

**STATISTICS CONCEPTS USED IN THE DOMAIN OF  
DATA SCIENCE AND MACHINE LEARNING**

XAVIER TANG

2020

# Contents

<b>1</b>	<b>Distribution</b>	<b>1</b>
1.1	Probability Mass Function (PMF)	1
1.2	Probability Density Function (PDF)	2
1.3	Cumulative Distributed Function (CDF)	2
1.4	Bernoulli Trials	2
1.5	Discrete variable: Binomial Distribution	3
1.6	Discrete variable: Bernoulli distribution	4
1.7	Discrete variable: Geometric Distribution	4
1.8	Discrete variable: Poisson Distribution	4
1.9	Continuous variable: Normal/Gaussian Distribution	4
1.10	Continuous variable: Exponential Distribution	4
<b>2</b>	<b>Metrics</b>	<b>5</b>
2.1	ROC	5
2.2	Standard Deviation	5
2.3	Variance	5
2.4	Confusion Matrix	5
2.4.1	Recall	5
2.4.2	Precision	5
2.4.3	F1	5

---

2.5	P value	5
2.6	T value	5
2.7	Z value	5
2.8	Correlation/Pearson	5
2.9	Covariance	5
<b>3</b>	<b>Testing</b>	<b>6</b>
3.1	T test	7
3.2	Chi-square test	7
3.3	Z test	7
3.4	A/B testing	7
3.4.1	Null hypothesis	7
3.4.2	Alternative hypothesis	7
3.4.3	Type I/II error	7
3.4.4	Statistical Power	7
3.5	Test of Significance	7
3.6	Hypothesis testing (one-way and two-way)	7
3.7	ANOVA	7
3.8	ANCOVA	7
3.9	One-sample/Two-sample bootstrap hypothesis test	7
3.10	Time series: p, d, q parameters, unit root and box test	7
<b>4</b>	<b>Thoerem</b>	<b>8</b>
4.1	Central Limit Theorem	8

---

4.2	Law of the large number	8
4.3	Naive Bayes Algorithm	8
4.4	Bayesian Statistics/Bayes Theorem	8
4.5	Sampling Theory	8
<b>5</b>	<b>General</b>	<b>9</b>
5.1	Confidence Interval	9
5.2	Conditional Probability	9
5.3	Normalisation	9
5.4	Standardisatio	9
5.5	Least-squared error	9
5.6	R-squared error	9
5.7	Mean-squared error	9
5.8	Inferential Statistics	9
5.9	Bias-variance trade off	9

# CHAPTER 1

## Distribution

This chapter deals with concepts mainly related to various probability distribution.

### 1.1 Probability Mass Function (PMF)

The Probability Mass Function (PMF) gives the set of probabilities of discrete outcome, e.g. discrete uniform PMF: roll one dice, each outcome is  $1/6$ .

More formally, a probability mass function (PMF) is a function that gives the probability that a discrete random variable is exactly equal to some value. In Eqn. (1.1) below, the PMF gives the probability of getting exactly  $k$  (discrete) successful Bernoulli trials.

Formal definition:

PMF is the probability distribution of a discrete random variable, and provides the possible values and associated probabilities. The probabilities associated with each possible values must be positive and sum up to 1. For all other values, the probabilities need to be 0.

## 1.2 Probability Density Function (PDF)

## 1.3 Cumulative Distributed Function (CDF)

## 1.4 Bernoulli Trials

This is a random experiment with exactly 2 possible outcomes. The probability of success is the same every time experiment is conducted. A similar analogy: Flipping a (possibly) biased coin, each coin has probability  $p$  of landing heads (success) and probability  $1 - p$  of landing tails (failure).

Closely related to a Bernoulli trial is a binomial experiment, which consists of a fixed number  $n$  of statistically independent Bernoulli trials, each with a probability of success  $p$ , and counts the number of success. The number  $k$  of success in  $n$  Bernoulli trials is Binomially distributed. The probability of exactly  $k$  success (out of  $n$ ) is given by the probability mass function:

$$P(k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (1.1)$$

where  $\binom{n}{k}$  is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \quad (1.2)$$

Example:

Let's say a bank made 100 mortgage loans. It is possible that anywhere between 0 and 100 of the loans will be defaulted upon. You would like to know the probability of getting a given number of defaults, given that the probability of a default is  $p = 0.05$ . To investigate this, you will do a simulation. You will perform 100 Bernoulli trials. Here, a success is a default. (Remember that the word 'success' just means that the Bernoulli trial evaluates to be True, i.e., did the loan recipient default?) You will do this for another 100 Bernoulli trials. And again and again until we have tried it 1000 times. Then, you will plot a histogram

describing the probability of the number of defaults. So we have performed 1000 times of 100 mortgage trials, the histogram should show a maximum at about 5 (because  $0.05 \times 100$ ) and probability is given in Eqn. (1.1).

Example:

Consider the simple experiment where a fair coin is tossed four times. Find the probability that exactly two of the tosses result in heads.

$$\begin{aligned} P(2) &= \binom{4}{2} p^2 (1-p)^{4-2} \\ &= 6 \times (0.5)^2 \times (0.5)^2 \\ &= \frac{3}{8} \end{aligned}$$

When multiple Bernoulli trials are performed, each with its own probability of success, these are sometimes referred to as Poisson trials.

## 1.5 Discrete variable: Binomial Distribution

The Binomial Distribution with parameters  $n$  and  $p$ , denoted  $\text{Bin}(n, p)$  is the discrete probability distribution of the number of successes in a sequence of  $n$  independent experiments, each asking a yes–no question, and each with its own boolean-valued outcome True (with probability  $p$ ), or failure (with probability  $q = 1 - p$ ). A single success or failure experiment is also called a Bernoulli trial.

The binomial distribution is frequently used to model the number of successes in a sample of size  $n$  drawn with replacement from a population of size  $N$ . If the sampling is carried out without replacement, then the draws are not independent and so the resulting distribution would be a hypergeometric distribution, not a binomial one. However, for  $N$  much larger than  $n$ , the binomial distribution remains a good approximation, and is widely used.

The PMF of Binomial Distribution is given in Eqn.(1.1). The CDF of Binomial Distribution can be expressed as:

$$\sum_{i=0}^k \binom{n}{i} p^i (1-p)^{n-i} \quad (1.3)$$

where we just add up all the probability for all the previous  $k$  values.

## 1.6 Discrete variable: Bernoulli distribution

The Bernoulli distribution is a special case of the binomial distribution, where a single trial is conducted (i.e., a binomial distribution with  $n = 1$ ). The PMF of Bernoulli distribution, over possible outcomes  $j$  is:

$$\begin{cases} p, \text{ if } j = 1 \\ q = 1 - p, \text{ if } j = 0 \end{cases} \quad (1.4)$$

## 1.7 Discrete variable: Geometric Distribution

## 1.8 Discrete variable: Poisson Distribution

## 1.9 Continuous variable: Normal/Gaussian Distribution

## 1.10 Continuous variable: Exponential Distribution



# CHAPTER 2

## Metrics

### 2.1 ROC

### 2.2 Standard Deviation

### 2.3 Variance

### 2.4 Confusion Matrix

#### 2.4.1 Recall

#### 2.4.2 Precision

#### 2.4.3 F1

### 2.5 P value

### 2.6 T value

### 2.7 Z value

### 2.8 Correlation/Pearson

### 2.9 Covariance

# CHAPTER 3

## Testing

### 3.1 T test

### 3.2 Chi-square test

### 3.3 Z test

### 3.4 A/B testing

#### 3.4.1 Null hypothesis

#### 3.4.2 Alternative hypothesis

#### 3.4.3 Type I/II error

#### 3.4.4 Statistical Power

### 3.5 Test of Significance

### 3.6 Hypothesis testing (one-way and two-way)

### 3.7 ANOVA

### 3.8 ANCOVA

### 3.9 One-sample/Two-sample bootstrap hypothesis test

### 3.10 Time series: p, d, q parameters, unit root and box test

# CHAPTER 4

## Thorem

4.1 Central Limit Theorem

4.2 Law of the large number

4.3 Naive Bayes Algorithm

4.4 Bayesian Statistics/Bayes Theorem

4.5 Sampling Theory

# CHAPTER 5

## General

5.1 Confidence Interval

5.2 Conditional Probability

5.3 Normalisation

5.4 Standardisation

5.5 Least-squared error

5.6 R-squared error

5.7 Mean-squared error

5.8 Inferential Statistics

5.9 Bias-variance trade off