Probability and Statistics notes

Aditya Mehrotra

April 2021

Contents

1 Probability

1.1 Basic mathematical definitions

Compliment: Given a set of of elements A where $A \subseteq S$ for some set of elements S, we define the compliment of A in S as the set of elements of S that are not in A. Mathematically, this is expressed as A^c or $S - A$.

Difference: The difference between sets A and B is all elements that are in A but not in B. Mathematically, this is expressed as $A - B$

Disjoint: A and B are disjoint if they have no common elements. Mathematically, this is when $A \cap B = \emptyset$

Product of sets: The cartesian product of two sets S and T, denoted by $S \times T$:

$$S \times T = \{(s, t) \mid s \in S, t \in T\}$$

Cardinality of a set: The number of elements in a set S, mathematically denoted by $|S|$

Inclusion exclusion principle:

$$|A \cup B| = |A| + |B| - |A \cap B|$$

1.2 Basic probability terminology

Experiment: A repeatable procedure with well-defined possible outcomes

Sample space: The set of all possible outcomes. This is denoted by Ω and sometimes by S.

Event: A subset of the sample space

Probability function: A function assigning a probability for each outcome. More specifically, for a discrete sample space S, a probability function P assigns each outcome $\omega \in S$ a number $P(\omega)$. This number is called the probability of ω.

P must satisfy two rules:

Rule 1: $0 \leq P(\omega) \leq 1$ (probabilities are between 0 and 1)

Rule 2: Given $S = \{\omega_1, \omega_2, \omega_3, ..., \omega_n\}$, $\sum_{j=1}^{n} P(\omega_j) = 1$.

(The sum of probabilities of all possible outcomes add to 1)

Additionally, the probability of an event $E \subseteq S$ is the sum of the probabilities of all outcomes in E. More specifically,

$$P(E) = \sum_{\omega \in E} P(\omega)$$

Discrete sample space: A listable sample space, can be finite or infinite.

1.3 Basic probability rules

For some events A, L and R contained in a sample space Ω:

$P(A^c) = 1 - P(A)$

If L and R are disjoint, then $P(L \cup R) = P(L) + P(R)$

If L and R are not disjoint, we have the inclusion-exclusion principle:

$$P(L \cup R) = P(L) + P(R) - P(L \cap R)$$

1.4 Conditional Probability

Conditional probability: The probability of A given B. Mathematically, this is expressed as $P(A|B)$. More rigorously, given events A and B, the conditional probability of A given B is,

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \textit{ provided that } P(B) \neq 0$$

This gives us the multiplication rule,

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Law of Total Probability: Suppose the sample space Ω is divided into n disjoint events, $B_1, B_2, ..., B_n$. Then for any event A:

$$P(A) = P(A \cap B_1) + P(A \cap B_2) + ... + P(A \cap B_n)$$

$$P(A) = P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + ... + P(A|B_n)P(B_n)$$

1.5 Independence

Independence: Two events are independent if knowledge that one occurred does not change the probability that the other occurred. Mathematically, A is independent of B if $P(A|B) = P(A)$. The formal definition of independence is, two events A and B are independent if

$$P(A \cap B) = P(B \cap A) = P(A) \cdot P(B)$$

Notice that by the multiplication rule, we have $P(A \cap B) = P(A|B) \cdot P(B)$ and $P(B \cap A) = P(B|A) \cdot P(A)$. Notice that $P(A \cap B) = P(B \cap A)$
Therefore,

1. If $P(B) \neq 0$, then A and B are independent if and only if $P(A|B) = P(A)$
2. If $P(A) \neq 0$, then A and B are independent if and only if $P(B|A) = P(B)$

1.6 Bayes' Theorem

Bayes' Theorem: For events A and B, Bayes' theorem says

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

This theorem tells us how to invert conditional probabilities.

1.7 Discrete random variables

Discrete Random Variable: Let Ω be a sample space. A discrete random variable is a function

$$X : \Omega \to \mathbb{R}$$

That takes a discrete set of values. X is called a "random" variable because it's value depends on the outcome of an experiment. We treat random variables as regular variables.

For any value a, we write $X = a$ to mean the event consisting of all outcomes ω with $X(\omega) = a$.

Example: A game with 2 dice.
If we roll two dice and record the outcomes as (i, j), we get a sample space of:

$$\Omega = \{(1,1), (1,2), (1,3), ..., (6,6)\} = \{(i,j) \mid i,j = 1, ..., 6\}$$

The probability function is $P(i,j) = \frac{1}{36}$
An example of a random variable Y here is:

$$Y(i,j) = i + j$$

In this case, the event $Y = 12$ is the set $\{(6,6),(6,6)\}$ as this is the set of all outcomes that sum to 12. Therefore, $P(Y = 12) = \frac{1}{6}$.
However, we can use another value, even a value that Y never takes. For example, for the event $Y = 100$, since Y never equals 100 then this is just the empty event:

$$Y = 100 = \{\} = \emptyset \quad P(Y = 100) = 0$$

Probability mass function The probability mass function (pmf) of a discrete random variable X is the function $p(a) = P(X = a)$

1. $0 \leq p(a) \leq 1$
2. a can be any number, if it is a value that X never takes, then $p(a) = 0$

Events and inequalities
Inequalities with random variables describe events. For a random variable X, $X \leq a$ is the set of all outcomes ω such that $X(\omega) \leq a$

Cumulative distribution function (cdf): The cumulative distribution function of a random variable X is the function F given by $F(a) = P(X \leq a)$. $F(a)$ is called the *cumulative* distribution function because $F(a)$ gives the total probability that accumulates by adding up the probabilities $p(b)$ as b runs from $-\infty$ to a, where $F(a))$ is defined for all values a.

The cdf function F has some properties:

1. F is non-decreasing. If $a \leq b$, then $F(a) \leq F(b)$. The cumulative probability $F(a)$ increases or remains constant as a increases.
2. $0 \leq F(a) \leq 1$. The accumulated probability is always between 0 and 1.
3. $\lim\limits_{a \to \infty} F(a) = 1$. As a gets very large, it becomes more and more certain that $X \leq a$.
4. $\lim\limits_{a \to -\infty} F(a) = 0$. As a gets very negative, it becomes more and more certain that $X > a$

1.8 Specific distributions

Bernoulli distributions: A random variable X has a Bernoulli distribution with parameter p if:

1. X takes the values 0 and 1
2. $P(X = 1) = p$ and $P(X = 0) = 1 - p$

We write $X \sim \text{Bernoulli}(p)$, which is read as "$X$ follows a Bernoulli distribution with parameter p"

Binomial distribution: The binomial distribution, written as $\text{Binomial}(n, p)$, models the number of successes in n independent $\text{Bernoulli}(p)$ trials. A single binomial trial consists of n Bernoulli trials. For example, for coin flips the sample space for a Bernoulli trial is $\{H, T\}$. The sample space for a binomial trial is all sequences of heads and tails of length n. Likewise, a Bernoulli random variable takes values 0 and 1 and a binomial random variable takes values $0, 1, 2, ..., n$.

Due to this unique property, for all $x \in \{0, 1, 2..., n\}$,

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Geometric Distribution: A random variable X follows a geometric distribution with parameter p if

1. X takes values 0, 1, 2, 3, ...
2. It's pmf is given by $p(k) = P(X = k) = (1-p)^k p$

Uniform distribution: The uniform distribution models any situation where all the outcomes are equally likely. Mathematically this is written as

$$X \sim \text{ uniform}(N)$$

X takes values $1, 2, 3, ..., N$, each with probability $\frac{1}{N}$

1.9 Discrete Random Variables: Expected value

Expected Value: Suppose X is a discrete random variable that takes values $x_1, x_2, ..., x_n$, with probabilities $p(x_1), p(x_2), ..., p(x_n)$. The expected value of X is denoted as $E(X)$ and mathematically defined by

$$E(X) = \sum_{j=1}^{n} p(x_j)x_j = p(x_1)x_1 + p(x_2)x_2) + ... + p(x_n)x_n$$

Note that the expected value can also be expressed as μ, or mean/average. Additionally, expected value provides a measure of the location or central tendency of a random variable.

Algebraic properties of $E(X)$:
$E(X)$ is linear. If X and Y are random variables on a sample space Ω, then

1. $$E(X + Y) = E(X) + E(Y)$$

2. If a and b are constants then

$$E(aX + b) = aE(X) + b$$

Change of variables formula: X is a discrete random variable taking values $x_1, x_2, ...$ and h is a function where $h(X)$ is a new random variable. Its expected value is

$$E(h(X)) = \sum_{j} h(x_j)p(x_j)$$