Probability Theory

SUZUKI, Atsushi

Discrete random variables

Definition

A random variable taking a value randomly in a discrete subset [1] of $\mathbb{R}$ (the set of real numbers) is called a ***discrete random variable***.

[1] Strictly speaking, “discrete” stands for “at most countable.” Here, we say a set is at most countable if and only if there exists a surjective map from the set of integers to the set.

Probability mass function (PMF)

When we consider a univariate discrete random variable taking a value in a discrete set $\mathscr{X} = \{x_1, x_2, \dots\} \subset \mathbb{R}$, we can completely understand the behaviour of X by knowing the probability of X taking a value x, where $x \in \mathscr{X}$. Hence, we define a function describing those probabilities.

Definition (probability mass function)

Let X be a discrete random variable taking a value in a discrete set $\mathscr{X} \subset \mathbb{R}$. We define the ***probability mass function (PMF)*** $P_X : \mathscr{X} \to [0, 1]$ of the random variable X by

$$P_X(x) := \Pr(X = x). \tag{1}$$

Multiple random variables

Example

- The prices of multiple stocks.
- The pixels of an image.
- The values at each time frame in a wave file.

When we consider multiple random variables, knowing each probability mass function is not sufficient to know their stochastic behaviour completely.

Knowing multiple random variables $\neq$ knowing multiple PMFs

If we have two discrete random variables X and Y, then just knowing each probability mass function is not sufficient.
Rather, what we need to know is the distribution of the **pair** (X, Y), which is called the ***joint distribution*** of the random variables X and Y.

Joint distribution and marginal distribution

In general, the ***joint distribution*** refers to the distribution of the tuple of multiple random variables. For example, if we have two random variables X and Y, the joint distribution refers to the distribution of the pair (X, Y). In contrast, when we consider multiple random variables, the distribution of a single random variable is called the ***marginal distribution*** of the random variable to distinguish it from the joint distribution.

Joint probability mass function (two variable cases)

If we have two discrete random variables X and Y, then just knowing each probability mass function is not sufficient. Rather, what we need to know is the probability of the pair (X, Y) taking every pair of values $(x, y) \in \mathcal{X} \times \mathcal{Y}$. That is, the following ***joint probability mass function (joint PMF)*** has all the information that we need.

Definition (two-variable Joint PMF)

Let X and Y be discrete random variables taking a value in discrete sets $\mathcal{X}$ and $\mathcal{Y}$, respectively, where $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$. We define the ***joint probability mass function (joint PMF)*** $P_{X,Y} : \mathcal{X} \times \mathcal{Y} \to [0, 1]$ of the pair of random variables X, Y by

$$P_{X,Y}(x, y) := \Pr(X = x, Y = y). \tag{2}$$

Joint probability mass function (general cases)

If we have m discrete random variables $X_1, X_2, \ldots, X_m$, then all we need to know is the following joint PMF.

Definition (Joint PMF (general cases))

Let $X_1, X_2, \ldots, X_m$ be discrete random variables taking a value in discrete sets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_m \subset \mathbb{R}$, respectively. We define the ***joint probability mass function (joint PMF)*** $P_{X_1, X_2, \ldots, X_m} : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m \to [0, 1]$ of random variables $X_1, X_2, \ldots, X_m$ by

$$P_{X_1, X_2, \ldots, X_m}(x_1, x_2, \ldots, x_m) := \Pr(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m). \quad (3)$$

Marginal PMF (two variable cases)

The joint PMF can tell us the PMFs of each discrete random variable, called ***marginal PMF***. For two discrete random variables X and Y that takes a value in $\mathscr{X}$ and $\mathscr{Y}$, respectively, suppose that the joint PMF is $P_{X,Y} : \mathscr{X} \times \mathscr{Y} \to [0,1]$. Then, the marginal PMFs P_X and P_Y are given by

$$P_X(x) = \sum_{y \in \mathscr{Y}} P_{X,Y}(x,y), \quad P_Y(y) = \sum_{x \in \mathscr{X}} P_{X,Y}(x,y), \tag{4}$$

Conditional distribution

If two random variables are “related,” then we get more precise information about a random variable’s distribution by knowing the value of the other random variable.
The ***conditional distribution*** is one of such

Histogram can represent a discrete random variable

Here, the value of any bin is nonnegative, and the sum of the values of all the bins is 1.

Summary statistics

Motivation: A probability mass function might have too much information to understand the behaviour of a random variable intuitively. Hence, we often want to calculate a single value (or a few values) that describes a distribution, called a ***descriptive statistic*** or ***summary statistic***[2].

[2] These words are often used to distinguish them from inferential statistics.

Summary statistics: examples

Central tendency measures give a representative value of the values that the random variable takes, e.g., ***expectation***, ***median***, ***mode***, etc.

Variability measures show how spread values the random variable takes, e.g., ***range***, ***variance***, ***standard deviation***, ***quartile deviation***.

Other measures e.g., kurtosis, skewness.

Central tendency measure 1: Expectation (mean)

Definition (Expectation)

The ***expectation*** of a discrete random variable X, denoted by $\mathbb{E}X$, $\mathbf{E}X$, $\langle X \rangle$, or $\overline{X}$, is the weighted mean of the values with the probability masses as weights. That is

$$\mathbb{E}X := \sum_{x \in \mathscr{X}} x P_X(x). \tag{5}$$

The expectation is also called the ***mean***.

Central tendency measure 2: Median

If a distribution takes some extremely large or small values, the expectation is significantly influenced by the probability of the random variable taking such values.
In such cases, some might want to use the ***median*** as a summary statistic. Roughly speaking, the median is defined so that the random variable is larger than the median in 50% probability and smaller than the median in 50% probability. The strict definition of the median is somewhat technical, so it is more important to understand intuition and how to calculate it.

Central tendency measure 2: Median (cont.)

Definition (The broader definition of the median)

Let $P : \mathbb{R} \to [0, 1]$ be the probability mass function of a univariate discrete random variable X. If a real value $m \in \mathbb{R}$ satisfies the following equation, then m is called a ***median*** of the distribution of X:

$$\Pr(X \leq m) \geq \frac{1}{2} \text{ and } \Pr(X \geq m) \geq \frac{1}{2}. \tag{6}$$

We can often see the above definition in the context of probability theory.

Central tendency measure 2: Median (cont.)

Definition (The narrower definition of the median)

Let $P : \mathbb{R} \rightarrow [0, 1]$ be the probability mass function of a univariate discrete random variable X. If a real value $m \in \mathbb{R}$ satisfies the following equation, then m is called a ***median*** of the distribution of X:

$$\Pr(X \leq m) \geq \frac{1}{2} \text{ and } \Pr(X \geq m) \geq \frac{1}{2}. \tag{7}$$

We can often see the above definition in the context of probability theory.

Variability measure: Variance

Variability measures show how spread values the random variable takes. If the value of a random variable tends to be far different from the expectation, then we would say that the random variable has large variability. Hence, we consider the deviation, defined as the difference $X - \mathbb{E}X$ between the random variable's value and its expectation (ignoring its sign).

Definition (Variance)

Let X be a random variable and assume that the expectation $\mathbb{E}X$ exists. Then, the ***variance*** $\mathbb{V}[X] \in \mathbb{R}_{\geq 0}$ is defined as the expectation of the squared deviation [3] $(X - \mathbb{E}X)^2$, that is,

$$\mathbb{V}[X] := \mathbb{E}(X - \mathbb{E}X)^2. \tag{8}$$

Variance of a discrete random variable

If X is a discrete random variable taking values in $\mathscr{X} \subset \mathbb{R}$,

$$\mathbb{V}[X] = \sum_{x \in \mathscr{X}} (x - \mathbb{E}X)^2 P_X(x). \tag{9}$$

Variability measure: Standard deviation

Variance's interpretation is somewhat tricky since it is not "linear." Specifically, the variance of $10X$ is 100 times as large as that of X. To make it "linear", we consider the square root of the variance, called the ***standard deviation*** of the random variable.

Definition (Standard deviation)

The ***standard deviation*** $\sigma[X] \in \mathbb{R}$ of the random variable X is defined as

$$\sigma[X] := \sqrt{\mathbb{V}[X]}. \tag{10}$$

As expected, $\sigma[cX] = |c|\sigma[X]$ for $c \in \mathbb{R}$.

Continuous random variables in real AI applications

Prices, images, internal states of neural networks, etc.

Representing a continuous random variable by a histogram

A histogram can show the probability of a random variable taking a value in fixed variables.

Limitation of a histogram

For any histogram, we have multiple random variables corresponding to the histogram. This implies that a histogram cannot uniquely specify a random variable.

Make the histogram finer

Observation: The finer bins a histogram has, the more precisely it can describe a random variable.

Probability density function (PDF): an infinitely precise histogram

Given a probability density function p, the probability of the random variable taking a value between a and b is given by the area bounded by the graph of $y = p(x)$ and $y = 0$ between $x = a$ and $x = b$.

Definite Integral

The (signed) area bounded by the graph of $y = p(x)$ and $y = 0$ between $x = a$ and $x = b$ is called the definite integral of p between a and b, which is denoted by

$$\int_a^b p(x)dx. \tag{11}$$

Other applications of definite integral

Consider a car whose velocity at time t is given by $v(t)$. Let the position of the car at time 0 be 0 then the position $x(t)$ at time t is given by

$$x(t) = \int_0^t v(t)\mathrm{d}t. \tag{12}$$

Expectation (mean) of a continuous random variable

If the probability density function of a random variable X is given by p, then the expectation of X is given by

$$\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x. \tag{13}$$

Cf.) The expectation of a discrete random variable X is given by

$$\sum_{x\in\mathcal{X}} xP(x), \tag{14}$$

where P is the probability mass function.

The expectation does not always exist

If the PDF of a random variable X is given by

$$p(x) = \frac{1}{1+x^2}, \tag{15}$$

then X does not have its expectation. Indeed, the improper integral

$$\int_{-\infty}^{+\infty} xp(x)\mathrm{d}x := \lim_{a\to-\infty}\int_a^c xp(x)\mathrm{d}x + \lim_{b\to+\infty}\int_c^b xp(x)\mathrm{d}x \tag{16}$$

diverges (both the first and second terms in the RHS diverge).

Variance and standard deviation of a continuous random variable

The variance of a random variable is given by the expectation of the square deviation; that is

$$\int_{-\infty}^{+\infty} (x-m)^2 p(x) \mathrm{d}x. \tag{17}$$

The standard deviation is given by the square root of the variance.

Calculating definite integrals

- Numerical integration
- Calculating analytically as the inverse operation of differentiation

Covariance

One principal question about the relation between two random variables X and Y is: "Do they tend to take large values simultaneously, or does one tend to be small when the other is large?"

To answer the question, we consider the product of $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$. $X - \mathbb{E}X$ and $Y - \mathbb{E}Y$ are positive if X and Y are relatively large, respectively.

Hence, if X and Y tend to take large values simultaneously, then the product $(X - \mathbb{E}X)(Y - \mathbb{E}Y)$ tend to be positive.

Conversely, if one tends to be small when the other is large, then the product $(X - \mathbb{E}X)(Y - \mathbb{E}Y)$ tend to be negative.

The above observation leads us to the definition of the ***covariance***.

Definition of the covariance

Definition

Let X and Y be random variables. Then, the ***covariance*** $\mathrm{Cov}(X,Y) \in \mathbb{R}$ between the two random variables X and Y is defined by

$$\mathrm{Cov}(X,Y) := \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]. \tag{18}$$

A positive covariance indicates that the two random variables tend to take relatively large values simultaneously. A negative covariance indicates that when one of the two takes a relatively large value, then the other tend to take a relatively small value.

Correlation

The covariance considers the scale of each random variable, not only the relation between them. Specifically, for $a, b \in \mathbb{R}$, we have that

$$\mathrm{Cov}(aX, bY) = ab\,\mathrm{Cov}(X, Y). \tag{19}$$

This implies that just multiplying the random variables by some factors changes the value of the correlation although the relation between aX and bY would be "qualitatively" the same as that of X and Y.
To see the "qualitative" relation between X and Y, we normalize it by dividing it by the covariance by the sum of the standard deviations of X and Y. The normalized covariance is called the ***correlation coefficient*** of X and Y.

Definition of the correlation coefficient

Definition (Correlation coefficient)

Let X and Y be random variables. The ***correlation coefficient*** $\operatorname{corr}[X, Y]$ between X and Y is given by

$$\operatorname{corr}[X, Y] := \frac{\operatorname{Cov}[X, Y]}{\sigma[X]\sigma[Y]}. \tag{20}$$

As expected, for positive real numbers a and b, we have that

$$\operatorname{corr}[aX, bY] = \operatorname{corr}[X, Y]. \tag{21}$$

Correlation $\neq$ Causality

If two random variables X and Y have a correlation, i.e., $\mathrm{corr}[X, Y] \neq 0$, you might expect that X is the cause of Y.
However, there are many possibilities behind the correlation, e.g.,

1. X is a cause of Y.
2. Y is a cause of X.
3. There exists a random variable Z that causes the both X and Y.
4. (When we estimate the correlation coefficient) There is no relation between X and Y but our estimation of the correlation coefficient is non-zero by estimation errors.

Hence, we cannot conclude that X is a cause of Y just by $\mathrm{corr}[X, Y] \neq 0$.

Calculating multi integral by iterated integral

We can calculate a multi-integral by an ***iterated integral***.

Theorem

Under some loose conditions[4], we have that

$$\begin{aligned}
&\iint_A p(x,y)\mathrm{d}x\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} 1_A(x,y)p(x,y)\mathrm{d}x\right]\mathrm{d}y \\
&= \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} 1_A(x,y)p(x,y)\mathrm{d}y\right]\mathrm{d}x.
\end{aligned} \tag{22}$$

[4]We refer the readers wanting to know the exact conditions to the Fubini-Tonelli theorem.

Sample and sample statistics

In real applications, we **rarely know the true distribution**, behind the data. On the other hand, we often **have many data points** that we can assume follow the same distribution (often independently). Such a series of data points is called ***sample*** of the distribution.
Statistics, data science, machine learning, etc., aim to **extract information about the true distribution from available data points. Sample statistics are the basis of those pieces of technology**.

Terminology: population and sample

In the context of statistics,

- The true distribution is often called the ***population***.
- A series of data points that we can assume follow the same distribution is called ***sample***. If it has many data points, we say that the sample is large, and if it has few data, we say that the sample is small.

Summary statistics and sample statistics

- ***Summary statistics*** aims to describe characteristics of a (known or true) distribution by a few values.
- ***Sample statistics*** aims to estimate some information about the true distribution from finite sample data.

We only have finite data points, so sample statistics are practically necessary to handle probability in real applications.

Sample mean

One principal summary statistic is the expectation.
For data points $X_1, X_2, \ldots, X_m$, we can easily calculate the ***sample mean***

$$m_m = \frac{1}{m}(X_1 + X_2 + \cdots + X_m), \tag{23}$$

the mean of the data points.
If we can assume that those data points are the values of random variables following the same distribution with a true mean μ, we expect m to approximate the true mean μ, which is unknown.
Is it correct? The answer is YES, according to the **law of large numbers**.

Law of large numbers

Theorem ((Strong) law of large numbers)

Let $X_1, X_2, \ldots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean of the distribution is $\mu \in \mathbb{R}$.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{24}$$

Then $\overline{X}_m$ converges to μ in probability 1.

Thus, the sample mean tells us some information about the unknown true distribution!

How the sample mean behaves?

The sample mean converges to the expectation. Now,

- How close to the expectation will the sample mean get as we increase the data points?
- What does the distribution of the sample mean look like?

The answer is

- The difference between the sample mean and the true expectation is proportional to the standard deviation σ of the true distribution and $\frac{1}{\sqrt{m}}$,
- With appropriate scaling, the distribution of the sample mean converges to a ***normal distribution***,

according to the ***central limit theorem***.

What is the normal distribution?

The normal distribution with a mean parameter $\mu \in \mathbb{R}$ and a variance parameter $\sigma^2 \in \mathbb{R}$ is a distribution represented by a PDF defined as follows:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x-\mu}{2\sigma^2}\right). \tag{25}$$

Central limit theorem (CLT)

Theorem (Central limit theorem (CLT))

Let $X_1, X_2, \ldots$ be an infinite sequence of independently and identically distributed (i.i.d.) random variables and assume that the mean and variance of the distribution are $\mu \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_{\geq 0}$, respectively.
Let $\overline{X}_m$ be the sample mean

$$\overline{X}_m := \frac{1}{m}(X_1 + X_2 + \cdots + X_m). \tag{26}$$

Then the CDF of $\sqrt{m}\frac{\overline{X}_m - \mu}{\sigma}$ converges to that of the standard normal distribution at any point in $\mathbb{R}$.

Estimation of a distribution

We have estimated the expectation only. In real applications, we might want to estimate the distribution itself. However, if the support of the distribution is infinite, it is not practical to determine a PMF or PDF from finite data points with no assumptions.
We often assume that the distribution is in a parametric model, which is a set of distributions parametrized by a few values.

Parametric model

Definition (A parametric model)

- ***A discrete parametric model*** on support $\mathcal{X} \subset \mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PMF $P : \mathcal{X} \times \Theta \to [0, 1]$ such that $P(\boldsymbol{x}; \boldsymbol{\theta})$ is a PMF on $\mathcal{X}$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.
- ***A continuous parametric model*** on support $\mathbb{R}^n$ is a pair of a parameter set $\Theta \subset \mathbb{R}^k$ and a parametrized PDF $p : \mathbb{R}^n \times \Theta \to \mathbb{R}_{\geq 0}$ such that $p(\boldsymbol{x}; \boldsymbol{\theta})$ is a PDF on $\mathbb{R}^n$ as a function of $\boldsymbol{x}$ for all $\boldsymbol{\theta} \in \Theta$.

Here, the nonnegative integer k is the dimension of the parameter.

When we have a parametric model, estimating a parameter corresponds to estimating a distribution.

Likelihood

To determine a parameter of a parametric model from data points, we quantify how "likely" the distribution indicated by a parameter is correct. When we have a PMF or PDF of a distribution, we simply define the value of the PMF or PDF of the data points as the "likelihood" of the distribution.

Definition (Likelihood)

Let $P(\cdot;\cdot)$ be a discrete parametric model with a parameter set Θ and $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m$ be values of data points.
Then the likelihood of $P(\cdot;\boldsymbol{\theta})$ is defined as

$$P(\boldsymbol{x}_1;\boldsymbol{\theta}) \cdot P(\boldsymbol{x}_2;\boldsymbol{\theta}) \cdot \cdots \cdot P(\boldsymbol{x}_m;\boldsymbol{\theta}). \tag{27}$$

Likewise, we can define the continuous version.

Maximum likelihood estimator

Once we define the likelihood of a distribution, all we need to do is to find a parameter that maximizes the likelihood.
The parameter vector that maximizes the likelihood is called the ***maximum likelihood estimator (MLE)***.

Why can we justify the maximum likelihood estimator (MLE)?

Similar to the sample mean, if data points are generated by a distribution indicated by a parameter vector in the parameter set of a parametric vector, the MLE has the following properties:

- ***Consistency***: The MLE converges to the true parameter.
- ***Asymptotic normality***: An appropriately scaled MLE's distribution converges to a normal distribution, and its error is proportional to $\frac{1}{\sqrt{m}}$.

Statistical test

In real applications (especially in medical applications), we need to judge from data whether a phenomenon happens or not.
Specifically, for some summary statistics θ, we want to judge from data points whether $\theta \in \mathscr{H}_1$ or not.
Statistical tests give us a framework to make such a judgement.

We cannot directly prove that “the hypothesis is correct.”

What we want to “prove” is the following statement: “if the data points’ values are $x_1, x_2, \ldots, x_m$, then $\theta \in \mathscr{H}_1$,” in some probability theory sense. However, in (frequentism) statistics, we cannot discuss the probability of a parameter being true given data points since a parameter is not a random variable.

In contrast, we can discuss the other direction, that is, given a parameter, we can discuss the probability of the random variables taking the given values.

So, we take the **contraposition** of the proposition that we originally wanted to prove.

Null hypothesis

The contraposition of "if the data points' values are $x_1, x_2, \ldots, x_m$, then $\theta \in \mathcal{H}_1$," is:
"If $\theta \notin \mathcal{H}_1$, then the data points' values are NOT $x_1, x_2, \ldots, x_m$."
The hypothesis $\mathcal{H}_0 := \mathcal{H} \setminus \mathcal{H}_1$ is called the null hypothesis. Here, $\mathcal{H}$ is the set of all the distributions that we assume is possible as a true distribution.

Terminology: rejecting and accepting a hypothesis

- We say that we ***reject*** a hypothesis when we conclude that the true distribution is not in the hypothesis.
- We say that we ***accept*** a hypothesis when we conclude that the true distribution is not in the hypothesis.

Test statistics and p-value

We consider a summary statistic of the empirical distribution. The summary statistic is a random variable since it is a function of the data points, which are random variables. Hence, the distribution of summary statistics is determined we assume a distribution of the data points.

For a distribution, if the value of the summary statistic is unlikely taken on the distribution, then we would conclude that the data points are not generated by the distribution.

How do we determine the unlikeliness of the value of the statistic?

As a criterion of the unlikeliness of the statistic's value, we consider the probability of the statistic taking a more extreme value [5]. The probability is called the ***p-value***. A small p-value indicates that the value of the statistic takes an extreme value.

[5] Hence, we need to define in which case the value of the statistic is extreme. Although it is intuitive for well-known cases, there does not seem to be a way to mathematically decide it.

Significance level

We reject a hypothesis consisting of a single distribution if the p-value of the distribution on the data points is small[6].
Now, how small should the threshold, called the ***significance level*** be?
There is no mathematical reason to determine it.
There is a convention to set the threshold at 0.05.

[6] We reject a hypothesis consisting of multiple distributions if we can reject the hypothesis consisting of any distribution in the original hypothesis

Significance level, false-positive, false-negative

The false-positive rate, the possibility of accepting the alternative hypothesis when the data points are generated by a distribution in the null hypothesis, is determined by the significance level.
So, is it better to use a smaller significance level?
The answer is NO. It is because it increases the false-negative rate, the possibility of failing to accept the alternative hypothesis when the data points are generated by a distribution in the alternative hypothesis.