Data Mining for Business Intelligence (IBM 312)

Sumit Kumar Yadav

Department of Management Studies

January 27, 2023

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??

Two ways of defining -

the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X <= x),$$

or the *probability density function* (pdf):

$$\rho(x)\, dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \to 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} \rho(t)\, dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??
Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X <= x),$$

or the *probability density function* (pdf):

$$\rho(x)\, dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \to 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} \rho(t)\, dt.$$

Continuous Random Variables

If a random variable X can take on any of a continuum of values, say, any value between 0 and 1, then we cannot define it by listing values x_i and giving the probability p_i that $X = x_i$; Why??
Two ways of defining -
the *cumulative distribution function*:

$$F(x) \equiv \text{Prob}(X <= x),$$

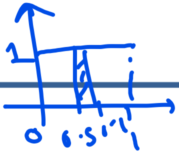
or the *probability density function* (pdf):

$$\rho(x)\, dx \equiv \text{Prob}(X \in [x, x + dx]) = F(x + dx) - F(x).$$

Letting $dx \to 0$, we find

$$\rho(x) = F'(x), \quad F(x) = \int_{-\infty}^{x} \rho(t)\, dt.$$

Expected Value



The expected value of a continuous random variable X is then defined by

$$E(X) = \int_{-\infty}^{\infty} x\underline{\rho(x)}\,dx.$$

Note that by definition, $\boxed{\int_{-\infty}^{\infty} \rho(x)\,dx = 1.}$ The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2\rho(x)\,dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

Expected Value

The expected value of a continuous random variable X is then defined by

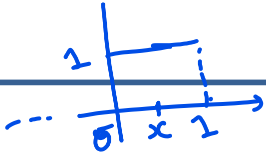
$$E(X) = \int_{-\infty}^{\infty} x\rho(x)\,dx.$$

Note that by definition, $\int_{-\infty}^{\infty} \rho(x)\,dx = 1$. The expected value of X^2 is

$$E(X^2) = \int_{-\infty}^{\infty} x^2\rho(x)\,dx,$$

and the variance is again defined as $E(X^2) - (E(X))^2$.

Uniform Distribution



$X \sim$

Example: Uniform Distribution in $[0, 1]$.

$P(X \leq x)$

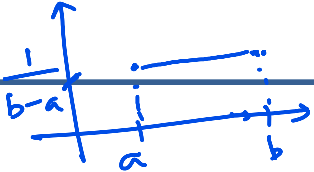
$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\rho(x)\,dx = \int_0^1 x\,dx = \frac{1}{2},$$

$$\mathrm{var}(X) = \int_0^1 x^2\,dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Uniform Distribution

$[a, b]$



Example: Uniform Distribution in $[0, 1]$.

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ x & \text{if } 0 \leq x \leq 1 \\ 1 & \text{if } x > 1 \end{cases}, \quad \rho(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } 0 \leq x \leq 1 \\ 0 & \text{if } x > 1 \end{cases}$$

$$E(X) = \int_{-\infty}^{\infty} x\rho(x)\, dx = \int_0^1 x\, dx = \frac{1}{2},$$

$$\frac{a+b}{2}$$

$$\operatorname{var}(X) = \int_0^1 x^2\, dx - \left(\frac{1}{2}\right)^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

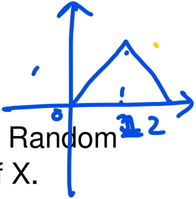
$$\frac{(b-a)^2}{12}$$

Practice Problem

$Z = X + Y$ $\quad (0,2)$

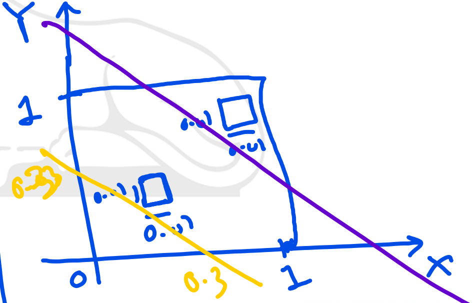
$P(Z \le z) = \frac{1}{2} z^2, \quad 0 < z < 1$

$P(Z \le z) = 1 - \frac{1}{2}(2-z)^2, \quad z > 1$



Random Variable X follows uniform distribution from [0,1], Random Variable Y follows same distribution and is independent of X. What is the distribution of X+Y?

$$f(z) = \begin{cases} z ; & 0 < z \le 1 \\ 2 - z ; & 1 < z \le 2 \end{cases}$$



Normal Distribution

$X \sim \text{Normal}(10, 25)$

Eg.– $\mu = 10, \sigma^2 = 25$

Example: Normal (Gaussian) Distribution, Mean μ, Variance σ^2.

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$\left[F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt \right]$$

Why is this weird density called normal?

$P(0 < X < 20) = ?? \; P(X < 20) - P(X < 0)$

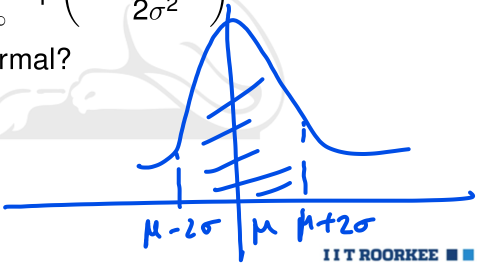
Normal Distribution

Example: Normal (Gaussian) Distribution, Mean μ, Variance σ^2.

$$\rho(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right),$$

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right) dt$$

Why is this weird density called normal?



Central Limit Theorem

Theorem

Let $\{X_k\}$ be a sequence of n mutually independent random variables having a common distribution, and mean (μ) and variance (σ^2) exists. Assuming, n to be large, the average of these random variables $\overline{X}$ follows approximately normal distribution with

1. *mean* $= \mu$
2. *variance* $= \frac{\sigma^2}{n}$

What is meant by large n? Typically, $n \geq 30$

Central Limit Theorem - Special Case

Theorem

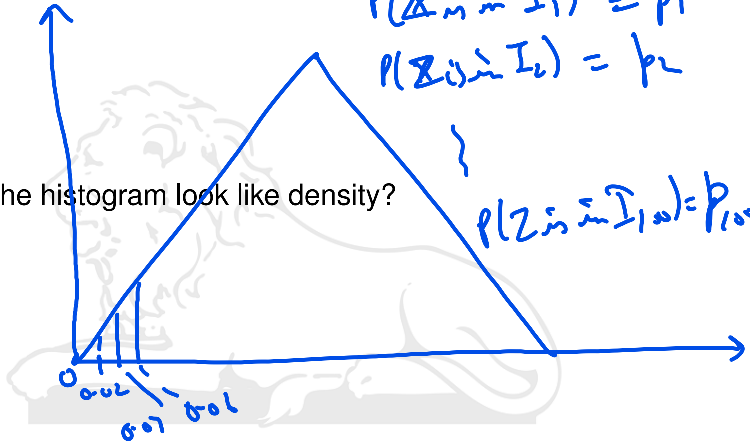
If the sample size is large, for WITH REPLACEMENT and independent sampling, the sample mean $\overline{X}$ is approximately normal with

1. *mean* = μ
2. *variance* = $\frac{\sigma^2}{n}$

What is meant by large n? Typically, $n \geq 30$

Simulation of Random numbers in Python

Why should the histogram look like density?



$P(Z \text{ is in } I_1) = p_1$

$P(Z \text{ is in } I_2) = p_2$

$\vdots$

$P(Z \text{ is in } I_{100}) = p_{100}$

Simulation of Random numbers in Python

Inverse Transform Method??