

# Basic Statistics and Random Process

Dihui Lai

dlai@wustl.edu

March 3, 2019

## CONTENT

- Introduction to Statistics: Random Variable
- Empirical View of Random Variable
- Common Probability Distributions
- Random Walk, i.i.d and Central Limit Theorem

# Introduction to Statistics: Random Variable

### 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"?  $1/6$ ?

### Example 2: Life time of a light bulb

A light bulb can go broken while use. The longer it is used, the more likely the bulb will break.

- The value does not need to be an integer, it can be 120 hours, 120.513 hours
- What's the probability of a light bulb breaks at the 100th hour?  
 $1/100$ ?

# 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 that an outcome is  $x$ .

# 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

$$P(x) = p_x, x = x_1, x_2, x_3, \dots$$

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

# Discrete Random Variable: Function and Expected Value

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

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



# Discrete Random Variable: Mean, 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 mean 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 mean

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

# 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 Expected Value

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

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

# Continuos Random Variable: Mean, Variance and Moments

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

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

The mean 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 mean

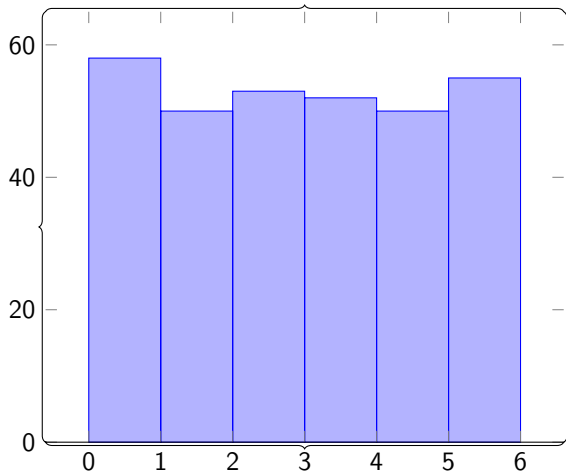
$$\sigma^2 = E[(X - \mu)^2]$$

## Empirical View of Random Variable

# Understand the Distribution of Discrete Data

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?

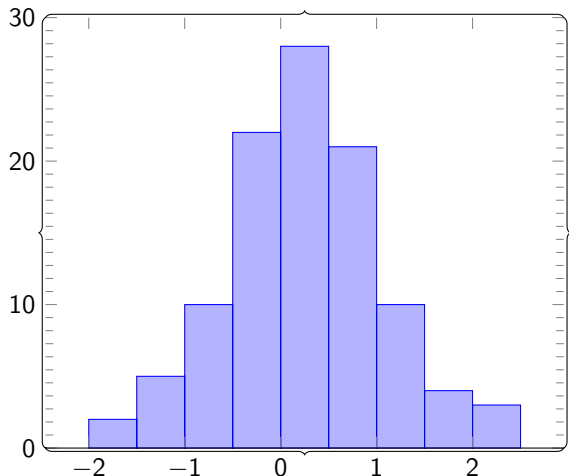
Count the number of occurrence for each value 1, 2, 3, 4, 5, 6



# 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 and Empirical Probability Distribution

- 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

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



# Calculate Mean using Empirical Probability Distribution

- Discrete random variable:

$$\mu = \sum_{i=1}^k x_i P(x_i) = \frac{\sum_{i=1}^k x_i N_i}{N}$$

This last term of the equation is the same as the arithmetic mean of the data points

# Calculate Mean using Empirical Probability Distribution

- Continuous random variable:

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

This last term of the equation is the approximately arithmetic mean of the data points because  $x_i n_i \approx \sum_{d \in [x_i - \Delta x, x_i + \Delta x)} d$ ,  $d \in [x_i - \Delta x, x_i + \Delta x)$  denotes the data points belong to the bin  $[x_i - \Delta x, x_i + \Delta x)$

# Histogram and Empirical Probability Distribution

Demo in Python

# Statistical Description of Data

logitude	latitude	houseAge	medianHouseValue	oceanProx
-122.23	37.88	41	452600.0	NEAR BAY
-122.22	37.86	21	358500.0	NEAR BAY
-122.24	37.85	52	352100.0	NEAR BAY
...	...	...		

Is "medianHouseValue" a random variable?

## Common Probability Distribution

# 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 mean of  $X$  is

$$E[X] = p$$

The variance of  $X$  is also

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

# Binomial Distribution

Consider a random event that is composed of  $n$  independent experiments, whose outcome could either be success (1) or failure (0). The probability of success is  $p$  and failure  $1 - p$ . 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 mean of  $X$  is

$$E[X] = np$$

The variance of  $X$  is also

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

# Poisson Distribution

Consider a random event, the number of occurrence within a given interval can be  $x = 0, 1, 2, 3 \dots$  (e.g. No. of car accidents occurs in a day in MO). The distribution of the discrete random variable  $X$  can be described as

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

The mean of  $X$  is

$$E[X] = \lambda$$

The variance of  $X$  is also

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



# Poisson Distribution

Consider a random event, the number of occurrence within a given interval can be  $x = 0, 1, 2, 3 \dots$  (e.g. No. of car accidents occurs in a day in MO). The distribution of the discrete random variable  $X$  can be described as

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

The mean of  $X$  is

$$E[X] = \lambda$$

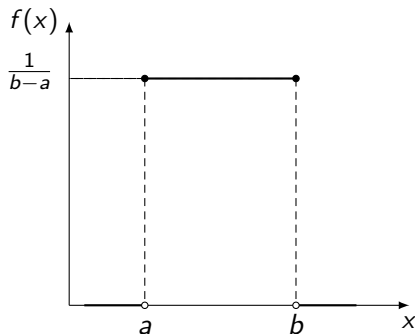
The variance of  $X$  is also

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

# 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  $\mu$  and  $\sigma^2$ . Its density function is given by

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

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

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

## Random Walk, i.i.d, Central Limit Theorem

# Random Walk

One-dimension random walk: a random process

- Starting at 0
- The movement at each step could be either  $+1$  or  $-1$ , of equal probability

# Random Walk

Define a random variable  $B$  that can take value  $+1$  or  $-1$  and have the following random distribution

$$P(b) = \begin{cases} 0.5 & \text{if } b = 1 \\ 0.5 & \text{if } b = -1 \end{cases}$$

The position of a random walk at  $t^{th}$  step is a random variable given by

$$Z_t = \sum_{i=1}^t B_i$$

# Sum of Independent and Identically Distributed (i.i.d)

Suppose we have  $n$  independent random variable  $X_1, X_2, X_3 \dots X_n$ , each have the same probability distribution. We say  $X_1, X_2, X_3 \dots X_n$  are independent and identically distributed (i.i.d). The sum of i.i.d random variables given by  $Z(n) = \sum_{i=1}^n X_i$  has the following properties

- $E[Z(n)] = n\mu$ ;
- $Var[Z(n)] = n\sigma^2$ ;
- When  $n$  is large, distribution of  $Z(n)/n$  is close to the normal distribution of mean  $\mu$  and variance  $\sigma^2/n$

Here,  $\mu$  and  $\sigma^2$  are the mean and variance of  $X_i$ , respectively.

# Random Walk

Since  $Z_t$  is the sum of i.i.d, of mean 0 and variance 1, when  $t$  is large,  $Z_t$  becomes a normal distribution of mean 0 and variance  $t$  (why?)