STAT 460/560 Class 1: Intro, overview, review of basic probability and random variables

Ben Bloem-Reddy

Reading: Chapters 1-2, [Was04]

1. Intro, logistics

2. Brief survey

3. Course overview and syllabus

4. Super speedy review of probability

Probability is a pre-requisite for this course. If anything in the rest of this class sheet is unfamiliar/unknown to you, you will probably struggle in STAT 460/560. *Everything that follows should be review.*

We will not get through everything on the sheet today. You should finish the activities and exercises before the next class to make sure that you're ready for this course.

5. Review of sample spaces and set notation

This really will be a review of basic probability. Recall that we use probability as a mathematical model for uncertainty in experiments. An experiment itself is modeled as:

- A sample space, Ω, which contains all of the possible *outcomes* $\omega \in \Omega$ of an experiment.
- Subsets of $E \subseteq \Omega$, called *events*.

Common examples include an experiment consisting of two coin flips ($\Omega = \{HH, HT, TH, TT\}$) or the measurement of temperature ($\Omega = \mathbb{R}_+ = [0, \infty)$).

> **Exercise 1.1.** Formalize a problem (experiment) of interest to you in terms of a sample space and outcomes, and describe two non-trivial events.

A set is just a collection of elements. Given two sets (or events) A and B, the *set difference* is $A \setminus B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \notin B\}$. Given a set A, its *complement* is $A^c = \{\omega \in \Omega : \omega \notin A\} = \Omega \setminus A$. Note that $A \setminus B = A \cap B^c$. The complement of Ω is the empty set, $\emptyset = \{\}$. The *union* of two events, A and B, is $A \cup B = \{\omega \in \Omega : \omega \in A \text{ or } \omega \in B \text{ or both}\}$. The *intersection* is $A \cap B = \{\omega \in \Omega : \omega \in A \text{ and } \omega \in B\}$. The notation $A \subset B$ indicates that the set A is contained in B. If A is finite, $|A|$ is the number of elements in it.

A sequence of sets, $A_1, A_2, \ldots$, is *disjoint* if $A_i \cap A_j = \emptyset$ for each $i \neq j$. A disjoint sequence forms a *partition* if also $\cup_{i \geq 1} A_i = \Omega$.

6. Axioms of probability

We won't worry about σ-algebras and measure theory in the class, though if you have encountered that material before then it's good to keep it in mind. I'll try to point out any potential technical difficulties we're glossing over in the course.

Using probability as a model of randomness stipulates that once we have our sample space Ω and collection of events $A \subset \Omega$, we also have a *probability measure* (or probability distribution), $\mathbb{P}$, that assigns a real number to each event A, denoted $\mathbb{P}(A)$. The *axioms of probability*, due to Kolmogorov, are conditions on $\mathbb{P}$:

Axiom 1 $\mathbb{P}(A) \geq 0$ for every event A.

Axiom 2 $\mathbb{P}(\Omega) = 1$.

Axiom 3 Countable additivity: If $A_1, A_2, \ldots$ are disjoint, then

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\ . \tag{1.1}$$

Activity 1.1. Prove the following properties of $\mathbb{P}$.

1. *Norming*: $\mathbb{P}(\emptyset) = 0$
2. *Finite additivity*: $A \cap B = \emptyset \Rightarrow \mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$
3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
4. *Monotonicity*: $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
5. *Inclusion/exclusion*: $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$

A sequence of sets is *monotone increasing* if $A_1 \subset A_2 \subset \cdots$. Its limit is defined as $\lim_{n\to\infty} A_n = \cup_{i\geq 1} A_i$. A sequence is *monotone decreasing* if $A_1 \supset A_2 \supset \cdots$. Its limit is $\lim_{n\to\infty} A_n = \cap_{i\geq 1} A_i$. Probability measures are continuous with respect to these kinds of limits.

Theorem 1.1. *If $A_1, A_2, \ldots$ is either monotone increasing or monotone decreasing then*

$$\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A)\ . \tag{1.2}$$

Exercise 1.2. Prove Theorem 1.1. (This is Theorem 1.8 in [Was04], where much of the proof is given. This activity therefore involves filling in the missing details of the monotone increasing direction, and also proving the monotone decreasing direction.)

7. Probability on discrete spaces

Recall that when Ω is finite (i.e., consists of a finite collection of elements) then the *uniform distribution* on Ω is defined by

$$\mathbb{P}(A) = \frac{|A|}{|\Omega|}\ , \quad A \subset \Omega\ . \tag{1.3}$$

8. Independent events

A set of events $\{A_i : i \in I\}$ is independent if

$$\mathbb{P}\left(\cap_{i\in I} A_i\right) = \prod_{i\in I} \mathbb{P}(A_i)\ . \tag{1.4}$$

Exercise 1.3. Exercise 1.11 in [Was04].

9. Random variables

We start with a sample space Ω, events, etc., as in last class. In general, a random variable is a mapping, or function, from Ω into some set in which the random variable takes values. In this class, we will deal almost exclusively with real-valued random variables, in which case our definition is as follows: a random variable is a mapping

$$X\colon \Omega \to \mathbb{R}\ , \tag{1.5}$$

that assigns a real number $X(\omega)$ to each $\omega \in \Omega$.[1] I find it helpful to keep in mind that a random variable is just a function that becomes random when we feed randomness into it.

As Wasserman notes, at some point it is common to stop mentioning the sample space Ω and work directly on the space(s) where our random variables take their values, but "the sample space is really there, lurking in the background." Things like dependence/independence don't work without an underlying sample space tying everything together, so it is necessary.

Given a random variable X and a subset $A \subset \mathbb{R}$, the *inverse image* is

$$X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}\ . \tag{1.6}$$

The probability distribution corresponding to X is

$$\mathbb{P}(X \in A) = \mathbb{P}(X^{-1}(A))\ . \tag{1.7}$$

Recall that the *indicator function* of a set $A \subset \Omega$ is

$$\mathbf{I}_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \notin A \end{cases}\ . \tag{1.8}$$

Observe that according to the definition of random variable, an indicator function on a subset of Ω is a random variable.

Example 1.1 (From [Was04], Example 2.4). Flip the same coin twice independently and let X be the number of heads. Explicitly, if Y_1, Y_2 are the outcomes of the coin flips, then

$$X(\omega) = \mathbf{I}_{\{H\}}(Y_1(\omega)) + \mathbf{I}_{\{H\}}(Y_2(\omega))\ . \tag{1.9}$$

X and its distribution are summarized as follows.

ω	$\mathbb{P}(\{\omega\})$	$X(\omega)$
TT	1/4	0
TH	1/4	1
HT	1/4	1
HH	1/4	2

$\mapsto$

x	$\mathbb{P}(X = x)$
0	1/4
1	1/2
2	1/4

Exercise 1.4. Continuing from Exercise 1.1, specify a random variable that occurs in your problem of interest. Is the distribution of the random variable easily calculated, as in Example 1.1?

10. Distribution functions, densities, etc.

For a random variable X that takes values in $\mathbb{R}$, there a few important functions that tell us everything we need to know about its distribution.

- The *cumulative distribution function*, or CDF, is the function $F_X\colon \mathbb{R} \to [0, 1]$, defined by

  $$F_X(x) = \mathbb{P}(X \le x)\ . \tag{1.10}$$

  Theorem 2.7 in [Was04] establishes that the CDF characterizes the distribution: If $F_X(x) = F_Y(x)$ for all $x \in \mathbb{R}$ then X and Y have the same distribution. Equality in distribution is denoted by $X \overset{\mathrm{d}}{=} Y$. Keep in mind that this is a statement about distributions, not about X and Y.

[1] Technically, a random variable must be measurable with respect to a σ-algebra on Ω and one on its range space.

Conversely, Theorem 2.8 in [Was04] says that if a function $F : \mathbb{R} \to [0, 1]$ "looks like" a distribution function (i.e., it is non-decreasing, right-continuous, and has the correct limits) then it is the CDF of *some* probability measure.

Lemma 2.15 in [Was04] collects some useful identities for computing probabilities from the CDF.

- The *quantile function*, or inverse CDF, is defined by

$$F^{-1}(q) = \inf\{x : F(x) > q\} . \tag{1.11}$$

  If F is continuous and strictly increasing then this is the functional inverse of F, i.e., $F^{-1}(q)$ is the unique real number x such that $F(x) = q$. If F has jumps and/or regions on which it is not increasing (i.e., it is flat) then some care must be taken. (Try computing the quantile function of the CDF in Figure 2.1 of [Was04].)

- If X takes countably many values $\{x_1, x_2, \dots\}$ then we say it is *discrete*, in which case the *probability mass function*, or PMF, is defined as $f_X(x) = \mathbb{P}(X = x)$. Thus, the CDF can be obtained as

$$F_X(x) = \sum_{x_i \le x} f_X(x_i) . \tag{1.12}$$

  The PMF uniquely characterizes the corresponding probability measure.

- X is said to be *continuous* if there is a function f_X such that $f_X(x) \ge 0$ for all $x \in \mathbb{R}$, $\int_{-\infty}^{\infty} f_X(x)dx = 1$, and

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx . \tag{1.13}$$

  The function f_X is called the *probability density function*, or PDF. Clearly, then,

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt , \tag{1.14}$$

  and $f_X(x) = \frac{dF_X}{dx}(x)$ at all points x at which F_X is differentiable.

  The PDF uniquely characterizes the corresponding probability measure.[2]

Activity 1.2. Exercise 2.4(a) of [Was04].

Exercise 1.5. Also find the quantile function for the previous activity.

Activity 1.3. Exercise 2.6 of [Was04].

Exercise 1.6. The *Poisson distribution* with parameter $\lambda > 0$ has PMF

$$f_X(x) = e^{-\lambda}\frac{\lambda^x}{x!} , \quad x \in \{0, 1, 2, \dots\} . \tag{1.15}$$

Write down the CDF, F_X, and show that $\lim_{x\to\infty} F_X(x) = 1$.

Exercise 1.7. The *Gamma distribution* with parameters $\alpha > 0, \beta > 0$ has PDF

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} , \quad x \in (0, \infty) . \tag{1.16}$$

[2] Technically this is only true up to sets of (Lebesgue) measure zero, so there may be many "different" PDFs that are probabilistically equivalent; all of the probabilistic calculations performed with them (CDF, expectation, etc.) will agree.

($\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt$ is the so-called Gamma function, which gives the distribution its name.)

Identify the names of the distributions corresponding to the following special cases:

- $\alpha = 1$, $\beta > 0$.
- $\alpha = p/2$, $\beta = 2$.

11. Multivariate and marginal distributions, independence

Let $X_1, \dots, X_n$ be random variables, and define $X = (X_1, \dots, X_n)$. X is called a *random vector*, or multivariate random variable. If the X_i's are discrete then the *joint PMF* is

$$f_X(x_1, \dots, x_n) = \mathbb{P}(X_1 = x_1, \dots, X_n = x_n) \; . \tag{1.17}$$

If the X_i's are continuous then the *joint PDF* is the function $f_X : \mathbb{R}^n \to \mathbb{R}_+$ satisfying:

- $f_X(x_1, \dots, x_n) \geq 0$ for all $(x_1, \dots, x_n) \in \mathbb{R}^n$;
- $\int_{\mathbb{R}^n} f_X(x_1, \dots, x_n) dx_1 \cdots dx_n = 1$;
- for any set $A \in \mathbb{R}^n$, $\mathbb{P}(X \in A) = \int_A f_X(x_1, \dots, x_n) dx_1 \cdots dx_n$.

As in the univariate case, a joint PMF/PDF is in unique correspondence with a probability measure on $\mathbb{R}^n$ (with the same measure-theoretic caveat about equivalence up to sets of measure zero).

For simplicity, we'll focus here on $n = 2$, the bivariate case, and write (X, Y) for the two random variables. From a joint PMF/PDF, we can obtain the marginal PMF/PDF of X by summing/integrating the joint PMF/PDF with respect to Y (over its entire support).

The random variables X and Y are said to be *independent* if for every $A, B \subset \mathbb{R}$,

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\mathbb{P}(Y \in B) \; . \tag{1.18}$$

If X and Y are independent, we write $X \perp\!\!\!\perp Y$.

Activity 1.4. [Was04], Exercise 2.10.

Conveniently, independence can be checked with the PMF/PDF.

Theorem 1.2 ([Was04], Theorem 2.30)**.** *Let X and Y have joint PDF $f_{X,Y}$. Then $X \perp\!\!\!\perp Y$ if and only if $f_{X,Y}(x, y) = f_X(x)f_Y(y)$ for all $(x, y) \in \mathbb{R}^2$.*

Exercise 1.8. Prove Theorem 1.2.

This generalizes to $n > 2$. If $X_1, \dots, X_n$ are independent and have the same marginal distribution with CDF F then we say that they are independent and identically distributed, or IID.

12. Conditional distributions*

Properly defining conditional distributions in general takes some sophisticated techniques from measure theory, but if we have a PMF or PDF then things work out without too much difficulty. Let $f_{X,Y}(x, y)$ be a joint PMF/PDF and $f_Y(y)$ the corresponding marginal PMF/PDF of Y. Then the *conditional PMF/PDF* of X given Y is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x, y)}{f_Y(y)} \; , \quad \text{if } f_Y(y) > 0 \; . \tag{1.19}$$

Exercise 1.9. Exercise 2.17, [Was04].

13. Transformations*

Given a random variable X with PMF/PDF f_X, how do we find the distribution of $Y = r(X)$, for some function r? Wasserman [Was04] breaks the general procedure down into three steps:

1. For each y, find the set $A_y = \{x : r(x) \leq y\}$.
2. Find the CDF
$$F_Y(y) = \int_{A_y} f_X(x)dx \; . \tag{1.20}$$
3. The PDF is $f_Y(y) = F'_Y(y)$.

If r is strictly monotone then it has a well-defined inverse $s = r^{-1}$, and the procedure simplifies into the formula

$$f_Y(y) = f_X(s(y)) \left| \frac{ds(y)}{dy} \right| \; . \tag{1.21}$$

Exercise 1.10. Let X and Y be independent random variables with PDFs f_X and f_Y, respectively. Let g and h be strictly monotone functions from $\mathbb{R}$ to $\mathbb{R}$. What is the joint PDF of $(g(X), h(Y))$?

Exercise 1.11. Exercise 2.21, [Was04].

14. The multivariate normal distribution

Suppose that X is a random vector of length k. X has a *multivariate normal distribution* with mean vector $\mu \in \mathbb{R}^k$ and symmetric, positive definite covariance matrix $\Sigma \in \mathbb{R}^{k \times k}$, denoted $X \sim \mathcal{N}(\mu, \Sigma)$, if it has PDF

$$f_X(x; \mu, \Sigma) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \right) \; . \tag{1.22}$$

Here, $|\Sigma|$ is the determinant of Σ.

There are two types of transformations we commonly encounter when dealing with normal distributions. The first is quite general.

Theorem 1.3. *Let $X \sim \mathcal{N}(\mu, \Sigma)$ in $\mathbb{R}^k$. Let $c \in \mathbb{R}^m$ be a vector and $B \in \mathbb{R}^{m \times k}$ be a matrix, so that $Y = c + BX$ is an affine transformation of X. Then $Y \sim \mathcal{N}(c + B\mu, B\Sigma B^T)$.*

The proof of this is straightforward but requires using the characteristic function.

As a special case, we can consider transforming an arbitrary normal random variable into a standard normal random variable, so that $\mu = 0$ and $\Sigma = \mathbb{I}$, the identity matrix. Since Σ is symmetric and positive definite, it can be factored as

$$\Sigma = U\Lambda U^T = (U\Lambda^{1/2})(U\Lambda^{1/2})^T = \Sigma^{1/2}\Sigma^{1/2} \tag{1.23}$$

where $U\Lambda U^T$ is the eigendecomposition with Λ a diagonal matrix of eigenvalues. In practice, the eigendecomposition is not used, but $\Sigma^{1/2}$ is obtained by Cholesky decomposition for numerical reasons.[3] This so-called square-root of Σ has some nice properties: $\Sigma^{1/2}$ is symmetric; $\Sigma^{1/2}\Sigma^{-1/2} = \Sigma^{-1/2}\Sigma^{1/2} = \mathbb{I}$; and $(\Sigma^{-1})^{1/2} = (\Sigma^{1/2})^{-1}$.

[3] The matrix obtained by Cholesky decomposition may differ from $U\Lambda^{1/2}$ but they will lead to equivalent results (in distribution).

Exercise 1.12. Show that if $X \sim \mathcal{N}(\mu, \Sigma)$ then $Z = \Sigma^{-1/2}(X - \mu) \sim \mathcal{N}(0, \mathbb{I})$.

See Chapter 16 of [JP04] for much more on the multivariate normal distribution.

References

[JP04] J. Jacod and P. Protter. *Probability Essentials*. 2nd. Springer-Verlag Berlin Heidelberg, 2004.

[Was04] L. Wasserman. *All of Statistics*. Springer New York, 2004.