

UNIT - 3

Sampling-Distributions, Parameter-Estimations, Hypothesis-Testing, Two-population, Linear Regression, Tests, Regression and Correlation, Block Chain



Sampling-Distributions



Special Distributions

Chi-Square DISTRIBUTION

A continuous random variable S is said to follow chi-square distribution with parameter (also called as the degree of freedom) n if its pdf is given by,

$$\Rightarrow f(s) = \frac{1}{2^{n/2}\Gamma(\frac{n}{2})} s^{\frac{n}{2}-1} e^{-\frac{s}{2}}, \quad s > 0, \quad n \text{ is a positive integer}$$

As such chi-square distribution is a one parameter family of distributions and the parameter is always a positive integer.

Notes:

- A chi-square variable S with parameter n is also denoted as χ_n^2 .
- If X is a standard normal variable, then X^2 is chi-square variable with parameter $n = 1$, i.e. X^2 is a χ_1^2 variable.
- If S is a chi-square variable with parameter n , then $E(S) = n$ and $Var(S) = 2n$.
- If Y is a χ_n^2 variable, then for sufficiently large n , the variable $\sqrt{2Y}$ is approximately normally distributed with mean $\sqrt{2n-1}$ and variance 1.

Student's t distribution

Suppose that X and Y are two independent random variables having standard normal distribution; and chi-square distribution with parameter k , respectively. As such $X \sim N(0, 1)$ and $Y \sim \chi_k^2$. Let us define,

$$\Rightarrow T = \frac{X}{\sqrt{Y/k}}.$$

Then T is called the Student's T variable with parameter k , and its pdf is given by,

$$\Rightarrow f_k(t) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{k}{2})\sqrt{\pi k}} \left(1 + \frac{t^2}{k}\right)^{-(k+1)/2}, -\infty < t < \infty$$

Notes:

- The graph of $f_k(t)$ is symmetrical. It resembles with the graph of normal distribution.
- We can show that $\lim_{k \rightarrow \infty} f_k(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

F distribution

Suppose that X and Y are two independent random variables having chi-square distributions with parameter k and l , respectively. As such $X \sim \chi_k^2$ and $Y \sim \chi_l^2$. Let us define,

$$\Rightarrow F = \frac{X/k}{Y/l} = \frac{lX}{kY} .$$

Then the pdf of variable F is given by,

$$\Rightarrow h_{k,l}(f) = \frac{\Gamma(\frac{k+l}{2})}{\Gamma(\frac{k}{2})\Gamma(\frac{l}{2})} \left(\frac{k}{l}\right)^{k/2} f^{\frac{k}{2}-1} \left(1 + \left(\frac{k}{l}\right)f\right)^{-(k+l)/2}, \quad f > 0$$

- This is called F distribution with parameters (k, l) . This is also called F distribution with degrees of freedom (k, l) .

RANDOM SAMPLE AND SAMPLING DISTRIBUTION

- Let X be a random variable, either discrete or continuous, following a distribution function $f(x)$. Then, (X_1, X_2, \dots, X_n) is a random sample of size n from the distribution of X if X_i , $i = 1, 2, \dots, n$ are independently and identically distributed with the same distribution function $f(x)$.
- We will, in this situation, have the joint distribution function of (X_1, X_2, \dots, X_n) as,
 - $f(x_1, x_2, \dots, x_n) = f(x_1)f(x_2) \dots f(x_n)$

STATISTIC / ESTIMATOR

- A function $T = T(X_1, X_2, \dots, X_n)$ of the random sample (X_1, X_2, \dots, X_n) is called a statistic.

Examples of Statistic:

- $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is a statistic called as *sample mean*
- $T_2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ is other statistic
- $T_3 = \max(X_1, X_2, \dots, X_n)$ is also a statistic
- $T_4 = \min(X_1, X_2, \dots, X_n)$ is yet another statistic
- $T_5 = T_4 - T_3$, also known as *range* is a statistic

One may note that since T is a function of (X_1, X_2, \dots, X_n) , T is also a random variable.



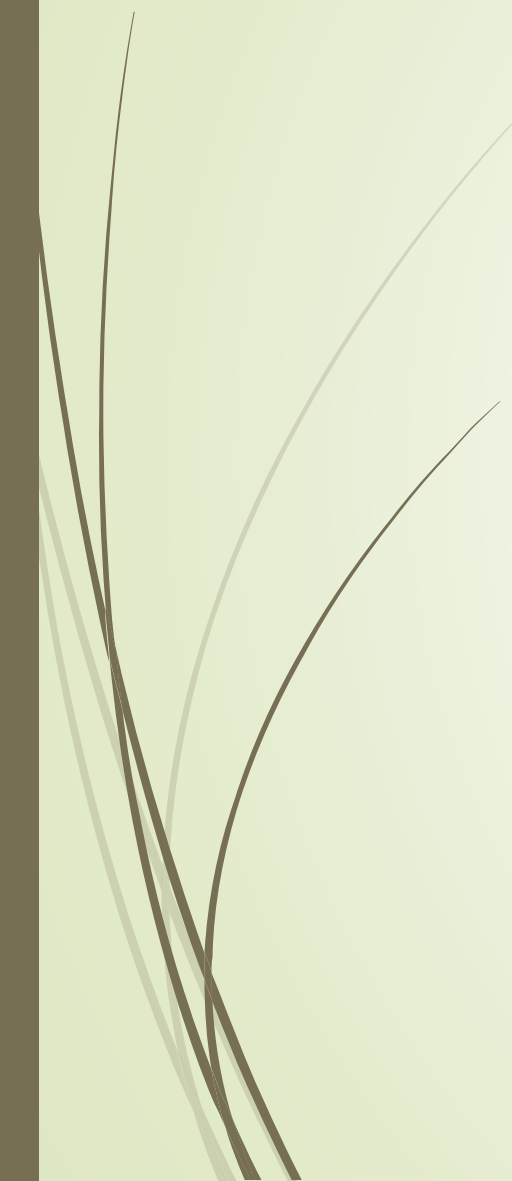
Sampling Distribution



The probability distribution of a statistic is called a sampling distribution.



Characteristics of Statistic / Estimator

- A statistic is defined in order to estimate the population parameters, for example sample mean is used to estimate the population mean, sample variance is used to estimate population variance etc.
 - Two important characteristics of Statistic / Estimator
 - Unbiasedness
 - Consistency
- 

Characteristics of Statistic / Estimator

Unbiased Statistic (Estimator)

An estimator $\hat{\theta}$ is **unbiased** for a parameter θ if its expectation equals the parameter,

$$\mathbf{E}(\hat{\theta}) = \theta$$

for all possible values of θ .

Bias of $\hat{\theta}$ is defined as $\text{Bias}(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)$.

Consistent Statistic (Estimator)

An estimator $\hat{\theta}$ is **consistent** for a parameter θ if the probability of its sampling error of any magnitude converges to 0 as the sample size increases to infinity. Stating it rigorously,

$$P \left\{ |\hat{\theta} - \theta| > \varepsilon \right\} \rightarrow 0 \text{ as } n \rightarrow \infty$$

for any $\varepsilon > 0$. That is, when we estimate θ from a large sample, the estimation error $|\hat{\theta} - \theta|$ is unlikely to exceed ε , and it does it with smaller and smaller probabilities as we increase the sample size.

Central Limit Theorem

- ▶ Let $X_1, X_2, \dots, X_n, \dots$ be a sequence of independent random variables with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Let us consider $X = X_1 + X_2 + \dots + X_n$. Then,
- ▶
$$Z = \frac{X - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$
- ▶ has approximately standard normal distribution.
- ▶ Approximation here is in the sense that when $n \rightarrow \infty$, then $Z \rightarrow$ Standard Normal Distribution.
- ▶ If $n \geq 30$ we say that this is a large sample otherwise we call it as a small sample.

CLT Demonstration using R

Distribution of Sample Mean when σ is known

- Let (X_1, X_2, \dots, X_n) be a random sample of size n taken from the variable X with distribution function $f(x)$. Let us also take $E(X) = \mu$ and $V(X) = \sigma^2$. Then, $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ follows approximately standard normal distribution.

- Here, we have $\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n) = (X_1/n + X_2/n + \dots + X_n/n)$

Also,

- $E\left(\frac{1}{n}X_i\right) = \frac{1}{n}\mu$, for all $i = 1, 2, 3, \dots, n$; and

- $V\left(\frac{1}{n}X_i\right) = \frac{1}{n^2}\sigma^2$, for all $i = 1, 2, 3, \dots, n$.

- Using central limit theorem, we deduce that following variable Z has approximately standard normal distribution:

$$Z = \frac{\bar{X} - \sum_{i=1}^n \frac{1}{n}\mu}{\sqrt{\sum_{i=1}^n \frac{1}{n^2}\sigma^2}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Sample mean is an unbiased estimate of population mean

- Let us show that sample mean $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ is an unbiased statistic of population mean μ .
- Here, let us consider that a sample (X_1, X_2, \dots, X_n) of size n is taken for a variable X following the distribution function $f(x)$ with mean μ and variance σ^2 .
- It means we are assuming that $E(X) = \mu$ and $E(X - (E(X)))^2 = E(X^2) - (E(X))^2 = \sigma^2$. We have to show that $E(\bar{X}) = \mu$.

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

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

- $$= \frac{1}{n} \sum_{i=1}^n E(X_i)$$

- $$= \frac{1}{n} \sum_{i=1}^n \mu$$

- $$= \mu$$

- As such, sample mean \bar{X} is an unbiased statistic of the population mean μ .

Sample mean is a consistent estimate of population mean

We can prove this using the Chebyshev's Inequality.

➤ Chebyshev's inequality is: $P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}$ [See its proof, homework]

This gives:

➤ $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2}$

OR

➤ $P(|\bar{X} - \mu| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}$

➤ The RHS tends to zero, as n tends to infinity.

➤ Hence, Sample mean is a consistent estimate of population mean.

Example

Suppose that height of an individual adult male is normally distributed with mean 68 inches and standard deviation 3 inches. A sample of 25 adult males is taken and their height is measured. Find the probability that sample mean will differ from the population mean by less than one inch. Also find $P(\bar{X} < 69.5)$.

Solution:

Example

Suppose that height of an individual adult male is normally distributed with mean 68 inches and standard deviation 3 inches. A sample of 25 adult males is taken and their height is measured. Find the probability that sample mean will differ from the population mean by less than one inch. Also find $P(\bar{X} < 69.5)$.

► **Solution:** Here, we have $\mu = 68$, $\sigma = 3$ and $n = 25$. We have to find $P(|\bar{X} - \mu| < 1)$.

►
$$P(|\bar{X} - \mu| < 1) = P(-1 < \bar{X} - \mu < 1) = P(\bar{X} - \mu < 1) - P(\bar{X} - \mu < -1)$$

►
$$= P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{1}{\sigma/\sqrt{n}}\right) - P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{-1}{\sigma/\sqrt{n}}\right) = P\left(Z < \frac{1}{\sigma/\sqrt{n}}\right) - P\left(Z < \frac{-1}{\sigma/\sqrt{n}}\right)$$

►
$$= P(Z < 1.67) - P(Z < -1.67) = \text{pnorm}(1.67) - \text{pnorm}(-1.67)$$

$$= 2 * \text{pnorm}(1.67) - 1 \text{ (Why?) } = 0.9051$$

► **As such, there are 90.51% chances that the sample mean will differ from the population mean by less than one inch.**

► Let us now find $P(\bar{X} < 69.5)$.

►
$$P(\bar{X} < 69.5) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{69.5 - 68}{3/5}\right) = P(Z < 2.5) = 0.9938$$

Definition of sample variance

Let X be a random variable, either discrete or continuous, following a distribution function $f(x)$. Let us take a random sample (X_1, X_2, \dots, X_n) of size n from X and define,

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

where, \bar{X} is the sample mean. Then, the statistic S^2 is called as sample variance. Its positive square root is called sample standard deviation.

Sample VAR is an unbiased estimate of population VAR

§ We define sample variance S^2 as:

$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ and we need to show that $E(S^2) = \sigma^2$ where $\sigma^2 = \text{pop. Variance}$

Let us consider that the random sample (X_1, X_2, \dots, X_n) is taken from the distribution of random variable X , having $E(X) = \mu$, and $V(X) = \sigma^2$.

So, $E(X_i) = \mu$ and $V(X_i) = \sigma^2$ for $i = 1, 2, \dots, n$.

We have that $(n-1)S^2 = \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n (X_i - \mu + \mu - \bar{X})^2$

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

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

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

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

$$\text{So, } E((n-1)S^2) = E\left(\sum_{i=1}^n (X_i - \mu)^2\right) - n \cdot E(\bar{X} - \mu)^2$$

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

$$= n \cdot \sigma^2 - n \cdot \sigma^2/n$$

$$= (n-1)\sigma^2$$

$$\text{OR, } (n-1)E(S^2) = (n-1) \cdot \sigma^2$$

$$\text{OR, } E(S^2) = \sigma^2$$

As such, S^2 is an unbiased estimate of σ^2 .

DISTRIBUTION OF SAMPLE VARIANCE

If X is normally distribution, then $\frac{(n-1)S^2}{\sigma^2}$ has chi-square distribution with parameter $(n - 1)$.

Example

Suppose that X has $N(0, 0.04)$ distribution. A sample $(X_1, X_2, \dots, X_{25})$ is taken from X . What is the probability that $\sum_{i=1}^{25} X_i^2$ exceeds 1.5 if sample mean is 0?

Example

Suppose that X has $N(0, 0.04)$ distribution. A sample $(X_1, X_2, \dots, X_{25})$ is taken from X . What is the probability that $\sum_{i=1}^{25} X_i^2$ exceeds 1.5 if sample mean is 0?

► Let us here use the fact that $\frac{(n-1)S^2}{\sigma^2}$ has chi-square distribution with parameter $(n-1)$. From the problem, we get $\mu = 0$, $\sigma^2 = 0.04$ and $n = 25$ and $\bar{x} = 0$. So, we have

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

►
$$= \frac{1}{(24)} \sum_{i=1}^{25} X_i^2 \quad \text{or, } 24S^2 = \sum_{i=1}^{25} X_i^2.$$

► We have to calculate, $P(\sum_{i=1}^{25} X_i^2 > 1.5) = P(24S^2 > 1.5)$

►
$$= 1 - P(24S^2 \leq 1.5) = 1 - P\left(\frac{24S^2}{\sigma^2} \leq \frac{1.5}{\sigma^2}\right)$$

►
$$= 1 - P(\chi_{24}^2 \leq 37.5) \approx 1 - 0.9625 = 0.0375$$

► Using R, $1 - P(\chi_{24}^2 \leq 37.5) = 1 - 0.9610182 = 0.0389818$

Distribution of Sample Mean when σ is not known

We know that,

- $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ has standard normal distribution $N(0, 1)$ and
- $\frac{(n-1)S^2}{\sigma^2}$ has chi-square distribution with parameter $(n - 1)$.

Using the definition of t-distribution

- $T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}}$ has t distribution with parameter $(n - 1)$

or, $T = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{S/\sigma} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ has t distribution with parameter $(n - 1)$.

- As such, we apply t distribution when population variance is not known.
 - The t distributions were discovered by William S. Gosset in 1908. Gosset was a statistician employed by the Guinness brewing company which had stipulated that he not publish under his own name. He therefore wrote under the pen name "Student".

Example

Suppose that height of an individual adult male is normally distributed with mean 68 inches. A sample of 25 adult males is taken and their height is measured. Find the probability that sample mean will differ from the population mean by less than one inch when sample variance s^2 is 4 inches².

Example

Suppose that height of an individual adult male is normally distributed with mean 68 inches. A sample of 25 adult males is taken and their height is measured. Find the probability that sample mean will differ from the population mean by less than one inch when sample variance s^2 is 4 inches².

- Here, we have $\mu = 68$, $n = 25$ and sample variance $s^2 = 4$. We have to find $P(|\bar{X} - \mu| < 1)$.
- $P(|\bar{X} - \mu| < 1) = P(-1 < \bar{X} - \mu < 1) = P(\bar{X} - \mu < 1) - P(\bar{X} - \mu < -1)$
- $= P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} < \frac{1}{s/\sqrt{n}}\right) - P\left(\frac{\bar{X} - \mu}{s/\sqrt{n}} < \frac{-1}{s/\sqrt{n}}\right)$
- $= P\left(T_{24} < \frac{1}{s/\sqrt{n}}\right) - P\left(T_{24} < \frac{-1}{s/\sqrt{n}}\right)$
- $= P(T_{24} < 2.5) - P(T_{24} < -2.5)$
- $= \text{pt}(2.5, 24) - \text{qt}(-2.5) = 0.9803458$
- [This will also be $= 2 * \text{qt}(2.5, 24) - 1$; t-distribution is symmetrical about y-axis]
 - This gives the desired probability.

Distribution of Ratio of Variances, when population variances are known

- Let us take two random samples of sizes n_1 and n_2 , respectively, from the two normal populations.
- Suppose that the variances of these two samples are S_1^2 and S_2^2 .
- Then, we have noted that $\frac{(n_1-1)S_1^2}{\sigma_1^2}$ is a chi-square variable with parameter $(n_1 - 1)$
- And $\frac{(n_2-1)S_2^2}{\sigma_2^2}$ is a chi-square variable with parameter $(n_2 - 1)$.
- As such $\frac{\frac{(n_1-1)S_1^2}{\sigma_1^2(n_1-1)}}{\frac{(n_2-1)S_2^2}{\sigma_2^2(n_2-1)}} = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$ has F distribution with parameters $((n_1 - 1), (n_2 - 1))$.
- F distribution is a 2-parameter family of distributions.



Thank You !