40.017 PROBABILITY & STATISTICS

An Introduction to Probability & Statistics

Michael Hoon

January 25, 2024

Contents

1 Set Theory

1.1 Sample Spaces

The mathematical framework for probability is built around *sets*. The *sample space* S of an experiment is the set of all possible outcomes of the experiment. An *event* A is a subset of S, and we say that A occurred if the actual outcome is in A.

1.2 Naive Definition of Probability

Let A be an event for an experiment with a finite sample space S. A naive probability of A is

$$\mathbb{P}_{\text{naive}}(A) = \frac{|A|}{|S|} = \frac{\text{number of outcomes favorable to A}}{\text{total number of outcomes}} \tag{1}$$

In general, the result about complements always holds:

$$\mathbb{P}_{\text{naive}}(A^c) = \frac{|A^c|}{|S|} = \frac{|S| - |A|}{|S|} = 1 - \frac{|A|}{|S|} = 1 - \mathbb{P}_{\text{naive}}(A)$$

An important factor about the naive definition is that it is restrictive in requiring S to be finite.

1.3 General Definition of Probability

Definition 1.1. A probability space consists of a sample space S and a probability function P which takes an event $A \subseteq S$ as input and returns $P(A)$, where $P(A) \in \mathbb{R},\ P(A) \in [0, 1]$. The function must satisfy the following axioms:

1. $\mathbb{P}(\emptyset) = 1,\ \mathbb{P}(S) = 1$
2. $\mathbb{P}(A) \geq 0$
3. If $A_1, A_2, \ldots$ are **disjoint events**, then:

$$\mathbb{P}\left(\bigcup_{j=1}^{\infty}\right) = \sum_{j=1}^{\infty} \mathbb{P}(A_j)$$

Disjoint events are **mutually exclusive** (i.e. $A_i \cap A_j = \emptyset\ \forall\ i \neq j$).

1.3.1 Properties of Probability

Theorem 1.2. Probability has the following properties, for any events A and B:

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
2. If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$
3. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

1.3.2 Inclusion-Exclusion Principle

For any events $A_1, \ldots A_n$,

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_i \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n) \tag{2}$$

For $n = 2$, we have a nicer result:

$$\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$$

1.4 Conditional Probability

Definition 1.3. If A and B are events with $\mathbb{P}(B) > 0$, then the *conditional probability* of A given B, denoted by $\mathbb{P}(A \mid B)$ is defined as:

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

Here A is the event whose uncertainty we want to update, and B is the evidence we observe. $\mathbb{P}(A)$ is the *prior* probability of A and $\mathbb{P}(A|B)$ is the *posterior* probability of A. (For any event A, $\mathbb{P}(A|A) = \frac{\mathbb{P}(A \cap A)}{\mathbb{P}(A)}$).

2 Derangement

A derangement is a permutation of the elements of a set in which no element appears in its original position. We use D_n to denote the number of derangements of n distinct objects.

2.1 Counting Derangements

We consider the number of ways in which n hats $(h_1, \ldots, h_n)$ can be returned to n people $(P_1, \ldots, P_n)$ such that no hat makes it back to its owner.

We obtain the recursive formula:

$$D_n = (n-1)(D_{n-1} + D_{n-2}),\ \forall\, n \geq 2 \tag{3}$$

With the initial conditions $D_1 = 0$ and $D_2 = 1$, we can use the formula to recursively compute D_n for any n.

There are various other expressions for D_n, equivalent to formula 3:

$$D_n = n! \sum_{i=0}^{n} \frac{(-1)^i}{i!},\ \forall\, n \geq 0 \tag{4}$$

2.1.1 Limiting Growth

From Equation 4, and the taylor series expansion for e:

$$e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!} \tag{5}$$

we substitute $x = -1$ and obtain the limiting value as $n \to \infty$:

$$\lim_{n\to\infty} \frac{D_n}{n!} = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{(-1)^i}{i!} = e^{-1} \approx 0.367879\ldots$$

This is the limit of the probability that a randomly selected permutation of a large number of objects is a derangement. The probability converges to this limit extremely quickly as n increases, which is why D_n is the nearest integer to $\frac{n!}{e}$.

3 Discrete Random Variables

We formally define a random variable:

Definition 3.1. Given an experiment with sample space S, a *random variable* (r.v.) is a function from the sample space S to the real numbers $\mathbb{R}$. It is common to denote random variables by capital letters.

Thus, a random variable X assigns a numerical value $X(s)$ to each possible outcome s of the experiment. The randomness comes from the fact that we have a random experiment (with Probabilities described by the probability function P); the mapping itself is deterministic.

There are two main types of random variables used in practice: *discrete* and *continuous* r.v.s.

Definition 3.2. A random variable X is said to be *discrete* if there is a finite list of values $a_1, a_2, \ldots, a_n$ or an infinite list of values $a_1, a_2, \ldots$ such that $\mathbb{P}(X = a_j \text{ for some } j) = 1$. If X is a discrete r.v., then the finite or countably infinite set of values x such that $P(X = x) > 0$ is called the *support* of X.

3.1 Binomial

3.2 Hypergeometric

If we have an urn filled with w white and b black balls, then drawing n balls out of the urn *with replacement* yields a $\text{Binom}(n, \frac{w}{(w+b)})$. If we instead sample *without replacement*, then the number of white balls follow a **Hypergeometric** distribution.

Theorem 3.3. If $X \sim \text{hypgeo}(n, j, k)$, then the PMF of X is:

$$\mathbb{P}(X = x) = \frac{\binom{j}{x}\binom{k}{n-x}}{\binom{j+k}{n}}$$

$\forall x \in \mathbb{Z}$ satisfying $0 \leq x \leq n$ and $0 \leq n - x \leq j$, and $P(X = x) = 0$ otherwise.

If j and k are large compared to n, then selection without replacement can be approximated by selection with replacement. In that case, the hypergeometric RV $X \sim \text{hypgeo}(n, j, k)$ can be approximated by a binomial RV $Y \sim \text{binomial}(n, p)$, where $p := \frac{j}{j+k}$ is the probability of selecting a black marble.

We can also write X as the sum of (dependent) Bernoulli random variables:

$$X = X_1 + X_2 + \cdots + X_n$$

where each X_i equals 1 if the ith selected marble is black, and 0 otherwise.

3.2.1 Hypergeometric Symmetry

Theorem 3.4. The $\text{hypergeo}(w, b, n)$ and $\text{hypergeo}(n, w + b - n, w)$ distributions are identical.

The proof follows from swapping the two sets of tags in the Hypergeometric story (white/black balls in urn) [3].

[3] The binomial and hypergeometric distributions are often confused. Note that in Binomial distributions, the Bernoulli trials are **independent**. The Bernoulli trials in Hypergeometric distribution are **dependent**, since the sampling is done *without replacement*.

3.3 Geometric

3.4 Negative Binomial

In a sequence of independent Bernoulli trials with success probability p, if X is the number of failures before the rth success, then X is said to have the Negative Binomial distribution with parameters r and p, denoted $X \sim \mathrm{NBin}(r, p)$.

Both the Binomial and Negative Binomial distributions are based on independent Bernoulli trials; they differ in the *stopping rule* and in what they are counting. The Negative Binomial counts the **number of failures until a fixed number of successes**.

Theorem 3.5. If $X \sim \mathrm{NBin}(r, p)$, then the PMF of X is

$$P(X = x) = \binom{x-1}{n-1}(1-p)^{x-n}p^n, \ \forall\, x \geq n$$

4