Contents

(All variables are real and one-dimensional unless otherwise specified.)

1. Probability of events

The probability of an **event** A is denoted by $P(A)$, and that of its **complement** $\bar{A}$ is $P(\bar{A}) = 1 - P(A)$.

1.1 Intersection of events

The probability for A to happen together with another event B is $P(A \cap B)$, which is also called their **joint probability**. If the occurrences of the two event do not influence each other, they are **independent** and satisfies the so-called "**product rule**".

$$P\left(A \cap B\right) \overset{\text{indep.}}{=} P\left(A\right) P\left(B\right)$$

However, it may get philosophically troublesome to determine "independence" . (Does a butterfly in Brazil cause a tornado in the United States?) Hence, as long as his data supports this equality, one may hedge to say that two events are **statistically independent** instead.

If A and B never happen together, they are called **mutually exclusive** and satisfies $P(A \cap B) \overset{\text{m.e.}}{=} 0$. A simple pair of mutually exclusive events is to get a head and to get a tail from flipping a coin.

1.2 Union of events

The probability for A or B to happen is calculated with the so-called "**sum rule**".

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

It may be generalized as the **inclusion-exclusion principle**: given a set of events $\{E_1, E_2, E_3, \ldots\}$, the probability for **at least one** to happen is

$$\begin{aligned} P\left(\bigcup_i E_i\right) &= \sum_i P(E_i) - \sum_{i<j} P(E_i \cap E_j) + \sum_{i<j<k} P(E_i \cap E_j \cap E_k) + \cdots \\ &= \sum_k \left[(-1)^{k-1} \sum_{i_1<\cdots<i_k} P\left(\bigcap_{j=1}^{k} E_{i_j}\right)\right] . \end{aligned}$$

If the events are pairwise independent, we may simplify the formula to

$$P\left(\bigcup_i E_i\right) \overset{\text{indep.}}{=} 1 - \prod_i \left[1 - P(E_i)\right] .$$

If the events are mutually exclusive, the terms form of intersection vanish and lead to

$$P\left(\bigcup_i E_i\right) \overset{\text{m.e.}}{=} \sum_i P(E_i) .$$

This is actually not a trivial statement but carries a deep philosophical meaning. I will talk about it in the coming tutorials.

1.3 Conditional probability

The probability for B to happen given that A has already happened is

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)} .$$

This is also called the **conditional probability** of B on A. If they are independent, the nominator can be factorized to give $P(B \mid A) \overset{\text{indep.}}{=} P(B)$, meaning that the occurrence of A does not alter the probability for B to occur at all. On the other hand, if they are mutually exclusive, one trivially obtains $P(B \mid A) \overset{\text{m.e.}}{=} 0$.

1.4 Association rules

We are interested in the **causality** between two events A and B: how likely does A cause B? Of course, since causality is a tough philosophical concept, we must restrict our definition:

A causes B if B happens when A happens.

Symbolically, we represent this idea with an **association rule** $A \Rightarrow B$, where A and B are now called the **antecedent** and the **consequence**. We do not care about how A exactly causes B. We just want to measure how good it is to predict "B will happen" when we see "A has happened". Various measures have been devised to assess it, and here are six measures that can be classified into two categories.

- **Measures of usefulness.** This kind of measure detects whether a rule $A \Rightarrow B$ is useful.

Support $P(A)$	Confidence $P(B \mid A)$	Rule power factor (RPF) $P(A \cap B)\, P(B \mid A)$
• A rule is useless if its support is low because it can be barely used.	• A rule with a higher confidence is intuitively more trustworthy. • However, confidence may be coincidentally high just because B happens a lot, regardless of whether A has happened or not.	• This is based on confidence but improved to solve the problem of coincidence. • If B occurs frequently without A, a rule's RPF remains low despite its high confidence.

- **Measures of interdependence.** This kind of measure detects the interdependence between A and B.
    - If it equals c, the events are independent and thus in no way causal.
    - If it is greater than c, they occur together more frequently than expected and thus suggests some kind of causality.
    - If it is less than c, they occur together less frequently than expected and thus suggests that one cause the other's complement.

Lift	Leverage	Conviction
$\dfrac{P(A \cap B)}{P(A)P(B)}$	$P(A \cap B) - P(A)P(B)$	$\dfrac{1 - P(B)}{1 - P(B \mid A)}$
$c = 1$	$c = 0$	$c = 1$

2. Distribution of random variables

A common type of event is to have observed a particular realization x of a random variable X. We would often like know how its probability $P(X = x)$ **distributes** over all possible x.

2.1 PMF, PDF, and CDF

If X is discrete, its **probability mass function** (PMF) is defined as

$$p_X(x) \equiv P(X = x)\,.$$

For example, the PMF of a dice's outcome D is
$$p_D(d) = \begin{cases} 1/6 & (d \in \{1,2,3,4,5,6\}) \\ 0 & (\text{otherwise}) \end{cases}.$$

However, the definition of PMF fails if X is continuous, because we cannot find an **exact real number** on the real number line. (There are, informally speaking, "infinitely many real numbers", so the probability to locate any particular real number goes to zero.) We need to adopt a different but similar concept for continuous random variables, namely **probability density function** (PDF). If X has a PDF $f_X(x)$, the following identity holds:

$$\int_a^b f_X(x)\mathrm{d}x \equiv P(a \le X \le b)\,.$$

In other words, the product $f_X(x)\mathrm{d}x$ may be regarded as the probability of observing $X \in [x, x + \mathrm{d}x]$. [Hence $f_X(x)$ alone is not probability but **probability density**.] As X definitely lies inside $(-\infty, \infty)$, the **normalization condition** $\int_{-\infty}^{\infty} f_X(x)dx \equiv 1$ must be true although $f_X(x)$ itself may exceed one or even diverge.

Finally, the **cumulative distribution function** (CDF) of X is defined as

$$F_X(x) \equiv P(X \le x) = \begin{cases} \displaystyle\sum_{x' \le x} p_X(x') & \text{(disc. } X\text{)} \\ \displaystyle\int_{-\infty}^{x} f_X(x')\mathrm{d}x' & \text{(cont. } X\text{)} \end{cases},$$

which must satisfy $F(-\infty) = 0$ and $F(+\infty) = 1$. Although it does not bring in new information, CDF is often more useful analytically because of its **monotonic** nature. For the continuous case, we can obtain $f_X(x) = F'_X(x)$ and know that $f_X(\pm\infty) = 0$.

2.2 Multivariate distributions

Now let us focus on continuous random variables, whereas you can easily rephrase the discussion below for discrete ones. We are often interested how a random variable's outcome correlates with others'. In such cases, we need to consider **multivariate distributions**, the simplest case of which contains only two random variables X and Y.

Joint distribution. Their **joint PDF** $f_{XY}(x, y)$ is defined to satisfy

$$\int_a^b \int_c^d f_{XY}(x,y)\mathrm{d}y\mathrm{d}x = P(X \in [a,b] \cap Y \in [c,d])\,.$$

There are unfortunately no general rules to construct $f_{XY}(x,y)$ from $f_X(x)$ and $f_Y(x)$ if X and Y are correlated, otherwise the PDF can be trivially factorized to become $f_{XY}(x,y) = f_X(x)f_Y(y)$.

Their **joint CDF** is defined like the one-variable case:

$$\begin{aligned} F_{XY}(x,y) &\equiv P(X \le x \cap Y \le y) \\ &= \int_{-\infty}^{x} \int_{-\infty}^{y} f_{XY}(x',y')\mathrm{d}y'\mathrm{d}x'\,, \end{aligned}$$

and it implies $f_{XY}\,(x,y) = \dfrac{\partial^2 F_{XY}}{\partial x \partial y}$. As the number of variables grows, it becomes more convenient to use the differential form of notation.

$$\begin{aligned} F_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n) &\equiv P\left(\bigcap_{i=1}^{n} X_i \le x_i\right) \\ \Rightarrow \quad f_{X_1X_2\ldots X_n}(x_1,x_2,\ldots,x_n) &\equiv \frac{\partial^n F_{X_1X_2\ldots X_n}}{\partial x_1 \partial x_2 \ldots \partial x_n} \end{aligned}$$

Marginal distribution. Sometimes we are given $f_{XY}(x,y)$, and we would like to extract $f_X(x)$ from it. Since $P(X=x) = P[X = x \cap Y \in (-\infty,\infty)]$, we obtain $\int_a^b f_X(x')\mathrm{d}x' = \int_a^b \int_{-\infty}^{\infty} f_{XY}(x',y')\mathrm{d}y'\mathrm{d}x'$ or simply

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\mathrm{d}y\,.$$

This kind of **deduced** PDF is also called a **marginal PDF**. Similarly, as $P(X \le x) = P[X \le x \cap Y \le \infty]$, the **marginal CDF** of X is

$$F_X(x) = F_{XY}\,(x, y = \infty)\ .$$

Conditional distribution. The **conditional PDF** of Y on X measures the probability density of Y given that $X = x$. It is defined as

$$f_Y(y \mid X = x) = \frac{f_{XY}(x, y)}{f_X(x)}$$

so that $P(Y \in [c, d] \mid X = x) = \int_c^d f_{Y|X}(y \mid x)\mathrm{d}y$. Here is a sloppy proof.

$$\begin{aligned}
P(Y \in [c, d] \mid X = x) &= \lim_{\delta \to 0} P(Y \in [c, d] \mid X = [x, x + \delta]) \\
&= \lim_{\delta \to 0} \frac{P(X = [x, x + \delta] \cap Y \in [c, d])}{P(X = [x, x + \delta])} \\
&= \lim_{\delta \to 0} \frac{\int_c^d \int_x^{x+\delta} f_{XY}(x', y)\mathrm{d}x'\mathrm{d}y}{F_X(x + \delta) - F_X(x)} \\
&= \lim_{\delta \to 0} \int_c^d \frac{\dfrac{\int_{-\infty}^{x+\delta} f_{XY}(x', y)\mathrm{d}x' - \int_{-\infty}^{x} f_{XY}(x', y)\mathrm{d}x'}{x + \delta - x}}{\dfrac{F_X(x + \delta) - F_X(x)}{x + \delta - x}}\mathrm{d}y \\
&= \int_c^d \frac{f_{XY}(x, y)}{f_X(x)}\mathrm{d}y \\
&= \int_c^d f_Y(y \mid X = x)\mathrm{d}y
\end{aligned}$$

From the fourth line to the fifth line, the nominator invokes the fundamental theorem of calculus.

2.3 Example: bivariate normal distribution

Let $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$ be two standard normal random variables with correlation r. Their joint PDF is given as

$$f_{XY}(x,y) = \frac{1}{2\pi\sqrt{1-r^2}} \exp\left[-\frac{x^2+y^2-2rxy}{2\left(1-r^2\right)}\right]$$

and may be referred to as the bivariate standard normal distribution. What is the conditional distribution of Y on X?

Solution. Dividing $f_{XY}(x,y)$ by $f_X(x) = \dfrac{e^{-x^2/2}}{\sqrt{2\pi}}$, we get

$$f_Y(y \mid X = x) = \frac{1}{\sqrt{2\pi\left(1-r^2\right)}} \exp\left[-\frac{(y-rx)^2}{2\left(1-r^2\right)}\right],$$

which turns out to be another normal distribution with mean rx and variance $1-r^2$. This fact may be alternatively written as

$$Y \mid X = x \sim \mathcal{N}\left(rx, 1-r^2\right).$$

3. Transformation of distribution

Given the PDF $f_X(x)$ of a continuous random variable X, we can calculate the PDF of an associated random variable $Y = g(X)$ with a general formula

$$f_Y(y) = \left| \frac{f_X(x)}{g'(x)} \right|_{x=g^{-1}(y)}$$

for y defined in the range of g so that $g^{-1}(y)$ is valid. Because it possesses a derivative g' and an inverse g^{-1}, the function g must be **continuous and one-to-one**; equivalently speaking, this means that g is **strictly monotonic**. If the function is not one-to-one, not only does it lack a proper inverse, but its derivative also hits zero somewhere (c.f. Rolle's theorem) and thus invalidates the general formula. Still, $f_Y(y)$ may be deduced with other approaches in this case.

Proof. Let us first assume that g is strictly increasing, so $g' > 0$.

$$F_Y(y) = P(Y \le y) \overset{g'>0}{=} P\left[X \le g^{-1}(y)\right] = F_X\left[g^{-1}(y)\right]$$

Then by the chain rule,

$$f_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X\left[g^{-1}(y)\right] = \left.\frac{\mathrm{d}F_X}{\mathrm{d}x}\frac{\mathrm{d}x}{\mathrm{d}y}\right|_{x=g^{-1}(y)} = \left.\frac{f_X(x)}{g'(x)}\right|_{x=g^{-1}(y)}.$$

Similarly, a strictly decreasing g with $g' < 0$ gives

$$F_Y(y) = P(Y \le y) \overset{g'<0}{=} P[(X \ge g^{-1}(y))] = 1 - F_X\left[g^{-1}(y)\right]$$

$$\Rightarrow f_Y(y) = -\left.\frac{f_X(x)}{g'(x)}\right|_{x=g^{-1}(y)}.$$

The assumption $g' < 0$ cancels the leading negative sign and makes $f_Y(y) \geq 0$. Hence, the two cases can be combined with an absolute sign.

3.1 Example: polynomial transformation

What is the PDF of $Y = X^p$ for $X \sim \mathcal{U}(0,1)$ with $p > 0$?

Solution. Because
$$f_X(x) = \begin{cases} 1 & (0 \leq x \leq 1) \\ 0 & (\text{otherwise}) \end{cases}$$

$$f_Y(y) = \begin{cases} \left|\dfrac{1}{px^{p-1}}\right| & (0 \leq y \leq 1) \\ 0 & (\text{otherwise}) \end{cases} = \begin{cases} \dfrac{1}{p} y^{1/p-1} & (0 \leq y \leq 1) \\ 0 & (\text{otherwise}) \end{cases}.$$

You may see that $f_Y(0) \to \infty$, but this divergence does not invalidate $f_Y(y)$. You may integrate it to see if it satisfies the normalization condition. For example, $p = 2$ yields $F_Y(y) = \sqrt{y}$, which indeed has a steep slope as $y \to 0^+$.

After learning how to find a random variable's distribution based on its definition, now let us state the problem in the other way around: how should we define a random variable so that it possesses a particular distribution?

4. Generation of random variables

Almost all programming languages provide a random number generator that returns a uniform random variable $X \sim \mathcal{U}(0, 1)$. We can generate random variables with any **strictly increasing** CDF $F_Y(y)$ by defining

$$Y = F_Y^{-1}(X) .$$

The monotonic condition is, again, necessary so that F_Y^{-1} exists.

Proof. The CDF of Y at an unknown realization of Y is also a random variable. Let us denote this random variable by $\tilde{Y} = F_Y(Y)$, and it is realized as $\tilde{Y} = \tilde{y}$ upon measuring Y. First, the new variable's range is fixed at $\tilde{y} \in [0, 1]$ because it is the output of the CDF F_Y. Then, consider the new variable's CDF $F_{\tilde{Y}}(\tilde{y})$: because $\tilde{y} \in [0, 1]$,

$$F_{\tilde{Y}}(\tilde{y}) = P\big(\tilde{Y} \leq \tilde{y}\big) = P\big[Y \leq F_Y^{-1}(\tilde{y})\big] = F_Y\left[F_Y^{-1}(\tilde{y})\right] = \tilde{y}$$

and thus $f_{\tilde{Y}}(y) = 1$. Consequently, $\tilde{Y} = F_Y(Y) \sim \mathcal{U}(0, 1)$ is isomorphic to X, i.e. the (pseudo-)random number by our programs. Hence, $Y = F_Y^{-1}(X)$ has the desired CDF $F_Y(y)$.

4.1 A simple example: translation

Express $Y \sim \mathcal{U}[a, b]$ in terms of $X \sim \mathcal{U}[0, 1]$.

Solution. For $y \in [a, b]$, $F_Y(y) = \dfrac{y - a}{b - a}$, so $Y = F_Y^{-1}(X) = (b - a)\, X + a$. (I guess you have learnt this skill since high school?)

4.2 Example: logistic distribution

A distribution is logistic if its CDF is a logistic function, which has an S-shaped curve. Fig. 1 shows the standard logistic function $L(x) = \frac{1}{1+e^{-x}}$. Express a logistically distributed random variable Y with $F_y(y) = L(y)$ in terms of $X \sim \mathcal{U}[0,1]$.

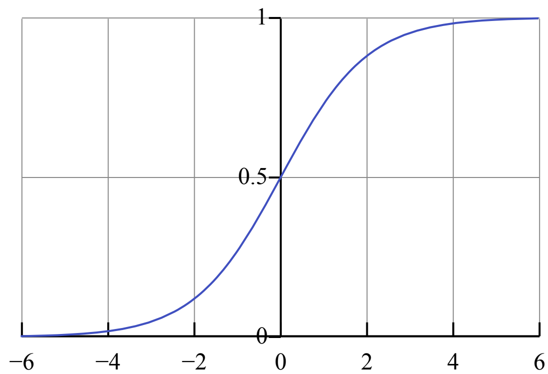

Fig. 1 $L(x) = \frac{1}{1+e^{-x}}$ against x. Retrieved from Wikimedia Commons.

Solution. $Y = L^{-1}(X) = -\ln\left(\frac{1}{X} - 1\right)$. You can verify this answer by plugging it into the general formula in Section 3. As a challenge, try to compare the mean and the variance of Y with those of a standard normal random variable $Z \sim \mathcal{N}(0,1)$. (They have the same mean, but Y has a larger variance $\sigma_Y^2 = \frac{\pi^2}{3} > 1$.)