DS203: Programming in Data Science
IE605: Engineering Statistics

Introduction to Probability and Statistics
Lecture 02

Manjesh K. Hanawal

21th August 2020

Previous Lecture:

- Sample Space and Events
- Axioms of probability
- Conditional probability
- Independence of probability
- Baye's formula

This Lecture:

- Random Variable (RVs)
- Discrete and Continuous RVs
- Cumulative density functions (CDFs)
- Probability Density functions (PDFs)
- Examples of discrete RVs
- Examples of Continuous RVs

Random Variables

In most experiments we would be interested in some function of outcomes and not the out come itself.

- Example 1: Tossing two coins We may be interested in the number of heads appeared. If we want at least one head, (H, H) and (H, T) are same
- Example 2: Rolling of two dice We may be interested in sum of the two outcomes and of the value of outcomes. If we want the sum to be 6 all $(1, 5), (5, 1), (2, 4), (4, 2), (3, 3)$ are same
- Example 3: Marks. You may be interested in what grades/points you receive and not the exact marks you score.

> 90	AA(10)
75-90	AB(9)
65-75	BB(8)

Random Variable Contd..

Roughly, random variable (X) is a real function on sample space

$$X : \Omega \to \mathbb{R}$$

Note: Formal definition of RV requires its inverse map to be measurable, but we do not go into this!

Example: Consider repeated throw of a coin. Your interest is in the number of tosses it takes to get head for the first time. How do you define a random value?

Ω	X
(H)	1
(T,H)	2
(T,T,H)	3
(T,T,T,H)	4
$\vdots$	$\vdots$

Probability of Random Variable

For any point $x \in \mathbb{R}$ and subset $\mathcal{A} \in \mathcal{R}$

- $\{X = x\} = \{w \in \Omega : X(w) = x\} \subset \Omega$
- $\{X \in \mathcal{A}\} = \{w \in \Omega : X(w) \in \mathcal{A}\} \subset \Omega$.

We can assign probabilities to these events.

- $P(\{X = x\}) = P_X(x)$
- $P(\{X \in \mathcal{A}\}) = P_X(\mathcal{A})$.

Example: Rolling two dice: Let X is the random variable which denotes the sum of the outcomes.

- $P_X(5) =$
- $P_X(\{4, 5\}) =$

Discrete vs Continuous RVs

Possible values taken by a random variable can be finite, countable or uncountable values.

- Discrete RV: Values taken are finite or countable
  - Sum of outcomes in rolling of two dice . $X \in \{1, 2, 3, 4, 5, 6\}$
  - Number of tosses till head appears. $X \in \{1, 2, 3, \ldots, \}$
- Continuous RV: Values taken are uncountable (to be made precise!)
  - Temperature of a room in Mumbai. $X \in [0, 40]$
  - Height of a person in cms. $X \in [50, 200]$
  - Price of a share. $X \in [p_{\min}, p_{\max}]$ for some $a < b$.

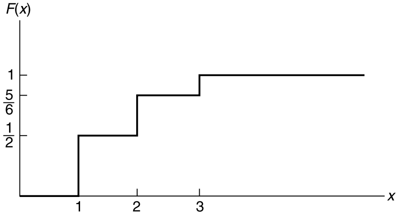
Cumulative Density function (CDF)

- CDF of a random variable X is a function $F_X : \mathbb{R} \to [0, 1]$, defined for any $x \in \mathbb{R}$ as

$$F_X(x) = P_X((-\infty, x]) = P(X \leq x).$$

- $F_X(x)$ denotes the probability that random variables takes value less than or equal to x

Example A random variable X takes values $1, 2, 3$ with probabilities $P_X(1) = \frac{1}{2}, P_X(2) = \frac{1}{3}, P_X(3) = \frac{1}{6}$



$$F_X(x) = \begin{cases} 0 & \text{if} \quad x < 1 \\ 1/2 & \text{if} \quad 1 \leq x < 2 \\ 5/6 & \text{if} \quad 2 \leq x < 3 \\ 1 & \text{if} \quad 3 \leq x \end{cases}$$

Properties of CDF

Properties of CDF: For any random variable X

- $F_x(x)$ is non-decreasing in x
- $\lim_{x\to\infty} F_X(x) = 1$
- $\lim_{x\to-\infty} F_X(x) = 0$
- $F_X(\cdot)$ is right continuous

Sketch:

- for any $x < y$. $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$
- X is finite! All values included as $X \to \infty$. None as $X \to -\infty$

All probability question about X can be answered from its CDF

- $P(x < X \leq y) = P(X \leq y) - P(X \leq x) = F(y) - F(x)$
- $P(X < x) = \lim_{h\to 0^+} F(x-h)$. ($h$ is decreasing to 0).
- $P(X < x)$ need not be equal to $P(X \leq x) = F(x)$ (right continuous!).

PMF and PDF

Probability Mass Function (PMF) of a Discrete RV

- Let discrete random varaible takes values $\{x_1, x_2, x_3, \ldots\}$
- $\{P(x_i), i = 1, 2, \ldots\}$ is called PMF of X. $\sum_i P(x_i) = 1$.
- $P(x_i)$ is the mass assigned to point x_i

Probability Density Function (PDF)

> Random variable X is continuous if there exists a non-negative function $f_X : \mathbb{R} \to \mathbb{R}_+$ such that for any $\mathcal{A} \in \mathbb{R}$
>
> $$P_X(\mathcal{A}) = \int_{x \in A} f_X(x) dx.$$

f_X is called the PDF function of X. Properties of PDF:

- $f_x(\cdot)$ is such that $\int_{-\infty}^{\infty} f_X(x) dx = P_X(X \in (-\infty, \infty)) = 1$
- $\mathcal{A} = [a, b]$, $P(a \leq X \leq b) = \int_a^b f_X(x) dx$.
- If $a = b$, $P(X = a) = \int_a^a f_X(x) dx = 0$. Probability that a continuous random value assuming a particular value is zero!

PDF properties continued

- $F_X(x) = P(X \leq x) = \int_{\infty}^{x} F_X(x)dx \implies \frac{d}{dx}F_X(x) = f(x)$
- $\mathcal{A} = \{a - \epsilon/2, a + \epsilon/2\}$ for some small $\epsilon > 0$
  $P(a + \epsilon/2 \leq X \leq a + \epsilon/2) = \int_{a-\epsilon/2}^{a+\epsilon/2} f_X(x)dx \sim \epsilon f_X(a)$. $f_X(x)$ is a measure of how likely random variable X will be near x.

Commonly Used Distributions

Discrete RVs

- Bernoulli
- Geometric
- Binomial
- Poisson
- Hypergeometric

Continuous RVs:

- Uniform
- Exponential
- Gaussian
- Rayleigh
- Gamma

Discrete RVs

Bernoulli, $X \sim Ber(p), p \in [0, 1]$

- X takes binary values, i,e., $\{0, 1\}$
- PMF: $P(X = 1) = p$ and $P(X = 0) = 1 - p$
- Examples: coin toss, any experiments involving binary values

Binomial, $X \sim Bin(n, p), p \in (0, 1], n \in \mathbb{N}$

- X takes value in $\{0, 1, 2, 3, \ldots, n\}$
- PMF: $P(X = i) = \binom{n}{i} p^i (1 - p)^{n-i}$, for $0 \leq i \leq n$
- Examples: Number of success in independent trials. What is the probability that 3 samples are classified correctly out of 5?

Discrete RVs Contd...

Geometric, $X \sim Geo(p), p \in (0,1]$

- X takes value in $\{1, 2, 3, 4, \ldots\}$
- PMF: $P(X = i) = (1-p)^{i-1}p$ for all $i \geq 1$
- Examples: Number of trials till success in independent trials. How many times I invest till profit is made?

Poisson, $X \sim Poi(\lambda), \lambda \geq 0$

- X takes value in $\{0, 1, 2, 3, 4, \ldots\}$
- PMF: $P(X = i) = \frac{e^{\lambda}\lambda^i}{i!}$ for all $i \geq 0$
- Examples: Used for counting. How many people visited a mall/airport/cinema today? How many cars on road today?

Continuous RVs

Uniform, $X \sim Unif(a, b), a, b \in \mathbb{R}$

- X takes value in $[a, b]$
- $$f_X(x) = \begin{cases} 1/(b-a) & \text{if } x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$
- Example: Height, weight, temperature. Often used when we do not have prior information.

Exponential, $X \sim Exp(\lambda), \lambda > 0$

- X takes value in $[0, \infty)$
- $$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$
- Example: Used to model life times. Time before a bulb fails. Time before the next customer/item arrives.

Continuous RVs contd.

Gaussian, $X \sim \mathcal{N}(\mu, \sigma), \mu \in \mathbb{R}, \sigma^2 > 0$

- X takes value in $(-\infty, \infty)$
- PDF:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\{-(x-\mu)^2/2\sigma^2\}, \text{ for } x \in (-\infty, \infty)$$

- Examples: Error and Noise modeling.

Rayleigh, $X \sim \mathit{Rayleigh}(\sigma^2), \sigma > 0$

- X takes value in $(0, \infty)$
- PDF:

$$f_X(x) = \begin{cases} (x/\sigma^2)\exp\{-r^2/(2\sigma^2)\} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

- Example: Envelop of noise. $X_1 \sim \mathcal{N}(0, \sigma^2)$ and $X_2 \sim \mathcal{N}(0, \sigma^2)$, Then $X = \sqrt{X_1^2 + X_2^2} \sim \mathit{Rayleigh}(\sigma^2)$, under some conditions (independence)

Other distributions

- Uniform distribution on finite set of elements
- Gamma (rainfall accumulated in a reservoir)
- Weibull (reliability and survival analysis)
- Laplace (speech recognition to model priors on DFT)

Expectation and Variances

Expectation: Many times, instead of actual value of experiment, we would be interested in expected/average/mean value. Expectation of random variable X is denoted as $E(X)$.

Discrete random variable X	Continuous random variable X
PMF $\{P_X(x_i), i = 1, 2, \ldots\}$	PDF f_X
$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$	$E(X) = \sum_{i=1}^{\infty} x_i P_X(x_i)$

Variance: How value of random variable varies around its mean. We measure variance, denoted $Var(X)$, as

$$Var(X) = E\left[(X - E(X))^2\right] = \begin{cases} \sum_{i=1}^{\infty} (x_i - E(X))^2 P_X(x_i) & \text{discrete} \\ \int_{-\infty}^{\infty} (x - E(X))^2 f_X(x) dx & \text{continuous} \end{cases}$$

Summary of Expectation and Variance of Distributions

Random Variable $X \sim$	Mean $E[X]$	Variance $Var(X)$
$Ber(p)$	p	$p(1-p)$
$Bin(n,p)$	np	$np(1-p)$
$Geo(n,p)$	$1/p$	$(1-p)/p^2$
$Poi(\lambda)$	λ	λ
$Uni(a,b)$	$(a+b)/2$	$(b-a)^2/12$
$Exp(\lambda)$	$1/\lambda$	$1/\lambda^2$
$\mathcal{N}(\mu,\sigma^2)$	μ	σ^2
$Rayleigh(\sigma^2)$	$\sigma\sqrt{\pi/2}$	$\sigma^2(1-\pi/2)$
$Gamma(n,\alpha)$	n/α	n/α^2