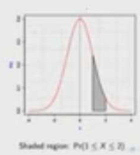
RANDOM VARIABLES AND PROBABILITY MASS/DENSITY FUNCTIONS

Random Variable

- A random variable (RV) is a map from sample space to a real line such that there is a unique real number corresponding to every outcome of sample space
  - eg. Coin toss sample space [H T] mapped to [0 1]. If the sample space outcomes are real valued no need for this mapping (eg. throw of a dice)
  - Allows numerical computations such as finding expected value of a RV
  - Discrete RV (throw of a dice or coin)
  - Continuous RV (sensor readings, time interval between failures)
  - Associated with the RV is also a probability measure

Probability Mass/Density Functions

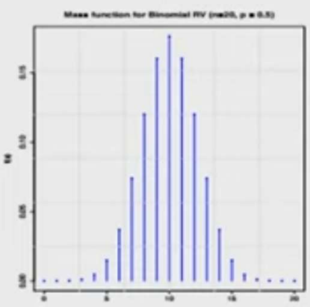
- For a discrete RV the probability mass function assigns a probability to every outcome in sample space
  - Sample space of RV (x) for a coin toss experiment: [0 1].
  - $P(x = 0) = 0.5$; $P(x = 1) = 0.5$
- For a continuous RV the probability density function $f(x)$ can be used to assign a probability to every interval on a real line
  - Continuous RV (x) can take any value $[-\infty, \infty]$
  - $P(a < x < b) = \int_a^b f(x)dx$ (Area under the curve)
  - Cumulative density function F(X)



$$F(b) = P(-\infty < x < b) = \int_{-\infty}^{b} f(x)dx$$

Binomial Mass Function

- Probability of obtaining k heads in n coin tosses with p the probability of obtaining a head in any toss
- RV x represents number of heads obtained
  - Sample space : [0, 1, …n]
  - $f(x = k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$
  - One outcome: $\underbrace{HH\ldots H}_{k \text{ times}}\underbrace{TT\ldots T}_{n-k \text{ times}}$
  - PMF characterized by one parameter p
  - For large n it tends to a Gaussian distribution



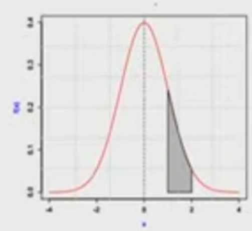
Binomial mass function for n = 20, p = 0.5

Gaussian or Normal Density Function

- Distribution used to characterize random errors in data

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - PDF characterized by two parameters μ and σ
  - Density function is symmetric
  - Standard normal distribution $\mu = 0$ and $\sigma = 1$



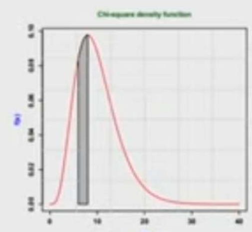
Shaded region: $\Pr(1 \leq X \leq 2)$

Gaussian density function for $\mu = 0$, $\sigma = 1$

GyanData Private Limited

Chi-square density function

- $f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma(\frac{n}{2})} x^{n/2-1} e^{-x/2}$
- Density is characterized by parameter n (degrees of freedom)
- Distribution of sum of squares of n independent standard normal RVs
- Distribution of sample variance



Shaded region: $\Pr(6 \leq X \leq 8)$

Moments of a pdf

- Similar to describing a function using derivatives, a pdf can be described by its moments
  - For continuous distributions
    - $E[x^k] = \int_{-\infty}^{\infty} x^k f(x)dx$
  - For discrete distributions
    - $E[x^k] = \sum_{i=1}^{N} x_i^k p(x_i)$
- Mean : $\mu = E[x]$
- Variance : $\sigma^2 = E[(x-\mu)^2] = \mathrm{E}[x^2] - \mu^2$
- Standard deviation = Square root of variance = σ

Properties of Gaussian RVs

- For a Gaussian RV x
  - Mean : $E[x] = \mu$
  - Variance : $E[(x-\mu)^2] = \sigma^2$
  - Symbolically $x \sim \mathcal{N}(\mu, \sigma^2)$
- Standard Gaussian RV $z \sim \mathcal{N}(0,1)$
- If $x \sim \mathcal{N}(\mu, \sigma^2)$ and $y = ax + b$ then
  - $y \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
- Standardization
  - If $x \sim \mathcal{N}(\mu, \sigma^2)$, then $z = \frac{(x-\mu)}{\sigma} \sim \mathcal{N}(0,1)$

Computation of Probability using R

- Function to compute probability given a value X
- Lower tail probability = $P(-\infty < x < X) = \int_{-\infty}^{X} f(x)dx$
- Functions p*norm*(X, mean, std, 'lower.tail' = TRUE/FALSE)
  - *norm* refers to the distribution and can be replaced by other distributions (chisq, exp, unif)
  - X is the value (limit)
  - Parameters of the distribution (eg. mean and std for normal distribution)
  - lower.tail = TRUE (default) to obtain lower tail probability and FALSE to obtain upper tail probability

Other functions in R

- Function to compute X given probability p
  - Function q*norm*(p, mean, std, 'lower.tail' = TRUE/FALSE)
  - Lower tail probability = $P(-\infty < x < X) = \int_{-\infty}^{X} f(x)dx = p$
- Function d*norm* to compute density function value
- Function r*norm* to generate random numbers from the distribution

Joint pdf of two RVs

- Joint pdf of two RVS x and y: f(x,y)
  - $P(x \leq a, y \leq b) = \int_{-\infty}^{b} \int_{-\infty}^{a} f(x,y)\,dx\,dy$
  - Covariance between x and y: $\sigma_{xy} = E[(x - \mu_x)(y - \mu_y)]$
  - Correlation between x and y: $\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$
- Two RVs x and y are uncorrelated if $\sigma_{xy} = 0$
- Two RVs x and y are independent if f(x,y) = f(x)f(y)

GyanData Private Limited

Multivariate Normal Distribution

- A vector of RVs $\mathbf{x} = [x_1 \quad x_2 \quad \cdots \quad x_n]^T$
- Multivariate Gaussian Distribution : $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
  - $E[\mathbf{x}] = \boldsymbol{\mu}$: Mean vector
  - $E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T] = \boldsymbol{\Sigma}$: Variance-covariance matrix
  - $f(\mathbf{x}) = \frac{1}{(2\pi)^{n/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}$: pdf
- Structure of Σ

$$\Sigma = \begin{bmatrix} \sigma_{x_1}^2 & \sigma_{x_1 x_2} & \cdots & \sigma_{x_1 x_n} \\ \sigma_{x_2 x_1} & \sigma_{x_2}^2 & \cdots & \\ \vdots & \vdots & \vdots & \\ \sigma_{x_n x_1} & \cdots & \cdots & \sigma_{x_n}^2 \end{bmatrix}$$