

POL502: Probability

Kosuke Imai
Department of Politics, Princeton University

December 12, 2003

1 Probability and Independence

To define probability, we rely on set theory we learned in the first chapter of this course. In particular, we consider an experiment (or trial) and its result called an outcome. Tossing a coin is a simple experiment anyone can do, but more complicated phenomena such as elections can also be considered as an experiment.

Definition 1 *The set of all possible outcomes of an experiment is called the sample space of the experiment and denoted by Ω . Any subset of Ω is called an event.*

That is, an event is any collection of possible outcomes of an experiment.

Definition 2 *A collection of subsets of Ω is called a sigma algebra (or sigma field) and denoted by \mathcal{F} if it satisfies the following properties*

1. If $\{A_n\}_{n=1}^{\infty}$ is a sequence of sets such that $A_n \in \mathcal{F}$ for any $n \in \mathbf{N}$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$.
2. If $A \in \mathcal{F}$, then $A^C \in \mathcal{F}$.
3. $\emptyset \in \mathcal{F}$.

The definition implies that $\bigcap_{n=1}^{\infty} A_n$ is a sigma algebra and that \emptyset and Ω belong to any sigma algebra (why?). Given a particular Ω , we have many sigma algebras the smallest of which is $\{\emptyset, \Omega\}$ and called a *trivial sigma algebra*. To sum up, any experiment is associated with a pair (Ω, \mathcal{F}) called a *measurable space*; where Ω is the set of all possible outcomes and \mathcal{F} contains all events whose occurrence we are interested. An example may help understand these concepts.

Example 1 *What is the sample space of flipping a fair coin twice? What is the sigma algebra which consists of all subsets of the sample space?*

Now, we are ready to define probability. The following is called *Probability Axiom* (or *Kolmogorov's Axiom*).

Axiom 1 *Given an experiment with a measurable space (Ω, \mathcal{F}) , a probability measure $P : \mathcal{F} \mapsto [0, 1]$ is a function satisfying*

1. $P(\emptyset) = 0$
2. $P(\Omega) = 1$

3. If $A_1, A_2, \dots \in \mathcal{F}$ is a collection of disjoint sets (i.e., $A_n \cap A_m = \emptyset$ for all pairs of n and m with $n \neq m$), then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$$

(Ω, \mathcal{F}, P) is called a probability space.

Given an event A , if $P(A) = 0$, then A is called *null*. If $P(A) = 1$, we say A occurs *almost surely*. Note that null events are not necessarily impossible. In fact, they occur all the time (why?).

Example 2 What is the probability that we get heads twice in the experiment of Example 1?

We derive some familiar (and maybe unfamiliar) properties of probability.

Theorem 1 (Probability) Let P be a probability measure and $A, B \in \mathcal{F}$.

1. $P(A^C) = 1 - P(A)$.
2. If $A \subset B$, then $P(B) = P(A) + P(B \setminus A) \geq P(A)$.
3. $P(A^C \cap B) = P(B) - P(A \cap B)$.
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Example 3 $P(A \cap B) \geq P(A) + P(B)$ is a special case of Bonferroni's inequality and can be used to bound the probability of a simultaneous event is unknown but the probabilities of the individual events are known.

We can extend these properties to a sequence of sets.

Theorem 2 (Probability and Sequence of Sets) Let P be a probability measure.

1. If $\{A_n\}_{n=1}^{\infty}$ is an increasing (decreasing) sequence of events such that $A_1 \subset A_2 \subset \dots$ ($A_1 \supset A_2 \supset \dots$), then for $A = \lim_{n \rightarrow \infty} A_n$ we have

$$P(A) = \lim_{n \rightarrow \infty} P(A_n)$$

2. (Boole's Inequality) If $\{A_n\}_{n=1}^{\infty}$ be a sequence of sets, then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} P(A_n)$$

Next, we study the conditional probability and independence.

Definition 3 If $P(B) > 0$, then the conditional probability that A occurs given that B occurs is defined as

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

The conditional probability can be very tricky as the following example shows.

Example 4 A couple is expecting twins.

1. In a ultrasound examination, the technician was only able to determine that one of the two was boy. What is the probability that both are boys?
2. During the delivery, the baby that was born first was a boy. What is the probability that both are boys?

Using similar reasoning, you should be able to find an answer of the famous Monte Hall problem.

Theorem 3 (Conditional Probability) For any events A and B such that $0 < P(B) < 1$, $P(A) = P(A | B)P(B) + P(A | B^C)P(B^C)$. More generally, let B_1, B_2, \dots, B_n be a partition of Ω such that $0 < P(B_i) < 1$ for all i , $B_i \cap B_j = \emptyset$ for all $i \neq j$, and $\Omega = \bigcup_{i=1}^{\infty} B_i$. Then, $P(A) = \sum_{i=1}^{\infty} P(A | B_i)P(B_i)$.

From this, one can derive the theorem named after Reverend Thomas Bayes.

Theorem 4 (Bayes' Rule) Let A_1, A_2, \dots be a partition of Ω and B be any set. Then, for all i ,

$$P(A_i | B) = \frac{P(B | A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B | A_j)P(A_j)}.$$

Note that the denominator is equal to $P(B)$.

Example 5 A particular district has 40% of republicans and 60% of democrats. The overall turnout was 40%, and we know that 45% of republicans voted. What is the probability of someone being a democrat given that she voted?

Definition 4 Two events A and B are said to be independent if $P(A \cap B) = P(A)P(B)$. More generally, a family of events $\{A_n : n \in \mathbf{N}\}$ is independent if $P(\bigcap_{n=1}^{\infty} A_n) = \prod_{n=1}^{\infty} P(A_n)$.

One can similarly define the concept of *conditional independence*: i.e., A and B are conditionally independent given C if $P(A \cap B | C) = P(A | C)P(B | C)$ provided $P(C) > 0$. The following theorems are easy to prove.

Theorem 5 (Independence) Let A and B be independent events. Then,

1. A and B^C (and A^C and B) are independent.
2. A^C and B^C are independent.

2 Random Variables and Probability Distributions

Often, we are more interested in some consequences of experiments than experiments themselves. For example, a gambler is more interested in how much they win or lose than the games they play. Formally, this is a function which maps the sample space into \mathbf{R} or its subset.

Definition 5 A random variable is a function $X : \Omega \mapsto \mathbf{R}$ satisfying $A(x) = \{\omega \in \Omega : X(\omega) \leq x\} \in \mathcal{F}$ for all $x \in \mathbf{R}$. Such a function is said to be \mathcal{F} -measurable.

After an experiment is done, the outcome $\omega \in \Omega$ is revealed and a random variable X takes some value. The distribution function of a random variable describes how likely it is for X to take a particular value.

Definition 6 The distribution function of a random variable X is the function $F : \mathbf{R} \mapsto [0, 1]$ given by $F(x) = P(A(x))$ where $A(x) = \{\omega \in \Omega : X(\omega) \leq x\}$ or equivalently $F(x) = P(X \leq x)$.

Now, we understand why the technical condition $A(x) \in \mathcal{F}$ in Definition 5 was necessary. We sometimes write F_X and P_X in order to emphasize these functions are defined for the random variable X . The two random variables, X and Y , are said to be *distributed identically* if $P(X \in A) = P(Y \in A)$ for any $A \in \mathcal{F}$. This implies in turn that $F_X(x) = F_Y(x)$ for any x . Finally, a distribution function has the following properties.

Theorem 6 (Distribution Function) A distribution function $F(x)$ of a random variable X satisfies the following properties.

1. $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow \infty} F(x) = 1$
2. If $x < y$, then $F(x) \leq F(y)$.
3. F is right-continuous: that is, $\lim_{x \downarrow c} F(x) = F(c)$ for any $c \in \mathbf{R}$.

Given this theorem, one can prove the following: $P(X > x) = 1 - F(x)$, $P(x < X \leq y) = F(y) - F(x)$, and $P(X = x) = F(x) - \lim_{y \uparrow x} F(y)$.

Example 6 Two examples of random variable and its distribution function.

1. **Bernoulli, Geometric.** In a coin toss experiment, a Bernoulli random variable can be defined as $X(\text{head}) = 1$ and $X(\text{tail}) = 0$. What is the distribution function? What about the distribution function of a random variable which represents the number of tosses required to get a head?
2. **Logistic.** A special case of the logistic distribution is given by $F(x) = \frac{1}{1+e^{-x}}$. Confirm that this satisfies Theorem 6.

One can classify random variables into two classes based on the probability function.

Definition 7 Let X be a random variable.

1. X is said to be discrete if it takes values in a countable subset $\{x_1, x_2, \dots\}$ of \mathbf{R} . The discrete random variable has probability mass function $f : \mathbf{R} \mapsto [0, 1]$ given by $f(x) = P(X = x)$.

2. X is said to be continuous if its distribution function can be expressed as

$$F(x) = \int_{-\infty}^x f(t) dt$$

for $x \in \mathbf{R}$ for some integrable function $f : \mathbf{R} \mapsto [0, \infty)$ called the probability density function.

We may write $f_X(x)$ to stress the role of X . For discrete distributions, the distribution function and probability mass function are related by

$$F(x) = \sum_{x_n \leq x} f(x_n) \quad \text{and} \quad f(x) = F(x) - \lim_{y \uparrow x} F(y)$$

The mass function has the following property $\sum f(x_n) = 1$.

Example 7 Two examples of discrete random variables. What real world phenomena can we model using these random variables?

1. **Binomial.** The sum of n identically distributed Bernoulli random variables with probability of success p is a Binomial random variable, which takes the values in the set $\{0, 1, 2, \dots, n\}$. The probability mass function with parameters p and n is

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

2. **Poisson.** A Poisson random variable X takes values in the set $\{0, 1, 2, \dots\}$ with the probability mass function and the parameter λ

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

3. **Poisson approximation to Binomial.** Show that if n is large and p is small, Poisson pmf can approximate Binomial pmf.

For continuous distributions, if F is differentiable at x the fundamental theorem of calculus implies $f(x) = F'(x)$. The density function has the following properties: $\int_{-\infty}^{\infty} f(x) dx = 1$, $P(X = x) = 0$ for all $x \in \mathbf{R}$, and $P(a \leq X \leq b) = \int_a^b f(x) dx$.

Example 8 Five examples of continuous distributions.

1. **Gamma, Exponential, χ^2 .** A gamma random variable takes non-negative values and has the following density function with the parameters $\alpha > 0$ (shape parameter), $\beta > 0$ (scale parameter),

$$f(x) = \frac{\beta^\alpha x^{\alpha-1}}{\Gamma(\alpha)} e^{-\beta x} \quad \text{where} \quad \Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-t} dt$$

The exponential distribution is a special case of the Gamma distribution with $\alpha = 1$, i.e., $f(x) = \beta e^{-\beta x}$. This distribution has “memoryless” property. Another important special case occurs when $\alpha = p/2$ and $\beta = 1/2$, and is called χ^2 distribution with p degrees of freedom.

2. **Beta, Uniform.** A beta random variable takes values in $[0, 1]$ and has the following density function with the parameters $\alpha, \beta > 0$

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \quad \text{where} \quad B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$$

The beta function is related to the gamma function by the identify

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)}$$

A uniform random variable X takes values in a closed interval, $[a, b]$, with the density function

$$f(x) = \frac{1}{b-a}$$

when $a = 0$ and $b = 1$, it is a special case of Beta distribution ($\alpha = \beta = 1$).

3. **Normal (Gaussian).** A Normal distribution has two parameters, mean μ and variance σ^2 ,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right]$$

If $\mu = 0$ and $\sigma^2 = 1$, then it is called the standard Normal distribution.

So far, we have considered a single random variable. However, the results can be extended to a random vector, a vector of multiple random variables. For example, we can think about an experiment where we throw two dice instead of one die at each time. Then, we need to define the *joint* probability mass (density) function for a random vector. For simplicity, we only consider bivariate distributions, but the same principle applies to multivariate distributions in general.

Definition 8 Let X and Y be random variables. The joint distribution function of X and Y is the function $F : \mathbf{R}^2 \mapsto [0, 1]$ defined by $F(x, y) = P(X \leq x, Y \leq y)$.

1. If (X, Y) is a discrete random vector, the joint probability mass function $f : \mathbf{R}^2 \mapsto \mathbf{R}$ is defined by $f(x, y) = P(X = x, Y = y)$.
2. If (X, Y) is a continuous random vector, the joint probability density function $f : \mathbf{R}^2 \mapsto \mathbf{R}$ is defined by

$$F(x, y) = \int_{-\infty}^y \int_{-\infty}^x f(s, t) ds dt$$

Example 9 Two examples of multivariate distributions.

1. **Multinomial.** An n -dimensional multinomial random vector $X = (X_1, \dots, X_n)$ has the following probability mass function.

$$f(x_1, x_2, \dots, x_n) = \frac{m!}{x_1! x_2! \dots x_n!} p_1^{x_1} p_2^{x_2} \dots p_n^{x_n}$$

where (p_1, p_2, \dots, p_n) is an $n \times 1$ vector of probabilities with $\sum_{i=1}^n p_i = 1$ and $m = \sum_{i=1}^n x_i$.

2. **Multivariate Normal.** An n -dimensional multivariate normal random vector $X = (X_1, \dots, X_n)$ with the following density function

$$f(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp \left[-\frac{1}{2} (\mathbf{x} - \mu)^\top \Sigma^{-1} (\mathbf{x} - \mu) \right]$$

where μ is an $n \times 1$ vector of mean and Σ is an $n \times n$ positive definite covariance matrix.

In addition to joint and marginal distributions, *conditional* distributions are often of interest.

Definition 9 Let X and Y be random variables with marginal probability mass (density) functions, $f_X(x)$ and $f_Y(y)$, and joint probability mass (density) functions, $f(x, y)$.

1. The conditional mass (density) function of Y given X is defined by

$$f(y | x) = \frac{f(x, y)}{f_X(x)}$$

2. X and Y are said to be independent if $f(x, y) = f_X(x)f_Y(y)$.

If X and Y are not independent, the direct way to obtain the marginal density of X from the joint density of (X, Y) is to integrate out Y . That is, $f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$. If Y is a discrete random variable, one needs to sum over Y . That is, $f_X(x) = \sum_{y \in \mathbf{R}} f(x, y)$. We end this section with the following examples.

Example 10 Consider two multivariate random variables defined in Example 9.

1. Let (X_1, X_2, \dots, X_n) be a multinomial random vector. Show that the marginal distribution of x_i for any $i \in \{1, \dots, n\}$ is a Binomial distribution. Also, show that $(X_1, X_2, \dots, X_{n-1})$ conditional on $X_n = x_n$ follows a multinomial distribution.
2. Rewrite the bivariate Normal density function using means μ_1, μ_2 , variances σ_1, σ_2 and the correlation ρ .

3 Expectations and Functions of Random Variables

We have studied the behavior of random variables. In this section, we are concerned about their expectation (or mean value, expected value).

Definition 10 Let X be a discrete (continuous) random variables with probability mass (density) function f . Then, the expectation of X is defined as

1. $E(X) = \sum_x xf(x)$ if X is discrete.
2. $E(X) = \int_{-\infty}^{\infty} xf(x) dx$ if X is continuous.

If $E(|X|) = \infty$, then we say the expectation $E(X)$ does not exist. One sometimes write E_X to emphasize that the expectation is taken with respect to a particular random variable X . The expectation has the following properties, all of which follow directly from the properties of summation and integral. In particular, the expectation is a linear operator (e.g., $1/E(X) \neq E(1/X)$).

Theorem 7 (Expectation) Let X and Y be random variables with probability mass (density) function f_X and f_Y , respectively. Assume that their expectations exist, and let g be any function.

1. $E[g(X)] = \sum_x g(x)f_X(x)$ if X is discrete, and $E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x) dx$ if X is continuous.
2. If $g_1(x) \geq g_2(x)$ for all x and any functions g_1 and g_2 , then $E[g_1(X)] \geq E[g_2(X)]$.
3. $E(aX + bY) = aE(X) + bE(Y)$ for any $a, b \in \mathbf{R}$, and in particular $E(a) = a$.

Example 11 Calculate the expectations of the following random variables if they exist.

1. Bernoulli random variable.
2. Binomial random variable.
3. Poisson random variable.
4. **Negative binomial.** A negative binomial random variable is defined as the number of failures of Bernoulli trials that is required to obtain r successes. Its probability mass function is

$$f(x) = \binom{r+x-1}{x} p^r (1-p)^x$$

where p is the probability of a success. A special case of negative binomial distribution when $r = 1$ is geometric distribution. Note that geometric distribution has the memoryless property that we studied for the exponential distribution: i.e, $P(X > s | X > t) = P(X > s - t)$ for $s > t$. You should be able to prove this by now.

5. Gamma random variable.
6. Beta random variable.
7. Normal random variable.
8. **Cauchy.** A Cauchy random variable takes a value in $(-\infty, \infty)$ with the following symmetric and bell-shaped density function.

$$f(x) = \frac{1}{\pi[1 + (x - \mu)^2]}$$

Using expectation, we can define the moments of a random variable.

Definition 11 Let X and Y be a random variable with their expectations μ_X and μ_Y .

1. If k is a positive integer, the k th moment m_k of X is defined to be $m_k = E(X^k)$. The k th central moment is defined as $\sigma^k = E[(X - \mu_X)^k]$. In particular, σ^2 is called variance and its square root σ is called standard deviation.
2. The covariance of X and Y is defined as $\text{cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$.
3. The correlation (coefficient) of X and Y is defined as $\rho_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$.

The following properties about the variances are worth memorizing.

Theorem 8 (Variances and Covariances) Let X and Y be random variables and $a, b \in \mathbf{R}$.

1. $\text{var}(aX + b) = a^2 \text{var}(X)$.
2. $\text{var}(aX + bY) = a^2 \text{var}(X) + b^2 \text{var}(Y) + 2\text{cov}(X, Y)$.
3. $\text{cov}(X, Y) = E(XY) - E(X)E(Y)$ and in particular $\text{var}(X) = E(X^2) - [E(X)]^2$.

Example 12 Find a variance of the random variables in Example 11.

One way to find moments is to consider moment generating functions.

Definition 12 Let X be a random variable with a distribution function F . The moment generating function of X is $M(t) = E_X(e^{tX})$ provided that this expectation exists for t in some neighborhood of 0.

Theorem 9 (Moment Generating Functions) If a random variable X has the moment generating function $M(t)$, then $E(X^n) = M^{(n)}(0)$ where $M^{(n)}(0)$ is the n th derivative of $M(t)$ evaluated at 0.

We end this section with a few useful theorems about expectation.

Theorem 10 (Independence and Expectation) If X and Y are independent random variables, then $E(XY) = E(X)E(Y)$. That is, both correlation and covariances are zero. More generally, $E[g(X)h(Y)] = E[g(X)]E[h(Y)]$ for any function $g(x)$ and $h(y)$.

Note that the converse is not true. One can also define conditional expectation and variance with respect to conditional distributions. They have the following useful properties.

Theorem 11 (Conditional Expectation and Conditional Variance) Let X and Y be random variables.

1. $E(X) = E[E(X | Y)]$
2. $\text{var}(X) = E[\text{var}(X | Y)] + \text{var}[E(X | Y)]$.

There are important inequalities involving the expectation. We study the following two.

Theorem 12 (Jensen's Inequality) Let X be a random variable. If g is a concave function, then $E[g(X)] \geq g(E(X))$

Theorem 13 (Hölder's Inequality) Let X and Y be random variables and $p, q > 1$ satisfying $1/p + 1/q = 1$. Then, $|E(XY)| \leq E(|XY|) \leq [E(|X|^p)]^{1/p} [E(|Y|^q)]^{1/q}$. When $p = q = 2$, the inequality is called Cauchy-Schwartz inequality. When $Y = 1$, it is Liapounov's inequality.

Theorem 14 (Chebychev's Inequality) Let X be a random variable and let $g(x)$ be a nonnegative function. Then, for any $\epsilon > 0$,

$$P(g(X) \geq \epsilon) \leq \frac{E[g(X)]}{\epsilon}$$

Example 13 Let X be any random variable with mean μ and variance σ^2 . Use Chebychev's inequality to show that $P(|X - \mu| \geq 2\sigma) \leq 1/4$.

Next, we study the functions of random variables and their distributions. For example, we want to answer the questions, what is the distribution of $Y = g(X)$ given a random variable X with distribution function $F(x)$?

Theorem 15 (Transformation of Univariate Random Variables) Let X be a continuous random variable with probability function F_X and probability density function $f_X(x)$. Define $Y = g(X)$ where g is a monotone function. Let also \mathcal{X} and \mathcal{Y} denote the support of distributions for X and Y , respectively.

1. If g is an increasing (decreasing) function, the probability function of Y is given by $F_Y(y) = F_X(g^{-1}(y))$ ($F_Y(y) = 1 - F_X(g^{-1}(y))$) for $y \in \mathcal{Y}$.
2. The probability density function of Y is given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{for } y \in \mathcal{Y} \\ 0 & \text{otherwise} \end{cases}$$

Example 14 Derive the probability function for the following transformations of a random variable.

1. **Uniform.** $Y = -\log(X)$ where $X \sim Unif(0, 1)$.
2. **Inverse Gamma.** $Y = 1/X$ where $X \sim Gamma(\alpha, \beta)$.

This rule can be generalized to the transformation of multivariate random variables.

Theorem 16 (Transformation of Bivariate Random Variables) Let (X, Y) be a vector of two continuous random variables with joint probability density function $f_{X,Y}(x, y)$. Consider a bijective transformation $U = g_1(X, Y)$ and $V = g_2(X, Y)$. Define the inverse of this transformation as $X = h_1(U, V)$ and $Y = h_2(U, V)$. Then, the joint probability density function for (U, V) is given by

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v)) |J|$$

where $J = \begin{vmatrix} \frac{dx}{du} & \frac{dx}{dv} \\ \frac{dy}{du} & \frac{dy}{dv} \end{vmatrix}$ is called Jacobian.

Example 15 Derive the distribution of the following random variables.

1. The joint distribution of $U = X + Y$ and $V = X - Y$ where both X and Y are independent standard normal random variables.
2. The joint distribution of $U = X + Y$ and $V = X/Y$ where both X and Y are independent exponential random variables.