1 Probability

1.1 Types of probability

Probability can be divided into two main types : **objective** and **subjective**.
The most common form of objective probability is **frequentist probability**.

Definition 1.1 (Frequentist probability)**.** Probability is only a property of a (hypothetical) repeated experiments.

The most common form of subjective probability is **Bayesian probability**.

Definition 1.2 (Bayesian probability)**.** Probability is an expression of our uncertainty and of our beliefs.

1.2 Basic concepts

There are 4 core concepts in probability theory: **sample space**, **event space**, **probability function**, **random variable**.

1.3 Sample space Ω

Definition 1.3 (Sample space)**.** The set of all possible outcomes of an experiment. Denoted by Ω. For example $\Omega = \{\mathrm{H}, \mathrm{T}\}$ for a coin flip.

Definition 1.4 (Discrete sample space)**.** A sample space is discrete if it is finite or countably infinite. E.g. $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a die roll.

Definition 1.5 (Continuous sample space)**.** A sample space is continuous if it is uncountably infinite. E.g. $\Omega = \mathbb{R}$ for a continuous random variable.

The examples summarized:

- $\Omega = \{1, 2, 3, 4, 5, 6\}$ for a die roll.
- $\Omega = \{\mathrm{H}, \mathrm{T}\}$ for two coin flips.
- $\Omega = \mathbb{R}$ for a continuous random variable.

1.4 Event space

Definition 1.6 (Event space E)**.** For a *discrete* sample space the corresponding event space is a the set of all subsets of the sample space AKA the *power set.* For a *continuous* sample space the event space is the set of all intervals.

Definition 1.7 (Event)**.** An event is something which has a probability. This can include *atomic events* or *compound events.*

Definition 1.8 (Atomic event)**.** An atomic event is a single outcome of an experiment. For example $\{1\}$ for a die roll.

Definition 1.9 (Compound event)**.** A compound event is an event with more than one sample point, such as the probability of rolling a 1 or a 2 in the case of a die roll.

For example in the case of a die roll:

- sample space $\Omega = \{1, 2, 3, 4, 5, 6\}$
- event space $E = \{\{\}, \{1\}, \{2\}, \{3\}, \dots, \{1, 2\}, \dots, \{1, 2, 3, 4, 5, 6\}\}$

1.5 Random variables

Definition 1.10 (Random variable X)**.** Intuitively a random variable is a way to describe events. It takes a sample space and describes some set of events.

Some examples of random variables assuming X = sample space of a die roll:

- $X = 1$ for the event of rolling a 1.
- $X > 3$ for the event of rolling a number greater than 3.
- $X = 2$ or 3 for the event of rolling a 2 or a 3.

Definition 1.11 (Probability function P)**.** A probability function is a function which takes a random variable (e.g. $X = 3$) and returns a probability, generally in decimal form between 0 and 1 (inclusive).

Some examples of probability functions assuming X = sample space of a die roll:

- $p(X = 1) = \frac{1}{6}$
- $p(X > 3) = \frac{1}{2}$
- $p(X = 2 \text{ or } 3) = \frac{2}{6} = \frac{1}{3}$

Definition 1.12 (Probability function of regular variable $p(X = x)$)**.** Probability functions can either be functions of well-defined events or functions of regular variables.

If we have an event space with only two atomic events $\{-1, 1\}$ then we can define a probability function for a variable x representing these events as follows:

$$p(X = x) = p(X) = p(x) = \begin{cases} \frac{1}{4} & \text{if } x = -1 \\ \frac{3}{4} & \text{if } x = 1 \end{cases}$$

In the case we have boolean random variables the notation is as follows:

$$p(X) = p(X = \text{true})$$
$$p(\neg X) = p(X = \text{false})$$

1.6 Probability vs Probability density

In the case of a discrete sample space we have probabilities assigned to each atomic event. For example 50% for a coin flip. In the case of a continuous sample space we have a probability density function (pdf) which gives the probability of a random variable being in a certain interval. For example the probability of a random variable being between 0 and 1.

Definition 1.13 (Probability density)**.** Probability density expresses the probability of a random variable being in a certain continous interval.

Some notes on probability density:

- In the case of a continuous sample space the probability of a random variable being in a single point is 0.
- The probability of a random variable being in a certain interval is the integral of the probability density function over that interval.

1.7 Joint probability

Definition 1.14 (Joint probability $p(X, Y)$)**.** Joint probability is the probability of two random variables being in a certain state at the same time. It can either be *instantiated* or *uninstantiated*.

To explain this we use an example teeth health and age, as usual first there is the sample space:

$$\text{Age} = \{\text{young}, \text{teen}, \text{old}\}$$
$$\text{Teeth} = \{\text{healthy}, \text{unhealthy}, \text{fake}\}$$

The sample space then consits of all possible combinations of these two random variables, which makes up 9 pairs. We expresses the uninstantiated probability distribution for all these combinations as follows

$$p(\text{Age}, \text{Teeth})$$

Definition 1.15 (Instantiated joint probability $p(X = x, Y = y)$)**.** Instantiated joint probability is the probability of two random variables being in a certain specific state at the same time.

For example $p(\text{Age} = \text{young}, \text{Teeth} = \text{healthy})$.

1.8 Marginal probability

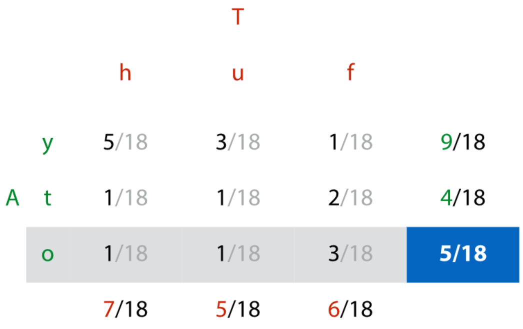
Definition 1.16 (Marginal probability $p(X)$)**.** The marginal probability of a random variable is the probability of that random variable being in a certain state. Calculating the marginal probability is called *marginalizing out* because we sum over the rows or columns of the variable we are interested in and write the results in the margins.

The formula for marginal probability is as follows:

$$p(X) = \sum_{y \in Y} p(X, Y = y)$$

Applying this to the example below for the marginal probability that $p(\text{Age} = \text{old})$ we get:

$$p(A = o) = p(o, T = h) + p(o, T = u) + p(o, T = f) = \frac{1}{18} + \frac{1}{18} + \frac{3}{18} = \frac{5}{18}$$



			T		
		h	u	f	
	y	5/18	3/18	1/18	9/18
A	t	1/18	1/18	2/18	4/18
	o	1/18	1/18	3/18	**5/18**
		7/18	5/18	6/18	

Figure 1: Marginal probability

1.9 Conditional probability

Definition 1.17 (Conditional probability $p(X \mid Y)$)**.** Conditional probability is the probability of a random variable being in a certain state given that another random variable is in a certain state.

The formula for conditional probability is as follows:

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{\underbrace{\sum_{x'} p(X = x', Y = y)}_{\text{marginal probability}}} = \frac{p(x, y)}{p(y)} \quad p(X = x) \Leftrightarrow p(x)$$

A useful formula form this is the one for joint probability as a function of conditional probability and marginal probability:

$$p(x, y) = p(x \mid y)p(y)$$

Applying the formula for the conditional probability to the example below we get:

$$P(T = f \mid A = y) = \frac{p(f, y)}{p(y)} = \frac{1}{9}$$

Marginalizing out in higher dimensions

Broadly speaking marginalizing out in higher dimensions for continous distributions is done similar to the discrete case but we use integrals instead of sums.

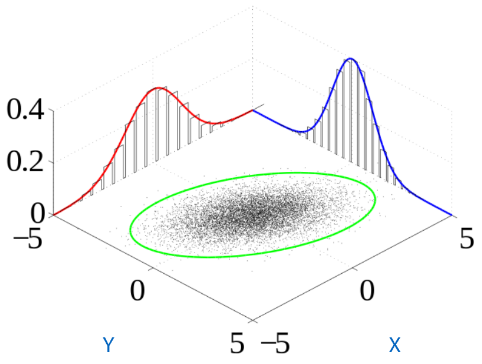


image source: By IkamusumeFan - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=30432580

Figure 2: marginal probability in higher dimensions

1.10 Independence

Definition 1.18 (Independence)**.** If two variables X and Y are independent then knowing Y will not change what we know about X.

Formally we can express independence as follows:

$$p(X \mid Y) = p(X)$$
$$\rightarrow P(X, Y) = p(X)p(Y)$$

Definition 1.19 (Conditional independence)**.** Conditional independence means that two variables can be dependent, but their dependence is entirely explained by a third variable.

Formally X and Y are conditionally independent given Z if:

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$$

As an example let say

$$\begin{aligned} X &= \text{Alice is home in time for dinner} \\ Y &= \text{Bob is home in time for dinner} \\ Z &= \text{A monster attacks the city} \end{aligned}$$

We assume in most cases the the time when Alice or Bob are home for dinner are independent. So knowing the time for Alice tells us nothing about the time for Bob and vv. But if a monster attacks the city then we assume this effects the time of both Alice and Bob, thus given that a monster attacks the city if we observe the time for Alice or Bob we can infer the time for the other person.
The key takeaways from this being that:

- Any dependence between X and Y is entirely explained by the monster attacking the city. Formally

$$p(X, Y \mid Z) = p(X \mid Z)p(Y \mid Z)$$

- If we know that the monster attacked the city then any further information about X or Y is redundant. Formally

$$p(X \mid Y, Z) = p(X \mid Z)$$

1.11 Bayes' rule

Definition 1.20 (Inversion problem)**.** The inversion problem is the problem of finding the probability of some hidden cause given an observed effect.

An example would be the situation where we call a restaurant to place an order yet noone picks up the phone. One line of reasoning you might follow is to

1. Assume that the restaurant burned down
2. Since the restaurant burned down there is a high probability noone would answer the phone

Here we are reasoning from the cause (the restaurant burned down) to the effect (noone answers the phone). More formally we are essentially saying that

$$p(\text{noone answers the phone} \mid \text{restaurant burned down})$$

is high and this lines up with our observation.

Though by the inversion problem we are interested in the opposite, we already know that noone answered the phone and we want to know the probability that the restaurant burned down. This is where Bayes' rule comes in. Which is the probability of the cause given the effect. Formally

$$p(\text{cause} \mid \text{effect}) = \frac{1}{p(\text{effect})} p(\text{effect} \mid \text{cause})p(\text{cause})$$

Bayes Example

Situation: Alice is late for dinner, we don't know if this is because a monster attacked the city, but we also can't say a monster didn't attack. So lets try and figure out the probability that a monster attacked (cause m) the city given that Alice is late for dinner (effect a).

$$\begin{aligned} p(m \mid a) &= \frac{p(a \mid m)p(m)}{p(a)} \\ &= \frac{p(a \mid m)p(m)}{\underbrace{p(a \mid t)p(t)}_{\text{traffic}} + \underbrace{p(a \mid m)p(m)}_{\text{monster}} + \underbrace{p(a \mid s)p(s)}_{\text{snowfall}}} \end{aligned}$$

Here we assume there are 3 possible causes to Alice being late, namely traffic, the monster attacking and snowfall. The denominator is the sum marginalizing out over all possible causes.

1.12 Expectation

Definition 1.21 (Expectation)**.** The expectation (expected value) is a way of expressing what you can expect to gain from a random process.

Formally for a given outcome function $f(x)$ and a probability function $p(x)$ the expectation is given by:

$$E[f(x)] = \sum_x f(x)p(x) = \int_x f(x)p(x)dx$$

Equivelantly we can also think of it as the sum of a finite list of $x_1, x_2, \dots, x_n$ outcomes with corresponding probabilities $p_1, p_2, \dots, p_n$:

$$E[f(x)] = x_1p_1 + x_2p_2 + \cdots + x_np_n$$

A die roll bet

Say for a bet you roll $n \leqslant 3$ you pay the corresponding amount of dollars. If you roll $n > 3$ you get the corresponding amount of dollars. We can define the expected value for this as

$$E[X] = (-3)\frac{1}{6} + (-2)\frac{1}{6} + (-1)\frac{1}{6} + (4)\frac{1}{6} + (5)\frac{1}{6} + (6)\frac{1}{6} = 1.5$$

Rules

There are some basic rules for manipulating expectations, all based mostly just on the same rules for manipulating sums and integrals since thats all expectations are formally.

$$\begin{aligned} E[c] &= c \\ E[cf(x)] &= cE[f(x)] \\ E[c + f(x)] &= c + E[f(x)] \\ E[f(x) + g(x)] &= E[f(x)] + E[g(x)] \end{aligned}$$

1.13 Common probability distributions

Bernoulli distribution

Definition 1.22 (Bernoulli distribution)**.** Any distribution with two outcomes. The probability of the first outcome is p and the probability of the second outcome is $1 - p$.

Some examples include:

- heads or tails
- guilty or innocent
- positive negative

Categorial distribution

Definition 1.23 (Categorial distribution)**.** Any distribution with some finite number (>2) of n outcomes. Where the probabilities of each outcome sum to 1. That is $p_1 + p_2 + \cdots + p_n = 1$. Hence we only need $n - 1$ probabilities to describe the distribution.

Some examples include:

- which team wins the next world cup
- hair color of a random person in ireland

Normal distribution

Definition 1.24 (Normal distribution)**.** The normal distribution is a continuous distribution with a bell-shaped curve. It is defined by two parameters, the mean μ and the standard deviation σ.

The parameters express

- μ the center of the distribution
- σ the spread of the distribution
- σ^2 the variance of the distribution

The distribution is good at expressing things that have a definite scale no matter the scale we are working at. For example no matter how many humans we look at most people will have a height clustered around the mean with fewer and fewer people having heights further and further from the mean.

Multivariate normal distribution

Definition 1.25 (Multivariate normal distribution)**.** The multivariate normal distribution is a generalization of the normal distribution to higher dimensions.