A Probability Primer
Math 6070, Spring 2013

Davar Khoshnevisan
University of Utah

January 23, 2013

Contents

1 Probabilities

Let $\mathcal{F}$ be a collection of sets. A *probability* P is a function, on $\mathcal{F}$, that has the following properties:

1. $\mathrm{P}(\varnothing) = 0$ and $\mathrm{P}(\Omega) = 1$;
2. If $A \subset B$ then $\mathrm{P}(A) \le \mathrm{P}(B)$;
3. (*Finite additivity*). If A and B are disjoint then $\mathrm{P}(A \cup B) = \mathrm{P}(A) + P(B)$;
4. For all $A, B \in \mathcal{F}$, $\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B) - \mathrm{P}(A \cap B)$;
5. (*Countable Additivity*). If $A_1, A_2, \ldots \in \mathcal{F}$ are disjoint, then $\mathrm{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$.

2 Distribution Functions

Let X denote a random variable. It *distribution function* is the function

$$F(x) = \mathrm{P}\{X \le x\}, \tag{1}$$

defined for all real numbers x. It has the following properties:

1. $\lim_{x \to -\infty} F(x) = 0$;
2. $\lim_{x \to \infty} F(x) = 1$;
3. F is right-continuous; i.e., $\lim_{x \downarrow y} F(x) = F(y)$, for all real y;
4. F has left-limits; i.e., $F(y-) := \lim_{x \uparrow y} F(x)$ exists for all real y. In fact, $F(y-) = \mathrm{P}\{X < y\}$;
5. F is non-decreasing; i.e., $F(x) \le F(y)$ whenever $x \le y$.

It is possible to prove that (1)–(5) are always valid for all what random variables X. There is also a converse. If F is a function that satisfies (1)–(5), then there exists a random variable X whose distribution function is F.

2.1 Discrete Random Variables

We will mostly study two classes of random variables: discrete, and continuous. We say that X is a *discrete* random variable if its possible values form a countable or finite set. In other words, X is discrete if and only if there exist $x_1, x_2, \ldots$ such that: $\mathrm{P}\{X = x_i \text{ for some } i \geq 1\} = 1$. In this case, we are interested in the *mass function* of X, defined as the function p such that

$$p(x_i) = \mathrm{P}\{X = x_i\} \quad (i \geq 1). \tag{2}$$

Implicitly, this means that $p(x) = 0$ if $x \neq x_i$ for some i. By countable additivity, $\sum_{i=1}^{\infty} p(x_i) = \sum_x p(x) = 1$. By countable additivity, the distribution function of F can be computed via the following: For all x,

$$F(x) = \sum_{y \leq x} p(y). \tag{3}$$

Occasionally, there are several random variables around and we identify the mass function of X by p_X to make the structure clear.

2.2 Continuous Random Variables

A random variable is said to be (absolutely) *continuous* if there exists a non-negative function f such that $\mathrm{P}\{X \in A\} = \int_A f(x)\, dx$ for all A. The function f is said to be the *density function* of X, and has the properties that:

1. $f(x) \geq 0$ for all x;
2. $\int_{-\infty}^{\infty} f(x)\, dx = 1$.

The distribution function of F can be computed via the following: For all x,

$$F(x) = \int_{\infty}^{x} f(y)\, dy. \tag{4}$$

By the fundamental theorem of calculus,

$$\frac{dF}{dx} = f. \tag{5}$$

Occasionally, there are several random variables around and we identify the density function of X by f_X to make the structure clear.

Continuous random variables have the peculiar property that $\mathrm{P}\{X = x\} = 0$ for all x. Equivalently, $F(x) = F(x-)$, so that F is continuous (not just right-continuous with left-limits).

3 Expectations

The (mathematical) *expectation* of a discrete random variable X is defined as

$$\mathrm{E}X = \sum_x xp(x), \tag{6}$$

where p is the mass function. Of course, this is well defined only if $\sum_x |x|p(x) < \infty$. In this case, we say that X is *integrable*. Occasionally, $\mathrm{E}X$ is also called the *moment*, *first moment*, or the *mean* of X.

Proposition 1 *For all functions g,*

$$\mathrm{E}g(X) = \sum_x g(x)p(x), \tag{7}$$

provided that $g(X)$ is integrable, and/or $\sum_x |g(x)|p(x) < \infty$.

This is not a trivial result if you read things carefully, which you should. Indeed, the definition of expectation implies that

$$\mathrm{E}g(X) = \sum_y y\mathrm{P}\{g(X) = y\} = \sum_y y p_{g(X)}(y). \tag{8}$$

The (mathematical) *expectation* of a continuous random variable X is defined as

$$\mathrm{E}X = \int_{-\infty}^{\infty} xf(x)\,dx, \tag{9}$$

where f is the density function. This is well defined when $\int_{-\infty}^{\infty} |x|f(x)\,dx$ is finite. In this case, we say that X is *integrable*. Some times, we write $\mathrm{E}[X]$ and/or $\mathrm{E}\{X\}$ and/or $\mathrm{E}(X)$ in place of $\mathrm{E}X$.

Proposition 2 *For all functions g,*

$$\mathrm{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,dx, \tag{10}$$

provided that $g(X)$ is integrable, and/or $\int_{-\infty}^{\infty} |g(x)|f(x)\,dx < \infty$.

As was the case in the discrete setting, this is not a trivial result if you read things carefully. Indeed, the definition of expectation implies that

$$\mathrm{E}g(X) = \int_{-\infty}^{\infty} yf_{g(X)}(y)\,dy. \tag{11}$$

Here is a result that is sometimes useful, and not so well-known to students of probability:

Proposition 3 *Let X be a non-negative integrable random variable with distribution function F. Then,*

$$\mathrm{E}X = \int_0^\infty (1 - F(x))\, dx. \tag{12}$$

Proof. Let us prove it for continuous random variables. The discrete case is proved similarly. We have

$$\int_0^\infty (1 - F(x))\, dx = \int_0^\infty \mathrm{P}\{X > x\}\, dx = \int_0^\infty \left(\int_x^\infty f(y)\, dy \right) dx. \tag{13}$$

Change the order of integration to find that

$$\int_0^\infty (1 - F(x))\, dx = \int_0^\infty \left(\int_0^y dx \right) f(y)\, dy = \int_0^\infty y f(y)\, dy. \tag{14}$$

Because $f(y) = 0$ for all $y < 0$, this proves the result. □

It is possible to prove that for all integrable random variables X and Y, and for all reals a and b,

$$\mathrm{E}[aX + bY] = a\mathrm{E}X + b\mathrm{E}Y. \tag{15}$$

This justifies the buzz-phrase, "expectation is a linear operation."

3.1 Moments

Note that any random variable X is integrable if and only if $\mathrm{E}|X| < \infty$. For all $r > 0$, the rth *moment* of X is $\mathrm{E}\{X^r\}$, provided that the rth *absolute moment* $\mathrm{E}\{|X|^r\}$ is finite.

In the discrete case,

$$\mathrm{E}[X^r] = \sum_x x^r p(x), \tag{16}$$

and in the continuous case,

$$\mathrm{E}[X^r] = \int_{-\infty}^\infty x^r f(X)\, dx. \tag{17}$$

When it makes sense, we can consider negative moments as well. For instance, if $X \geq 0$, then $\mathrm{E}[X^r]$ makes sense for $r < 0$ as well, but it may be infinite.

Proposition 4 *If $r > 0$ and X is a non-negative random variable with $\mathrm{E}[X^r] < \infty$, then*

$$\mathrm{E}[X^r] = r \int_0^\infty x^{r-1}(1 - F(x))\, dx. \tag{18}$$

Proof. When $r = 1$ this is Proposition 3. The proof works similarly. For instance, when X is continuous,

$$\begin{aligned} \mathrm{E}[X^r] &= \int_0^\infty x^r f(x)\,dx = \int_0^\infty \left(r \int_0^x y^{r-1}\,dy \right) f(x)\,dx \\ &= r \int_0^\infty y^{r-1} \left(\int_y^\infty f(x)\,dx \right) dy = r \int_0^\infty y^{r-1} \mathrm{P}\{X > y\}\,dy. \end{aligned} \tag{19}$$

This verifies the proposition in the continuous case. □

A quantity of interest to us is the *variance* of X. If is defined as

$$\mathrm{Var}X = \mathrm{E}\left[(X - \mathrm{E}X)^2 \right], \tag{20}$$

and is equal to

$$\mathrm{Var}X = \mathrm{E}[X^2] - (\mathrm{E}X)^2. \tag{21}$$

Variance is finite if and only if X has two finite moments.

3.2 A (Very) Partial List of Discrete Distributions

You are expected to be familar with the following discrete distributions:

1. Binomial $(n\,, p)$. Here, $0 < p < 1$ and $n = 1, 2, \ldots$ are fixed, and the mass function is

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \qquad \text{if } x = 0, \ldots, n. \tag{22}$$

   - $\mathrm{E}X = np$ and $\mathrm{Var}X = np(1-p)$.
   - The binomial $(1\,, p)$ distribution is also known as Bernoulli (p).

2. Poisson (λ). Here, $\lambda > 0$ is fixed, and the mass function is:

$$p(x) = \frac{e^{-\lambda}\lambda^x}{x!} \qquad x = 0, 1, 2, \ldots. \tag{23}$$

   - $\mathrm{E}X = \lambda$ and $\mathrm{Var}X = \lambda$.

3. Negative binomial $(n\,, p)$. Here, $0 < p < 1$ and $n = 1, 2, \ldots$ are fixed, and the mass function is:

$$p(x) = \binom{x-1}{n-1} p^n (1-p)^{x-n} \qquad x = n, n+1, \ldots. \tag{24}$$

   - $\mathrm{E}X = n/p$ and $\mathrm{Var}X = n(1-p)/p^2$.

3.3 A (Very) Partial List of Continuous Distributions

You are expected to be familar with the following continuous distributions:

1. Uniform $(a\,,b)$. Here, $-\infty < a < b < \infty$ are fixed, and the density function is

$$f(x) = \frac{1}{b-a} \qquad \text{if } a \le x \le b. \tag{25}$$

   - $\mathrm{E}X = (a+b)/2$ and $\mathrm{Var}X = (b-a)^2/12$.

2. Gamma $(\alpha\,,\beta)$. Here, $\alpha, \beta > 0$ are fixed, and the density function is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \qquad -\infty < x < \infty. \tag{26}$$

   Here, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}\,dt$ is the (Euler) gamma function. It is defined for all $\alpha > 0$, and has the property that $\Gamma(1+\alpha) = \alpha\Gamma(\alpha)$. Also, $\Gamma(1+n) = n!$ for all integers $n \ge 0$, whereas $\Gamma(1/2) = \sqrt{\pi}$.

   - $\mathrm{E}X = \alpha/\beta$ and $\mathrm{Var}X = \alpha/\beta^2$.
   - Gamma $(1\,,\beta)$ is also known as Exp (β). [The *Exponential distribution.*]
   - When $n \ge 1$ is an integer, Gamma $(n/2\,,1/2)$ is also known as $\chi^2(n)$. [The *chi-squared* distribution with n *degrees of freedom.*]

3. $N(\mu\,,\sigma^2)$. [*The normal distribution*] Here, $-\infty < \mu < \infty$ and $\sigma > 0$ are fixed, and the density function is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \qquad -\infty < x < \infty. \tag{27}$$

   - $\mathrm{E}X = \mu$ and $\mathrm{Var}X = \sigma^2$.
   - $N(0\,,1)$ is called the *standard normal* distribution.
   - We have the distributional identity, $\mu + \sigma N(0\,,1) = N(\mu\,,\sigma^2)$. Equivalently,

$$\frac{N(\mu\,,\sigma^2) - \mu}{\sigma} = N(0\,,1). \tag{28}$$

   - The distribution function of a $N(0\,,1)$ is an important object, and is *always* denoted by Φ. That is, for all $-\infty < a < \infty$,

$$\Phi(a) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\,dx. \tag{29}$$

4 Random Vectors

Let $X_1, \ldots, X_n$ be random variables. Then, $\boldsymbol{X} := (X_1, \ldots, X_n)$ is a *random vector.*

4.1 Distribution Functions

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be an N-dimensional random vector. Its *distribution function* is defined by

$$F(x_1, \ldots, x_n) = \mathrm{P}\left\{X_1 \leq x_1\,,\ \ldots\ ,\ X_n \leq x_n\right\}, \tag{30}$$

valid for all real numbers $x_1, \ldots, x_n$.

If $X_1, \ldots, X_n$ are all discrete, then we say that $\boldsymbol{X}$ is discrete. On the other hand, we say that $\boldsymbol{X}$ is (absolutely) *continuous* when there exists a non-negative function f, of n variables, such that for all n-dimensional sets A,

$$\mathrm{P}\{\boldsymbol{X} \in A\} = \int \cdots \int_A f(x_1, \ldots, x_n)\, dx_1\, \ldots\, dx_n. \tag{31}$$

The function f is called the *density function* of $\boldsymbol{X}$. It is also called the *joint density function* of $X_1, \ldots, X_n$.

Note, in particular, that

$$F(x_1, \ldots, x_n) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_n} f(u_1, \ldots, u_n)\, du_n \cdots du_1. \tag{32}$$

By the fundamental theorem of calculus,

$$\frac{\partial^n F}{\partial x_1 \partial x_2 \ldots \partial x_n} = f. \tag{33}$$

4.2 Expectations

If g is a real-valued function of n variables, then

$$\mathrm{E}g(X_1, \ldots, X_n) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \ldots, x_n) f(x_1, \ldots, x_n)\, dx_1\, \ldots\, dx_n. \tag{34}$$

An important special case is when $n = 2$ and $g(x_1, x_2) = x_1 x_2$. In this case, we obtain

$$\mathrm{E}[X_1 X_2] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} u_1 u_2 f(u_1\,, u_2)\, du_1\, du_2. \tag{35}$$

The *covariance* between X_1 and X_2 is defined as

$$\mathrm{Cov}(X_1\,, X_2) := \mathrm{E}\left[(X_1 - \mathrm{E}X_1)\,(X_2 - \mathrm{E}X_2)\right]. \tag{36}$$

It turns out that

$$\mathrm{Cov}(X_1\,, X_2) = \mathrm{E}[X_1 X_2] - \mathrm{E}[X_1]\mathrm{E}[X_2]. \tag{37}$$

This is well defined if both X_1 and X_2 have two finite moments. In this case, the *correlation* between X_1 and X_2 is

$$\rho(X_1\,, X_2) := \frac{\mathrm{Cov}(X_1\,, X_2)}{\sqrt{\mathrm{Var}X_1 \cdot \mathrm{Var}X_2}}, \tag{38}$$

provided that $0 < \mathrm{Var}X_1, \mathrm{Var}X_2 < \infty$.

The *expectation* of $\boldsymbol{X} = (X_1, \ldots, X_n)$ is defined as the vector $\mathrm{E}\boldsymbol{X}$ whose jth coordinate is $\mathrm{E}X_j$.

Given a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$, its *covariance matrix* is defined as $\boldsymbol{C} = (C_{ij})_{1 \le i,j \le n}$, where $C_{ij} := \mathrm{Cov}(X_i\, X_j)$. This makes sense provided that the X_i's have two finite moments.

Lemma 5 *Every covariance matrix $\boldsymbol{C}$ is positive semi-definite. That is, $\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x} \ge 0$ for all $\boldsymbol{x} \in \mathbf{R}^n$. Conversely, every positive semi-definite $(n \times n)$ matrix is the covariance matrix of some random vector.*

4.3 Multivariate Normals

Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ be an n-dimensional vector, and $\boldsymbol{C}$ an $(n \times n)$-dimensional matrix that is *positive definite*. The latter means that $\boldsymbol{x}'\boldsymbol{C}\boldsymbol{x} > 0$ for all non-zero vectors $\boldsymbol{x} = (x_1, \ldots, x_n)$. This implies, for instance, that $\boldsymbol{C}$ is invertible, and the inverse is also positive definite.

We say that $\boldsymbol{X} = (X_1, \ldots, X_n)$ has the *multivariate normal distribution* $N_n(\boldsymbol{\mu}\,, \boldsymbol{C})$ if the density function of $\boldsymbol{X}$ is

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{2\pi \det \boldsymbol{C}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}, \tag{39}$$

for all $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbf{R}^n$.

- $\mathrm{E}\boldsymbol{X} = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{C}$.
- $\boldsymbol{X} \sim N_n(\boldsymbol{\mu}\,, \boldsymbol{C})$ if and only if there exists a positive definite matrix $\boldsymbol{A}$, and n i.i.d. standard normals $Z_1, \ldots, Z_n$ such that $\boldsymbol{X} = \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{Z}$. In addition, $\boldsymbol{A}'\boldsymbol{A} = \boldsymbol{C}$.

When $n = 2$, a multivariate normal is called a *bivariate normal*.

Warning. Suppose X and Y are each normally distributed. Then it is *not* true in general that $(X\,, Y)$ is bivariate normal. A similar caveat holds for the n-dimensional case.

5 Independence

Random variables $X_1, \ldots, X_n$ are (statistically) *independent* if

$$\mathrm{P}\left\{X_1 \in A_1\,, \ldots, X_n \in A_n\right\} = \mathrm{P}\left\{X_1 \in A_1\right\} \times \cdots \times \mathrm{P}\left\{X_n \in A_n\right\}, \tag{40}$$

for all one-dimensional sets $A_1, \ldots, A_n$. It can be shown that $X_1, \ldots, X_n$ are independent if and only if for all real numbers $x_1, \ldots, x_n$,

$$\mathrm{P}\left\{X_1 \le x_1\,, \ldots, X_n \le x_n\right\} = \mathrm{P}\left\{X_1 \le x_1\right\} \times \cdots \times \mathrm{P}\left\{X_n \le x_n\right\}. \tag{41}$$

That is, the coordinates of $\boldsymbol{X} = (X_1, \ldots, X_n)$ are independent if and only if $F_{\boldsymbol{X}}(x_1, \ldots, x_n) = F_{X_1}(x_1) \cdots F_{X_n}(x_n)$. Another equivalent formulation of independence is this: For all functions $g_1, \ldots, g_n$ such that $g_i(X_i)$ is integrable,

$$\mathrm{E}\left[g(X_1) \times \ldots \times g(X_n)\right] = \mathrm{E}[g_1(X_1)] \times \cdots \times \mathrm{E}[g_n(X_n)]. \tag{42}$$

A ready consequence is this: If X_1 and X_2 are independent, then they are *uncorrelated* provided that their correlation exists. Uncorrelated means that $\rho(X_1\,, X_2) = 0$. This is equivalent to $\mathrm{Cov}(X_1\,, X_2) = 0$.

If $X_1, \ldots, X_n$ are (pairwise) uncorrelated with two finite moments, then

$$\mathrm{Var}(X_1 + \cdots + X_n) = \mathrm{Var}X_1 + \cdots + \mathrm{Var}X_n. \tag{43}$$

Significantly, this is true when the X_i's are independent. In general, the formula is messier:

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathrm{Var}X_i + 2 \mathop{\sum\sum}_{1 \le i < j \le n} \mathrm{Cov}(X_i\,, X_j). \tag{44}$$

In general, uncorrelated random variables are not *independent*. An exception is made for multivariate normals.

Theorem 6 *Suppose* $(\boldsymbol{X}\,, \boldsymbol{Y}) \sim N_{n+k}(\boldsymbol{\mu}\,, \boldsymbol{C})$*, where* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are respectively* n*-dimensional and* k*-dimensional random vectors. Then:*

1. $\boldsymbol{X}$ *is multivariate normal.*
2. $\boldsymbol{Y}$ *is multivariate normal.*
3. *If* $\mathrm{E}X_iY_j = 0$ *for all* i, j*, then* $\boldsymbol{X}$ *and* $\boldsymbol{Y}$ *are independent.*

For example, suppose $(X\,, Y)$ is bivariate normal. Then, X and Y are normally distributed. If, in addition, $\mathrm{Cov}(X\,, Y) = 0$ then X and Y are independent.

6 Convergence Criteria

Let $X_1, X_2, \ldots$ be a countably-infinite sequence of random variables. There are several ways to make sense of the statement that $X_n \to X$ for a random variable X. We need a few of these criteria.

6.1 Convergence in Distribution

We say that X_n converges to X *in distribution* if

$$F_{X_n}(x) \to F_X(x), \tag{45}$$

for all $x \in \mathbf{R}$ at which F_X is continuous. We write this as $X_n \stackrel{d}{\to} X$.

Very often, F_X is continuous. In such cases, $X_n \stackrel{d}{\to} X$ if and only if $F_{X_n}(x) \to F_X(x)$ for all x. Note that if $X_n \stackrel{d}{\to} X$ and X has a continuous distribution then also

$$\mathrm{P}\{a \le X_n \le b\} \to \mathrm{P}\{a \le X \le b\}, \tag{46}$$

for all $a < b$.

Similarly, we say that the random vectors $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ converge in distribution to the random vector $\boldsymbol{X}$ when $F_{\boldsymbol{X}_n}(\boldsymbol{a}) \to F_{\boldsymbol{X}}(\boldsymbol{a})$ for all $\boldsymbol{a}$ at which $F_{\boldsymbol{X}}$ is continuous. This convergence is also denoted by $\boldsymbol{X}_n \stackrel{d}{\to} \boldsymbol{X}$.

6.2 Convergence in Probability

We say that X_n converges to X *in probability* if for all $\epsilon > 0$,

$$\mathrm{P}\left\{|X_n - X| > \epsilon\right\} \to 0. \tag{47}$$

We denote this by $X_n \stackrel{\mathrm{P}}{\to} X$.

It is the case that if $X_n \stackrel{\mathrm{P}}{\to} X$ then $X_n \stackrel{d}{\to} X$, but the converse is patently false. There is one exception to this rule.

Lemma 7 *Suppose $X_n \stackrel{d}{\to} c$ where c is a non-random constant. Then, $X_n \stackrel{\mathrm{P}}{\to} c$.*

Proof. Fix $\epsilon > 0$. Then,

$$\mathrm{P}\{|X_n - c| \le \epsilon\} \ge \mathrm{P}\{c - \epsilon < X_n \le c + \epsilon\} = F_{X_n}(c+\epsilon) - F_{X_n}(c-\epsilon). \tag{48}$$

But $F_c(x) = 0$ if $x < c$, and $F_c(x) = 1$ if $x \ge c$. Therefore, F_c is continuous at $c \pm \epsilon$, whence we have $F_{X_n}(c+\epsilon) - F_{X_n}(c-\epsilon) \to F_c(c+\epsilon) - F_c(c-\epsilon) = 1$. This proves that $\mathrm{P}\{|X_n - c| \le \epsilon\} \to 1$, which is another way to write the lemma. $\square$

Similar considerations lead us to the following.

Theorem 8 (Slutsky's theorem) *Suppose $X_n \stackrel{d}{\to} X$ and $Y_n \stackrel{d}{\to} c$ for a constant c. If g is a continuous function of two variables, then $g(X_n, Y_n) \stackrel{d}{\to} g(X, c)$. [For instance, try $g(x, y) = ax + by$, $g(x, y) = xye^x$, etc.]*

When c is a random variable this is no longer valid in general.

7 Moment Generating Functions

We say that X has a *moment generating function* if there exists $t_0 > 0$ such that

$$M(t) := M_X(t) = \mathrm{E}[e^{tX}] \text{ is finite for all } t \in [-t_0, t_0]. \tag{49}$$

If this condition is met, then M is the moment generating function of X.

If and when it exists, the moment generating function of X determines its entire distribution. Here is a more precise statement.

Theorem 9 (Uniqueness) *Suppose X and Y have moment generating functions, and $M_X(t) = M_Y(t)$ for all t sufficiently close to 0. Then, X and Y have the same distribution.*

7.1 Some Examples

1. Binomial $(n\,,p)$. Then, $M(t)$ exists for all $-\infty < t < \infty$, and
$$M(t) = \left(1 - p + pe^t\right)^n. \tag{50}$$

2. Poisson (λ). Then, $M(t)$ exists for all $-\infty < t < \infty$, and
$$M(t) = e^{\lambda(e^t - 1)}. \tag{51}$$

3. Negative Binomial $(n\,,p)$. Then, $M(t)$ exists if and only if $-\infty < t < |\log(1-p)|$. In that case, we have also that
$$M(t) = \left(\frac{pe^t}{1-(1-p)e^t}\right)^n. \tag{52}$$

4. Uniform $(a\,,b)$. Then, $M(t)$ exists for all $-\infty < t < \infty$, and
$$M(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}. \tag{53}$$

5. Gamma $(\alpha\,,\beta)$. Then, $M(t)$ exists if and only if $-\infty < t < \beta$. In that case, we have also that
$$M(t) = \left(\frac{\beta}{\beta - t}\right)^\alpha. \tag{54}$$
Set $\alpha = 1$ to find the moment generating function of an exponential (β). Set $\alpha = n/2$ and $\beta = 1/2$—for a positive integer n—to obtain the moment generating function of a chi-squared (n).

6. $N(\mu\,,\sigma^2)$. The moment generating function exists for all $-\infty < t < \infty$. Moreover,
$$M(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right). \tag{55}$$

7.2 Properties

Beside the uniqueness theorem, moment generating functions have two more properties that are of interest in mathematical statistics.

Theorem 10 (Convergence Theorem) *Suppose $X_1, X_2, \ldots$ is a sequence of random variables whose moment generating functions all exists in an interval $[-t_0\,,t_0]$ around the origin. Suppose also that for all $t \in [-t_0\,,t_0]$, $M_{X_n}(t) \to M_X(t)$ as $n \to \infty$, where M is the moment generating function of a random variable X. Then, $X_n \xrightarrow{d} X$.*

Example 11 (Law of Rare Events) Let X_n have the $\mathrm{Bin}(n\,,\lambda/n)$ distribution, where $\lambda > 0$ is independent of n. Then, for all $-\infty < t < \infty$,

$$M_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n. \tag{56}$$

We claim that for all real numbers c,

$$\left(1 + \frac{c}{n}\right)^n \to e^c \text{ as } n \to \infty. \tag{57}$$

Let us take this for granted for the time being. Then, it follows at once that

$$M_{X_n}(t) \to \exp\left(-\lambda + \lambda e^t\right) = e^{\lambda(e^t - 1)}. \tag{58}$$

That is,

$$\mathrm{Bin}\ (n\,,\lambda/n) \xrightarrow{d} \mathrm{Poisson}\ (\lambda). \tag{59}$$

This is Poisson's "law of rare events" (also known as "the law of small numbers").

Now we wrap up this example by verifying (57). Let $f(x) = (1+x)^n$, and Taylor-expand it to find that $f(x) = 1 + nx + \frac{1}{2}n(n-1)x^2 + \cdots$. Replace x by c/n, and compute to find that

$$\left(1 + \frac{c}{n}\right)^n = 1 + c + \frac{(n-1)c^2}{2n} + \cdots \to \sum_{j=0}^{\infty} \frac{c^j}{j!}, \tag{60}$$

and this is the Taylor-series expansion of e^c. [There is a little bit more one has to do to justify the limiting procedure.]

The second property of moment generating functions is that if and when it exists for a random variable X, then all moments of X exist, and can be computed from M_X.

Theorem 12 (Moment-Generating Property) *Suppose X has a finite moment generating function in a neighborhood of the origin. Then,* $\mathrm{E}(|X|^n)$ *exists for all n, and $M^{(n)}(0) = \mathrm{E}[X^n]$, where $f^{(n)}(x)$ denotes the nth derivative of function f at x.*

Example 13 Let X be a $N(\mu\,,1)$ random variable. Then we know that $M(t) = \exp(\mu t + \frac{1}{2}t^2)$. Consequently,

$$M'(t) = (\mu + t)e^{\mu t + (t^2/2)}, \quad \text{and} \quad M''(t) = \left[1 + (\mu + t)^2\right]e^{\mu t + (t^2/2)} \tag{61}$$

Set $t = 0$ to find that $\mathrm{E}X = M'(0) = \mu$ and $\mathrm{E}[X^2] = M''(0) = 1 + \mu^2$, so that $\mathrm{Var}X = \mathrm{E}[X^2] - (\mathrm{E}X)^2 = 1$.

8 Characteristic Functions

The *characteristic function* of a random variable X is the function

$$\phi(t) := \mathrm{E}\left[e^{itX}\right] \qquad -\infty < t < \infty. \tag{62}$$

Here, the "i" refers to the complex unit, $i = \sqrt{-1}$. We may write ϕ as ϕ_X, for example, when there are several random variables around.

In practice, you often treat e^{itX} as if it were a real exponential. However, the correct way to think of this definition is via the Euler formula, $e^{i\theta} = \cos\theta + i\sin\theta$ for all real numbers θ. Thus,

$$\phi(t) = \mathrm{E}[\cos(tX)] + i\mathrm{E}[\sin(tX)]. \tag{63}$$

If X has a moment generating function M, then it can be shown that $M(it) = \phi(t)$. [This uses the technique of "analytic continuation" from complex analysis.] In other words, the naive replacement of t by it does what one may guess it would. However, one advantage of working with ϕ is that *it is always well-defined*. The reason is that $|\cos(tX)| \le 1$ and $|\sin(tX)| \le 1$, so that the expectations in (63) exist. In addition to having this advantage, ϕ shares most of the properties of M as well! For example,

Theorem 14 *The following hold:*

1. ***(Uniqueness Theorem)*** *Suppose there exists $t_0 > 0$ such that for all $t \in (-t_0\,, t_0)$, $\phi_X(t) = \phi_Y(t)$. Then X and Y have the same distribution.*

2. ***(Convergence Theorem)*** *If $\phi_{X_n}(t) \to \phi_X(t)$ for all $t \in (-t_0\,, t_0)$, then $X_n \stackrel{d}{\to} X$. Conversely, if $X_n \stackrel{d}{\to} X$, then $\phi_{X_n}(t) \to \phi_X(t)$ for all t.*

8.1 Some Examples

1. Binomial $(n\,, p)$. Then,

$$\phi(t) = M(it) = \left(1 - p + pe^{it}\right)^n. \tag{64}$$

2. Poisson (λ). Then,

$$\phi(t) = M(it) = e^{\lambda(e^{it}-1)}. \tag{65}$$

3. Negative Binomial $(n\,, p)$. Then,

$$\phi(t) = M(it) = \left(\frac{pe^{it}}{1-(1-p)e^{it}}\right)^n. \tag{66}$$

4. Uniform $(a\,, b)$. Then,

$$\phi(t) = M(it) = \frac{e^{itb} - e^{ita}}{t(b-a)}. \tag{67}$$

5. Gamma $(\alpha\,,\beta)$. Then,

$$\phi(t) = M(it) = \left(\frac{\beta}{\beta - it}\right)^{\alpha}. \tag{68}$$

6. $N(\mu\,,\sigma^2)$. Then, because $(it)^2 = -t^2$,

$$\phi(t) = M(it) = \exp\left(i\mu t - \frac{\sigma^2 t^2}{2}\right). \tag{69}$$

9 Classical Limit Theorems

9.1 The Central Limit Theorem

Theorem 15 (The CLT) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with two finite moments. Let $\mu := \mathrm{E}X_1$ and $\sigma^2 = \mathrm{Var}X_1$. Then,*

$$\frac{\sum_{j=1}^{n} X_j - n\mu}{\sigma\sqrt{n}} \xrightarrow{d} N(0\,,1). \tag{70}$$

Elementary probability texts prove this by appealing to the convergence theorem for moment generating functions. This approach does not work if we know only that X_1 has two finite moments, however. However, by using characteristic functions, we can relax the assumptions to the finite mean and variance case, as stated.

Proof of the CLT. Define

$$T_n := \frac{\sum_{j=1}^{n} X_j - n\mu}{\sigma\sqrt{n}}. \tag{71}$$

Then,

$$\begin{aligned} \phi_{T_n}(t) &= \mathrm{E}\left[\prod_{j=1}^{n} \exp\left(it\left(\frac{X_j - \mu}{\sigma\sqrt{n}}\right)\right)\right] \\ &= \prod_{j=1}^{n} \mathrm{E}\left[\exp\left(it\left(\frac{X_j - \mu}{\sigma\sqrt{n}}\right)\right)\right], \end{aligned} \tag{72}$$

thanks to independence; see (42) on page 7. Let $Y_j := (X_j - \mu)/\sigma$ denote the standardization of X_j. Then, it follows that

$$\phi_{T_n}(t) = \prod_{j=1}^{n} \phi_{Y_j}\left(t/\sqrt{n}\right) = \left[\phi_{Y_1}\left(t/\sqrt{n}\right)\right]^n, \tag{73}$$

because the Y_j's are i.i.d. Recall the Taylor expansion, $e^{ix} = 1 + ix - \frac{1}{2}x^2 + \cdots$, and write $\phi_{Y_1}(s)$ as $\mathrm{E}[e^{itY_1}] = 1 + it\mathrm{E}Y_1 - \frac{1}{2}t^2\mathrm{E}[Y_1^2] + \cdots = 1 - \frac{1}{2}t^2 + \cdots$. Thus,

$$\phi_{T_n}(t) = \left[1 - \frac{t^2}{2n} + \cdots\right]^n \to e^{-t^2/2}. \tag{74}$$

See (57) on page 9. Because $e^{-t^2/2}$ is the characteristic function of $N(0\,,1)$, this and the convergence theorem (Theorem 15 on page 11) together prove the CLT. □

The CLT has a multidimensional counterpart as well. Here is the statement.

Theorem 16 *Let $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots$ be i.i.d. k-dimensional random vectors with mean vector $\boldsymbol{\mu} := \mathrm{E}\boldsymbol{X}_1$ and covariance matrix $\boldsymbol{Q} := \mathrm{Cov}\boldsymbol{X}$. If $\boldsymbol{Q}$ is non-singular, then*

$$\frac{\sum_{j=1}^n \boldsymbol{X}_j - n\boldsymbol{\mu}}{\sqrt{n}} \xrightarrow{d} N_k(\boldsymbol{0}\,, \boldsymbol{Q}). \tag{75}$$

9.2 (Weak) Law of Large Numbers

Theorem 17 (Law of Large Numbers) *Suppose $X_1, X_2, \ldots$ are i.i.d. and have a finite first moment. Let $\mu := \mathrm{E}X_1$. Then,*

$$\frac{\sum_{j=1}^n X_j}{n} \xrightarrow{\mathrm{P}} \mu. \tag{76}$$

Proof. We will prove this in case there is also a finite variance. The general case is beyond the scope of these notes. Thanks to the CLT (Theorem 15, page 11), $(X_1 + \cdots + X_n)/n$ converges in distribution to μ. Slutsky's theorem (Theorem 8, page 8) proves that convergence holds also in probability. □

9.3 Variance Stabilization

Let $X_1, X_2, \ldots$ be i.i.d. with $\mu = \mathrm{E}X_1$ and $\sigma^2 = \mathrm{Var}X_1$ both defined and finite. Define the partial sums,

$$S_n := X_1 + \cdots + X_n. \tag{77}$$

We know that: (i) $S_n \approx n\mu$ in probability; and (ii) $(S_n - n\mu) \stackrel{d}{\approx} N(0\,, n\sigma^2)$. Now use Taylor expansions: For any smooth function h,

$$h(S_n/n) \approx h(\mu) + \left(\frac{S_n}{n} - \mu\right) h'(\mu), \tag{78}$$

in probability. By the CLT, $(S_n/n) - \mu \stackrel{d}{\approx} N(0\,, \sigma^2/n)$. Therefore, Slutsky's theorem (Theorem 8, page 8) proves that

$$\sqrt{n}\left[h\left(\frac{S_n}{n}\right) - h(\mu)\right] \xrightarrow{d} N\left(0\,, \sigma^2|h'(\mu)|^2\right). \tag{79}$$

[Technical conditions: h' should be continuously-differentiable in a neighborhood of μ.]

9.4 Refinements to the CLT

There are many refinements to the CLT. Here is a particularly well-known one. It gives a description of the farthest the distribution function of normalized sums is from the normal.

Theorem 18 (Berry–Esseen) *If* $\rho := \mathrm{E}\{|X_1|^3\} < \infty$*, then*

$$\max_{-\infty<a<\infty} \left| \mathrm{P}\left\{ \frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \le a \right\} - \Phi(a) \right| \le \frac{3\rho}{\sigma^3\sqrt{n}}. \tag{80}$$

10 Conditional Expectations

Let us begin by recalling some basic notions of conditioning from elementary probability. Throughout this section, X denotes a random variable and $\boldsymbol{Y} := (Y_1, \ldots, Y_n)$ an n-dimensional random vector.

10.1 Conditional Probabilities and Densities

If $X, Y_1, \ldots, Y_n$ are all discrete random variables, then the conditional mass function of X, given that $\boldsymbol{Y} = \boldsymbol{y}$, is

$$p_{X|\boldsymbol{Y}}(x \,|\, \boldsymbol{y}) := \frac{\mathrm{P}\{X = x, Y_1 = y_1, \ldots, Y_n = y_n\}}{\mathrm{P}\{Y_1 = y_1, \ldots, Y_n = y_n\}}, \tag{81}$$

provided that $\mathrm{P}\{\boldsymbol{Y} = \boldsymbol{y}\} > 0$. This is a bona fide mass function [as a function of the variable x] for every fixed choice of $\boldsymbol{y}$. [It doesn't make sense to worry about its behavior in the variables $y_1, \ldots, y_n$.]

Similarly, if the distribution of $(X, Y_1, \ldots, Y_n)$ is absolutely continuous, then the conditional density function of X, given that $\boldsymbol{Y} = \boldsymbol{y}$, is

$$f_{X|\boldsymbol{Y}}(x \,|\, \boldsymbol{y}) := \frac{f_{X,\boldsymbol{Y}}(x, y_1, \ldots, y_n)}{f_{\boldsymbol{Y}}(y_1, \ldots, y_n)}, \tag{82}$$

provided that the observed value $\boldsymbol{y}$ is such that the joint density $f_{X,\boldsymbol{Y}}$ of the random vector $(X, \boldsymbol{Y})$ satisfies

$$f_{\boldsymbol{Y}}(y_1, \ldots, y_n) > 0. \tag{83}$$

Note that (83) is entirely possible, though $\mathrm{P}\{\boldsymbol{Y} = \boldsymbol{y}\} = 0$ simply because $\boldsymbol{Y}$ has an absolutely continuous distribution. Condition (83) is quite natural in the following sense: Let $\mathbb{B}$ denote the collection of all n-dimensional vectors $\boldsymbol{y}$ such that $f_{\boldsymbol{Y}}(y_1, \ldots, y_n) = 0$. Then,

$$\mathrm{P}\{\boldsymbol{Y} \in \mathbb{B}\} = \int_{\mathbb{B}} f_{\boldsymbol{Y}}(y_1, \ldots, y_n)\, dy_1 \cdots dy_n = 0. \tag{84}$$

In other words, we do not have to worry about defining $f_{X|\boldsymbol{Y}}(x \,|\boldsymbol{y})$ when $\boldsymbol{y}$ is not in $\mathbb{B}$.

10.2 Conditional Expectations

If we have observed that $\boldsymbol{Y} = \boldsymbol{y}$, for a known vector $\boldsymbol{y} = (y_1, \ldots, y_n)$, then the best linear predictor of X is the [classical] conditional expectation

$$\mathrm{E}(X \mid \boldsymbol{Y} = \boldsymbol{y}) := \begin{cases} \sum_x x\mathrm{P}\{X = x \mid \boldsymbol{Y} = \boldsymbol{y}\} & \text{if } (X, \boldsymbol{Y}) \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_{X|\boldsymbol{Y}}(x \mid \boldsymbol{y})\, dx & \text{if } (X, \boldsymbol{Y}) \text{ has a joint pdf.} \end{cases} \tag{85}$$

The preceding assumes tacitly that the sum/integral converges absolutely. More generally, we have for any nice function φ,

$$\mathrm{E}(\varphi(X) \mid \boldsymbol{Y} = \boldsymbol{y}) := \begin{cases} \sum_x \varphi(x)\mathrm{P}\{X = x \mid \boldsymbol{Y} = \boldsymbol{y}\} & \text{if discrete,} \\ \int_{-\infty}^{\infty} \varphi(x) f_{X|\boldsymbol{Y}}(x \mid \boldsymbol{y})\, dx & \text{if joint pdf exists,} \end{cases} \tag{86}$$

provided that the sum/integral converges absolutely. The preceding is in fact a theorem, but a careful statement requires writing too many technical details from integration theory.

10.3 An Intuitive Interpretation

The basic use of conditional expectations is this: If we observe that $\boldsymbol{Y} = \boldsymbol{y}$, then we predict X, based only on our observation that $\boldsymbol{Y} = \boldsymbol{y}$, as $\mathrm{E}(X \mid \boldsymbol{Y} = \boldsymbol{y})$.

Example 19 We perform 10 independent Bernoulli trials [$p :=$ probability of success per trial]. Let X denote the total number of successes. We know that X has a $\mathrm{Bin}(10, p)$ distribution. If $Y :=$ the total number of successes in the first 5 trials, then you should check that $\mathrm{E}(X \mid Y = 0) = 5p$. More generally, $\mathrm{E}(X \mid Y = y) = y + 5p$ for all $y \in \{0, \ldots, 5\}$.

The previous example shows you that it is frequently more convenient to use a slightly different form of conditional expectations: We write $\mathrm{E}(X \mid \boldsymbol{Y})$ for the random variable whose value is $\mathrm{E}(X \mid \boldsymbol{Y} = \boldsymbol{y})$ when we observe that $\boldsymbol{Y} = \boldsymbol{y}$. In the previous example, this definition translates to the following computation: $\mathrm{E}(X \mid Y) = Y + 5p$. This ought to make very good sense to you, before you read on!

The classical Bayes' formula for conditional probabilities has an analogue for conditional expectations. Suppose $(X, \boldsymbol{Y})$ has a joint density function $f_{X,\boldsymbol{Y}}$.

Then,

$$
\begin{aligned}
\mathrm{E}(X) &= \int_{-\infty}^{\infty} x f_X(x)\,dx \\
&= \int_{-\infty}^{\infty} x \left(\int_{\mathbf{R}^n} f_{X,\boldsymbol{Y}}(x\,, y_1\,, \ldots, y_n)\, d\boldsymbol{y} \right) dx \\
&= \int_{-\infty}^{\infty} x \left(\int_{\mathbf{R}^n} f_{X|\boldsymbol{Y}}(x \mid \boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y} \right) dx \\
&= \int_{\mathbf{R}^n} \left(\int_{-\infty}^{\infty} x f_{X|\boldsymbol{Y}}(x \mid \boldsymbol{y})\, dx \right) f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y} \\
&= \int_{\mathbf{R}^n} \mathrm{E}(X \mid \boldsymbol{Y} = \boldsymbol{y}) f_{\boldsymbol{Y}}(\boldsymbol{y})\, d\boldsymbol{y} \\
&= \mathrm{E}\left\{\mathrm{E}(X \mid \boldsymbol{Y})\right\}.
\end{aligned} \tag{87}
$$

This is always true. That is, we always have

$$
\mathrm{E}(X) = \mathrm{E}\left\{\mathrm{E}(X \mid \boldsymbol{Y})\right\}, \tag{88}
$$

provided that $\mathrm{E}|X| < \infty$.