

Foundation of Analytics: Lecture 2

Dihui Lai

dlai@wustl.edu

September 13, 2020

CONTENT

- Discrete Random Variable
- Continuous Random Variable
- Demo

Example 1: Roll a dice

There six possible outcome of rolling a dice i.e. "1", "2", "3", ... "6".

- If I roll a dice 60 times, how many times do you get "1"?
- What is the probability of getting "1"? $1/6$?

Dice-Rolling: Expectation, Variance etc.

- Unique values: 1, 2, 3, 4, 5, 6
- min:1; max:6
- expectation: $\frac{\sum x_i}{N}$?
- variance: ?

Random Variable

A **random variable** X can take different values with certain probability. To understand a random variable, we need to consider two things:

- The possible outcome value of an experiment: x
- The probability of an outcome is x : $P(x)$.

Discrete Random Variable

A discrete random variable X

- Can take k possible values $x_1, x_2, x_3 \dots x_k$
- Each with probability of $p_1, p_2, p_3 \dots p_k$. For simplicity, we denote the probabilities using a probability mass function

$$P(x_i) = p_i, i = 1, 2, 3, \dots k$$

- The probabilities for all possible values sum up to be 1, i.e.
$$\sum_{i=1}^k p_i = 1$$

Histogram/Distribution of Discrete Variables

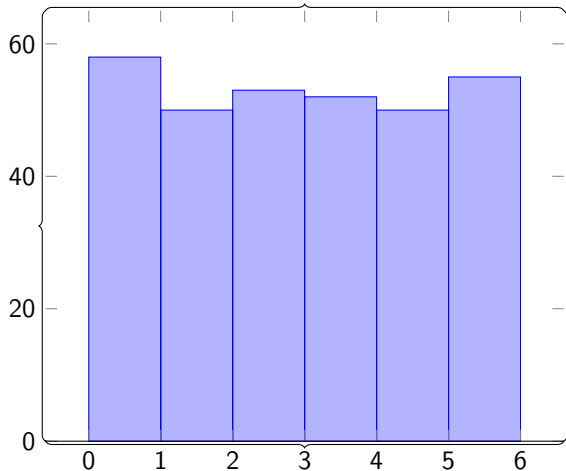
- For data of discrete values e.g. $x^T = [x^1, x^2, x^3 \dots, x^n]$,
- Find the unique value of the data
- Count the number of data occurred at each discrete value N_i . The total number of data points $N = \sum_i N_i$. The empirical probability mass function can be estimated by

$$P(x_i) = \frac{N_i}{N}$$

Dice-Rolling: Distribution

Given a series of data, $[1, 2, 1, 3, 4, 6, 6, 4, 5, 5, \dots]$. What can you tell about the underlying story? Is it from a dice-rolling process?

Count the number of occurrence for each value 1, 2, 3, 4, 5, 6



Discrete Random Variable: Function and Expectation

The expectation of a function, $g(X)$ is given by

$$E[g(X)] = \sum_{i=1}^k g(x_i)P(x_i)$$

Discrete Random Variable: Expectation, Variance and Moments

In special case when $g(X) = X^n$, we have the n^{th} raw moment of X

$$E[X^n] = \sum_{i=1}^k x_i^n P(x_i)$$

The expectation of X is the 1^{st} raw moment of X

$$\mu = E[X]$$

The variance of X is the 2^{nd} moment of X about the expectation

$$\sigma^2 = E[(X - \mu)^2]$$

Bernoulli Distribution

Consider a random variable X that can take value 1 with probability p and 0 with probability $1 - p$.

$$P(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

The expectation of X is

$$E[X] = p$$

The variance of X is

$$E[(X - \mu)^2] = p(1 - p)$$

Joint Distribution and Algebra of Random Variables

If we create a random variable from two random variables:

$$Z = X + Y$$

Distribution: $f(z)$

Expectation:

$$E[Z] = E[X + Y] = E[X] + E[Y] = E[Y] + E[X]$$

Variance:

$$Var[Z] = Var[X + Y] = Var[X] + 2Cov[X, Y] + Var[Y]$$

Covariance is defined as

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])]$$

Algebra of Multiple Random Variables

Expectation:

$$E\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n E(X_i)$$

Variance:

$$Var\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n Var(X_i) + \sum_{i,j} Cov(X_i, X_j)$$

Binomial Distribution

Consider a random event whose outcome is the summation of n independent Bernoulli distribution. The distribution of the corresponding random variable X can be described as

$$P(x) = \binom{n}{k} p^x q^{n-x}, x = 0, 1, 2, \dots, n$$

The expectation of X is

$$E[X] = np$$

The variance of X is

$$E[(X - \mu)^2] = np(1 - p)$$

Poisson Distribution

Consider a random event, the probability of 1 occurrence within a unit time is p . What's the probability distribution of events occurrence within time interval of τ (e.g. No. of car accidents occurs in a day in MO).

The distribution of the discrete random variable X is

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!}, x = 0, 1, 2, \dots$$

The expectation of X is

$$E[X] = \lambda$$

λ is the average number of events per interval, $\lambda = p\tau$

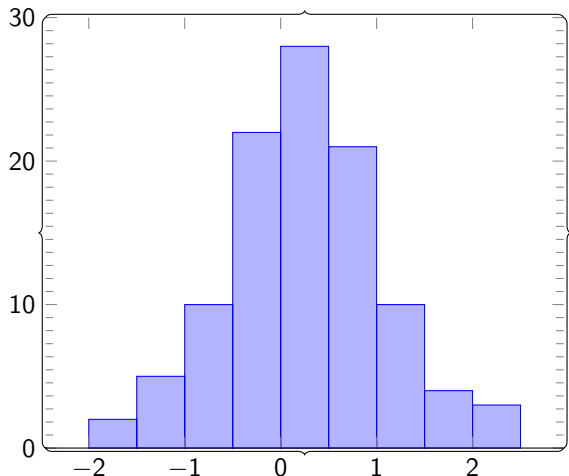
The variance of X is

$$E[(X - \mu)^2] = \lambda$$

Understand the Distribution of Continuous Data

How about real value data, $[-1.407, 0.412, -1.198, 1.552, \dots]$?

Count the number of data points that falls into the intervals of $[-2, -1.5), [-1.5, -1.0), \dots, [0, 0.5), [0.5, 1) \dots$



Histogram/Empirical Distribution of Continuous Variables

- For data of discrete values, count the number of data occurred at each discrete value N_i . The total number of data points $N = \sum_i N_i$. The empirical probability mass function is given by

$$P(x_i) = \frac{N_i}{N}$$

- For data of continuous values, define k equal-sized-bins (e.g. $[x_i - \Delta x, x_i + \Delta x), i = 1, 2, 3, \dots, k$). Count the number of data belong to each bin N_i , the total number of data points $N = \sum_i^k N_i$. The empirical probability distribution is given by

$$P(x_i - \Delta x \leq x < x_i + \Delta x) = \frac{N_i}{N}$$

Continuos Random Variable

A continous random variable X

- Can have a range of values e.g. $(-\infty, +\infty)$, $[0, 1)$, $[0, +\infty)$
- The probability that $a \leq x \leq b$ is defined as

$$P(a \leq x \leq b) = \int_a^b f(x)dx$$

where $f(x)$ is the probability density function. Note: $f(x)$ is not probability

- The pdf $f(x)$ has to satisfy the following property

$$P(-\infty \leq x \leq +\infty) = \int_{-\infty}^{+\infty} f(x)dx = 1$$

Continuos Random Variable: Function and Expectation

If we denote a function of a random variable as $g(X)$, the expectation of $g(X)$ is given by

$$E[g(X)] = \int_{-\infty}^{+\infty} g(x)f(x)dx$$

Continuos Random Variable: Expectation, Variance and Moments

In a special case, when $g(X) = X^n$, the expectation of $g(X)$ is called the n^{th} raw moment of X

$$E[X^n] = \int_{-\infty}^{+\infty} x^n f(x) dx$$

The expectation of X is the 1^{st} raw moment of X

$$\mu = E[X]$$

The variance of X is the 2^{nd} moment of X about the expectation

$$\sigma^2 = E[(X - \mu)^2]$$

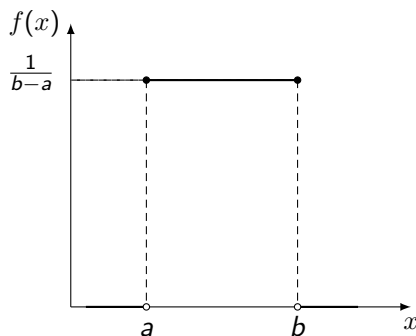
Calculate Expectation using Empirical Probability Distribution

$$\begin{aligned}
 \mu &= \sum_{i=1}^k x_i f(x_i - \Delta x \leq x < x_i + \Delta x)(2\Delta x) \\
 &= \sum_{i=1}^k x_i P(x_i - \Delta x \leq x < x_i + \Delta x) \\
 &= \frac{\sum_{i=1}^k x_i N_i}{N}
 \end{aligned}$$

Uniform Distribution

A uniform distribution is given by

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



Gaussian Distribution

A continuous random variable Z is called a standard normal if

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

The probability of $z \leq z_0$ is given by

$$P(Z \leq z_0) = \int_{-\infty}^{z_0} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

Let $X = \mu + \sigma Z$. Then X is a normal distribution with parameters μ and σ^2 . Its density function is given by

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

The expectation of X : $E[X] = \mu$

The variance of X : $E[(X - \mu)^2] = \sigma^2$