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

XAVIER TANG

2020

Contents

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. (1.3) below, the PMF gives the probability of getting exactly k (discrete) successful Bernoulli trails.

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.

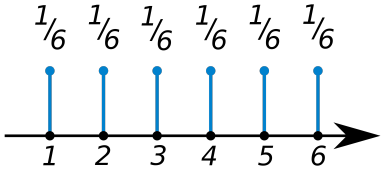

Figure 1.1: The probability mass function of a fair die. All the values of this function must be non-negative and sum up to 1

1.2 Probability Density Function (PDF)

The Probability Density Function (PDF) of a continuous random variable, is a function whose value at any given sample (or point) in the sample space (the set of possible values taken by the random variable) can be interpreted as providing a relative likelihood that the value of the random variable would equal that sample.

In a more precise sense, the PDF is used to specify the probability of the random variable falling within a particular range of values, as opposed to taking on any one value. This probability is given by the integral of this variable's PDF over that range—that is, it is given by the area under the density function but above the horizontal axis and between the lowest and greatest values of the range. The probability density function is nonnegative everywhere, and its integral over the entire space is equal to 1. A PDF must be integrated over an interval to yield a probability, which is different from PMF (other than continuous VS discrete random variable).

More formally, the PDF is most commonly associated with absolutely continuous univariate distribution, a random variable X has PDF f_X and the probability of this variable taking values between a and b, i.e. $a \leq X \leq b$ will be:

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx \tag{1.1}$$

If F_X is the cumulative distribution function (CDF) of X, then:

$$F_X(x) = \int_{-\infty}^{x} f_X(u)du \tag{1.2}$$

1.3 Cumulative Distribution Function (CDF)

The Cumulative Distribution Function (CDF) of a real-valued random variable X (continuous or discrete), evaluated at x, is the probability that X will take a value less than or equal to

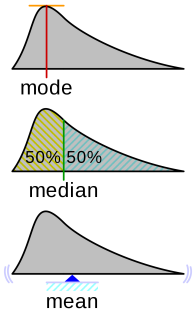

Figure 1.2: Geometric visualisation of the mode, median and mean of an arbitrary probability density function.

x. In the case of a scalar continuous distribution, it gives the area under the probability density function (PDF) from $-\infty$ to x.

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 trails, 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 (PMF):

$$f(k,n,p) = P(k;n.p) = P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.3}$$

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

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.4}$$

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 0.05×100) and probability is given in Eqn. (1.3).

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.(1.3). The CDF of Binomial Distribution can be expressed as:

$$F(k; n, p) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \tag{1.5}$$

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:

$$f(k; p) = \begin{cases} p, \text{if } j = 1 \\ q = 1 - p, \text{if } j = 0 \end{cases} \tag{1.6}$$

1.7 Discrete variable: Geometric Distribution

The Geometric Distribution is either of two discrete probability distribution described below:

1) The probability distribution of the number X of Bernoulli Trials needed to get one success, supported on the set {1, 2, 3, ... }.

This geometric distribution, sometimes denoted $Geo(p)$, gives the probability that the first occurrence of success requires k independent trials, each with success probability (Bernoulli Trial). The Probability Mass Function (PMF) is (probability that the kth trial (out of k trials) is the first success):

$$P(X = k) = (1 - p)^{k-1}p, k = 1, 2, 3, ... \tag{1.7}$$

and the Cumulative Distribution Function (CDF) is:

$$P(X \leq k) = 1 - (1 - p)^k \tag{1.8}$$

This distribution is used for modelling the number of trials up to and including the first success.

2) The probability distribution of the number $Y = X - 1$ of failures before the first success, supported on the set {0, 1, 2, 4, ... }.

This form of geometric distribution is used for modelling the number of failures until the first success:

$$P(X = k) = (1 - p)^k p, k = 0, 1, 2, 3, ... \tag{1.9}$$

and the Cumulative Distribution Function (CDF) is:

$$P(X \leq k) = 1 - (1 - p)^{k+1} \tag{1.10}$$

The geometric distribution is an appropriate model if the following assumption is true:

1) The phenomenon being modelled is a sequence of independent trials.

2) There are only two possible outcomes for each trial.

3) The probability of success p is the same for each trial.

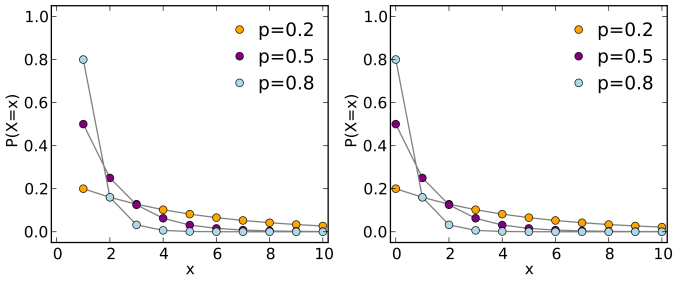

Figure 1.3: Probability Mass Function (PMF) of geometric distribution for case 1) (left) and case 2) (right).

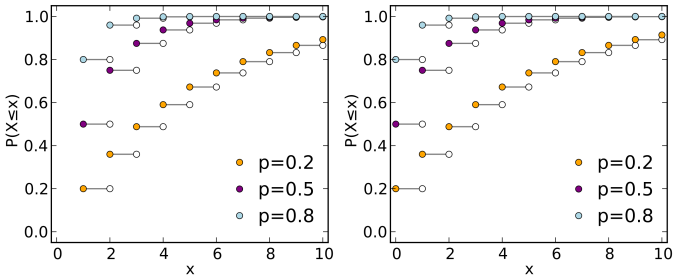

Figure 1.4: Cumulative Distribution Function (CDF) of geometric distribution for case 1) (left) and case 2) (right).

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

Thoerem

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 Standardisatio

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