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

XAVIER TANG

Ver 0.1

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.

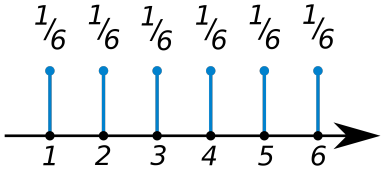

Figure 1.1: The probability mass function of a fair die. All the values of this function must be non-negative and sum up to 1

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

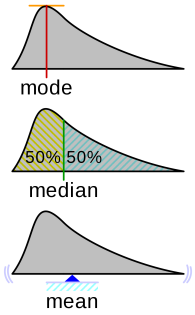

Figure 1.2: Geometric visualisation of the mode, median and mean of an arbitrary probability density function.

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, Bin(n, p)

The Binomial Distribution with parameters n and p, denoted 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, Ber(p)

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, Geo(p)

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.

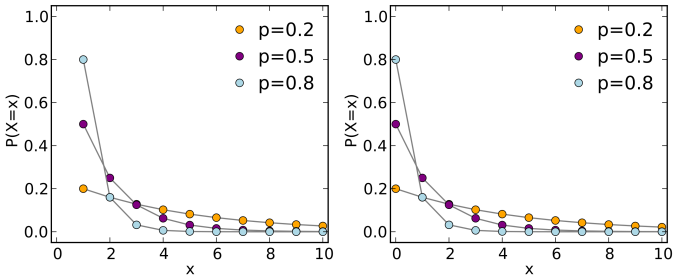

Figure 1.3: Probability Mass Function (PMF) of geometric distribution for case 1) (left) and case 2) (right).

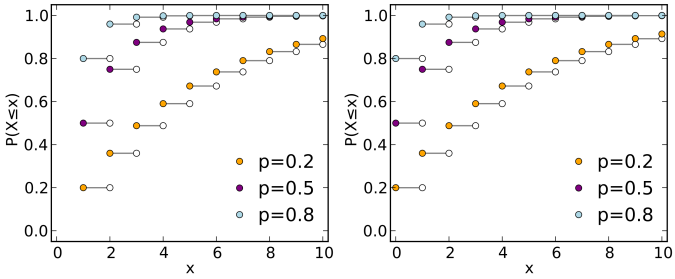

Figure 1.4: Cumulative Distribution Function (CDF) of geometric distribution for case 1) (left) and case 2) (right).

1.8 Discrete variable: Poisson Distribution, Pois(λ)

The Poisson Distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of times or space if these events occur with a known constant mean rate and independently of the time since the last event.

For instance, an individual keeping track of the amount of mail they receive each day may notice that they receive an average number of 4 letters per day. If receiving any particular piece of mail does not affect the arrival times of future pieces of mail, i.e., if pieces of mail from a wide range of sources arrive independently of one another, then a reasonable assumption is that the number of pieces of mail received in a day obeys a Poisson distribution.

The Poisson Distribution is popular for modelling the number of times an event occurs in an interval of time or space. A discrete random variable X is said to have a Poisson Distribution with parameter $\lambda > 0$, if, for $k = 0, 1, 2, ...$, the Probability Mass Function (PMF) of X is given by:

$$f(k; \lambda) = P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} = P(\text{k events in interval}) \tag{1.11}$$

where λ is equal to the expected value of X and also its variance.

The Cumulative Distribution Functions (CDF) is:

$$F(k; \lambda) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!} \tag{1.12}$$

The Poisson Distribution is an appropriate model if the following assumptions are true:
1) k is the number of times an event occurs in an interval and k can take values 0, 1, 2, ...
2) The occurrence of one event does not affect the probability that a second event will occur. That is, events occur independently.
3) The average rate at which events occur is independent of any occurrences. For simplicity, this is usually assumed to be constant, but may in practice vary with time.
4) Two events cannot occur at exactly the same instant; instead, at each very small sub-interval exactly one event either occurs or does not occur.

The Poisson Distribution is also the limit of a binomial distribution, , for which the probability of success p for each trial equals to $\frac{\lambda}{\text{num. of trials}}$, as the number of trials goes to infinity.
If the number of Bernoulli trials goes to infinity (or every large), then Binomial distribution can be converted into Poisson distribution. (p will be small, it is a rare event.)

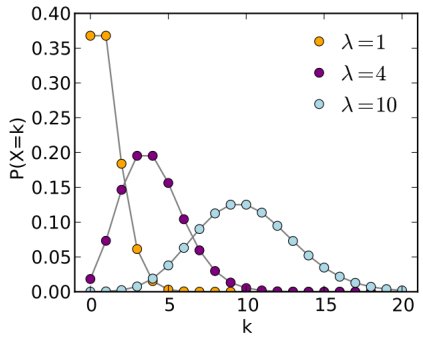

Figure 1.5: Probability Mass Function (PMF) of Poisson distribution.

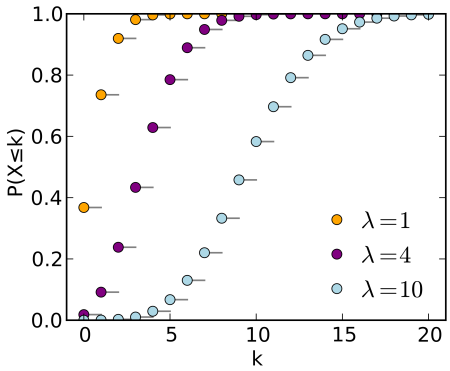

Figure 1.6: Cumulative Distribution Function (CDF) of Poisson distribution.

1.9 Continuous variable: Normal Distribution, $\mathcal{N}(\mu, \sigma^2)$

A Normal (Gaussian) distribution is a type of continuous probability distribution for a real-valued random variable. The general form of its Probability Density Function (PDF) is: The Cumulative Distribution Functions (CDF) is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{1.13}$$

where μ is the mean or expectation of the distribution (and also its median and mode), while σ is the standard deviation. The variance of the distribution is σ^2. Note that these μ and σ is associated with Normal Distribution and not the one computed directly from the data. Note also that Normal Distribution are affected by outliers because probability of any event happen more than 4 standard deviations from the mean is very low.

Normal distributions are often used in the natural and social sciences to represent real-valued random variables whose distributions are not known. Their importance is partly due to the Central Limit Theorem (CLT). CLT establishes that in some situations, when independent random variables are added, their properly normalized sum tends toward a normal distribution (informally a bell curve), even if the original variables themselves are not normally distributed.

The Cumulative Distribution Function (CDF) for Normal Distribution is:

$$F(x) = \frac{1}{2}[1 + \text{erf}\,(\frac{x-\mu}{\sigma\sqrt{2}})] \tag{1.14}$$

where erf is the error function defined as:

$$\text{erf}\; z = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt \tag{1.15}$$

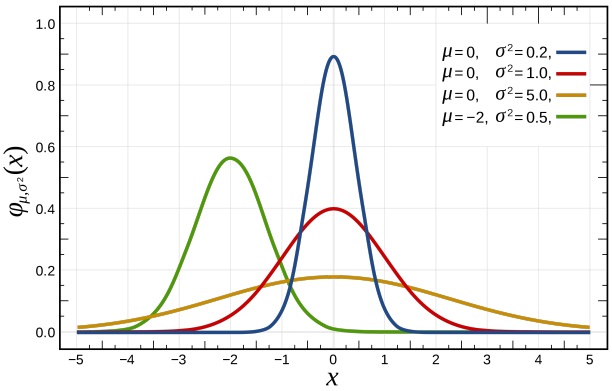

Figure 1.7: Probability Density Function (PDF) of Normal distribution.

About 68% of values drawn from a normal distribution are within one standard deviation σ away from the mean μ. About 95% of the values lie within two standard deviations and about 99.7% are within three standard deviations. This is known as the 68-95-99.7 rule.

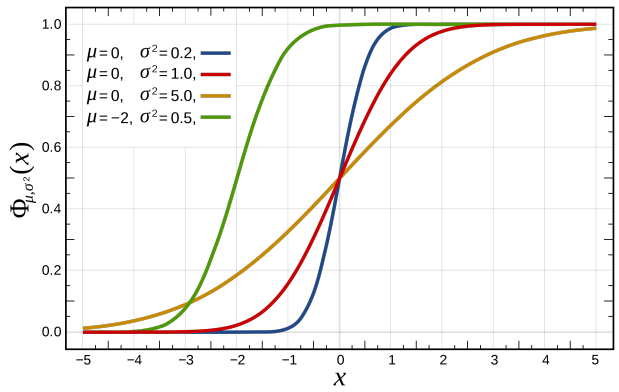

Figure 1.8: Cumulative Distribution Function (CDF) of Normal distribution.

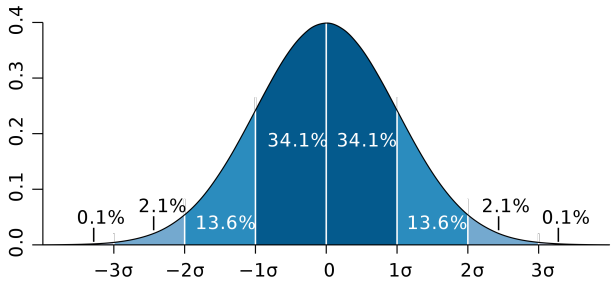

Figure 1.9: Normal distribution with standard deviation

1.10 Continuous variable: Exponential Distribution, Exp(λ)

Poisson process:

the timing of the next event is completely independent of when the previous event happened (memoryless).

The number of arrivals of a Poisson process in a given amount of time is Poisson Distributed.

The waiting time between arrivals of a Poisson process is Exponentially Distributed.

The Probability Density Function (PDF) of an exponential distribution is:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{1.16}$$

where λ is the rate parameter used in the Poisson Distribution.

The Cumulative Distribution Function (CDF) is given by:

$$F(x;\lambda) = \begin{cases} 1 - e^{-\lambda x}, x \geq 0 \\ 0, x < 0 \end{cases} \tag{1.17}$$

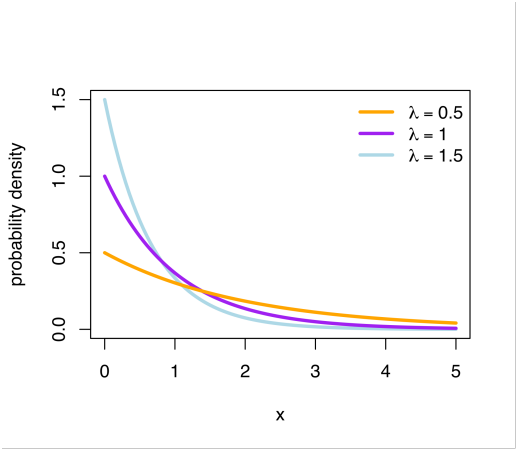

Figure 1.10: Probability Density Function (PDF) of Exponential Distribution.

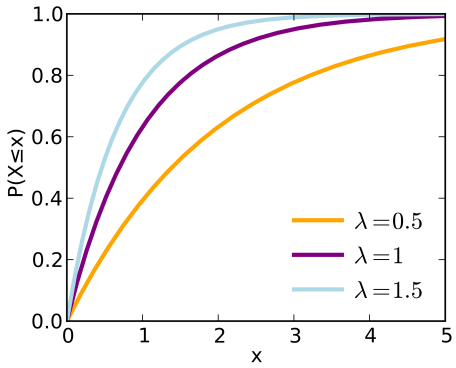

Figure 1.11: Cumulative Distribution Function (CDF)of Exponential Distribution

Chapter 2

Metrics

2.1 Variance, σ^2 or Var(X)

Variance is the expectation of the squared deviation of a random variable from its mean. Note that the unit of variance is the square of the variable's unit.

1) discrete random variable

This is for discrete random variable (applicable to most dataset), if the generator of random variable X is discrete with Probability Mass Function (PMF) that maps value x_i to probability p_i, (certain x_i and p_i pairs can be same value due to same x value in the dataset) then the variance will be:

$$Var(X) = \sum_{i=1}^{n} p_i(x_i - \mu)^2 \tag{2.1}$$

where μ is the expected value:

$$\mu = \sum_{i=1}^{n} p_i x_i \tag{2.2}$$

If the each value in the n data points are equally likely, then the variance will be:

$$Var(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{2.3}$$

2) continuous random variable

If the random variable X has a Probability Density Function (PDF) $f(x)$, then the variance will be:

$$Var(X) = \int_{\mathbb{R}} (x - \mu)^2 f(x) dx \tag{2.4}$$

where μ is the expected value:

$$\mu = \int_{\mathbb{R}} x f(x) dx \tag{2.5}$$

2.2 Standard Deviation, σ

The standard deviation is a measure of the amount of variation or dispersion of a set of values. For both discrete and continuous random variable, the standard deviations are $\sqrt{\text{variance}}$ and note that this quantity has the same physical units as the random variable.

2.3 Standard Error

The standard error is a type of standard deviation for the distribution of the means.

There will be, of course, different means for different samples (from the same population), this is called "sampling distribution of the mean". This variance between the means of different samples can be estimated by the standard deviation of this sampling distribution and it is the standard error of the estimate of the mean. Standard error measures the precision of the estimate of the sample mean. The standard error is strictly dependent on the sample size and thus the standard error falls as the sample size increases. It makes total sense if you think about it, the bigger the sample, the closer the sample mean is to the population mean and thus the estimate of it is closer to the actual value.

$$\text{Standard Error} = \frac{\sigma}{\sqrt{n}} \tag{2.6}$$

where σ is the standard deviation of the population (although sometimes population standard deviation is unknown, we can replace it with sample standard deviation as an estimate) and n is the size (number of observations) of the sample.

2.4 Confusion Matrix

The confusion matrix is a table specifically for the problem of statistical classification. A typical confusion matrix table is shown below:

		Actual class	
		P	N
Predicted class	P	TP	FP
	N	FN	TN

Figure 2.1: Confusion Matrix

where:

TP : True Positive or Hit

TN : True Negative or Correct Rejection

FP : False Positive or Type I error

FN : False Negative or Type II error

A more graphical way of seeing this is shown in Fig. (2.2).

2.4.1 Recall

Recall is also called Sensitivity or True Positive Rate (TPR), it measures how many relevant items are selected (among the total numbers of relevant elements):

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2.7}$$

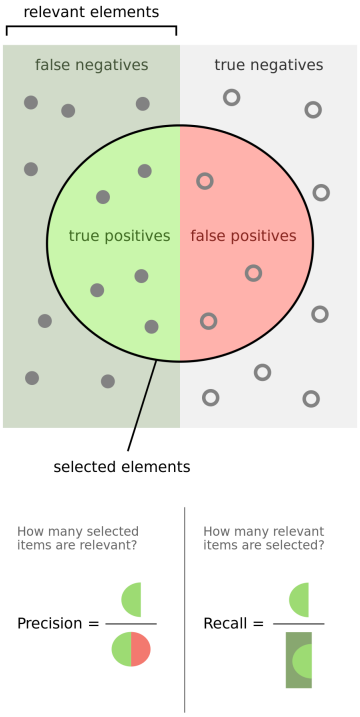

Figure 2.2: Graphical representation of confusion matrix

2.4.2 Precision

Precision is also called Positive Predictive Value (PPV), it measures how many selected items are relevant:

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{2.8}$$

2.4.3 F_1

The F_1 score is a measure of a test's accuracy. It is basically the harmonic mean of the precision and recall. The highest possible F_1 score is 1, indicating perfect precision and recall, and the lowest possible score is 0, if either precision or recall is 0.

$$F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{2.9}$$

2.5 ROC

2.6 Correlation/Pearson

2.7 Covariance

2.8 Confidence Interval

2.9 Confidence Level

2.10 P value

2.11 T value

2.12 Z value

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