LECTURE 1: REVIEW OF BASIC PROBABILITY

Michal Kolesár[*]

August 22, 2024

1. PROBABILITY SPACES

REFERENCE Casella and Berger (2002, CB from hereon), Chapter 1.1–1.3, or Hogg, McKean, and Craig (2019, HMC from hereon) Chapter 1.1–1.4. Hansen (2022, H22 from hereon), Chapter 1.

To define formally what we mean by a random variable, we need to begin by introducing the concept of a probability space.

A probability space is a triple $(\Omega, \mathcal{A}, P)$, where:

1. Ω is called a *sample space*. It is any non-empty set ($\mathbb{R}^k, \mathbb{N}, [0,1], \{1, \dots, K\}$ etc), finite, countable, or uncountable. Elements $\omega \in \Omega$ are referred to as outcomes of a random experiment. In each random experiment, there is one realized outcome ω. Subsets A of Ω are called *events*: they either happen or they do not. We say that event A happens if the realized outcome ω belongs to A, i.e. $\omega \in A$. Event A does not occur if $\omega \notin A$.

2. $\mathcal{A}$ is some family of subsets of Ω, called *σ-algebra* (or *σ-field*). That is, $\mathcal{A}$ is a collection of events. It needs to satisfy three properties: it needs to contain Ω, and must be closed under complements and countable unions: $\Omega \in \mathcal{A}$, $A \in \mathcal{A} \implies A^C \in \mathcal{A}$, and $A_1, A_2, \dots \in \mathcal{A} \implies (\cup_{i=1}^{\infty} A_i) \in \mathcal{A}$.

3. P is a function from $\mathcal{A}$ into $[0,1]$, i.e. $P\colon \mathcal{A} \to [0,1]$, called a *probability measure*. For each event $A \in \mathcal{A}$, $P(A)$ gives the probability of the event happening.

   The function P needs to satisfy three probability axioms (also called Kolmogorov axioms): it needs to be non-negative ($P(A) \geq 0$ for all $A \in \mathcal{A}$), assign probability one to Ω ($P(\Omega) = 1$), and be σ-additive (countably additive), meaning if $A_1, A_2, \dots \in \mathcal{A}$ is a countable collection of mutually exclusive events, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$.

[*] Email: mkolesar@princeton.edu.

These definitions raise two immediate questions. First, why restrict P to only be countably additive? The answer is immediate: we know that for a uniform random variable, $P(X = x) = 0$, and thus $\sum_{x \in [0,1]} P(X = x) \neq P(X \in [0,1])$.

Second, why restrict P to be defined only on $\mathcal{A}$, and not all subsets of Ω? Here the answer is more technical: it turns out it is not possible in general. For instance, it is impossible to come up with a function P that assigns probabilities to *all* subsets of the real line, and also satisfies Kolmogorov axioms. This is both surprising and annoying, and it leads to a quite formidable mathematical apparatus called measure theory.

Digression. A classic counterexample is as follows. Let x be equivalent to y if $x - y$ is rational. This equivalence relation partitions the unit interval. Let H be a set that contains exactly one element from each equivalence class (we construct it using the axiom of choice). Then the interval $(0,1]$ equals the union $\cup_{\{r \in [0,1):\, r \text{ rational}\}} (H \oplus r)$, where $\oplus$ denotes a shift, $a \oplus r = a + r$ if $a + r < 1$ and $a + r - 1$ otherwise. Further, since H contains exactly one element from each class, the sets $H \oplus r$ are disjoint, and hence, for a uniform random variable, $P(X \in (0,1]) = \cup_{r \in [0,1), r \text{ rational}} P(X \in H \oplus r) = \cup_{r \in [0,1), r \text{ rational}} P(X \in H)$, where the last inequality follows by shift invariance. But while the left-hand side (LHS) equals 1, the right-hand side (RHS) must equal 0 (if $P(X \in H) = 0$) or ∞ (if $P(X \in H) > 0$). ⊠

The upshot is that while we could take $\mathcal{A}$ to consist of all subsets of Ω if Ω is countable, the σ-algebra when Ω is uncountable must necessarily exclude some subset of Ω. Events not in $\mathcal{A}$ are called *not measurable*: we cannot assign probabilities to them.

In what follows, though, we will largely ignore issues of measurability, since it has virtually no impact on results that are of interest to econometricians or economists, and mostly forget about the σ-algebra $\mathcal{A}$.[1] It is mentioned here only for completeness. Why are these issues of little practical impact? Solovay (1970) proved that it is not possible to construct non-measurable subsets of $\mathbb{R}^d$ in without invoking the axiom of choice (more precisely non-Lebesgue measurable sets). This means that all subsets of $\mathbb{R}^d$ that arise "in practice" are in the Lebesgue σ-algebra. If this is the case, why does the formidable apparatus exist? There are two main reasons

1. Some elegance. For instance, we can have one definition for expected value covering all random variables: discrete, continuous, mixtures of discrete and continuous variables, or those that are neither (more on those below); we can also have a unified way of talking about the "information" contained in collections of events or collections of random variables.

2. It ensures that once you do start working with complicated objects (think Brownian motion etc), everything is still coherent.

Example 1. To make the concept of probability triples concrete, think about defining the probability space and events for

1. Tossing three coins

[1] In fact, even if we did bother with it, the σ-algebras for uncountable subsets are never defined directly, instead, we typically take a simple collection, such as the collection of all intervals, for which we can assign probability measures, and then argue that the probability definition extends to the smallest σ-algebra containing all intervals.

2. Waiting time for a bus in seconds

3. Angle of the minute hand on a clock when this class is over ☒

Digression. The definition of a probability measure doesn't tell us what probability *means*. There are two common interpretations. The first is that probabilities are the relative long-run frequency of outcomes (say coin tosses). This is not helpful for questions such as "Will this lecture go over time?" The second view that addresses this issue is that probability is subjective or Bayesian; it measures degrees of belief. The probability axioms then ensure that these beliefs are consistent. For now, these differences in interpretations don't really matter; they will become more important once we move from probability onto statistics. ☒

2. RANDOM VARIABLES

REFERENCE CB, Chapters 1.4–1.6 and 2.1, or HMC, Chapter 1.5–1.7, H22 Chapter 2.1–2.3, 2.7–2.12.

2.1. *Basic Definitions*

A random variable X is a function from the sample space Ω to the real line, $X\colon \Omega \to \mathbb{R}$ (technically the function needs to be measurable, which is a complication that we ignore). In other words, for each outcome $\omega \in \Omega$, we have realization $X(\omega)$ of random variable X. The probability that X takes on a value in a set A is denoted $P_X(X \in A) = P(\{\omega \in \Omega\colon X(\omega) \in A\})$. When it doesn't cause confusion, we will abuse notation, and simply write $P(X \in A)$, dropping the $_X$ subscript.

Example. In the tossing three heads experiment, let X be the random variable that denotes the total number of heads. In the waiting time example, we can let $X(\omega) = \omega$. Or set X to be the waiting time in minutes. ☒

Each random variable has a *cumulative distribution function* (CDF), $F_X\colon \mathbb{R} \to [0, 1]$, defined as $F_X(x) = P_X(X \le x)$. Here $P_X(X \le x) = P(\{\omega \in \Omega\colon X(\omega) \le x\})$ denotes the probability that $X \le x$, using the notation above. To emphasize X has CDF F_X, we write $X \sim F_X$.

It's not hard to show that F_X has three properties:

1. it's non-decreasing;

2. $\lim_{x\to\infty} F_X(x) = 1$ and $\lim_{x\to-\infty} F_X(x) = 0$; and

3. it's right-continuous: $\lim_{t\downarrow 0} F_X(x + t) = F_X(x)$ for all x.

In fact, this also holds in reverse: for any function F that satisfies these three properties, one can construct a random variable whose distribution is F (Durrett 2019, Theorem 1.2.2). Also, the CDF uniquely determines completely characterizes X: everything there is to know about various probabilities P_X is contained in the CDF. (Why is this not obvious?)

Based on the shape of the CDF, we can distinguish between 3 types of random variables:

DISCRETE F_X is constant except for a countable number of points (i.e. F_X is a step function).

Thus, a discrete random variable X can be characterized by a list of possible values, $\mathcal{X} = \{x_1, x_2, \dots\}$, and their probabilities, $p = \{p_1, p_2, \dots\}$, where p_i denotes the probability that X will take value x_i, i.e. $p_i = P(X = x_i)$ for all $i = 1, 2, \dots$, such that $\sum_{i=1}^{\infty} p_i = 1$ and $p_i \geq 0$ for all i. Then the CDF of X is given by $F_X(t) = \sum_{j\colon x_j \leq t} p_j$.

CONTINUOUS X is continuous if its CDF can be written as $F_X(x) = \int_{-\infty}^{x} f_X(t)\,\mathrm{d}t$, for some function $f_X\colon \mathbb{R} \to \mathbb{R}$, called the probability density function (PDF) (or just *density*).

- If X is continuous, then F_X must be differentiable almost everywhere with $dF_X(t)/dt = f_X(t)$. If F_X is continuously differentiable, then X is continuous.
- At continuity points of f_X, it must be that $f_X(t) = dF_X(t)/dt$ by the fundamental theorem of calculus.
- Note that by Kolmogorov axioms, $\int_{-\infty}^{\infty} f_X(s)ds = 1$ and $f_X(s) \geq 0$ for all $s \in \mathbb{R}$. In fact, the reverse holds: for any function f_X satisfying these two properties, one can construct a random variable X such that X has PDF f_X.
- $P(X = x) = \int_x^x f_X(t)\,\mathrm{d}t = 0$ for a continuous random variable. This property "explains" why we need to restrict the third probability axiom to collections of countable events.

MIXED A random variable is referred to as *mixed* if it is not discrete and not continuous. As an example, consider censoring: rather than observing some underlying continuously distributed random variable X, we observe $\max\{0, X\}$.

Digression. In analysis, functions F_X that can be written as an integral of another function are called absolutely continuous. Not all continuous distribution functions are absolutely continuous (see wikipedia article "Cantor Function", or Romano and Siegel (1986, Example 2.3)). Thus, not all variables can be written as mixtures of a continuous and discrete random variables. However, functions F_X that are continuous but not absolutely continuous are rather pathological and play little role in econometrics, at least in the one-dimensional case. In multidimensional cases, they can arise if a random vector is actually a function of a lower-dimensional random vector, such as $(X, -X)$. ⊠

Example. Is $F(x) = 1 - e^{-x}$ a CDF?[2] ⊠

[2] Here e is the exponential function. I am sure you've seen it before, but you have probably not seen is the following definition, which has nothing to do with this course. Suppose I ask you to grade your own homework: I collect it, shuffle it, and then give each student a random homework. Of course, no student should grade their own homework: what's the probability that this will indeed be the case? It's $\sum_{k=0}^{n}(-1)^k/k!$. Define the limit to be $1/e$.

2.2. *Quantiles*

If CDF F of some random variable X is strictly increasing and continuous then it has an inverse, $q(x) = F^{-1}(x)$. It is defined for all $x \in (0,1)$. Note that

$$P(X \leq q(x)) = P(X \leq F^{-1}(x)) = F(F^{-1}(x)) = x$$

for all $x \in (0,1)$. Therefore, $q(x)$ is called the *x-quantile* of X. It is such a number that random variable X takes a value smaller or equal to this number with probability x. If F is not strictly increasing or continuous, then we define $q(x)$ as a generalized inverse of F, i.e. $q(x) = \inf\{t \in \mathbb{R} \colon F(t) \geq x\}$ for all $x \in (0,1)$. In other words, $q(x)$ is a number such that $F(q(x) + \varepsilon) \geq x$ and $F(q(x) - \varepsilon) < x$ for any $\varepsilon > 0$. As an exercise, check that $P(X \leq q(x)) \geq x$.

2.3. *Transformations of Random Variables*

Suppose we have random variable X, and are interested in some real-valued function $g(X)$ of this random variable. To study $g(X)$, we can define another random variable $Y = g(X)$. The CDF of Y can be calculated as

$$F_Y(t) = P(Y \leq t) = P(g(X) \leq t) = P(X \in g^{-1}(-\infty, t]),$$

where g^{-1} is the (possibly set-valued) inverse of g. The set $g^{-1}(-\infty, t]$ consists of all $s \in \mathbb{R}$ such that $g(s) \in (-\infty, t]$, i.e. $g(s) \leq t$. If g is strictly increasing, then this simplifies to

$$F_Y(t) = P(X \leq g^{-1}(t)) = F_X(g^{-1}(t)) \tag{1}$$

for all $t \in g(\mathbb{R})$.

What about the density of Y?

Lemma 1 (Change of variable formula). Suppose that X is continuous with PDF f_X, and support $\mathcal{X} = \{x \in \mathbb{R} \colon f_X(x) > 0\}$. *Suppose that $g \colon \mathcal{X} \to \mathbb{R}$ is one-to-one, differentiable and that its derivative g' does not vanish. Then $Y = g(X)$ is also continuous with density*

$$f_Y(t) = \frac{f_X(g^{-1}(t))}{|g'(g^{-1}(t))|}$$

Proof. Since g is continuous, it must be either strictly increasing or strictly decreasing. If it's increasing, then differentiating eq. (1) and using the chain rule yields

$$f_Y(t) = \frac{dF_X(g^{-1}(t))}{dt} = f_X(g^{-1}(t))\frac{dg^{-1}(t)}{dt} = f_X(g^{-1}(t))\frac{1}{g'(g^{-1}(t))}.$$

If g is decreasing, differentiate $F_Y(t) = 1 - F_X(g^{-1}(t))$. Here we use the fact that if g is differentiable and $g' \neq 0$, then g^{-1} is differentiable. □

Example. Suppose that $g(x) = x^2$. Can you show that the PDF of Y equals $f_Y(y) = (f_X(-\sqrt{y}) + f_X(\sqrt{y}))/2\sqrt{y}$, for $y > 0$ and 0 otherwise? If you get stuck, check Example 2.1.7 in CB. As an example, suppose that the PDF of x is $(2\pi)^{-1/2}e^{-x^2/2}$. What is the PDF of y? Do you recognize the distribution of y? ☒

Example. An important type of function is a linear transformation. If $Y = aX + b$ for some $a, b \in \mathbb{R}$, and X is continuous, then Y is also continuous with $f_Y(t) = f_X((t - b)/a)/|a|$. If $a > 0$, then $F_Y(t) = P(aX + b \leq t) = F((t - b)/a)$. ☒

Example (Example 2.1.4 in CB). Suppose X has support $(0, 1)$, and PDF $f_X(x) = 1$. Consider the random variable $Y = -\log(X)/\lambda$. Then the change-of-variable formula implies that its PDF is $f_Y(y) = \lambda e^{-\lambda y}$. Do you recognize the distributions of X and Y? ☒

3. EXPECTATIONS

REFERENCE CB, Chapter 2.2–2.3, or HMC, Chapter 1.8–1.9, H22 Chapter 2.5–2.6, 2.13–2.17, 2.23

Informally, the expected value of some random variable can be interpreted as its average value. Formally, if X is a discrete random variable such that $\sum_i |x_i| p_i < \infty$, then its expectation is defined as

$$E[X] = \sum_i x_i p_i.$$

If X is continuous, and $\int_{-\infty}^{\infty} |x| f_X(x) \,\mathrm{d}x < \infty$, then

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) \,\mathrm{d}x.$$

If $\sum_i |x_i| p_i$ or $\int_{-\infty}^{\infty} |x| f_X(X) \,\mathrm{d}x$ are infinite, then we say the expectation does not exist.

If we're interested in the expectation of the random variable $g(X)$, for some function $g \colon \mathbb{R} \to \mathbb{R}$, then in the discrete case

$$E[g(X)] = \sum_i g(x_i) p_i,$$

and in the continuous case

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) \,\mathrm{d}x,$$

provided the expectations exist. Computing the expectation in this way is often much easier than computing it as $\int_{-\infty}^{\infty} y f_Y(y) \,\mathrm{d}y$, where f_Y is the PDF of $Y = g(X)$, which requires working out f_Y first.

For some particular functions g, the expectations are so common that they have earned themselves special names:

- mean: $g(x) = x$, $E[X]$;
- second moment: $g(x) = x^2$, $E[X^2]$;
- variance: $g(x) = (x - E[X])^2$, $E[(X - E[X])^2]$;
- k-th moment: $g(x) = x^k$, $E[X^k]$;
- k-th central moment: $g(x) = (x - E[X])^k$, $E[(X - EX)^k]$; and
- moment generating function (MGF) $g_t(x) = e^{tx}$, $E[e^{tX}]$. If the moment generating function exists for t in the neighborhood of zero, then all moments exist, and $E[X^k]$ equals the kth derivative of $E[e^{tx}]$ wrt t, evaluated at $t = 0$. The converse is not true (classic counterexample: $Y = e^X$, where X is normal, has all moments, but not an MGF; see Romano and Siegel (1986, Example 3.11)). If two random variables have the same MGF, then they have the same distribution.

The variance of a random variable X is commonly denoted by $V(X)$, or $\mathrm{var}(X)$.

Digression. Is knowing all moments $E[X^k]$ of a random variable enough to determine its distribution? The answer is no, unless the support of X is bounded. A classic counterexample, due to Heyde (1963), is as follows. Consider the PDFs $f_X(x) = (2\pi)^{-1/2} \cdot e^{-(\log x)^2/2}/x$, $x \geq 0$ (this is the PDF of a log-normal distribution) and $f_Y(x) = f_X(x) \cdot (1 + \sin(2\pi \log x))$. The PDFs are clearly different, so if $X \sim f_X(x)$ and $Y \sim f_Y(x)$, then X and Y have different distributions. It can be shown that if $X \sim f_X(x)$, all moments exist. Now, using the transformation $y = \log x - k$, we get:

$$\begin{aligned} E[Y^k] &= \int_0^\infty x^k f_X(x) \cdot (1 + \sin(2\pi \log x))\, \mathrm{d}x \\ &= E[X^k] + \int_0^\infty x^{k-1} (2\pi)^{-1/2} e^{-(\log x)^2/2} \sin(2\pi \log x)\, \mathrm{d}x \\ &= E[X^k] + (2\pi)^{-1/2} e^{k^2/2} \int_{-\infty}^\infty e^{-y^2/2} \sin(2\pi(y + k))\, \mathrm{d}y. \end{aligned}$$

Now, since the sine function is odd and $\sin(2\pi k + a) = \sin(a)$ for all integers k, the integrand is odd, and hence the integral is zero. Thus, $E[Y^k] = E[X^k]$ for all k. ⊠

Example 2 (St. Petersburg Paradox). You're offered to play the following game. We toss a coin until it turns up heads. If this happens on the kth toss, you get 2^k dollars. How much should you pay to participate? The expected value of the payout X is infinite: $E[X] = \sum_{k=1}^\infty 2^k \cdot 2^{-k} = \sum_{k=1}^\infty 1 = \infty$. This shows you the importance of having concave utility, and of having hurricane insurance. ⊠

3.1. *Properties of expectation*

1. The most useful property is linearity: if X and Y are two random variables whose expectations exist, and a and b are two constants, then $E[aX + bY] = aE[X] + bE[Y]$. This holds because sums and integrals are linear.
2. For any constant a (non-random), $E[a] = a$.

3. If X is a random variable, then $\operatorname{var}(X) = E[X^2] - (E[X])^2$. Indeed,

$$\begin{aligned}\operatorname{var}(X) &= E[(X - E[X])^2] = E[X^2 - 2XE[X] + (E[X])^2] \\ &= E[X^2] - E[2XE[X]] + E[(E[X])^2] = E[X^2] - 2E[X]E[X] + (E[X])^2 \\ &= E[X^2] - (E[X])^2.\end{aligned}$$

4. If X is a random variable and a, b are constant, then $\operatorname{var}(aX + b) = a^2 \operatorname{var}(X)$.

4. BIVARIATE (MULTIVARIATE) DISTRIBUTIONS

REFERENCE CB, Chapter 4.1–4.3, and 4.5–4.6, HMC, Chapter 2, H22, Chapter 4.

Several random variables are defined in analogy to the scalar case: a k-dimensional random vector $X = (X_1, \dots, X_k)$ is a (measurable) function from Ω to $\mathbb{R}^k$.

4.1. *Joint, marginal, conditional*

If X and Y are two random variables, then

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y) = P(\{\omega \colon X(\omega) \leq x\} \cap \{\omega \colon Y(\omega) \leq y\})$$

denotes their joint CDF. (X, Y) is a continuous random vector if for some function $f_{X,Y}$ called the *joint PDF*, $F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t)\, \mathrm{d}t\mathrm{d}s$.

As in the scalar case, at continuity points of $f_{X,Y}(x, y)$,

$$f_{X,Y}(x, y) = \frac{\partial^2 F_{X,Y}(x, y)}{\partial x \partial y}.$$

From the joint PDF $f_{X,Y}$ one can calculate the PDF of, say, X. Indeed,

$$F_X(x) = P(X \leq x) = \lim_{y \to \infty} F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f_{X,Y}(s, t)\, \mathrm{d}t\mathrm{d}s.$$

Therefore $f_X(s) = \int_{-\infty}^{\infty} f_{X,Y}(s, t) dt$. The PDF of X is called *marginal* if we want to emphasize that it comes from a joint PDF of X and Y. Two random variables are independent iff $F_{X,Y}(x, y) = F_X(x) F_Y(y)$. In the continuous case, independence is equivalent to (check!) $f_{X,Y}(x, y) = f_X(x) f_Y(x)$.

Example. Let $f_{X,Y}(x, y) = (x + y)\, \mathbb{1}\{0 \leq x \leq 1\}\, \mathbb{1}\{0 \leq y \leq 1\}$. Then you can check that $f_X(x) = \mathbb{1}\{0 \leq x \leq 1\}(x + 1/2)$, and that X and Y are not independent. ⊠

If X and Y have a joint PDF, then we can define a *conditional* PDF of Y given $X = x$

(for x such that $f_X(x) > 0$):

$$f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{f_X(x)}. \tag{2}$$

- Since $P(X = x) = 0$, this definition cannot be directly be motivated by the definition of conditional probability. However, one way to argue for this formula is to condition instead on the event $x - h < X < x + h$ for $h > 0$ and then to make limiting arguments as $h \to 0$. In any case, you should check that the resulting PDF is well-defined, since $f_{Y|X}(y \mid x) \geq 0$ and $\int_{-\infty}^{\infty} f_{Y|X}(y \mid x)\,\mathrm{d}y = 1$.
- The conditional CDF is $F_{Y|X}(y) = \int_{-\infty}^{y} f_{Y|X}(u \mid x)\,\mathrm{d}u$, and most definitely *not* $F_{Y|X}(y) = F_{X,Y}(x,y)/F_X(x)$.

Conditional probability is a full characterization of how Y is distributed for any given $X = x$. The probability that $Y \in A$ for some set A given that $X = x$ can be calculated as $P(Y \in A \mid X = x) = \int_A f_{Y|X}(y \mid x)dy$. We can calculate the conditional expectation of Y given $X = x$ similarly: $E[Y \mid X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y \mid x)dy$. As an exercise, think how we can define the conditional distribution of Y given $X = x$ if X and Y are discrete random variables.

One extremely useful property of a conditional expectation is *the law of iterated expectations*: for any random variables X and Y,

$$E[E[Y \mid X]] = E[Y].$$

Another useful identity (some statisticians call it EVVE's law, and call the law of iterated expectations Adam's law) is

$$\operatorname{var}(Y) = E[V(Y \mid X)] + V(E[Y \mid X]).$$

Example (Optimal Forecasting). Suppose you want to predict Y based on the information seeing $X = x$. Let $h(x)$ be the forecast, so that $Y - h(x)$ is the forecast error. Suppose we are interested in minimizing expected squared loss, i.e. $E[(Y - h(x))^2 \mid X = x]$. How should we choose $h(x)$? The answer is that the optimal h is given by $h^*(x) = E[Y \mid X = x]$. The proof is a problem-set exercise. A mind-blowing fact is that this property fully characterizes the conditional expectation: indeed, one way to define the conditional expectation is to define it as a function $h^*(x)$ which minimizes the forecast error. ☒

4.2. *More on independence*

If X, Y are continuous, you can check that X and Y are independent iff $f_{Y|X}(y \mid x) = f_Y(y)$ for all $x \in \mathbb{R}$, i.e. if the marginal PDF of Y equals conditional PDF Y given $X = x$ for all $x \in \mathbb{R}$. If X and Y are independent, then $g(X)$ and $f(Y)$ are also independent for any functions $g\colon \mathbb{R} \to \mathbb{R}$ and $f\colon \mathbb{R} \to \mathbb{R}$. In addition, if X and Y are independent, then

$E[XY] = E[X]E[Y]$. Indeed,

$$\begin{aligned} E[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_{X,Y}(x,y)dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xyf_X(x)f_Y(y)dxdy \\ &= \int_{-\infty}^{\infty} xf_X(x)dx \int_{-\infty}^{\infty} yf_Y(y)dy \\ &= E[X]E[Y]. \end{aligned}$$

The converse is not true, as discussed in Remark 2 below.

4.3. *Covariance*

For any two random variables X and Y we can define covariance as

$$\operatorname{cov}(X,Y) = E[(X - E[X])(Y - E[Y])].$$

As an exercise, check that $\operatorname{cov}(X,Y) = E[XY] - E[X]E[Y]$.

Covariances have several useful properties:

1. $\operatorname{cov}(X,Y) = 0$ whenever X and Y are independent.
2. $\operatorname{cov}(aX,bY) = ab\operatorname{cov}(X,Y)$ for any random variables X and Y and any constants a and b
3. $\operatorname{cov}(X+a,Y) = \operatorname{cov}(X,Y)$ for any random variables X and Y and any constant a
4. $\operatorname{cov}(X,Y) = \operatorname{cov}(Y,X)$ for any random variables X and Y
5. $|\operatorname{cov}(X,Y)| \leq \sqrt{\operatorname{var}(X)\operatorname{var}(Y)}$ for any random variables X and Y
6. $\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y)$ for any random variables X and Y
7. $\operatorname{var}(\sum_{i=1}^n X_i) = \sum_{i=1}^n \operatorname{var}(X_i)$ whenever $X_1, \ldots, X_n$ are independent

Remark 2. Following up on the remark in Section 4.2, the converse to property item 1 not true. Heuristically, covariance measures the strength of *linear* relationship between X and Y, and for independence, we need that there is *no* relationship between X and Y, linear or nonlinear. In between these properties is the property of *mean-independence*: we say that Y is mean-independent of X if $E[Y \mid X] = E[Y]$. This means that X is useless for predicting the first moment of Y. This property is weaker than independence, since X may still be useful for predicting other features of the distribution of Y, such as its variance.

To prove property 5, consider random variable $X - aY$ with $a = \operatorname{cov}(X, Y)/\operatorname{var}(Y)$. On the one hand, $\operatorname{var}(X - aY) \geq 0$. On the other hand,

$$\begin{aligned}\operatorname{var}(X - aY) &= \operatorname{var}(X) - 2a\operatorname{cov}(X, Y) + a^2\operatorname{var}(Y) \\ &= \operatorname{var}(X) - 2\frac{\operatorname{cov}(X, Y)^2}{\operatorname{var}(Y)} + \frac{\operatorname{cov}(X, Y)^2}{\operatorname{var}(Y)}.\end{aligned}$$

Thus, the last expression is nonnegative as well. Multiplying it by $\operatorname{var}(Y)$ yields the result. A shorter proof uses the Cauchy-Schwarz inequality—try it as an exercise.

The correlation of two random variables X and Y is defined by

$$\operatorname{corr}(X, Y) = \operatorname{cov}(X, Y)/\sqrt{\operatorname{var}(X)\operatorname{var}(Y)}.$$

By property 5 above, $|\operatorname{corr}(X, Y)| \leq 1$. $|\operatorname{corr}(X, Y)| = 1$ iff then X and Y are linearly dependent, i.e. there exist constants a and $b \neq 0$ such that $P(X = a + bY) = 1$.

Tukey (1954) famously asked: "Does anyone know when the correlation coefficient is useful, as opposed to when it is used? If so, why not tell us?" I am with Tukey in that I think a regression coefficient from a bivariate regression of Y on X is a more useful way of scaling covariance to make it more interpretable (you'll learn all about that in the second part of this course).

4.4. *Multivariate mean and covariance matrix*

Let X be a random vector with dimension k, $X = (X_1, \ldots, X_k)'$ (here the superscript $'$ denotes transpose, so that X is a column vector). The mean of X is defined as

$$\mu = E[X] = \begin{pmatrix} E[X_1] \\ \vdots \\ E[X_k] \end{pmatrix}.$$

For a random $k \times p$ matrix Y, the definition is similar.

The $k \times k$ matrix $\Sigma = E[XX'] - E[X]E[X]'$ is called the covariance matrix. The (i, j) element Σ_{ij} equals $\operatorname{cov}(X_i, X_j)$. You can check that if A is a non-random matrix, then $E[AX] = AE[X]$ and the covariance matrix of AX is given by $A\Sigma A'$.

4.5. *Transformations*

Suppose $X \in \mathbb{R}^k$ is continuous with density f_X. What's the density of $Y = g(X)$? Using a multivariate version of the change-of-variables formula from calculus (see Chapter 4.3

in CB or Chapter 2.2 in HMC), we obtain the formula

$$f_Y(t) = f_X(g^{-1}(t))|J|,$$

where $|J|$ is the absolute value of the Jacobian determinant J of the inverse transformation, that is the absolute value of the determinant of the matrix with (i,j) element equal to $\partial x_i / \partial y_j$.

Example. Suppose $Y = HX$, where H is a non-singular matrix. Then $|J| = |\det(H^{-1})| = |\det(H)|^{-1}$, so $f_Y(t) = f_X(H^{-1}t)|\det(H)|^{-1}$. ☒

5. EXAMPLES OF RANDOM VARIABLES

REFERENCE CB, Chapter 3.1–3.3, or HMC, Chapter 3.1–3.4, 3.6, H22, Chapter 3.

Discrete random variables:

BERNOULLI We say that a random variable X has a Bernoulli distribution with parameter p, denoted $X \sim \text{Bernoulli}(p)$, if it takes values from $\mathcal{X} = \{0,1\}$, $P\{X = 0\} = 1 - p$ and $P\{X = 1\} = p$. Its expectation $E[X] = 1 \cdot p + 0 \cdot (1-p) = p$. Its second moment $E[X^2] = 1^2 \cdot p + 0^2 \cdot (1-p) = p$. Thus, its variance is $\text{var}(X) = E[X^2] - (E[X])^2 = p - p^2 = p(1-p)$.

BINOMIAL If we have i.i.d. (independent and identically distributed) variables $X_i \sim \text{Bernoulli}(p)$, $i = 1, \dots, n$, then $Y = \sum_{i=1}^n X_i$ is Binomial with parameters n and p, and

$$P(Y = k) = \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k}.$$

POISSON X has a Poisson distribution with parameter λ, denoted $X \sim \text{Poisson}(\lambda)$, if it takes values from $\mathcal{X} = \{0,1,2,\dots\}$ and $P(X = j) = e^{-\lambda}\lambda^j / j!$. As an exercise, check that $E[X] = \lambda$ and $\text{var}(X) = \lambda$.

It is often used for modeling "successes" that occur over intervals of time (number of buses that arrive, number of people entering an obscure google search etc.).

Continuous random variables:

UNIFORM A random variable X has a uniform distribution with parameters a, b, denoted $X \sim U(a,b)$, if its density is $f_X(x) = 1/(b-a)$ for $x \in (a,b)$ and $f_X(x) = 0$ otherwise.

NORMAL X has a normal distribution with mean μ and variance σ^2, denoted $X \sim \mathcal{N}(\mu, \sigma^2)$, if its density is

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

for all $x \in \mathbb{R}$. Its expectation $E[X] = \mu$ and its variance $\operatorname{var}(X) = \sigma^2$ (you can check). As an exercise, check that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = (X - \mu)/\sigma \sim \mathcal{N}(0, 1)$. Y is said to have a standard normal distribution. It is known that the CDF of $\mathcal{N}(\mu, \sigma^2)$ is not analytic, i.e. it can not be written as a composition of simple functions. However, there exist tables that give its approximate values. The CDF of a standard normal distribution is commonly denoted by Φ: if $Y \sim \mathcal{N}(0, 1)$, then $F_Y(t) = P(Y \leq t) = \Phi(t)$.

CHI-SQUARED If $X_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, k$ are i.i.d., then $Y = \sum_{i=1}^{k} X_i^2$ is distributed chi-squared with k degrees of freedom, denoted $Y \sim \chi^2(k)$. You should check that $E[Y] = k$ and $\operatorname{var}(Y) = 2k$. Its PDF is given by $f(x) = x^{p/2-1}e^{-x/2}/(\Gamma(p/2)2^{p/2})$ if $x > 0$ and 0 otherwise, where $\Gamma\colon \mathbb{R}_+ \to \mathbb{R}_+$ is the gamma function defined by $\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}\,\mathrm{d}x$. For integer t, $\Gamma(t) = (t-1)!$. The gamma function generalizes the factorial function to non-integer arguments.

STUDENT T-DISTRIBUTION If $Z \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2(k)$, and Z and Y are independent, then $T = Z/\sqrt{Y/k}$ is distributed student t with k degrees of freedom, denoted $T \sim t(k)$

F DISTRIBUTION If $H \sim \chi^2(p)$ and $Y \sim \chi^2(k)$ are independent, then the distribution of the random variable

$$F = \frac{H/p}{Y/k}$$

is called the *F-distribution*, (sometimes *Snedecor's F distribution* or the *Fisher–Snedecor distribution*) with (p, k) degrees of freedom, denoted $F \sim F(p, k)$. Note that if $T \sim t(k)$, then $T^2 \sim F(1, p)$, and $1/T^2 \sim F(p, 1)$.

6. MULTIVARIATE NORMAL

REFERENCE HMC, Chapter 3.5, H22, Chapter 5.7–5.8.

Surprisingly, it's not really possible to give a direct definition of *multivariate normal distribution*. An indirect definition is that a p-dimensional random vector X is multivariate normal iff the one dimensional variables $a'X$ are normally distributed for all vectors $a \in \mathbb{R}^k$. It follows that each X_i is also (marginally) normally distributed. We use the notation $X \sim \mathcal{N}(\mu, \Sigma)$, or $X \sim \mathcal{N}_k(\mu, \Sigma)$ to emphasize the dimension, where $\mu = E[X]$ denotes its mean and $\Sigma = E[XX'] - \mu\mu'$ is the covariance matrix.

Digression. Why is it not sufficient to define an n-vector X to be normal if each X_i is normal? Here is a general answer. Let $C(x, y)$ be some joint probability distribution on the unit square $[0, 1]^2$ such that the marginals of C are uniform (a general fancy name for any such distribution is a *copula*). Then, since the marginals are assumed to be uniform, $C(x, 1) = x$ and $C(1, y) = y$.

Let $\Phi(x)$ denote the CDF of a standard normal distribution, and define the joint CDF of the random vector $(X, Y)'$ by $F(x, y) = C(\Phi(x), \Phi(y))$. Then the marginal distributions are standard normal, $P(X \leq x) = \Phi(x)$ and $P(Y \leq y) = \Phi(y)$ since $P(X \leq x) = \lim_{y\to\infty} C(\Phi(x), \Phi(y)) = C(\Phi(x), 1) = \Phi(x)$. The joint PDF is given by

$$f_{X,Y}(x,y) = \phi(x)\phi(y)c(\Phi(x),\Phi(y)),$$

where $\phi(x)$ is the density of a standard normal, and c is the density of C. This multivariate density only coincides with the multivariate normal density defined above if for some positive semi-definite matrix Σ,

$$c_\Sigma(x,y) = \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2} \begin{pmatrix} \Phi^{-1}(x) \\ \Phi^{-1}(y) \end{pmatrix}' (\Sigma^{-1} - I) \begin{pmatrix} \Phi^{-1}(x) \\ \Phi^{-1}(y) \end{pmatrix} \right).$$

In this case, C is called a *Gaussian copula*. For any other choice of c, such as

$$c(x,y) = 2(I(0 \le x, y \le 1/2) + I(1/2 < x, y \le 1)),$$

the random vector (X,Y) is *not* multivariate normal. In this case, the joint density becomes

$$f_{X,Y}(x,y) = \begin{cases} 2\phi(x)\phi(y) & \text{if } x \text{ and } y \text{ have the same sign,} \\ 0 & \text{otherwise.} \end{cases}$$

In this case, the marginals of X and Y are normal, X and Y are positively correlated, but the joint distribution is not normal. If fact, by being clever about the copula C, one can generate an arbitrary dependence structure between X and Y, and yet preserve the normality of the marginal distributions. ☒

A normal distribution has several useful properties:

1. If Σ is positive definite (remember that the $k \times k$ matrix Σ is positive definite if $a'\Sigma a > 0$ for any non-zero $k \times 1$ vector a), then one can show that the multivariate PDF is given by
$$f_X(x) = \frac{\exp(-(x-\mu)'\Sigma^{-1}(x-\mu)/2)}{(2\pi)^{n/2}\sqrt{\det(\Sigma)}}.$$
for any $k \times 1$ vector x. Some texts define the multivariate normal through its PDF, but this has the disadvantage that it rules out the case in which Σ is singular.

2. If X and Y are uncorrelated and jointly normal, then X and Y are independent. As an exercise, check this statement. Conversely, if X and Y are independent and marginally normal, then they are jointly normal.

3. If $X \sim \mathcal{N}(\mu, \Sigma)$ is a $k \times 1$ dimensional normal vector, and A is a fixed $p \times k$ matrix, then $Y = AX$ is a normal $p \times 1$ vector: $Y \sim \mathcal{N}(A\mu, A\Sigma A')$.

4. Special case of 3: If $X \in \mathcal{N}(\mu, \Sigma) \in \mathbb{R}^2$ with $\Sigma_{12} = 0$ (so that X_1 and X_2 are independent), then $X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \Sigma_{11} + \Sigma_{22})$

Remark 3. For property 2, the statement is not true if X and Y are merely uncorrelated and marginally normal. For example, set $Y = WX$, where W is Rademacher and independent of $X \sim \mathcal{N}(0,1)$ (i.e. $P(W=1) = 1/2$ and $P(W=-1) = 1/2$). Then $Y \sim \mathcal{N}(0,1)$ and $\text{cov}(X,Y) = E[WX^2] - E[WX]E[X] = 0 - 0 = 0$, but clearly Y and X are not independent. The problem is that $P(X+Y=0) = 1/2$, so they are not jointly normal.

6.1. *Conditional distribution*

Another useful property of a normal distribution is that its conditional distribution is normal as well. If

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12} & \Sigma_{22} \end{pmatrix} \right),$$

then

$$X_1 \mid X_2 = x_2 \sim \mathcal{N}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2)\ ,\ \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}),$$

provided Σ_{22} is full rank. If X_1 and X_2 are both random variables (as opposed to random vectors), then the conditional mean becomes $E[X_1 \mid X_2 = x_2] = \mu_1 + \operatorname{cov}(X_1, X_2)(x_2 - \mu_2)/\operatorname{var}(X_2)$.

Proof. One has several options. The first is brute force application of the definition of conditional PDF in eq. (2). We give a cleaner proof. Consider the joint distribution of

$$\begin{pmatrix} W \\ X_2 \end{pmatrix} = \begin{pmatrix} I & -\Sigma_{12}\Sigma_{22}^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$$

It follows from property 3 that W and X_2 are joint normal with zero covariance, and that $E[W] = \mu_1 - \Sigma_{12}\Sigma_{22}^{-1}\mu_2$, and $\operatorname{var}(W) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Hence, by property 2, W and X_2 are independent. Therefore,

$$W + \Sigma_{12}\Sigma_{22}^{-1}X_2 \mid X_2 = x_2 \sim W + \Sigma_{12}\Sigma_{22}^{-1}x_2$$

Since $X_1 = W + \Sigma_{12}\Sigma_{22}^{-1}X_2$, this gives the result. □

Digression. This sort of proof strategy turns out to be fairly general. In fact, as you'll discover in the second part of the course, the variable W is the residual from the (population) regression of the (vector) "outcome" X_1 onto a vector of "predictors" X_2. That's how I came up with the "consider": we know that residuals are uncorrelated with the predictors by definition. ☒

REFERENCES

Casella, George, and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Pacific Grove, CA: Duxbury/Thomson Learning.

Durrett, Rick. 2019. *Probability: Theory and Examples.* 5th ed. New York, NY: Cambridge University Press. https://doi.org/doi.org/10.1017/9781108591034.

Hansen, Bruce E. 2022. *Probability and Statistics for Economists.* Princeton, NJ: Princeton University Press.

Heyde, Christopher Charles. 1963. "On a Property of the Lognormal Distribution." *Journal of the Royal Statistical Society,* Series B (Methodological), 25, no. 2 (July): 392–393. https://doi.org/10.1111/j.2517-6161.1963.tb00521.x.

Hogg, Robert V., Joseph W. McKean, and Allen T. Craig. 2019. *Introduction to Mathematical Statistics.* 8th ed. Boston, MA: Pearson.

Romano, Joseph P., and Andrew F. Siegel. 1986. *Counterexamples in Probability and Statistics.* New York, NY: Routledge. https://doi.org/10.1201/9781315140421.

Solovay, Robert M. 1970. "A Model of Set-Theory in Which Every Set of Reals Is Lebesgue Measurable." *Annals of Mathematics* 92, no. 1 (July): 1–56. https://doi.org/10.2307/1970696.

Tukey, John W. 1954. "Unsolved Problems of Experimental Statistics." *Journal of the American Statistical Association* 49, no. 268 (December): 706–731. https://doi.org/10.2307/2281535.