Statistical Foundation of Data Science (MA 589)

Lecture Slides

Topic 05: Sampling Distributions based on Normal Population

Univariate Normal Distribution

Definition 5.1: A continuous random variable X is said to have a univariate normal distribution if the PDF of X is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \text{ for all } x \in \mathbb{R},$$

where $\mu \in \mathbb{R}$ and $\sigma > 0$.

Notation: $X \sim N(\mu, \sigma^2)$.

Theorem 5.1: If $X \sim N(\mu, \sigma^2)$, all moments of X exist. In particular, $E(X)$ and $Var(X)$ exist, and they are given by $E(X) = \mu$ and $Var(X) = \sigma^2$.

Remark 5.1: This means that a normal distribution is completely specified by its mean and variance.

Bivariate Normal

Definition 5.2: A two dimensional random vector $\boldsymbol{X} = (X_1,\ X_2)$ is said to have a bivariate normal distribution if $aX_1 + bX_2$ is a univariate normal for all $(a,\ b) \in \mathbb{R}^2 \setminus (0,0)$.

Remark 5.2: If $\boldsymbol{X}$ has bivariate normal distribution, then each of X_1 and X_2 is univariate normal. Hence $E(X_1)$, $E(X_2)$, $Var(X_1)$, $Var(X_2)$, and $Cov(X_1,\ X_2)$ exist.

Notation: $\boldsymbol{\mu} = E(\boldsymbol{X}) = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$ and $\Sigma = Var(\boldsymbol{X}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$.

Theorem 5.2: Let $\boldsymbol{X}$ be a bivariate normal random vector. If $\boldsymbol{\mu} = E(\boldsymbol{X})$ and $\Sigma = Var(\boldsymbol{X})$, then for any fixed $\boldsymbol{u} = (a,\ b) \in \mathbb{R}^2 \setminus (0,0)$, $\boldsymbol{u}'\boldsymbol{X} \sim N(\boldsymbol{u}'\boldsymbol{\mu}, \boldsymbol{u}'\Sigma\boldsymbol{u})$.

Bivariate Normal

Theorem 5.3: Let $\boldsymbol{X}$ be a bivariate normal random vector, then $M_{\boldsymbol{X}}(\boldsymbol{t}) = e^{\boldsymbol{t}'\mu + \frac{1}{2}\boldsymbol{t}'\Sigma\boldsymbol{t}}$ for all $\boldsymbol{t} \in \mathbb{R}^2$.

Remark 5.3: Thus the bivariate normal distribution is completely specified by the mean vector $\boldsymbol{\mu}$ and the variance-covariance matrix Σ.

Notation: $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$.

Corollary 5.1: If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$, then $X_1 \sim N(\mu_1, \sigma_{11})$ and $X_2 \sim N(\mu_2, \sigma_{22})$.

Remark 5.4: The converse of the above theorem is not true.

Remark 5.5: If $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ and $Cov(X_1, X_2) = 0$, then X_1 and X_2 are independent.

Probability Density Function

Theorem 5.4: Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then, for all $\boldsymbol{x} \in \mathbb{R}^2$, $\boldsymbol{X}$ has a joint PDF given by

$$\begin{aligned} f(\boldsymbol{x}) &= \frac{1}{2\pi|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\} \\ &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{A(x_1, x_2, \mu_1, \mu_2, \sigma_1, \sigma_2, \rho)}, \end{aligned}$$

where $\sigma_1 = \sqrt{\sigma_{11}}$, $\sigma_2 = \sqrt{\sigma_{22}}$, ρ is correlation coefficient between X_1 and X_2, and

$$A = -\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right\}.$$

Conditional Probability Density Function

Theorem 5.5: Let $\boldsymbol{X} \sim N_2(\boldsymbol{\mu}, \Sigma)$ be such that Σ is invertible, then

1. for all $x_2 \in \mathbb{R}$, the conditional PDF of X_1 given $X_2 = x_2$ is given by

$$f_{X_1|X_2}(x_1|x_2) = \frac{1}{\sigma_{1|2}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_1 - \mu_{1|2}}{\sigma_{1|2}}\right)^2\right] \quad \text{for } x_1 \in \mathbb{R},$$

   where $\mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ and $\sigma_{1|2}^2 = \sigma_1^2(1 - \rho^2)$.

2. $X_1|X_2 = x_2 \sim N\left(\mu_{1|2}, \sigma_{1|2}^2\right)$.
3. $E(X_1|X_2 = x_2) = \mu_{1|2} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(x_2 - \mu_2)$ for all $x_2 \in \mathbb{R}$.
4. $Var(X_1|X_2 = x_2) = \sigma_{1|2}^2 = \sigma_1^2(1 - \rho^2)$ for all $x_2 \in \mathbb{R}$. Hence the conditional variance does not depend on x_2.

Dist. of Sample Mean and Variance

Theorem 5.6: Let X_1, X_2, $\ldots$ X_n be i.i.d. $N(0,1)$ random variables. Then $\sum_{i=1}^n X_i^2 \sim Gamma(n/2, 1/2) \equiv \chi_n^2$.

Theorem 5.7: Let X_1, X_2, $\ldots$ X_n be i.i.d. $N(\mu,\sigma^2)$ random variables. Then $\overline{X} \sim N(\mu,\, \sigma^2/n)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$, and $\overline{X}$ and S^2 are independently distributed. Here $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$.

Student's t Distribution

Theorem 5.8: Let X and Y are two independent random variables with $X \sim N(0, 1)$ and $Y \sim \chi^2_\nu$. Then, for all $t \in \mathbb{R}$, the PDF of the random variable $T = \frac{X}{\sqrt{Y/\nu}}$ is

$$f_T(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}.$$

Remark 5.6: A random variable T is said to have Student's t distribution with degrees of freedom ν if the PDF of the random variable is $f_T(\cdot)$ as given above. We will use the notation $T \sim t_\nu$ to mean that the RV T has a t distribution with ν degrees of freedom.

Corollary 5.2: Let $X_1, X_2, \ldots X_n$ be i.i.d. $N(\mu, \sigma^2)$ random variables. Then

$$\sqrt{n}\,\frac{\overline{X} - \mu}{S} \sim t_{n-1},$$

where S^2 is the sample variance.

F Distribution

Theorem 5.9: Let X and Y be two independent RVs with $X \sim \chi^2_{d_1}$ and $Y \sim \chi^2_{d_2}$. Then, for all $x > 0$, the PDF of the RV $F = \frac{X/d_1}{Y/d_2}$ is

$$f_F(x) = \frac{1}{\mathrm{B}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)} \left(\frac{d_1}{d_2}\right)^{\frac{d_1}{2}} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}} .$$

Remark 5.7: A random variable F is said to have F distribution with degrees of freedoms d_1 and d_2 if the PDF of the random variable is $f_F(\cdot)$ as given above. We will use the notation $F \sim F_{d_1, d_2}$ to mean that the RV F has a F distribution with d_1 and d_2 degrees of freedoms.

F Distribution

Corollary 5.3: Let $X_1, X_2, \ldots X_n$ be i.i.d. $N(\mu_1, \sigma_1^2)$ RVs. Let $Y_1, Y_2, \ldots Y_m$ be i.i.d. $N(\mu_2, \sigma_2^2)$ RVs. Also, assume that X_i's and Y_j's are independent RVs. Then

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

where S_1^2 and S_2^2 are sample variances based on X_i's and Y_j's respectively.