Proability and Statistics

MA20205

Bibhas Adhikari

Autumn 2022-23, IIT Kharagpur

Lecture 17
November 1, 2022

Sampling distributions

Statistic If $X_1, X_2, \ldots, X_n$ is a random sample from a population then a statistic is a function of X_is.

Sampling distributions

Statistic If $X_1, X_2, \ldots, X_n$ is a random sample from a population then a statistic is a function of X_is.

Sampling distribution If the distribution of the population is known then often it is possible to find the probability distribution of the statistics associated with the population. If T is such a statistic then the distribution of T is called the sampling distribution of T.

Sampling distributions

Student's t-distribution This distribution was discovered by W. S. Gosset (1876-1936) who published his work under the pseudonym of student. Thus it is known as Student's t-distribution. This distribution is a generalization of Cauchy distribution and normal distribution.

Sampling distributions

Student's t-distribution This distribution was discovered by W. S. Gosset (1876-1936) who published his work under the pseudonym of student. Thus it is known as Student's t-distribution. This distribution is a generalization of Cauchy distribution and normal distribution.

A continuous random variable X is said to have a t-distribution with ν degrees of freedom if the pdf is of the form

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

where $\nu > 0$.

Sampling distributions

Student's t-distribution This distribution was discovered by W. S. Gosset (1876-1936) who published his work under the pseudonym of student. Thus it is known as Student's t-distribution. This distribution is a generalization of Cauchy distribution and normal distribution.

A continuous random variable X is said to have a t-distribution with ν degrees of freedom if the pdf is of the form

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

where $\nu > 0$.

If X has t-distribution then we write $X \sim t(\nu)$

Sampling distributions

Setting $\nu = 1$, the pdf of t-distribution becomes

$$f(x;1) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty,$$

which is the pdf of Cauchy distribution.

Sampling distributions

Setting $\nu = 1$, the pdf of t-distribution becomes

$$f(x;1) = \frac{1}{\pi(1+x^2)}, -\infty < x < \infty,$$

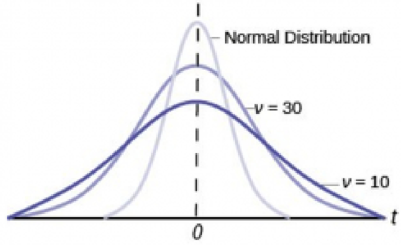
which is the pdf of Cauchy distribution.

Setting $\nu \to \infty$ then

$$\lim_{\nu \to \infty} f(x;\nu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, -\infty < x < \infty$$

which is the pdf of standard normal distribution.

Sampling distributions



Sampling distributions

Properties of t-distribution If $X \sim t(\nu)$ then

$$E(X) = \begin{cases} 0, & \text{if } \nu \geq 2 \\ \text{DNE}, & \text{if } \nu = 1 \end{cases}$$

and

$$Var(X) = \begin{cases} \frac{\nu}{\nu-2}, & \text{if } \nu \geq 3 \\ \text{DNE}, & \text{if } \nu = 1, 2 \end{cases}$$

where DNE means does not exist.

Sampling distributions

Properties of t-distribution If $X \sim t(\nu)$ then

$$E(X) = \begin{cases} 0, & \text{if } \nu \geq 2 \\ \text{DNE}, & \text{if } \nu = 1 \end{cases}$$

and

$$Var(X) = \begin{cases} \frac{\nu}{\nu-2}, & \text{if } \nu \geq 3 \\ \text{DNE}, & \text{if } \nu = 1, 2 \end{cases}$$

where DNE means does not exist.

If $Z \sim \mathcal{N}(0, 1)$ and $U \sim \chi^2(\nu)$ and in addition, Z and U are independent, then the random variable W defined by

$$W = \frac{Z}{\frac{U}{\nu}}$$

has a t-distribution with ν degrees of freedom.

Sampling distributions

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ be a random sample from the population X, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

Sampling distributions

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ be a random sample from the population X, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

Proof Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Thus $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$.

Sampling distributions

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ be a random sample from the population X, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

Proof Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Thus $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1)$.

Further we know

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Sampling distributions

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $X_1, X_2, \ldots, X_n$ be a random sample from the population X, then

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t(n-1)$$

Proof Since $X_i \sim \mathcal{N}(\mu, \sigma^2)$, $\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. Thus $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0, 1)$.

Further we know

$$\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2(n-1).$$

Hence

$$\frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{(n-1)S^2}{(n-1)\sigma^2}}} \sim t(n-1) \text{ (by previous result)}$$

Sampling distributions

(Snedecor's) F-distribution The F-distribution was named in the honr of Ronald Fisher by george Snedecor.

Sampling distributions

(Snedecor's) *F*-distribution The *F*-distribution was named in the honr of Ronald Fisher by george Snedecor.

A continuous random variable X is said to have a F-distribution with degrees of freedom ν_1 and ν_2 if its pdf is of the form

$$f(x;\nu_1,\nu_2) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}x\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}}, & \text{if } 0 \leq x < \infty \\ 0, & \text{otherwise} \end{cases}$$

where $\nu_1, \nu_2 > 0$.

Sampling distributions

(Snedecor's) F-distribution The F-distribution was named in the honr of Ronald Fisher by george Snedecor.

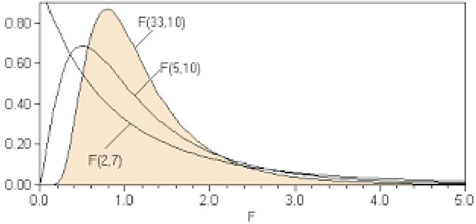
A continuous random variable X is said to have a F-distribution with degrees of freedom ν_1 and ν_2 if its pdf is of the form

$$f(x;\nu_1,\nu_2) = \begin{cases} \dfrac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}} x^{\frac{\nu_1}{2}-1}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)\left(1+\frac{\nu_1}{\nu_2}x\right)^{\left(\frac{\nu_1+\nu_2}{2}\right)}}, & \text{if } 0 \leq x < \infty \\ 0, & \text{otherwise} \end{cases}$$

where $\nu_1, \nu_2 > 0$.

Then we write $X \sim F(\nu_1, \nu_2)$.

Sampling distributions



Sampling distributions

Properties of F-distribution If $X \sim F(\nu_1, \nu_2)$ then

$$E(X) = \begin{cases} \frac{\nu_1}{\nu_2 - 2}, & \text{if } \nu_2 \geq 3 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2 \end{cases} \qquad Var(X) = \begin{cases} \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, & \text{if } \nu_2 \geq 5 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2, 3, 4 \end{cases}$$

where DNE means does not exist.

Sampling distributions

Properties of F-distribution If $X \sim F(\nu_1, \nu_2)$ then

$$E(X) = \begin{cases} \frac{\nu_1}{\nu_2 - 2}, & \text{if } \nu_2 \geq 3 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2 \end{cases} \qquad Var(X) = \begin{cases} \frac{2\nu_2^2(\nu_1 + \nu_2 - 2)}{\nu_1(\nu_2 - 2)^2(\nu_2 - 4)}, & \text{if } \nu_2 \geq 5 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2, 3, 4 \end{cases}$$

where DNE means does not exist.

If $X \sim F(\nu_1, \nu_2)$ then $\frac{1}{X} \sim F(\nu_2, \nu_1)$ (which is helpful to calculate probabilities like $P(X \leq a)$ from F-table)

Sampling distributions

Properties of F-distribution If $X \sim F(\nu_1, \nu_2)$ then

$$E(X) = \begin{cases} \frac{\nu_1}{\nu_2 - 2}, & \text{if } \nu_2 \geq 3 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2 \end{cases} \qquad Var(X) = \begin{cases} \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}, & \text{if } \nu_2 \geq 5 \\ \text{DNE}, & \text{if } \nu_2 = 1, 2, 3, 4 \end{cases}$$

where DNE means does not exist.

If $X \sim F(\nu_1, \nu_2)$ then $\frac{1}{X} \sim F(\nu_2, \nu_1)$ (which is helpful to calculate probabilities like $P(X \leq a)$ from F-table)

If $U \sim \chi^2(\nu_1)$ and $V \sim \chi^2(\nu_2)$, and they are independent then

$$\frac{\frac{U}{\nu_1}}{\frac{V}{\nu_2}} \sim F(\nu_1, \nu_2)$$

Sampling distributions

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_1, \ldots, X_n$ be random sample of size n from a population X. Let $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $Y_1, \ldots, Y_m$ be a random sample of size m from the population Y. Then the statistic

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n-1, m-1),$$

where S_1^2 and S_2^2 are sample variances of the samples X_is and Y_js respectively.

Sampling distributions

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_1, \ldots, X_n$ be random sample of size n from a population X. Let $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $Y_1, \ldots, Y_m$ be a random sample of size m from the population Y. Then the statistic

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n-1, m-1),$$

where S_1^2 and S_2^2 are sample variances of the samples X_is and Y_js respectively.

Proof Since $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $(n-1)\frac{S_1^2}{\sigma_1^2} \sim \chi^2(n-1)$. Similarly, since $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $(m-1)\frac{S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$

Sampling distributions

Let $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_1, \ldots, X_n$ be random sample of size n from a population X. Let $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and $Y_1, \ldots, Y_m$ be a random sample of size m from the population Y. Then the statistic

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \sim F(n-1, m-1),$$

where S_1^2 and S_2^2 are sample variances of the samples X_is and Y_js respectively.

Proof Since $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$, $(n-1)\frac{S_1^2}{\sigma_1^2} \sim \chi^2(n-1)$. Similarly, since $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$, $(m-1)\frac{S_2^2}{\sigma_2^2} \sim \chi^2(m-1)$ Therefore

$$\frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} = \frac{\frac{(n-1)S_1^2}{(n-1)\sigma_1^2}}{\frac{(m-1)S_2^2}{(m-1)\sigma_2^2}} \sim F(n-1, m-1).$$