Notes On Basic Probability

Zichen Wang

Winter 2022

intro The study of probability theory starts from defining probability models and stating some basic axioms that all probability models obey.

1 Probability Models

A probability model $(\Omega, \mathcal{E}, \mathbb{P})$ consists of the sample space, the collection of all events, and the probability function.

1.1 Sample Space

A sample space Ω is the set of all possible outcomes of an experiment. It must be *mutually exclusive*, which means that no two outcomes in the sample space happen at one experiment. It must also be *collectively exhaustive*, which means there exist one outcome in the sample space for every experiment.

1.2 Event

An event E is a subset of the sample space. When the outcome of an experiment is an element of the event, we say that the event *occurs*.

1.3 Probability Function

A Probability Function $\mathbb{P} : \mathcal{E} \to [0, 1]$ assigns a number between 0 and 1 to every event. Note that an event with probability 0 is still possible to occur (consider a point in a square).

2 Probability Axioms

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A) \geq 0$
- $\mathbb{P}\big(\bigcup A_i\big) = \sum \mathbb{P}(A_i)$

intro This pages studies the properties of probability distributions. The easiest case is when all outcomes are discrete and have equal probability. More generally follows the study of various relations between the sets of outcomes (events).

3 Discrete Uniform Distribution

When all outcomes are equally likely, the probability of an event is reduced to a counting problem, where

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}$$

4 Conditional Probability

4.1 Definition

The conditional probability $\mathbb{P}(A|B)$ is the probability of A given that B occurs. Assume $\mathbb{P}(B) \neq 0$,then

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

Essentially, conditioning changes the sample space but preserves the relative likelihoods of the experiment outcomes. Thus properties in relation to unions and intersections in the original sample space should still hold after conditioning.

4.2 Law of Total Probability

$A_1, ..., A_n$ is a partition of the sample space, then

$$\mathbb{P}(B) = \sum \mathbb{P}(B \cap A_i) = \sum \mathbb{P}(B|A_i)\mathbb{P}(A_i)$$

4.3 Baye's Rule

$A_1, ..., A_n$ is a partition of the sample space, then

$$\mathbb{P}(A_k|B) = \frac{\mathbb{P}(A_k \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B|A_k)\mathbb{P}(A_k)}{\sum \mathbb{P}(B|A_i)\mathbb{P}(A_i)}$$

5 Independence of Events

Two events A and B are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

or equivalently,

$$\mathbb{P}(A|B) = \mathbb{P}(A)$$

In general, n events $A_1, ..., A_n$ are independent if

$$\mathbb{P}(\cap A_j) = \prod \mathbb{P}(A_j) \text{ for any } \{A_j\} \subset \{A_i\}$$

Alternatively, n events $A_1, ..., A_n$ are *pairwise* independent if ,

$$\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i) \cdot \mathbb{P}(A_j) \text{ for any } A_i, A_j$$

intro This page introduces the concept of discrete random variables and develops various tools to describe the probability distributions under measurement. Two specific probability models are mentioned for their unique importance.

6 Discrete Random Variable

6.1 Probability Mass Function

The probability mass function (pmf) describes the probability that the experiment outcome, under measurement, equals a certain value.

$$p_X(x) = \mathbb{P}(X = x) = \mathbb{P}(\omega \in \Omega : X(\omega) = x)$$

If follows that

- $p_X(x) \geq 0$
- $\sum_{x \in X(\Omega)} p_X(x) = 1$

Joint PMFs generalizes to the situation with two or more random variables. Similar properties hold as the following

- $\sum \sum p_{X,Y}(x, y) = 1$
- $p_X(x) = \sum_y p_{X,Y}(x, y)$
- $p_{X|Y}(x|y) = \mathbb{P}(X = x|Y = y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$

6.2 Independence of Random Variables

With the notion of pmf, one can now define that two random variables X, Y are independent if

$$p_{X,Y}(x, y) = p_X(x) \cdot p_Y(y) \text{ for any x, y}$$

Note that this definition follows directly from the third property of joint pmf together with

$$p_{X|Y}(x|y) = p_X(x)$$

6.3 Expectation

The expected value of a random value is the weighted average of all the experiment outcomes under measurement. Formally, it is defined as

$$\mathbb{E}[X] = \sum x p_X(x)$$

6.3.1 Functions of Random Variables

A function of a random variable is still a random variable. It can be calculated as

$$\mathbb{E}[f(X)] = \sum_{y \in f(X(\Omega))} y \cdot p_{f(X)}(y)$$

or equivalently,

$$\mathbb{E}[f(X)] = \sum_{x \in X(\Omega)} f(x) \cdot p_X(x)$$

6.3.2 Linearity of Expectation

In general, it is not true that $\mathbb{E}[f(X)] = f(\mathbb{E}[X])$. However, the equation holds when f is a linear transformation.

$$\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}[X] + \beta\mathbb{E}[Y]$$

Furthermore, when X and Y are independent,

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

6.3.3 Law of Total Probability

The law of total probability states that given $A_1, ..., A_n$ as a partition of the sample space, then

$$\mathbb{E}[X] = \sum (\mathbb{P}(A_i) \cdot \mathbb{E}[X|A_i])$$

6.4 Variance

The variance of a random variable describes how far the experiment outcomes, under measurement, spread away from the mean. Formally, it is defined as

$$Var(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

It can also be proved that

$$Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Some properties of variance

- $Var(aX) = a^2 Var(X)$
- $Var(X + b) = Var(X)$
- if X and Y are independent, then $Var(X + Y) = Var(X) + Var(Y)$

7 Geometric Distribution

Link to Geometric Distribution, Wikipedia

8 Binomial Distribution

Link to Binomial Distribution, Wikipedia

9 Continuous Random Variable

9.1 Probability Density Function

Comparable to pmf in the discrete case, the probability density function (pdf) of a continuous random variable enable the calculation over an interval

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

If follows that

- $f_X(x) \geq 0$
- $\int_{-\infty}^{\infty} f_X(x)dx = 1$

Generalizing into two random variables, the calculation becomes

$$\mathbb{P}((X,Y) \in S) = \iint_S f_{X,Y}(x,y)dxdy$$

Some properties regarding multiple pdf

- $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
- $f_{X|Y}(x|y) = \dfrac{f_{X,Y}(x,y)}{f_Y(y)}$

Two random variables are *independent* if

$$f_{X,Y}(x,y) = f_X(x) \cdot f_Y(y) \text{ for any x, y}$$

9.2 Expectation

The expectation of a continuous random variable is defined as

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx$$

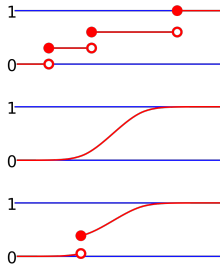
Note that all properties of expectation and variance that hold under discrete random variable still hold under continuous random variable.

9.3 Cumulative Distribution Function

The cumulative distribution function sums all probabilities less or equal to a value. It is well-defined for both discrete and continuous random variables.

$$\begin{aligned} F_X(x) = \mathbb{P}(X \leq x) &= \sum_{k \leq x} p_X(k) \text{ for discrete case} \\ &= \int_{-\infty}^{\infty} f_X(t)dt \text{ for continuous case} \end{aligned}$$

The plot for discrete cdf consists of intervals of constants connected by jump discontinuities, while the plot for continuous cdf is continuous. There could also be the case where the cdf is combination of the two. In either cases, the function monotonically increases to 1.



10 Normal Distribution

Link to Normal Distribution, Wikipedia

11 Derived Distribution

Operations over known random variables give new distributions called derived distributions. Often times it is helpful to derive a closed form of the the probability density function of the derived distribution.

11.1 Discrete Case

If the original probability distribution X is discrete, the derived distribution Y would also be discrete. Thus

$$p_Y(y) = \mathbb{P}(g(X) = y) = \sum_{g(x)=y} p_X(x)$$

11.2 Continuous Case

In the continuous case, one first calculate the cdf of the derived distribution and then take its derivative to get the pdf. To calculate the cdf, rewrite the cdf in term of the original distribution. Suppose the derived random variable $Y = g(X)$, then

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$$

$$f_Y(y) = \frac{d}{dy} F_Y(y)$$

Specifically, when the new distribution is derived after a *linear transformation* $Y = aX + b$, it follows that

$$f_Y(y) = \frac{1}{|a|} f_X(\frac{y-b}{a})$$

Another special case is when g is *monotonic*, where

$$f_Y(y) = f_X(x) \cdot \frac{1}{|g'(x)|}$$

The *convolution* of two random variables takes $W = X + Y$ with

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x) f_Y(w - x) dx$$

12 Covariance

The joint pmf of two normal distributions would still produce a bell-shape distribution, but in three dimensional space. The covariance describes the situation when the two random variables are dependent of each other, causing the bell-shape distribution to skew. By definition,

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}[X]) \cdot (Y - \mathbb{E}[Y])]$$

An easier way to calculate the covariance would be

$$cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

When the covariance is positive, the distribution skews in top-right to bottom-left direction. Conversely, when the covariance is negative, the distribution skews in top-left to bottom-right direction.
The *correlation coefficient* is the standardized version of the covariance with no unit

$$\rho = \mathbb{E}\left[\frac{X - \mathbb{E}[X]}{\sigma_X} \cdot \frac{Y - \mathbb{E}[Y]}{\sigma_Y}\right] = \frac{cov(X, Y)}{\sigma_X \sigma_Y}$$

It follows that

- $-1 \leq \rho \leq 1$
- $\rho = 1 \Leftrightarrow (X - \mathbb{E}[X]) = c(Y - \mathbb{E}[Y])$
- $\rho = 0 \Leftarrow$ *independence*

.