40.017 PROBABILITY & STATISTICS

Lecture Notes

Michael Hoon

February 1, 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 $X \sim \text{NBin}(r, p)$.

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 $X \sim \text{NBin}(r, p)$, then the PMF of X is

$$P(X = x) = \binom{x-1}{n-1}(1-p)^{x-n}p^n, \ \forall\, x \geq n \tag{6}$$

4 Law of Large numbers

Assume that we have i.i.d. $X_1, X_2, \ldots$ with finite mean μ and finite variance σ^2. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Definition 4.1. The (Weak) Law of Large Numbers (LLN) says that as n grows, the sample mean $\bar{X}_n$ converges to the true mean μ. Mathematically,

$$\forall \epsilon > 0, \ \mathbb{P}(|\bar{X}_n - \mu < \epsilon) = 1, \text{ as } n \to \infty \tag{7}$$

For any positive margin ϵ, as n gets arbitrarily large, the probability that $\bar{X}_n$ is within ϵ of μ approaches 1.

Note that the LLN does not contradict the fact that a coin is memoryless (in the repeated coin toss experiment). The LLN states that the proportion of Heads converges to $\frac{1}{2}$, but this does not imply that after a long string of Heads, the coin is "due" for a Tails to "*balance things out*". Rather, the convergence takes place through *swamping*: past tosses are swamped by the infinitely many tosses that are yet to come.

4.1 Inequalities

The inequalities in this section provide bounds on the probability of an r.v. taking on an 'extreme' value in the right or left rail of a distribution.

4.1.1 Markov's Inequality

Definition 4.2. Let X be any random variable that takes only non-negative values, that is, $\mathbb{P}(X < 0) = 0$. Then for any constant $a > 0$, we have:

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a} \tag{8}$$

For an intuitive interpretation, let X be the income of a randomly selected individual from a population. Taking $a = \mathbb{E}(X)$, Markov's Inequality says that $\mathbb{P}(\mathbb{X} \geq \nvDash\mathbb{E}(\mathbb{X})) \leq \frac{\nVdash}{\nvDash}$. i.e., it is impossible for more than half the population to make at least twice the average income.

4.1.2 Chebyshev's Inequality

Gives general bounds for the probability of being k standard deviations (SD) away from the mean.

Definition 4.3. Let Y be any random variable with mean $\mu < \infty$ and variance $\sigma^2 > 0$. Then for any constant $k > 0$, we have:

$$\mathbb{P}(|Y - \mu| \geq k\sigma) \leq \frac{1}{k^2} \tag{9}$$

5 Central Limit Theorem

Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance σ^2.

Definition 5.1. The CLT states that for large n, the distribution of $\bar{X}_n$ after standardisation approaches a standard Normal distribution. By standardisation, we mean that we subtract μ, the mean of $\bar{X}_n$, and divide by $\frac{\sigma}{\sqrt{n}}$, the standard deviation of $\bar{X}_n$.

$$\lim_{n\to\infty} \mathbb{P}(\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \leq x) = \Phi(x) \tag{10}$$

which is the cdf of the standard normal. Informally, when n is large (≥ 30), then $\bar{X}_n$ and $\sum_{i=1}^{n} X_i$ can each be approximated by a normal RV with the same mean and variance; the actual distribution of X_i becomes irrelevant:

$$\bar{X}_n \approx N(\mu, \frac{\sigma^2}{n}), \qquad \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)$$

6 Moments

6.1 Interpreting Moments

Definition 6.1. Let X be an r.v. with mean μ and variance σ^2. For any positive integer n, the n^{th} moment of X is $\mathbb{E}(X^n)$, the n^{th} central moment is $\mathbb{E}((X - \mu)^n)$.

In particular, the mean is the first moment and the variance is the second central moment.

6.2 Moment Generating Functions

A moment generating function, as its name suggests, is a generating function that encodes the **moments** of a distribution. Starting with an infinite sequence $(a_0, a_1, a_2, \dots)$, we 'condense' or 'store' it as a single function g, the generating function of the sequence:

$$\sum_{n=0}^{\infty} a_n \frac{t^n}{n!} := g(t)$$

Definition 6.2. When we take $a_n = \mathbb{E}(X^n)$, the resulting generating function is known as the **moment generating function (MGF)** of X, and is denoted by $M_X(t)$.

The MGF of X can be computed as an expected value:

Note that $M_X(0) = 1$ for any valid MGF.

6.3 Formulas & Theorems

Some important formulas for the MGF of X:

$$\boxed{M_X(t) = \mathbb{E}(e^{tX})} \tag{11}$$

where if X is **discrete** with pmf f, then

$$M_X(t) = \sum_{allx_i} e^{tx_i} f(x_i) \tag{12}$$

and if X is **continuous** with pdf f, then

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f(x)\, \mathrm{d}x \tag{13}$$

Theorem 6.3. Given the MGF of X, we can get the n^{th} moment of X by evaluating the n^{th} derivative of the MGF at 0:

$$\boxed{\mathbb{E}(X^n) = M_X^{(N)}(0)} \tag{14}$$

Theorem 6.4. If X and Y are independent, then the MGF of $X + Y$ is the product of the individual MGFs:

$$M_{X+Y}(t) = M_X(t)M_Y(t) \tag{15}$$

This is true because if X and Y are independent, then $\mathbb{E}(e^{t(X+Y)}) = \mathbb{E}(e^{tX})\mathbb{E}(e^{tY})$

Theorem 6.5. If two random variables have the same MGF, then they have the same distribution (same cdf, equivalently, same pdf or pmf) [a].

[a]For this to apply, the MGF needs to exist in an open interval around $t = 0$

6.4 Examples (Discrete)

6.4.1 Binomial MGF

We have $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$. The MGF can be found by:

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} \binom{n}{x} (\underbrace{e^t p}_{a})^x (\underbrace{1-p}_{b})^{n-x} \\ &= (e^t p + 1 - p)^n \end{aligned} \tag{16}$$

by using the fact that

$$\sum_x \binom{n}{x} a^x b^{n-x} = (a+b)^n$$

and from which we can obtain $\mathbb{E}(X) = M_X'(0) = n\,\overbrace{(e^t p + 1 - p)^{n-1} \cdot e^t p \mid_{t=0}}^{p} = np$

6.4.2 Poisson MGF

For a Poisson r.v., where $X \sim \text{Poisson}(\lambda)$ We have $f(x) = e^{-\lambda}\frac{\lambda^x}{x!}$. Then,

$$
\begin{aligned}
M_X(t) &= \sum_{x=0}^{\infty} e^{tx} e^{-\lambda} \frac{\lambda^x}{x!} \qquad (17) \\
&= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(e^t \lambda)^x}{x!} \\
&= e^{-\lambda} e^{e^t \lambda} \\
&= e^{e^t \lambda - \lambda} \\
&= e^{\lambda(e^t - 1)}
\end{aligned}
$$

We can now find:

$$M'_X(t) = e^{\lambda(e^t-1)}(\lambda e^t)$$

and therefore

$$M'_X(0) = e^0(\lambda e^0) = \lambda$$

6.5 Examples (Continuous)

6.5.1 Standard Normal

If $Z \sim \mathcal{N}(0,1)$ is a standard normal r.v., then $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$. For continuous distributions, we need to use the infinite integral:

$$
\begin{aligned}
M_Z(t) = \mathbb{E}(e^{tZ}) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{tx} e^{-\frac{x^2}{2}} \, \mathrm{d}x \qquad (18) \\
&= \qquad (19)
\end{aligned}
$$

References

Some references used in these notes:

Introduction to Probability, Joe Blitzstein & Jessica Hwang.