STATISTICS

CHRISTIAN REMLING

1. Review of probability

We start out with a quick review of probability theory. A *probability measure* P (or just *probability,* in short) is a function on subsets of a *sample space* Ω with the following properties:

(1) $0 \leq P(A) \leq 1, \quad P(\Omega) = 1$;
(2) $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

The subsets $A, B \subseteq \Omega$ are often called *events* in this context.

Say you roll a fair die. Then you would describe this random experiment with the help of the sample space $\Omega = \{1, 2, \ldots, 6\}$ and the probability measure $P(\{j\}) = 1/6$ for $j = 1, 2, \ldots, 6$. Strictly speaking, I haven't specified a full probability measure yet: so far, $P(A)$ has been defined only for one element subsets of Ω, not for general A. However, additivity (= property (2)) of course forces us to set $P(A) = |A|/6$ for arbitrary $A \subseteq \Omega$, and this P does have properties (1), (2).

More generally, this procedure lets us define probability measures on arbitrary finite or countable sample spaces $\Omega = \{\omega_1, \omega_2, \ldots\}$: suppose we are given numbers p_n, $0 \leq p_n \leq 1$, with $\sum p_n = 1$. Then

$$P(A) = \sum_{n:\omega_n \in A} p_n \tag{1.1}$$

is a probability measure. The dice example above is of this type, with $p_n = 1/6$ for $n = 1, 2, \ldots, 6$. Such a finite or countable Ω is called a *discrete* sample space.

Exercise 1.1. Show that conversely, every probability P on a discrete sample space is of the type (1.1) for suitable p_n's. So what exactly are the p_n's equal to?

A different type of example is obtained if we take $\Omega = [0, 1]$, say, and

$$P(A) = \int_A dx.$$

This is also called the uniform probability measure on $[0, 1]$. More generally, we can take an integrable function $f(x) \geq 0$ with $\int_{-\infty}^{\infty} f(x)\, dx =$

1 and define

$$P(A) = \int_A f(x)\,dx \tag{1.2}$$

on the sample space $\Omega = \mathbb{R}$. The previous example can then be viewed as a special case of this, with $f = \chi_{[0,1]}$, the indicator function of the unit interval. We refer to sample spaces of this type as *continuous* sample spaces. The function f from (1.2) is called a *density.*

I mention in passing that if one wants to develop the theory fully rigorously, then certain technical issues arise here, which have to be addressed. For example, $P(A)$ is in fact not defined for all $A \subseteq \Omega$ by (1.2) because we cannot integrate over totally arbitrary subsets of $\mathbb{R}$. Thus in the continuous case, P will not be defined on *all* subsets of Ω but only on so-called *measurable sets.* Since all moderately reasonable sets are measurable, we can safely ignore these issues here. Also, to obtain a reasonably well-behaved theory, one has to replace (2) by a stronger variant, called *σ-additivity:*

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n) \qquad \text{if } A_m \cap A_n = \emptyset \text{ for } m \neq n$$

None of this matters for our purposes, so I'll leave the matter at that.

The central notion of probability theory is that of a *random variable.* A random variable is, by definition, a function $X : \Omega \to \mathbb{R}$. For example, on the sample space $\Omega = \{1, 2, \ldots, 6\}$, the function $X(\omega) = 2\omega + 3$ is a random variable, and so is $X(\omega) = \sin e^{\omega}$. These examples look rather contrived and not particularly relevant for the average gambler; typically, one would perhaps be more interested in a random variable such as

$$X(\omega) = \begin{cases} 0 & \omega \text{ odd} \\ 1 & \omega \text{ even} \end{cases}.$$

Given a random variable X, the probabilities of events of the form $\{\omega \in \Omega : X(\omega) \in B\}$, with $B \subseteq \mathbb{R}$, become especially relevant. In fact, X induces a new probability measure P_X on the new sample space $\Omega_X = \mathbb{R}$ in this way. More precisely, we define

$$P_X(B) = P(X \in B) \qquad (B \subseteq \mathbb{R}).$$

Here, we use self-explanatory (and very common) notation on the right-hand side: a condition as the argument of P (such as $X(\omega) \in B$) really refers to the event that is defined by this condition. In other words, a more explicit (and formally correct) version of the right-hand side would have been $P(\{\omega \in \Omega : X(\omega) \in B\})$.

The new probability measure P_X on $\mathbb{R}$ is called the *distribution* of X. The closely related function $F : \mathbb{R} \to \mathbb{R}$, $F(x) = P_X((-\infty, x]) = P(X \leq x)$ is called the *cumulative distribution function* of X. In probability theory and even more so in statistics, we are almost always interested in random variables; typically, we are told what their distribution is, and since this is good enough to work out the probabilities of events involving these random variables, the original sample space can then completely disappear in the background. Very frequently, one does not even bother to tell what it is.

Again, we will only consider *discrete* or *continuous* distributions here: in the first case, the distribution P_X is concentrated on a finite or countable set $\{x_1, x_2, \ldots\} \subseteq \mathbb{R}$, and, as above, it suffices to specify the numbers $p_n = P_X(\{x_n\}) = P(X = x_n)$ to determine P_X completely. In the second case, there is a density f so that

$$P(X \in B) = \int_B f(x)\, dx.$$

Exercise 1.2. Let X be a continuous random variable with density f and cumulative distribution function F. Show that $f = F'$.

The *expectation* or *expected value* EX of a random variable X can be defined as follows:

$$EX = \sum_n x_n P(X = x_n) \quad \text{(discrete)},$$
$$EX = \int_{-\infty}^{\infty} x f(x)\, dx \quad \text{(continuous)}$$

In both cases, we only define EX if the sum or integral converges *absolutely*. If EX^2 also converges, in this sense, then we can also define the *variance* of X:

$$\mathrm{Var}(X) = E(X - EX)^2$$

We first take the square here, then do the expectation. This matters:

Exercise 1.3. (a) Give an example of a random variable with $(EX)^2 \neq EX^2$.
(b) In fact, show that $\mathrm{Var}(X) = EX^2 - (EX)^2$ for any random variable.
(c) Conclude that $(EX)^2 \leq EX^2$.

The variance measures the average squared deviation from the expected value. If we take the square root of this, $\sigma = \sqrt{\mathrm{Var}(X)}$, then we obtain a rough measure of how far from its average value X will typically be. We call σ the *standard deviation* of X.

It is not so clear at first what good squaring and later taking the square root can do here, and indeed a more obvious way to introduce a typical deviation would have been $E|X - EX|$. This, however, is mathematically inconvenient: the absolute value is not as easily handled in algebraic manipulations as squares and square roots. It turns out that the actual definition of σ is much more useful.

Exercise 1.4. Consider $X(\omega) = \omega$ on the dice example sample space. Work out σ_X and $E|X - EX|$; in particular, show that these numbers are distinct.

Events $A, B \subseteq \Omega$ are called *independent* if $P(A \cap B) = P(A)P(B)$. This identity can be given a more intuitive form if we make use of *conditional probabilities,* defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad \text{if } P(B) \neq 0.$$

So if $P(B) \neq 0$, then A, B are independent precisely if $P(A) = P(A|B)$. If $P(A) \neq 0$ as well, then this is also equivalent to $P(B) = P(B|A)$. In this form, the condition catches the intuitive meaning of independence perfectly: A, B are independent precisely if the additional knowledge that B has occurred does not force us to change the probability for A (and vice versa).

Exercise 1.5. Show that A, B are independent precisely if A, B^c are independent.

Exercise 1.6. Show that if $P(B) = 0$ or $P(B) = 1$, then A, B are independent for any A. Also, convince yourself that this result is intuitively plausible.

Exercise 1.7. Consider the events $A = \{1, 3, 5\}$, $B = \{1, 2, 3\}$, $C = \{1, 6\}$ on the dice sample space. Which pairs are independent? Check your formal answers against your intuitive understanding.

Exercise 1.8. For what events A is A independent of itself (that is, A, A are independent)? Please also try to understand your answer intuitively.

More generally, we call a collection of (possibly more than two, perhaps infinitely many) events A_j *independent* if

$$P(A_{j_1} \cap A_{j_2} \cap \ldots \cap A_{j_N}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_N})$$

for any selection $A_{j_1}, \ldots, A_{j_N}$.

We call random variables X, Y *independent* if all pairs of events of the form $X \in A$, $Y \in B$, with $A, B \subseteq \mathbb{R}$ are independent. So,

intuitively, X, Y are independent if knowledge about the values of Y does not influence the distribution of X at all, and conversely. Say you toss a coin twice, and you let X and Y be the outcome of the first and second coin toss, respectively. Then it just has to be true that X, Y are independent. Let's try to set this up in a more precise way. A natural sample space is $\Omega = \{(x, y) : x, y = 0, 1\}$, with Laplace probability on this space (and we also identify $0, 1 \leftrightarrow H, T$). Then we are interested in $X(\omega) = \omega_1$, $Y(\omega) = \omega_2$, where we now write $\omega = (\omega_1, \omega_2) \in \Omega$, with $\omega_j = 0, 1$.

Exercise 1.9. Verify formally that X, Y are indeed independent.

Independence for a collection of random variables X_j is now defined in the expected way, by considering arbitrary selections of events of the form $X_{j_k} \in A_k$, as above.

If X, Y are independent, then $EXY = EX \cdot EY$ and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.

Exercise 1.10. Derive the second identity from the first one. Also, give examples that show that both identities will (in general) fail if X, Y are not assumed independent.

Exercise 1.11. Consider $\Omega = [0, 1]$ with uniform probability measure. Define random variables X_j as follows: cut $[0, 1]$ into 2^j intervals of equal length 2^{-j}, and then alternate between the values ± 1, starting with -1 (draw a picture; these functions X_j are called the *Rademacher functions*).
(a) Show that the X_j, $j = 1, 2, 3, \ldots$ are independent, identically distributed random variables with distribution $P(X_1 = -1) = P(X_1 = 1) = 1/2$.
(b) Find another function $Y : \Omega \to \{-1, 1\}$ that also has the same distribution, but is not one the X_j's. Can you in fact find a Y such that all of $Y, X_1, X_2, \ldots$ are still independent?

Some distributions occur so frequently that they deserve special names. A *Bernoulli* random variable X takes only two values, let's say 0 and 1. We call S *binomially distributed* with parameters n, p (in symbols: $S \sim B(n, p)$) if S takes the values $0, 1, \ldots, n$ and

$$P(S = k) = \binom{n}{k} p^k (1 - p)^{n-k}. \tag{1.3}$$

This is not as arbitrary as it perhaps looks at first sight. One natural way to produce a $B(n, p)$-distributed random variable is to start out with n independent, identically distributed (this is usually abbreviated

"iid") Bernoulli random variables $X_1, \ldots, X_n$, with $P(X_1 = 1) = p$, so $P(X_1 = 0) = 1 - p$. Then we set

$$S = X_1 + \ldots + X_n. \tag{1.4}$$

We can interpret S as counting the number of successes in a string of n independent Bernoulli experiments with probability of success equal to p in each individual trial (and by *success* we now of course mean that $X_j = 1$).

Let's prove this claim that the S from (1.4) is $B(n, p)$-distributed. Actually, before we do this, let's review another piece of general theory: For arbitrary random variables $Z_1, \ldots, Z_n$, it would not be possible to obtain the distribution of the sum $Y = Z_1 + \ldots + Z_n$ (or other functions $Y = g(Z_1, \ldots, Z_n)$) from the individual distributions of the Z_j. Rather, what we need is the *joint distribution* $P(Z_1 \in A_1, \ldots, Z_n \in A_n)$; this can be viewed as a probability measure $P_{Z_1,\ldots,Z_n}$ on the new sample space $\mathbb{R}^n$.

Exercise 1.12. Give a simple example that demonstrates that the joint distribution can not be obtained from the individual distributions. *Suggestion:* Take $\Omega = \{(x, y) : x, y = 0, 1\}$ and X, Y as in the example above. Compare this pair of random variables with the pair X, X. Show that the individual distributions agree, but (of course) the joint distributions of X, Y and X, X are (very) different.

However, this is not an issue here because we also assumed that the X_j are independent, and with this extra assumption, we have

$$P(X_1 \in A_1, \ldots, X_n \in A_n) = P(X_1 \in A_1) \cdots P(X_n \in A_n)$$

(independence is essentially defined by this condition), so we do obtain the joint distribution from the individual distributions here. In fact, since the X_j are also identically distributed, the right-hand side equals $P(X_1 \in A_1) \cdots P(X_1 \in A_n)$.

Now let's return to the problem at hand, namely the distribution of S from (1.4). Notice that $S = k$ precisely if $X_j = 1$ for exactly k values of $j = 1, 2, \ldots, n$, and then of course $X_j = 0$ for the remaining $n - k$ values. Each specific sequence of values of this type has probability $p^k(1 - p)^{n-k}$, and now (1.3), for the S from (1.4), follows by counting the $0, 1$ strings of length n with exactly k ones. These are determined by the slots where we put the ones, so there are $\binom{n}{k}$ such strings.

The binomial distribution comes up frequently; for example, if you toss a biased coin n times, then this random variable S describes the total number of 1's.

Our second example is the *normal distribution.* This is the most important distribution of them all. The normal distribution is a continuous distribution. We say that a random variable X is normally distributed with mean μ and standard deviation σ if X has density

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

Again, we sometimes write this as $X \sim N(\mu, \sigma)$. The normal distribution with $\mu = 0$, $\sigma = 1$ is called the *standard normal distribution.*

Exercise 1.13. Use the evaluation

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2}\, dx = 1$$

to show that f is indeed a density and $EX = \mu$, $\mathrm{Var}(X) = \sigma^2$ if $X \sim N(\mu, \sigma)$.

Unlike the binomial (and to some extent, the Poisson) distribution, the normal distribution is important not because it describes certain specific basic models but because it has a universal property: an arbitrary random experiment, repeated many times and suitably rescaled, approaches a normal distribution if the individual experiments are independent. This statement is known as the *central limit theorem.*

To formulate this precisely, let the X_j be iid random variables. The common distribution is almost completely arbitrary, we only need to assume that EX_1^2 exists. Let $S_n = X_1 + \ldots + X_n$, and let's abbreviate $\mu = EX_1$, $\sigma^2 = \mathrm{Var}(X_1)$. Then $ES_n = n\mu$ and, by independence, $\mathrm{Var}(S_n) = n\sigma^2$. We cannot really expect S_n to converge in any sense as $n \to \infty$ (typically, S_n should take values at about distance $\sigma\sqrt{n}$ from $n\mu$), but we can consider the rescaled version

$$\frac{S_n - n\mu}{\sqrt{n}\sigma},$$

which has expectation zero and variance one. Indeed, this is approximately $N(0,1)$-distributed:

Theorem 1.1 (Central Limit Theorem)**.** *For $a \le b$, we have that*

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sqrt{n}\sigma} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

Exercise 1.14. To entertain yourself, you toss a coin 40,000 times. Use the CLT to approximately compute the probability that the coin will come up heads at most 20,100 times. (Express this as an integral of the standard normal density, or look up the numerical value in a table.)