P. Probability[1]

1. DISCRETE RANDOM VARIABLES

1.1 Probability laws.

Suppose a repeatable event can have n distinct possible numerical outcomes $x_1, \ldots, x_n$. For example, the event could be the tossing of a die, with its face numbers 1, 2, 3, 4, 5, 6 as the possible outcomes; in symbols, $x_i = i, \quad i = 1, \ldots, 6$.

We assign to each possible outcome x_i a number $P(x_i)$, the *probability that x_i is the outcome of the event*. P can be thought of as a function whose domain is the finite set $\{x_1, \ldots, x_n\}$ and whose values lie between 0 and 1.

For instance, in the die-tossing example, if the die is fair (i.e., unloaded), then $P(i) = 1/6$, for $i = 1, \ldots, 6$. If the die is loaded, however, the values $P(i)$ will not all be equal. To say for instance that $P(2) = .4$ means that in a large number N of trials (i.e., rollings of the die), the face 2 will come up in about 4/10 (or 40%) of them. More precisely, we have

Definition 1.1A *Given a repeatable event with a set $\{x_1, \ldots, x_n\}$ of n possible numerical outcomes, the associated* **probability function** *P is defined to be the function having domain $\{x_1, \ldots, x_n\}$ and values*

$$P(x_i) = \lim_{N \to \infty} \frac{\#\text{ of times } x_i \text{ is the outcome in } N \text{ trials}}{N} .$$

Several important properties of probability functions are given by the Range, Addition, and Multiplication Laws.

Range Law. *(i)* $0 \le P(x_i) \le 1$ *for all* x_i .
(ii) x_i *is never the outcome* $\Rightarrow \; P(x_i) = 0$; $\quad x_i$ *is always the outcome* $\Rightarrow \; P(x_i) = 1$.
(iii) $P(x_1) + \ldots + P(x_n) = 1$.

Proof. The first two follow easily from Definition 1.1A; to prove (iii), in N trials, we have

$$(\#\text{ of outcomes } x_1) + \cdots + (\#\text{ of outcomes } x_n) \;=\; N,$$

so that dividing both sides by N and taking the limit as $N \to \infty$ gives

$$\lim_{N \to \infty} \frac{\#\text{ of outcomes } x_1}{N} + \cdots + \lim_{N \to \infty} \frac{\#\text{ of outcomes } x_n}{N} \;=\; 1,$$

which is (iii), in view of Definition 1.1A.

To state the other two probability laws, we need to extend the meaning of P so it assigns values also to combinations of outcomes. To illustrate once again with the die-tossing:

$P(2 \text{ or } 4)$ is the probability that the outcome of a trial will be either 2 or 4;

$P(2, \text{ then } 4)$ is the probability that two trials will have successive outcomes 2 and 4.

We shall continue to use the single word "outcome" to describe these combinations, and denote them in symbols by capital letters, writing for instance "the outcomes $A = (2 \text{ or } 4)$ and $B = (2, \text{ then } 4)$."

Addition law. *If* $x_i \ne x_j$, $P(x_i \text{ or } x_j) = P(x_i) + P(x_j)$;
if x_i, x_j, x_k *are distinct,* $P(x_i \text{ or } x_j \text{ or } x_k) = P(x_i) + P(x_j) + P(x_k)$, *and so on.*

These and their extensions to more outcomes are proved just like (iii) above.

We say two successive trials are **independent** if the outcome of the first has no influence on the outcome of the second: successive rolls of a die, for example. By contrast, succes-

[1] This section, with the accompanying exercises and solutions, is based on Notes by Frank Morgan Jr.

sive drawings without replacement of a card from a pack (outcomes: R, B) would not be independent, since if you first draw a red card, $P(R) < \frac{1}{2} < P(B)$ for the second drawing.

Multiplication law. *If A_1 and A_2 are outcomes for two independent trials, then*

$$P(A_1, \textit{then } A_2) = P(A_1) \cdot P(A_2).$$

The formula extends to N successive trials, if they are independent.

To see this intuitively for two, say $P(A_1) = 1/3$, and $P(A_2) = 1/4$. Then in a large number of trial pairs, about 1/3 will have A_1 as their first outcome, and of these, about 1/4 will have A_2 as their second outcome, so about $\frac{1}{3} \cdot \frac{1}{4} = \frac{1}{12}$ of the trial pairs will have outcome $(A_1, \text{ then } A_2)$.

Example 1.1A A fair die is tossed three times. What is the probability of getting an odd number each time?

Solution. Let A denote the outcome "odd number". By the addition law,

$$P(A) = P(1 \text{ or } 3 \text{ or } 5) = P(1) + P(3) + P(5) = \tfrac{1}{6} + \tfrac{1}{6} + \tfrac{1}{6} = 1/2 \quad ,$$

and by the (extended) multiplication law, $P(A, \text{ then } A, \text{ then } A) = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 1/8$.

Most of our probability work will be expressed using the idea of *random variable*, which summarizes and generalizes what we have done so far.

Definition 1.1B *A* **finite random variable** *X for a repeatable trial is given by*

(a) a finite set $\{x_1, \ldots, x_n\}$ of numerical values that X can take on;

(b) A probability function $P(X)$ whose domain is $\{x_1, \ldots, x_n\}$ and whose values $P(x_i)$ are given by Definition 1.1A.

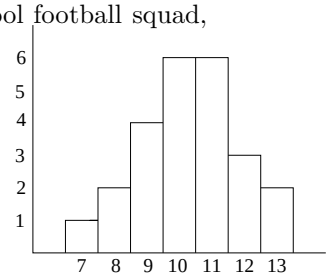
A finite random variable X is thus something like an ordinary real variable x, except that it can take on only a finite set of values (unlike x, which usually takes on all real values) and in addition with each such value x_i, there is an associated probability $P(x_i)$ that X takes on the value x_i.

Sometimes $P(x_i)$ is written $P(X = x_i)$ for emphasis or clarity, or to remind you of the symbol being used for the random variable..

The probability function $P(X)$ associated with the finite random variable X always satisfies the three laws: Range Law, Addition Law, and Multiplication Law, since these follow directly from Definition 1.1A.

Example 1.1B If we tabulate the shoe sizes of 24 men on a high school football squad, the results might be (the histogram summarizes the data):

shoe size :	7	8	9	10	11	12	13
no. of players :	1	2	4	6	6	3	2



Here the "shoesize" random variable S takes on values from the set $\{7, 8, \ldots, 13\}$, and the associated probability function $P(S)$ is defined by

$$P(S = 7) = \frac{1}{24}, \qquad P(9) = \frac{1}{6}, \qquad \text{and in general:} \quad P(i) = \frac{\#\text{ players wearing size } i}{24}\,.$$

1.2 Expectation. To get a quick feel for what a random variable X is telling you, it is natural to ask what its "average value" is; the official designation for this is the *mean* or *expected value* or *expectation* of X, in symbols, $m(X)$ or $\mathbf{E}(X)$. Treating this intuitively for the moment, for tosses of a fair die,

$$\text{expectation} \;=\; \text{average value} \;=\; \frac{1+2+3+4+5+6}{6} \;=\; 3.5$$

(Note that, like the 2.1 children in the average American family, the average value of X won't usually be one of its possible values.) We will learn more by calculating the average shoe size in Example 1.1B, a.k.a. the expectation of the shoesize random variable S.

$$\begin{aligned}\mathbf{E}(S) \;&=\; \frac{7+8+8+9+9+9+9+\ldots+13+13}{24}\\ &= \frac{7+8\cdot 2+9\cdot 4+\ldots+13\cdot 2}{24}\\ &= 7\cdot\frac{1}{24}+8\cdot\frac{2}{24}+9\cdot\frac{4}{24}+\ldots+13\cdot\frac{2}{24},\end{aligned}$$

which can be written more expressively as

$$= 7\cdot P(7)+8\cdot P(8)+9\cdot P(9)+\cdots+13\cdot P(13)\approx 10.3\ .$$

The reasoning leading up to this last line leads to the definition of expectation:

Definition 1.2 *The* **expectation** *(or* **mean** *or* **expected value***) of the finite random variable X having values $x_1,\ldots,x_n$ is defined to be*

$$\mathbf{E}(X) \;=\; x_1P(x_1)+\ldots+x_nP(x_n) \;=\; \sum_1^n x_iP(x_i)\ . \tag{2}$$

Example 1.2 A die is loaded so that 6 comes up twice as often as the other five numbers. Describe the associated random variable Y, and calculate its expectation.

Solution. Y takes on the values $1,\ldots,6$.

To determine its probability function $P(Y)$, we have

$$P(1)=\ldots=P(5)=\frac{1}{2}P(6),\quad \text{from the loading information;}$$
$$P(1)+P(2)+P(3)+P(4)+P(5)+P(6)=1,\quad \text{by the range law (iii);}$$

Substituting from the first line and adding, we get

$$5\cdot\frac{1}{2}P(6)+P(6)=1,\ \text{whence}\ P(6)=\frac{2}{7},\quad P(1)=\ldots=P(5)=\frac{1}{7}.$$

Using these values, and the definition (2), we get for the expectation

$$\mathbf{E}(Y)=\frac{1+2+3+4+5+6\cdot 2}{7}=\frac{27}{7}=3\tfrac{6}{7}\ ,$$

a little more weighted toward 6 than the $\mathbf{E}(X)=3.5$ calculated for the fair die.

We can think of the expectation as a *weighted average* of the values $x_1,\ldots,x_n$. The ordinary average would simply add up the values and divide by n (which gives each of them the equal weight $\frac{1}{n}$). Instead, we weight each x_i, multiplying it by the likelihood of its occurrence, and calculate this weighted average to get the expectation.

1.3 Discrete Random Variables. There are interesting random variables which take on an infinity of values. Our work above can be easily extended to these, provided we assume that the set of possible values can be arranged in a list $x_1 < x_2 < \ldots < x_n < \ldots$. (For example, if all the values are positive integers, they can be listed in this way; however, if the values are all the real numbers in some interval of positive length, no such list exists.)

In the definition, we will also include the case of finitely many values.

Definition 1.3 *A* **discrete random variable** X *for a repeatable trial consists of*

(a) a finite or infinite list $x_1 < x_2 < x_3 < \ldots < x_n < \ldots$ *of numerical values that* X *can take on;*

(b) A probability function $P(X)$*, i.e., a function whose domain is the set* $\{x_1, \ldots, x_n, \ldots\}$ *and whose values* $P(x_i)$ *are given by Definition 1.1A.*

As before, it follows from Definition 1.1A and the three probability laws (which remain valid when the list of x_i is infinite) that

$$0 \le P(x_i) \le 1 \qquad \text{and} \qquad P(x_1) + P(x_2) + \ldots + P(x_n) + \ldots = 1. \tag{3}$$

The new feature here is that if the list of possible numerical outcomes is infinite, then in (3) the sum is to be interpreted as meaning that the infinite series of probabilities converges to the sum 1. Similarly, by analogy to the finite case, we define the **expectation** $\mathbf{E}(X)$ to be the sum of the following series (if it converges):

$$\mathbf{E}(X) \;=\; \sum_1^\infty x_i P(x_i)\ . \tag{4}$$

Example 1.3 A fair coin is tossed until it comes up heads for the first time. Let X be the random variable telling the number of tosses required. Find $\mathbf{E}(X)$, the average number of tosses required if you make the trial many times.

Solution. The possible values of X are the positive integers. The corresponding probabilities $P(n)$ are given by:

n :	1	2	3	4	...
toss pattern:	H	TH	TTH	$TTTH$	...
$P(n)$:	1/2	1/4	1/8	1/16	...

For example, if $n = 3$, there are 8 possible patterns for the three tosses (two possibilities for each toss), all equally likely, but only the pattern TTH produces the value 3 for X. Or you can use the multiplication law for probabilities: the outcomes of the three tosses are independent events, so the probability of getting T on the first toss, T on the second, and H on the third is $\frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{8}$. According to (4),

$$\mathbf{E}(X) = \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{4}{16} + \ldots \;=\; \sum_1^\infty \frac{n}{2^n}\ .$$

To sum the series, think of its successive terms as the successive sums of the horizontal rows in the following pattern (only the first four rows are shown):

$$\begin{array}{cccc} 1/2 & & & \\ 1/4 & 1/4 & & \\ 1/8 & 1/8 & 1/8 & \\ 1/16 & 1/16 & 1/16 & 1/16 \end{array}$$

If we sum up the terms in this pattern by columns instead, the first column sums to 1, the second to 1/2 (since its terms are just half those of the first column), the third similarly to 1/4, so that adding up the sums of the successive columns gives

$$\mathbf{E}(X) \;=\; 1 + 1/2 + 1/4 + \ldots \;=\; 2\ .$$

1.4 The Poisson Random Variable This is a discrete random variable associated with events which occur sparsely, but whose frequency has to be estimated and allowed for: defects in a manufacturing process, errors in transmission or recording of data, requests for 100-year eggs at a suburban Chinese restaurant, etc.

Definition 1.4 *The* **Poisson random variable** *has as its domain the non-negative integers $k = 0, 1, 2, \ldots$, with associated probability function $P(X)$ given by*

$$P(X = k) = C\,\frac{m^k}{k!}, \qquad \text{where } m > 0 \text{ is fixed and } C = e^{-m}. \tag{5}$$

The value of C is dictated by the fact that the probabilities must sum to 1:

$$\sum_0^\infty C\,\frac{m^k}{k!} \;= C\sum_0^\infty \frac{m^k}{k!} = C\,e^m = 1\ .$$

The interpretation of the parameter m in the definition is given by

Theorem 1.4 *If X is a Poisson random variable with parameter m, then $\mathbf{E}(X) = m$.*

Proof. We use (4) and (5), dividing both sides by C to make calculating neater and dropping the first term (since $k = 0$):

$$\begin{aligned}\frac{1}{C}\mathbf{E}(X) &= 1\cdot\frac{m}{1!} + 2\cdot\frac{m^2}{2!} + 3\cdot\frac{m^3}{3!} + 4\cdot\frac{m^4}{4!} + \ldots\\ &= m\left(1 + \frac{m}{1!} + \frac{m^2}{2!} + \frac{m^3}{3!} + \ldots\right);\\ \mathbf{E}(X) &= C\,m\,e^m = m\ .\end{aligned}$$

Example 1.4 On average, two people during the lunchtime hour show up at the Wilbur box office to buy tickets for "Death of a Salesperson". What's the probability that in that hour: (a) no one shows up? (b) more than two show up?

Solution. Since box office visits seem infrequent, we take the number of lunchtime hour buyers to be a Poisson random variable X with mean (i.e., average value, or parameter value) $m = 2$. Using (5), (note that $0! = 1$),

(a) $P(X = 0) = e^{-2}\cdot\dfrac{2^0}{0!} = e^{-2} \approx .14$;

(b) $P(0) + P(1) + P(2) + P(X > 2) = 1$, since these exhaust the possibilities. Therefore

$$\begin{aligned}P(X > 2) &= 1 - e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!}\right)\\ &= 1 - e^{-2}(1 + 2 + 2)\\ &= 1 - 5e^{-2} \approx .32\end{aligned}$$

Exercises: Section 8A

2. CONTINUOUS RANDOM VARIABLES

2.1 Probability densities and distribution functions

The variables (like x, y, or t) used in calculus are usually not discrete, but rather *continuous* — that is, their values do not come from a finite or infinite list or numbers, but rather can be any real number, or any real number in some interval. For instance, if x is the height in inches of an M.I.T. undergraduate, its values could be any real number in say the interval $[30, 84]$, though the extreme values would be unlikely.

We want to turn x into a random variable X, that is, assign probabilities to the values of x. We can do this in two ways.

a) Since as a practical matter, height can only be measured to the nearest half inch, we can declare the possible values of x to be 30, 30.5, 31, ..., 83.5, 84, thus turning it into a discrete variable. Then by making a histogram of the actual student heights (as we did for the football-players' shoe sizes), we can assign a probability to each one of these possible heights, and thus create a discrete random variable X having x as its associated ordinary variable.

While this will be in some sense accurate, it has the disadvantage that we will not be able to use the procedures of calculus to study it: all calculations will have to be done by computer, and we might not see the general principles.

b) The other way is to find a continuous differentiable function $h(x)$ that is a good approximation to the histogram. Then

$$\text{no. of students such that } a \le X \le b \quad \approx \quad \int_a^b h(x)\,dx \tag{6}$$

To convert to probabilities, divide by the number N of students; letting $f(x) = \dfrac{h(x)}{N}$,

$$P(a \le X \le b) \quad = \quad \frac{\text{no. of students such that } a \le X \le b}{N} \quad \approx \quad \int_a^b f(x)\,dx. \tag{7}$$

Definition 2.1A *We define a* **continuous random variable** X *to be a variable* x *whose values lie in* $(-\infty, \infty)$*, together with a* **probability density** *function* $f(x)$*, such that*

$$f(x) \ge 0 \quad \textit{for all } x, \qquad \int_{-\infty}^{\infty} f(x)\,dx = 1\ , \tag{8a}$$

and for any two numbers $a < b$*,*

$$P(a \le X \le b) \ = \ \int_a^b f(x)\,dx\ . \tag{8b}$$

Sometimes the interval in which the random variable X takes its values (the domain of $f(x)$) is not the whole x-axis, but some smaller finite interval $[x_1, x_2]$, or half-infinite interval $[x_1, \infty)$. Then $[a, b]$ will lie inside this interval, and the second integral in (8a) will have limits x_1 and x_2 (resp. x_1, ∞).

In $P(a \le X \le b)$, one can replace $\le$ by $<$ without altering the value, and one can write x instead of X and it will be accepted (grudgingly).

The density function $f(x)$ is usually continuous, but it can have a finite number of discontinuities (or even an infinite number, as long as it is still integrable).

For calculating probabilities, (8b) shows that we have to know the area under the density curve $y = f(x)$. According to the first and second fundamental theorems of calculus,

$$(9) \qquad P(a \le X \le b) = \int_a^b f(x)\,dx = F(b) - F(a), \qquad \text{where} \qquad F(x) = \int_{-\infty}^{x} f(t)\,dt\ .$$

Definition 2.1B *The* **distribution function** *for the random variable X, or for its associated density function $f(x)$ is the function*

$$(10) \qquad F(x) \ = \ \int_{-\infty}^{x} f(t)\,dt\ .$$

The distribution function $F(x)$ is a particular antiderivative of the density $f(x)$, and as (9) shows, it is the principal tool in calculating probabilities. If $F(x)$ is not an elementary function, (10) shows it can be calculated by numerical integration. Users obtain its values from tables, calculators, or mathematics programs. From (8a) and (10), we see that

$$(11) \qquad F(x) \text{ is increasing;} \qquad \lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to\infty} F(x) = 1\ .$$

If the domain of X is the interval $[x_1, x_2]$, then in (10) the lower limit of the integral will be x_1, and in (11), the $-\infty$ and ∞ will be replaced respectively by x_1 and x_2.

Definition 2.1C *If X is a continuous random variable with $-\infty < x < \infty$ and probability density $f(x)$, we define its* **expectation** *by*

$$(12) \qquad \mathbf{E}(X) = \int_{-\infty}^{\infty} x\, f(x)\,dx$$

This is the continuous analog of the definition (4) of the expectation of a discrete random variable — the summation is replaced by integration, and the range of x-values is now $(-\infty, \infty)$, rather than the discrete list $x_1, x_2, \ldots$. As with discrete random variables, the expectation $E(X)$ can be thought of as the "average value" of X.

The expectation integral (12) will have limits x_1, x_2 if $[x_1, x_2]$ is the domain of X.

Example 2.1 Uniform random variables.

What is the simplest random variable?

Among the *finite* random variables X of section 1, it is the **uniform** random variable — the one in which all of the values $x_1, \ldots, x_n$ turn up with the same probability: $P(x_i) = 1/n$. Tossing a fair die gives an example.

On the other hand, if X is an *infinite* discrete random variable, i.e., has an infinite list $x_1, x_2, \ldots$ of possible values, it cannot be uniform. For if $P(x_i)$ has the same positive constant value k for all i, we would have $\sum_1^\infty P(i) = \infty$, whereas we know the probabilities have to sum to 1.

A similar situation holds for the continuous case. If the values of the continuous random variable X all lie in the finite interval $[x_1, x_2]$, then we call it *uniform* if its density function is constant. Since (8a) tells us the total area under the probability density curve must be 1, we have for the **continuous uniform random variable on** $[x_1, x_2]$:

$$f(x) \;=\; \frac{1}{x_2 - x_1} \quad \textit{(density)}; \qquad\qquad P(a \le X \le b) \;=\; \frac{b-a}{x_2 - x_1}, \quad [a,b] \subseteq [x_1, x_2].$$

As the integral (9) shows, for the continuous uniform random variable X, it is still the case that $P(X = x_0) = 0$ for any single value $x = x_0$: it is not the probability that is a positive constant, but rather the associated probability density.

Analogously to the discrete case, if the range of a continuous random variable X is an interval of infinite length, X cannot be uniform.

2.2 Exponential Random Variable. This is the continuous random variable whose domain is $[0, \infty)$, and whose density function is

$$f(x) = \frac{e^{-x/m}}{m}, \qquad x \ge 0\ . \tag{13}$$

exponential density

It depends on a single positive parameter m which is chosen to fit the model, and turns out to be the *expectation* of X.

Theorem 2.2 *For the exponential random variable (13) with parameter m, the distribution and expectation are respectively,*

$$F(x) = 1 - e^{-x/m}\ ; \qquad\qquad \mathbf{E}(X) = m\ . \tag{14}$$

Proof. From the definitions (10) and (12), we have

$$F(x) \;=\; \int_0^x \frac{e^{-t/m}}{m}\, dt = -e^{-t/m}\Big]_0^x = -e^{-x/m} + 1\ ;$$

and using integration by parts,

$$\begin{aligned}\int_0^\infty \frac{x\, e^{-x/m}}{m}\, dx &= -x\, e^{-x/m}\Big]_0^\infty - \int_0^\infty e^{-x/m}\, dx \\ &= \quad 0 \;-\; 0 \quad - e^{-x/m}(-m)\Big]_0^\infty \;=\; m\ .\end{aligned}$$

Example 2.2 A certain radioactive substance with a long half-life emits a β-particle on the average every 10 seconds. What's the probability of waiting more than a minute for the next emission?

Solution. Radioactive waiting times are typically modeled by an exponential random variable X, whose variable x represents the time to the next emission. Here the average value of x is 10, so $\mathbf{E}(X) = m = 10$, and we calculate from scratch:

$$P(x > 60) = \int_{60}^\infty \frac{e^{-x/10}}{10}\, dx = -e^{-x/10}\Big]_{60}^\infty = e^{-6} \approx .002,$$

or using (14),

$$=\; 1 - P(x < 60) \;=\; 1 - F(60) \;=\; e^{-60/10}\ .$$

Thus the probability of waiting more than a minute is .2%, to one significant figure.

Exercises: Section 8B

3. STANDARD DEVIATION

3.1 Variance and Standard Deviation.

The mean or expectation of the random variable X can be thought of as a parameter associated with X which summarizes in a single number $m = \mathbf{E}(X)$ some important information about X — its "average value".

In this section we define a second parameter, called the **standard deviation** of X, which measures in a single number σ ("sigma") how widely spread out the values of X are around their average value, weighting them according to their probability of occurring. If X is a continuous random variable, its probability density function will tend to look like a gentle hill centered around its mean if σ is large, but like a sudden peak arising from a relatively flat landscape if σ is small. If X is a discrete random variable, its histogram will have a similar appearance according to whether σ is large or small.

Mathematically, it is a little easier to work with σ^2 rather than σ itself. This quantity is called the **variance** of the random variable X, and denoted as above by σ^2. Its definition runs as follows. For convenience, we include the definition (4) of expectation given earlier.

Definition 3.1A *For a discrete random variable X, having values $x_1, x_2, \ldots$ we define*

$$m = \sum_i x_i P(x_i) \qquad \textit{(mean, or expectation, of } X\textit{)} \tag{4}$$

$$\sigma^2 = \sum_i (x_i - m)^2 P(x_i) = \sum_i x_i^2 P(x_i) - m^2 \quad \textit{(variance of } X\textit{)} \tag{15}$$

$$\sigma = \sqrt{\sigma^2} \qquad \textit{(standard deviation of } X\textit{)}$$

Remarks. The sums will be from 1 to n or from 1 to ∞ according to whether X is respectively a finite or infinite discrete random variable.

The two sums in (15) are equal, since $(x_i - m)^2 = x_i^2 - 2mx_i + m^2$, and therefore

$$\begin{aligned}\sum_1^\infty (x_i - m)^2 P(x_i) &= \sum_i x^2 P(x_i) - 2m \sum_i x_i P(x_i) + m^2 \sum_i P(x_i) \\ &= \sum_i x^2 P(x_i) - 2m \cdot m + m^2 \cdot 1 \\ &= \sum_i x^2 P(x_i) - m^2,\end{aligned}$$

using the definition (4) of expectation, and the fact that the probabilities have to sum to 1.

Why do we give the variance in (15) using two forms for the sum?

The first form of the sum shows that variance is always *non-negative* (since $(x_i - m)^2$ and $P(x_i)$ are), so that defining σ as its (positive) square root is a legal operation. The first sum also shows the significance of the variance: if the x_i that are far from the mean m occur with high probability — in other words, if the spread is large — then the variance will be large.

The second form of the sum in (15) leads to a simpler calculation, assuming that the mean m will be calculated first.

Example 3.1 Find the standard deviation of the shoe-sizes in Example 1.1B.

Solution. Using the second sum in (15), with $m = 10.3$ (see p.3, line 13), we have

$$\begin{aligned}\sigma^2 &= 7^2 \cdot P(7) + 8^2 \cdot P(8) + 9^2 \cdot P(9) + \ldots + 13^2 \cdot P(13) \ - \ (10.3)^2 \\ &= 7^2 \cdot \frac{1}{24} + 8^2 \cdot \frac{2}{24} + 9^2 \cdot \frac{4}{24} + \ldots + 13^2 \cdot \frac{2}{24} \ - \ (10.3)^2 \\ &= 2.118 \ ; \\ \sigma &= \sqrt{2.118} \approx 1.45 = 1.5, \quad \text{to one decimal place.}\end{aligned}$$

The definition of variance and standard deviation for a continuous random variable is just the natural analog of the above; again, we include the definition (12) of "mean":

Definition 3.1B *For a continuous random variable X with domain $(-\infty, \infty)$ and probability density $f(x)$, we define*

$$(12) \qquad m \ = \ \int_{-\infty}^{\infty} x\, f(x)\, dx \qquad \textit{(mean, or expectation, of } X\textit{)}$$

$$(16) \qquad \sigma^2 \ = \ \int_{-\infty}^{\infty} (x-m)^2 f(x)\, dx = \int_{-\infty}^{\infty} x^2 f(x)\, dx - m^2 \qquad \textit{(variance of } X\textit{);}$$

$$\sigma \ = \ \sqrt{\sigma^2} \qquad \textit{(standard deviation of } X\textit{)}$$

The two integrals in (16) are equal by an argument which is the natural analog of the one we gave for the sums in (15); and just as for the sums, the right- hand integral is usually easier to calculate, but it is the left-hand integral that shows the variance is non-negative, and therefore has a square root.

Theorem 3.1 For the *Poisson random variable X with parameter m*, (cf. (5)),

$$(17) \qquad \mathbf{E}(X) = m, \qquad \sigma(X) = \sqrt{m}\ .$$

For the *exponential random variable X with parameter m*, (cf. (13)),

$$(18) \qquad \mathbf{E}(X) = m, \qquad \sigma(X) = m\ .$$

Proof. The values of the two expectations were calculated in Theorems 1.4 and 2.2, respectively. The calculation of the standard deviations are left as exercises.

Exercises: Section 8C

4. NORMAL RANDOM VARIABLES

4.1 The Standard Normal Random Variable. This is the most widely-used continuous random variable — it's the one whose associated density function has the infamous "Curve" as its graph. We will use the notations Z for this random variable, z for the associated numerical variable z, $\phi(z)$ for its density function, and $\Phi(z)$ for its distribution function.

Though it has many applications, perhaps the most basic is to describe how a large number of measurements of some physical quantity are distributed about the true value.

For example, suppose we use a large number of cheap thermometers to measure the temperature (in ^oC) of a container of melting ice. The true value is 0, but some will read higher, some lower. Most will read between -1 and 1; if the quality control at the thermometer factory is just at the right level of incompetence, the histogram of the thermometer readings will be approximated by the density function below, and the collection of readings will be a typical set of outcomes for the associated random variable Z.

Definition 4.1 *The* **standard normal random variable** *Z is the continuous random variable having as its density, distribution, and associated probabilities (in z, w-coordinates)*

$$w = \phi(z) = \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2} \qquad \textit{(standard normal density)} \tag{19}$$

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2}\, dt; \qquad \textit{(standard normal distribution)} \tag{20}$$

$$P(a \le Z \le b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz = \Phi(b) - \Phi(a)\ . \tag{21}$$

The factor $1/\sqrt{2\pi}$ is used in the density function (19) so that the total area under $\phi(z)$ will equal 1. This follows from the statements in (22) below:

- the value of the integral on the left is "well-known" (it will be calculated in 18.02);
- the value of the middle integral then follows by setting $z = t\sqrt{2}$;
- the value of the integral on the right follows since the integrand is an even function.

$$\int_0^\infty e^{-t^2} dt = \frac{\sqrt{\pi}}{2} \quad \Rightarrow \quad \int_0^\infty e^{-z^2/2} dz = \sqrt{\frac{\pi}{2}} \quad \Rightarrow \quad \int_{-\infty}^\infty e^{-z^2/2} dz = \sqrt{2\pi}\ . \tag{22}$$

The last equality shows that the total area under the graph of the density function (19) is 1, as required by (8a).

Since $\Phi(z)$ is not an elementary function, it must be calculated by numerical integration; this means in practice that its values are obtained from tables, by pressing calculator buttons, or from computer programs like Matlab, Maple, or Mathematica.

The table at the top of the next page gives values of the standard normal distribution $\Phi(Z)$ to four decimal places, for $z = 0$ to $z = 3$, in increments of .1 . You will need these for the problems to calculate numerical values of probability, by using (21). Values for $z < 0$ can be obtained by symmetry, as explained after the table.

Table of values for $\Phi(Z)$, $Z \geq 0$

z:	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	
$\Phi(z)$:	.5000	.5398	.5793	.6179	.6554	.6915	.7257	.7580	.7881	.8159	
z:	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	
$\Phi(z)$:	.8413	.8643	.8849	.9032	.9192	.9332	.9452	.9554	.9641	.9713	
z:	2.0	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
$\Phi(z)$:	.9772	.9821	.9861	.9893	.9918	.9938	.9953	.9965	.9974	.9981	.9987

To get values of $\Phi(z)$ for negative z, we use the symmetry of $\phi(z)$ about $z = 0$. Still assuming $z \geq 0$, we have

$$\Phi(-z) \;=\; \int_{-\infty}^{-z} \phi(t)\,dt \;=\; \int_{z}^{\infty} \phi(t)\,dt \;=\; 1 - \Phi(z) \tag{23}$$

The most frequent application is to intervals symmetric around $z = 0$; using (21) and (23),

$$P(-a < Z < a) \;=\; \Phi(a) - \Phi(-a) \;=\; 2\Phi(a) - 1\ . \tag{24}$$

Using the table for $z = 1.0$ and $z = 2.0$, this gives the approximate values most widely used in statistics (cf. Simmons, bottom p. 419 for a picture showing these values visually):

$$P(-1 < Z < 1) \;=\; .68 \;\approx\; 2/3; \qquad\qquad P(-2 < Z < 2) \;=\; .95\ . \tag{25}$$

The intervals in (25) are chosen because of their relation to the mean and standard deviation of Z:

Theorem 4.1 *The mean and standard deviation of the standard normal random variable Z are respectively*

$$\mathbf{E}(Z) \;=\; 0; \qquad\qquad \sigma(Z) \;=\; 1\ . \tag{26}$$

Proof. The integral (12) giving $\mathbf{E}(Z)$ has the integrand $\frac{1}{\sqrt{2\pi}} z\, e^{-z^2/2}$ which is an odd function; therefore its integral over the symmetric interval $(-\infty, \infty)$ is zero.

The integral for the variance can be calculated using integration by parts; the result is 1, so $\sigma = \sqrt{1} = 1$. We leave the details as an exercise.

Thus the values in (25) give the respective probabilities that Z is within one or two standard deviations from its mean value.

4.2 The Normal Random Variable.

To be adaptable to any situation, we need to modify the standard normal density function $w = \phi(z)$ given by (19) in two ways.

First Modification. Suppose our factory starts producing less accurate thermometers. The measurements of the melting ice should then still be clustered around 0, but spread out more, though the density function (19) should still have the same general properties and shape as before. To achieve this, we stretch the z-axis by a positive scale factor which we will call σ, since it will turn out to be the standard deviation of the new density function. (In this example $\sigma > 1$, though in general it need not be). Using variables, the way to do this stretching is to make a linear change of variable from z to x_1, say:

$$z = \frac{x_1}{\sigma} \qquad \text{(so that } x_1 = \sigma \text{ corresponds to } z = 1) \ .$$

At the same time, shrink the w-axis by the factor σ (the reason will be given in a moment) by making the change of variable $w = \sigma\, y$ (so that $w = \sigma$ corresponds to $y = 1$). Under these two operations, the standard normal density function $\phi(z)$ then becomes in x_1, y-coordinates,

$$y = \frac{1}{\sigma}\,\phi(x_1/\sigma), \qquad i.e., \qquad \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(x_1/\sigma)^2} \ . \tag{27}$$

Remark. We shrink the w-axis by $1/\sigma$ so the total area under (19) will still be 1.

This can be proved formally by changing the variable in the integral for the area, but to see it intuitively, imagine the area under $\phi(z)$ divided into equal tiny squares of side ϵ say, and area ϵ^2. Stretching the z-axis and shrinking the w-axis by the same factor σ turns each ϵ-square into a rectangle of horizontal side $\sigma\epsilon$ and vertical side ϵ/σ, whose area is still ϵ^2.

Since the stretching and shrinking changes each little square into a rectangle, but does not change its area or the total number of little squares or rectangles, it does not change the total area under $\phi(z)$: it started out 1 and remains 1 after the z-stretching and w-shrinking.

Second Modification. If we use these same thermometers to measure some other temperature m, the graph of the density function should be moved so that its maximum lies over the point m; to achieve this, we make the further change of variable $x_1 = x - m$ (so that $x_1 = 0$ corresponds to $x = m$), giving as our final result:

Definition-Theorem 4.2A *Under the change of variables* $z = \dfrac{x-m}{\sigma}, \quad w = \sigma\, y$ *, the standard normal density function (19) becomes, in* x, y*-coordinates,*

$$y = \phi_{m,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-m)^2/\sigma^2}, \qquad \textbf{the normal density function} \tag{28}$$

and its associated normal distribution function is

$$\Phi_{m,\sigma}(x) \ = \ \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}(t-m)^2/\sigma^2} dt \qquad \textbf{the normal distribution} \tag{29}$$

The normal density function (28) (as well as its associated random variable $X_{m,\sigma}$ and normal distribution function (29)) thus depends on two parameters m and σ, which turn out to be the mean and standard deviation of $X_{m,\sigma}$.

Theorem 4.2B For the normal density function (19) $\phi_{m,\sigma}(x) = \dfrac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-m)^2/\sigma^2}$,

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} x\, e^{-\frac{1}{2}(x-m)^2/\sigma^2} dx = \ m \ ;$$
$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-m)^2\, e^{-\frac{1}{2}(x-m)^2/\sigma^2} dx \ = \ \sigma^2 \ .$$

Proofs. Use the change of variables in Theorem 4.2A to simplify both integrals; you'll need (20) for the first, and integration by parts for the second. It's a good exercise in changing variables and integration techniques; if you get stuck, it's worked out in Simmons, p.418 (using t instead of z, and a slightly different change of variable, but the calculations are basically the same).

The Simmons Calculus has in 12.5, p.419 several graphs illustrating how the normal density curve looks for several different values of the standard deviation σ and the mean m.

Probability calculations for $X = X_{m,\sigma}$ and the normal distribution $\Phi_{m,\sigma}$ are done using the table on p. 11 for the standard $\Phi(Z)$ and the change of variable in Theorem 4.2A:

$$z = \frac{x-m}{\sigma}; \qquad\qquad x = \sigma z + m \ .$$

This gives the basic formula relating the normal and standard distributions:

$$P(a < X < b) \ = \ P\left(\frac{a-m}{\sigma} < Z < \frac{b-m}{\sigma}\right) , \tag{30}$$

or using the change of variable in the other direction, (25) turns into the rule-of-thumb formulas most frequently used in statistical applications:

$$P(m-\sigma < X < m+\sigma) \ = \ .68 \ \approx \ 2/3 \tag{31}$$

$$P(m-2\sigma < X < m+2\sigma) \ = \ .95 \ . \tag{32}$$

Example 4.2A If the height in inches of a typical male Bostonian is a normal random variable H having a mean of 68 and a standard deviation of 4, what percentage will be able to walk through a 6'4" doorframe without having to stoop?

Solution. The doorframe is 76 inches. Since $76 = 68 + 2\cdot 4$, the door height is two standard deviations from the mean, so that (32) tells us

$$P(H > 76) \ \approx \ \frac{1}{2}(.05) = .025;$$

thus about $2\frac{1}{2}\%$ will have to stoop, so that about $97\frac{1}{2}\%$ will not have to.

Example 4.2B In the end-of-term evaluation forms, a calculus lecturer had an average rating of 4.5, with $\sigma = 1.5$. What fraction of the students gave the rating:

a) 4 or 5? *b*) between 3 and 6?

Solution. For (a), we use (30); letting $X = X_{4.5,1.5}$, we have

$$\begin{aligned} P(4 < X < 5) &= P\left(-\frac{.5}{1.5} < Z < \frac{.5}{1.5}\right) \\ &= 2\Phi(1/3) - 1 \;=\; .26 \;\approx\; 1/4, \quad \text{by (24) and the table for } \Phi(Z). \end{aligned}$$

For (b), we can use (31), since 3 and 6 are one standard deviation away from 4.5:

$$P(3 < X < 6) \;=\; .68 \;\approx\; 2/3\ .$$

Exercises: Section 8D

5. CENTRAL LIMIT THEOREM AND APPLICATIONS

5.1 Central limit theorem. This theorem is one of the main applications of the normal distribution. It's the basis for the polling and sampling techniques which try to predict how a population thinks, looks, and behaves from questioning just a few people.

Suppose X is a random variable — discrete or continuous — giving the possible outcomes of some experiment, with the associated probabilities for each outcome (or the probability density, if it's a continuous random variable). It has a mean m and standard deviation σ.

Common sense, or what is often called the "law of averages", says that though the outcome of any particular trial of the experiment may not come out close to m, if you repeat the experiment many times and take the average of the results, that ought to be close to the true mean. In fact, if you didn't know in advance what the mean was, that would be a good way of determining it: the more trials you average in, the closer your answer should be to the mean.

The central limit theorem makes this idea more precise. In it, one can think of X as the random variable describing the numerical outcome of an experiment, and $X_1, \ldots, X_n$ as the identical random variables describing respectively the outcomes from the first trial of the experiment, the second, and so on down to the n-th trial. Then $\bar{X}$ is the new random variable whose outcome is the average of the results of these n trials.

Theorem 5.1 The Central Limit Theorem. *Let X be a discrete or continuous random variable, with mean m and standard deviation σ, and let $X_1, X_2, \ldots, X_n$ be n copies of X. Then if n is large,*

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \tag{33}$$

is approximately a normal random variable, whose mean and standard deviation are

$$E(\bar{X}) = m, \qquad \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}\ . \tag{34}$$

The definition of $\bar{X}$ in the theorem is incomplete, since we haven't said how to assign probabilities or a probability density to the values of $\bar{X}$, based on those of X. However, it is clear from the experimental trial viewpoint that these probabilities or density function do exist — we would need to know them if we were trying to prove the Central Limit Theorem, but they aren't necessary if we just want to use it.

The theorem has two points:

• even though X may not be a normal random variable, $\bar{X}$ will be, if n is large enough, at least approximately, so that we can apply the results of section 4 to it;

• when n is large, the standard deviation of $\bar{X}$ is small, which says that the graph of the density function for $\bar{X}$ has most of its area in a sharp peak around the mean; so the average of n trials will with high probability be close to the true mean.

Example 5.1A *"Press 0 to get a representative; this call may be monitored for quality"...*

The average time a Sleep-Tight Motel representative spends on the telephone taking a reservation is 5 minutes, with a standard deviation of 2 minutes.

One outsourced representative, Dora Chatemup, was monitored for 100 calls, and spent an average of 5.5 minutes per call. What are the chances that she was following the guidelines in the Motel's How-to-Take-Reservations Manual, and just happened to get an unusually large number of hard-to-satisfy callers?

Solution. Let X be the random variable giving the length of a call; according to the data given, $m(X) = 5$, and $\sigma(X) = 2$.

Let $\bar{X}$ be the average of 100 calls; then $m(\bar{X}) = 5$, and $\sigma(\bar{X}) = 2/\sqrt{100} = .2$, by the Central Limit Theorem. Using (30) and the table for $\Phi(Z)$,

$$P(\bar{X} \geq 5.5) \;=\; P\left(Z \geq \frac{5.5-5}{.2}\right) \;=\; P(Z \geq 2.5) = 1 - \Phi(2.5) \;=\; .0062,$$

so the probability is only around $\frac{1}{2}\%$ that her average time of 5.5 minutes was due just to the luck of the draw (i.e., chance).

Example 5.1B We want to test whether a psychic has Extra-Sensory Perception by seeing if he can guess whether the next number emitted by a random number generator will be even or odd. How many numbers would he have to guess at, for a 51% correct guess rate to have a 1% probability of being due to chance alone?

Solution. Let X be the random variable which has the value 1 if the guess is correct, 0 otherwise. Assuming ESP doesn't exist, the associated probabilities are $P(1) = P(0) = .5$, the mean and standard deviation are (using (15)):

$$m(X) = .5, \qquad \sigma^2(X) = 1^2(.5) - (.5)^2 = .25, \qquad \sigma(X) = .5\ .$$

If n guesses are made and averaged to get the random variable $\bar{X}$, the Central Limit Theorem says that if n is large,

$$m(\bar{X}) = .5\ ; \qquad \sigma(\bar{X}) = \frac{.5}{\sqrt{n}}$$

Using (30),

$$P(\bar{X} \geq .51) = P\left(Z \geq \frac{.01\sqrt{n}}{.5}\right) = P(Z \geq .02\sqrt{n}) = 1 - \Phi(.02\sqrt{n})$$

When is this probability $\leq 1\%$? From the table, calculating roughly,

$$1 - \Phi(.02\sqrt{n}) = .01 \quad \Rightarrow \quad .02\sqrt{n} > 2.4 \quad \Rightarrow \quad n > (120)^2 \;=\; 14,400.$$

So a score of 51% correct out of say 15,000 guesses would have only a 1% probability of being due to chance alone – or as one says, would confirm the existence of ESP at the 99% confidence level (unless the psychic cheated, something professional magicians have been better at detecting than professional scientists.)

5.2 Polling a random sample An application of statistics that everyone is familiar with is *polling* — selecting a random sample of the population, asking who they favor in an election, and from this concluding how the population as a whole will vote. How does one determine how large the sample should be, and report the confidence level of the poll — what the margin of error is, and perhaps what the chances of error are?

Assume there are two candidates — Hack and Shifty. Let p denote the unknown fraction of the voting population that favors Hack; then $1-p$ will be the fraction that favors Shifty. Our aim is to find upper and lower limits on the value of p, with say 95% probability that our limits are correct. How many voters will we have to poll to achieve this?

Consider the "polling" random variable defined by $X = \begin{cases} 1, & \text{Hack favored, } P(1) = p; \\ 0, & \text{Shifty favored, } P(0) = 1-p \end{cases}$. Using (15), we calculate its mean, variance, and standard deviation to be

$$m(X) = p, \qquad \sigma^2(X) = p - p^2, \qquad \sigma = \sqrt{p-p^2} \tag{35}$$

If we think of $X_1, X_2, \dots, X_n$, as the random variables representing the outcomes from asking the first voter, the second, and so on up to the n-th in the sample, the average of these $\bar{X}$ represents the fraction of this sample of size n that favors Hack. By the Central Limit Theorem

$$m(\bar{X}) = p, \qquad \bar{\sigma} = \sigma(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

Therefore using (32), we get

$$P\big(m(\bar{X}) - 2\bar{\sigma} < \bar{X} < m(\bar{X}) + 2\bar{\sigma}\big) = .95\ , \tag{36a}$$

or rewriting this to make it more compact,

$$P\big(|\bar{X} - p| < 2\bar{\sigma}\big) = .95\ . \tag{36b}$$

Or in the informal way the result is usually said (remember that p is what we want to know, but the value of $\bar{X}$ is the experimental outcome of polling the sample of size n)

$$p = \bar{X} \pm 2\bar{\sigma} \quad \text{with 95\% confidence}\ . \tag{36c}$$

We don't know how big $\bar{\sigma} = \sigma/\sqrt{n}$ is, since it depends on $\sigma = \sqrt{p-p^2}$, and the value of p is unknown. However, we can estimate it: by elementary calculus, $p - p^2$ has its maximum point at $(\frac{1}{2}, \frac{1}{4})$; using this, and then the Central Limit Theorem,

$$p - p^2 \le \frac{1}{4} \quad \Rightarrow \quad \sigma = \sqrt{p-p^2} \le \frac{1}{2} \quad \Rightarrow \quad \bar{\sigma} \le \frac{1}{2\sqrt{n}}\ . \tag{37}$$

Since the probability in (36) — the confidence level in (36c) — can only rise if we replace $2\bar{\sigma}$ by a larger number, (36c) and (37) imply our final answer:

$$p = \bar{X} \pm \frac{1}{\sqrt{n}} \quad \text{with 95\% confidence}\ . \tag{38}$$

Example 5.1A 100 entering first-graders picked at random from the seven Cambridge elementary schools were asked if they believed in the tooth fairy. 45 said "yes", 55 said "no". With 95% confidence, what fraction p of Cambridge first-graders are believers?

Solution. Using (38) with $n = 100$, we get $p = .45 \pm .10$; in other words,

$$.35 < p < .55 \qquad \text{with 95\% confidence,}$$

or as the local news would report it, "at the 95% confidence level, 45% of the first-graders are believers, with a 10% margin for error." (They'd probably leave out the 95% confidence clause.)

Example 5.1B How large a sample would be needed to use the results of a poll to predict with 95% confidence the percentage of voters favoring Hack, with a 1% margin of error?

Solution. Again using (38), we want $\sqrt{n} = 100$, so $n = 10,000$.

Remark. The real difficulty in this work is in obtaining a sample that is truly random, and not biased by neighborhood, social class, race, religion, etc. There is an equal problem in getting reliable answers, which even if honestly given can depend on how the question is phrased, the tone of voice, and even the expression on the pollster's face, if it's insufficiently poker.

Exercises: Section 8E