Lecture Slides for

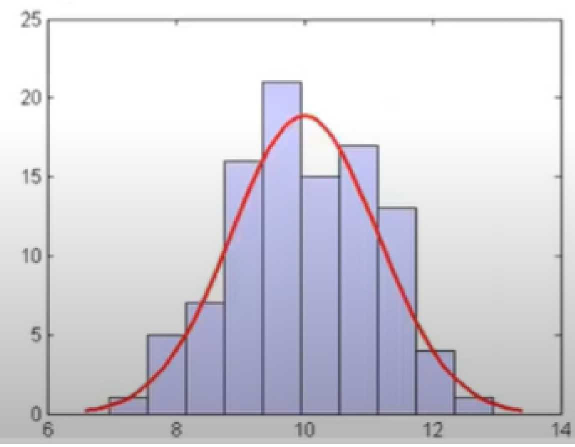
INTRODUCTION TO DATA ANALYTICS: RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Dr. LALIT KUMAR SINGH

Probability distributions

Probability distributions

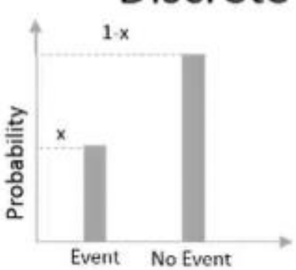
- Why do we need to talk about probability distributions. What does it have to do with Data?
- Remember the histogram?

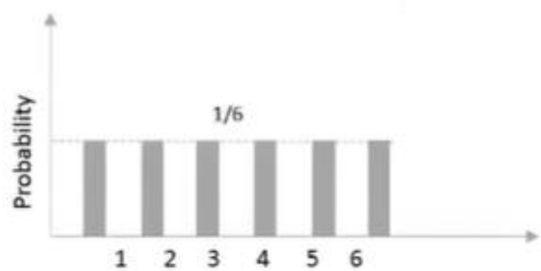


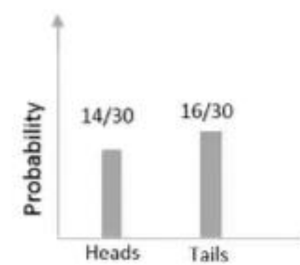
Random Variables

Random Variables

- Random Variable: A variable whose value is subject to variations due to randomness.
- The mathematical function describing this randomness (the probabilities for the set of possible values a random variable can take) is called a probability distribution.
- Continuous and Discrete probability density functions
  - Discrete



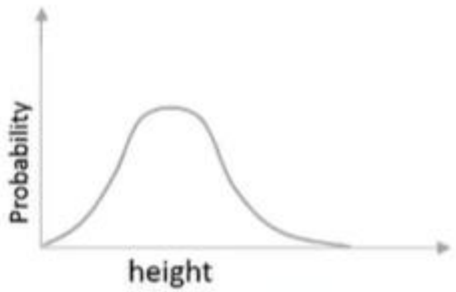




Random Variables

Random variables

- Continuous Distributions



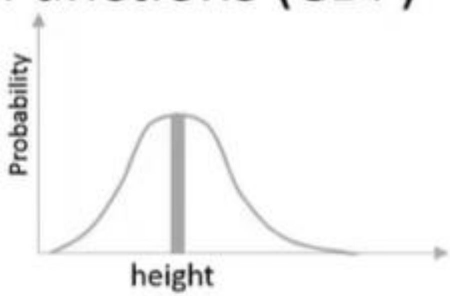
- Probability of certain height

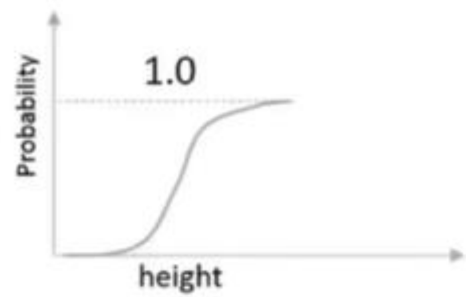
- Total Probability of all outcomes

Random Variables

Random Variables

- Probability Density functions (PDFs) and Cumulative Density Functions (CDF)





PDF

CDF

- Going from PDF to CDF and vice versa

Common distributions

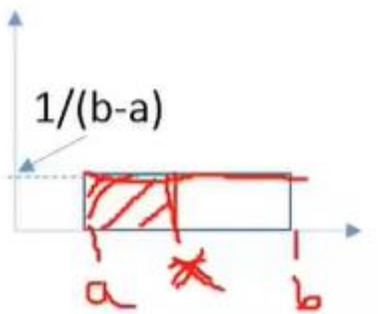
- Uniform
  - Discrete
    - The six sided dice, coin toss
    - Formula for pdf: $f(X=x)=\frac{1}{k}$ for all x that belongs to a specific set with k elements

    And $f(X=x)=0$ for all other values of x.
  - Continuous
    - Number of seconds past the minute
    - Exact age of a randomly selected person between the ages of 50-60
    - Formula for PDF:

$$f(x)=\begin{cases}\frac{1}{b-a} & \text{for } a \le x \le b \\ 0 & \text{for } x<a \text{ and } x>b\end{cases}$$

    - What is the CDF, mean and Variance?

CDF = $\frac{x-a}{b-a}$



$\le b$

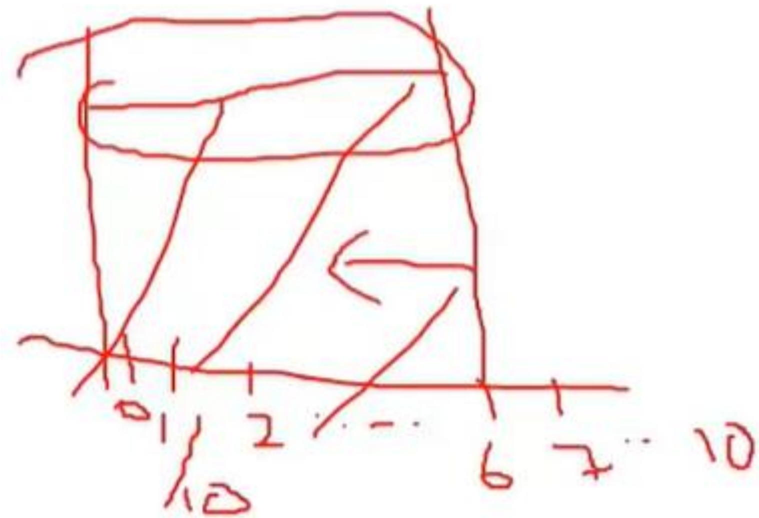
$$\frac{x-a}{b-a}$$

Mean = $\frac{1}{2}(b+a)$

Variance = $\frac{1}{12}(b-a)^2$

Common distributions

- Binomial
  - What is it + Example: Toy problem
  - Example Real-world: Probability of 3 out of 10 mergers. Probability of there being 5 defective products in a batch of 20.
  - Formula for PMF: $\binom{n}{k} p^{k}(1-p)^{n-k}$
  - Formula for CDF is just the summation
  - It is more useful for small n’s
  - Mean: np, variance: np(1-p)



Common distributions

- Poisson
  - Discrete distribution that signifies the probability of ‘x’ occurrences of a certain event over a certain period of time or space.
  - Examples: Number of defaults per month, Number of banks per square kilometre.
  - PMF (not PDF) $\frac{\lambda^{k}}{k!} e^{-\lambda}$
  - Mean and variance are λ (lambda >0)

Common distributions

- Geometric
  - Number of attempts before an event
  - The interarrival distribution counterpart of a binomial. The coin toss case (uniform, binomial, geometric)
  - PMF $(1-p)^{k-1}p$
  - CDF $1-(1-p)^{k}$
  - Mean is $\frac{1}{p}$, and variance $\frac{1-p}{p^2}$

Common distributions

- Exponential
  - The interarrival times of the Poisson distribution
  - The continuous version of the geometric distribution
  - Memoryless
  - PDF: $\lambda e^{-\lambda x}$, where lambda>0
  - CDF: 1-$e^{-\lambda x}$
  - Mean: $\frac{1}{\lambda}$
  - Variance: $\frac{1}{\lambda^2}$

Common distributions

- Parallels to the Binomial, Exponential, Geometric

	Interarival Distribution	Count per unit interarrival distribution
Discrete Interarrival	Geometric	Binomial
Continuous interarrival	Exponential	Poisson

	Continuous Distribution
	Discrete Distribution

Working with distributions

- Going from PDF to CDF (continuous)

$$F(x) = \int_{-\infty}^{x} f(x)\, dx$$

- Going from CDF to PDF (continuous)

$$f(x) = \frac{d}{dx} F(x)$$

- Mean

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} \qquad E[x] = \sum_{i=1}^{\infty} p_i x_i \qquad E[x] = \int_{-\infty}^{\infty} x f(x) dx$$

- Variance/Standard deviation

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{n} (x_i - \bar{x})^2} \qquad \sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x - \mu)^2\, dx =} \qquad \sigma = \sqrt{\int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2}$$