in\right\} \to 1 \text{ as } n \to \infty.$$

Week law of large number

Let $X_1, X_2, \ldots, X_n$ are independently distributed random variables with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$, $i = 1, 2, \ldots, n$. The law of large number states that for any small positive number $\in$,

$$P\{|\bar{x} - \mu| \leq \in\} \to 1 \text{ as } n \to \infty$$

where $\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$ and $\mu = \frac{\mu_1 + \mu_2 + \cdots + \mu_n}{n}$.