Notes on Probability

Allin Cottrell*

1 Probability: the classical approach

The classical approach to probability, which has roots in the study of games of chance, is based on the assumption that we can identify a class of *equiprobable individual outcomes* of a random experiment. For example, the outcomes heads and tails are equiprobable when a fair coin is tossed; the outcomes 1, 2, 3, 4, 5, and 6 are all equiprobable when a fair die is rolled.

Under these conditions the probability of any event, A, is the number of outcomes that correspond to A, which we'll write as n_A, divided by the total possible outcomes, n. In other words it's the proportion of the total outcomes for which A occurs.

$$0 \leq P(A) = \frac{n_A}{n} \leq 1 \tag{1}$$

For example, let A be the event of getting an even number when rolling a fair die. There are three outcomes corresponding to this event, namely 2, 4 and 6, out of a total of six possible outcomes, so the probability $P(A) = \frac{3}{6} = \frac{1}{2}$.

2 Complementary probabilities

This is simple but important. If the probability of some event A is $P(A)$ then the probability that event A does *not* occur, written $P(\neg A)$, must be

$$P(\neg A) = 1 - P(A).$$

For example, if the chance of rain for tomorrow is 80 percent, the chance that it doesn't rain tomorrow must be 20 percent. When trying to compute a given probability, it is sometimes *much* easier to compute the complementary probability first, then subtract from 1 to get the desired answer.

This principle can be justified on the classical approach as follows. Let $n_{\neg A}$ denote the number of outcomes that do not correspond to event A. Since every outcome either corresponds to event A or does not, we have $n = n_A + n_{\neg A}$, or $n_{\neg A} = n - n_A$. But then, from first principles,

$$P(\neg A) = \frac{n_{\neg A}}{n} = \frac{n - n_A}{n} = \frac{n}{n} - \frac{n_A}{n} = 1 - P(A)$$

3 Addition rule

The addition rule provides a means of calculating the probability of $A \cup B$ (read "A or B"), that is, the probability that either of two events occurs.

With equiprobable individual outcomes, we have $P(A) = \frac{n_A}{n}$ and $P(B) = \frac{n_B}{n}$. As a first approximation, to find $P(A \cup B)$ we need to add together the number of outcomes corresponding to event A and the number corresponding to B, then divide by n. But if the two events intersect, $n_A + n_B$ will overstate the number of outcomes corresponding to $A \cup B$: specifically, the outcomes contained in the intersection of the events will

---

be double-counted (see Figure 1). Therefore we must subtract (once) the number of outcomes contained in the intersection, which we'll write as n_{AB}. Thus

$$P(A \cup B) = \frac{n_A + n_B - n_{AB}}{n} = \frac{n_A}{n} + \frac{n_B}{n} - \frac{n_{AB}}{n} = P(A) + P(B) - \frac{n_{AB}}{n} \tag{2}$$
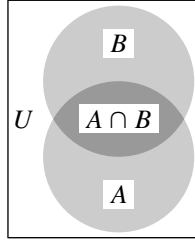

Figure 1: Illustrating the addition rule

Consider the last term on the right in equation (2) above, $\frac{n_{AB}}{n}$. This represents the fraction of the total outcomes that correspond to the intersection of A and B. In other words, it's the probability that A and B both occur, $P(A \cap B)$. Thus the equation above can be put into its final form as:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \tag{3}$$

Example: Find the probability of drawing a spade (A) or a king (B) from a deck of cards. There are 52 cards, 13 spades, 4 kings, and one king of spades ($A \cap B$), so

$$P(A \cup B) = \frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52}$$

4   Multiplication rule

It is obvious that

$$\frac{n_{AB}}{n} \equiv \frac{n_A}{n} \times \frac{n_{AB}}{n_A} \tag{4}$$

(the right-hand side is obtained by multiplying the left-hand side by $n_A/n_A = 1$).

Let's see what identity (4) is saying. On the left-hand side we have $n_{AB}/n = P(A \cap B)$, or the probability that A and B both occur. On the right-hand side we have, first, $n_A/n = P(A)$, which represents the "marginal" (or unconditional) probability of A. The second term is n_{AB}/n_A: this represents the number of outcomes corresponding to ($A \cap B$) over the number of outcomes corresponding to A. Think about this expression. It is nothing other than the *conditional probability*, $P(B|A)$. This can be read as "the probability of B given A". We are taking A as "given" by putting only the outcomes in A into the denominator. In the numerator are the outcomes in A that are also in B. The ratio is then the proportion of the outcomes in A that are also in B, or (assuming equiprobable individual outcomes), $P(B|A)$. Thus equation (4) can be re-written as

$$P(A \cap B) = P(A) \times P(B|A) \tag{5}$$

which is the general form of the multiplication rule for joint probabilities. In the special case where the events A and B are *independent*, the conditional probability $P(B|A)$ equals the marginal probability $P(B)$ and the rule simplifies to

$$P(A \cap B) = P(A) \times P(B)$$

2

5 Conditional and marginal probabilities: further note

Consider the following pseudo-facts: the chance of snow on Friday is 10 percent; if it snows, there's only a 5 percent chance that all students are present in ECN 215; if it doesn't snow there's an 80 percent chance that all students show up. These points are tabulated below:

	conditional on:	
	snow ($P = 0.1$)	no snow ($P = 0.9$)
P(all present)	0.05	0.8

Suppose we want to know the "overall" probabiility that all students are present. Apply the multiplication rule down the two columns: first, the probability of snow AND all present is $0.05 \times 0.1 = 0.005$; second, the probability of no snow AND all present is $0.8 \times 0.9 = 0.72$. Then apply the addition rule: the probability of (all present AND snow) OR (all present AND no snow) is $0.005 + 0.72 = 0.725$. This "overall" probability is what we call the *marginal* probability, and perhaps you can see how it comes by this name: it's the number you get, on the edge or margin of the sort of table shown above, if you multiply out the conditional probabilities times the probabilities of the events on which we're conditioning, then add up across the events on which we're conditioning.

In general,

$$P(A) = \sum_{i=1}^{N} P(A|E_i) \times P(E_i)$$

where $E_1, \ldots, E_N$ represent N mutually exclusive and jointly exhaustive events (one and only one of them will occur, like "snow" and "no snow" above).

The following point concerning conditional probabilities is particularly important. In general,

$$P(A|B) \neq P(B|A)$$

That is, the probability of *A* given *B* is generally *not* the same as the probability of *B* given *A*. Back to cards for an illustration: let *A* denote drawing a king and *B* denote drawing a face card, on a single draw from a well shuffled deck. It's clear that $P(A|B) = \frac{1}{4}$ while $P(B|A) = 1$.

6 Generalizing from the classical model

We have introduced probability in terms of the classical approach, which involves counting and manipulating the number of equiprobable outcomes corresponding to events of interest. Although it is *easiest* to justify the addition and multiplication rules in classical terms, these rules generalize: they apply to any probabilities, however derived (e.g. by observation of relative frequencies over time, or by "expert judgment", rather than by counting outcomes). They are in the nature of *consistency conditions*. Thus if I believe (on whatever grounds) that $P(A) = .60$ and $P(B) = .30$, and that events *A* and *B* are independent, then in consistency I am bound to believe that $P(A \cap B) = .60 \times .30 = .18$. If I am inconsistent (do not follow the above rules) there is a particular economic symptom. Provided I'm willing to make bets based on my probability judgments, it will be possible to set up a series of bets that I'm *sure* to lose, on average over time. This is known as the "Dutch book argument"; it was developed by Frank Ramsey.

7 Introducing probability distributions: discrete random variables

Up till now we've spoken only of the probability of *events* of one kind or another. In econometrics we're generally more concerned with the probability *distributions* of variables of interest. To approach this topic we'll first make a distinction between *discrete* and *continuous* random variables. A discrete random variable is one that can take on a finite set of distinct values depending on the outcome of some random experiment. A simple example would be the number that appears uppermost when a die is rolled. Such a variable is generally the outcome of a *counting*

x_i	$P(X = x_i)$	$x_i P(X = x_i)$
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$
6	$\frac{1}{6}$	$\frac{6}{6}$
$\sum$	$\frac{6}{6} = 1$	$\frac{21}{6} = 3.5 = E(X)$

Table 1: Probability distribution for rolling a fair die

operation. A continuous random variable can take on any real value (within a specified range, perhaps), for example the weight of a randomly selected individual. Here the variable is generally the outcome of some sort of measurement, rather than counting.

We start with the simpler case of discrete random variables. Here, the probability distribution for some random variable, X, is a mapping from the possible values of X to the probability that X takes on each of those values. The mapping may be represented by a mathematical function or a table.

Consider the example shown in Table 1. The first column shows the possible values of X (= the number appearing uppermost when a die is rolled), and the second shows the probability for each value. If the die is fair, the probability is the same for all x_i. This is known as the *uniform* or *rectangular* distribution. It is graphed in Figure 2. One key feature of the table is that the sum of the entries in the probabilities column equals 1.0: this is necessarily true of any discrete probability distribution:

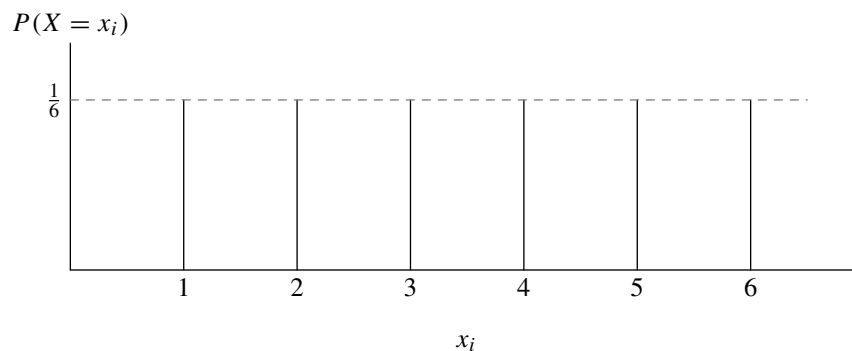$$\sum_{i=1}^{N} P(X = x_i) = 1.0$$

Figure 2: Graph of probability distribution for one die

The third column in Table 1 shows the product, value of variable times probability that the value occurs. The summation of this column gives the *expected value* or mean of the random variable. You should be familiar, I hope, with three *measures of central tendency*: mean, median and mode. The median is the middle value when the data are sorted by size; the mode is the value that occurs with greatest frequency or probability; and the mean is as just

described:

$$E(X) \equiv \mu_X = \sum_{i=1}^{N} x_i P(X = x_i) \tag{6}$$

Note that in this example the expected value is *not* the value that occurs with greatest probability: the probability of getting 3.5 on any single die-roll is zero. Rather, the expected value is the expected *average* if the random experiment is repeated many times.

Besides the expected value, another key feature of any probability distribution is the *degree of dispersion* of the distribution around its mean. This is usually measured by the *variance* of the distribution (or the square root of the variance, which is called the *standard deviation*). The mean can be described as the probability-weighted sum of the possible values of the random variable; in the same terms, the variance is the probability-weighted sum of the squared deviations of the possible values of the random variable from its mean, or

$$\text{Var}(X) \equiv \sigma_X^2 = \sum_{i=1}^{N} [x_i - E(X)]^2 P(X = x_i) \tag{7}$$

Table 2 shows the calculation of the variance for the die-rolling example. $\text{Var}(X) = 2.917$ so the standard deviation is $\sqrt{2.917} = 1.708$.

Another way of expressing the variance is to say it's the expected value of the squared deviation from the mean—since "expected value" just means, probability-weighted sum (for a discrete random variable):

$$\text{Var}(X) = E[X - E(X)]^2 \tag{8}$$

Equation (8) can be manipulated as follows:

$$\begin{aligned} \text{Var}(X) &= E\left[X^2 - 2XE(X) + E(X)^2\right] \\ &= E(X^2) - 2E[XE(X)] + [E(X)]^2 \\ &= E(X^2) - 2[E(X)]^2 + [E(X)]^2 \end{aligned}$$

So that, finally:

$$\text{Var}(X) = E(X^2) - [E(X)]^2 \tag{9}$$

Or in words again, the variance of X is the *expectation of the square* of X minus the *square of the expectation* of X. This underlines the fact that, in general,

$$E(X^2) \neq [E(X)]^2$$

(these two terms are equal only if and only if the distribution has a variance of zero, or in other words is *degenerate*). In the die-rolling example, we know that $[E(X)]^2 = 3.5^2 = 12.25$. $E(X^2)$, on the other hand, $= \frac{1}{6} \times 1^2 + \frac{1}{6} \times 2^2 + \cdots + \frac{1}{6} \times 6^2 = 15.167$. The variance then equals $15.167 - 12.25 = 2.917$, as calculated in Table 2.

Two dice

Let's try a slightly more interesting example than the single die roll. Let X represent the average of the two numbers appearing uppermost when two fair dice are rolled. The exercise is to determine the probability distribution of X, and to use this to find the mean and variance of X.

To start, we can set out the *sample space*, or in other words the full set of possible outcomes for the pair of dice.

x_i	$P(X = x_i)$	$x_i - E(X)$	$[x_i - E(X)]^2$	$[x_i - E(X)]^2 P(X = x_i)$
1	$\frac{1}{6}$	-2.5	6.25	1.0417
2	$\frac{1}{6}$	-1.5	2.25	0.3750
3	$\frac{1}{6}$	-0.5	0.25	0.0833
4	$\frac{1}{6}$	$+0.5$	0.25	0.0833
5	$\frac{1}{6}$	$+1.5$	2.25	0.3750
6	$\frac{1}{6}$	$+2.5$	6.25	1.0417
$\sum$	1	0		$2.917 = \text{Var}(X)$

Table 2: Variance calculation: rolling a fair die

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

We can then set out the averages (X values) corresponding to these outcomes

1.0	1.5	2.0	2.5	3.0	3.5
1.5	2.0	2.5	3.0	3.5	4.0
2.0	2.5	3.0	3.5	4.0	4.5
2.5	3.0	3.5	4.0	4.5	5.0
3.0	3.5	4.0	4.5	5.0	5.5
3.5	4.0	4.5	5.0	5.5	6.0

Now over to you. *Construct a table similar to the combination of Tables 1 and 2 above, which enables you to show the distribution and calculate the expected value and variance.* Using a spreadsheet program is probably a good idea. (Note that the distribution in this case is not uniform: for instance $X = 1.0$ occurs with probability $\frac{1}{36}$ while $X = 1.5$ occurs with probability $\frac{2}{36}$.)

8 Measures of association: covariance and correlation

The concept of variance, for a single variable, generalizes to provide the concept of *covariance* for two variables. The covariance of X and Y is the expected value of the cross-product, deviation of X from its mean times deviation of Y from its mean.

$$\text{Cov}(X, Y) = \sigma_{XY} = E\{[X - E(X)][Y - E(Y)]\} \tag{10}$$

Given N observations on X and Y, the covariance may be calculated as

$$\text{Cov}(X, Y) = \frac{1}{N} \sum_{i=1}^{N} [x_i - E(X)][y_j - E(Y)]$$

Covariance can take on any value, positive, negative or zero. It provides a measure of the linear association between X and Y. The logic can be seen if Y is graphed against X with the axes set to intersect at the point $(E(X), E(Y))$ as in Figure 3. The points display an upward-sloping linear association. This will be expressed in
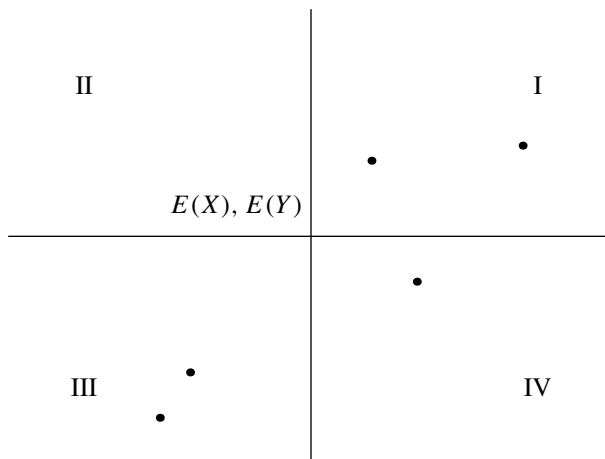
6

Figure 3: Positive covariance of X and Y

a positive covariance: most of the points lie in quadrants I and III, where the deviations of X and Y from their respective means are of the same sign, and hence the cross products are positive. If the points were scattered evenly in all four quadrants then positive and negative cross-products would tend to cancel and the covariance would be close to zero.

The correlation coefficient for two variables X and Y, written ρ_{XY}, is a scaled version of covariance: we divide through by the product of the standard deviations of the two variables.

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \tag{11}$$

The resulting measure lies between -1 (indicating a perfect downward-sloping linear association) and $+1$ (perfect upward-sloping linear relationship).

9 Distribution concepts for continuous random variables

For a simple example of a continuous random variable, consider the clock face plus spinner shown below. Let the random variable $X =$ the number towards which the spinner points when it comes to rest.
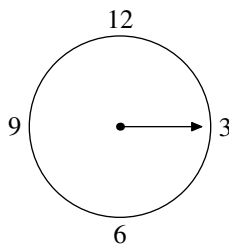
What is the probability that $X = 3.00\dot{0}$? Applying the classical rule, this is one out of an infinite number of equiprobable outcomes, so the probability is zero. And this goes for any outcome that is specified in full precision. Unlike the cases examined earlier, we can't draw up a useful mapping from specific values of the random variable to the probability that those values occur. The "counting of outcomes" approach is not going to work.

Try another angle. The spinner must end up pointing *somewhere* in the range 0 to 12. So we can map from the full circumference of the clock face to a probability measure of 1.0. While we can't usefully count individual

outcomes, we can think in terms of *fractions* of that total measure. For instance, if the spinner is "fair" there ought to be a probability of $\frac{1}{4}$ that it ends up pointing into the range 0 to 3, a probability of $\frac{1}{6}$ that it points into the range 7 to 9, and so on. We can work this up into the idea of a *cumulative distribution function* or cdf, which can be written as

$$F(x) = P(X < x) \tag{12}$$

For the spinner example, $F(x) = 0$ for all $x < 0$. It equals 0.25 for $x = 3$, 0.50 for $x = 6$ and so on, yielding the graph shown in Figure 4. A cumulative probability of 1 is reached where $x = 12$.
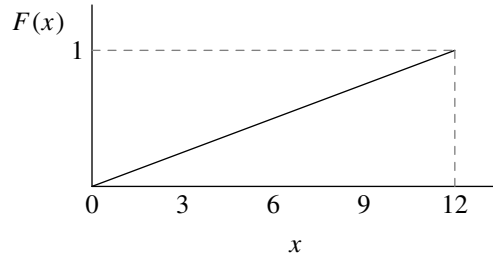

Figure 4: cdf for spinner

Another essential concept when dealing with continuous random variables is the *probability density function* or pdf. This is defined as the derivative of the cdf with respect to x, and is usually written as $f(x)$:

$$f(x) = \frac{d}{dx} F(x) \tag{13}$$

For the spinner example the cdf is a straight line, as we saw above, so its derivative (slope) is a constant. By inspection of Figure 4 we can see the constant value is $\frac{1}{12}$. Thus the pdf is as shown in Figure 5.
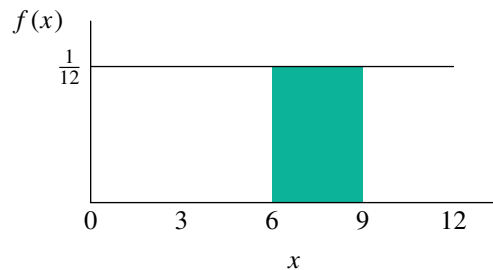

Figure 5: pdf for spinner

The pdf has a height of $\frac{1}{12}$ at $x = 6$. This does *not* mean that the probability of $X = 6$ is $\frac{1}{12}$ (as we know, it's zero). Rather, we use the pdf in this way: we can determine the probability of X falling into any given range by taking the integral of the pdf over that interval (i.e. finding the area under the curve between the specified values).

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\, dx \tag{14}$$

This is illustrated in Figure 5: the probability that X falls into the range 6 to 9 is the area under the pdf between 6 and 9, namely $\frac{3}{12}$.

10   Central limit theorem and Gaussian distribution

The graph of the spinner's pdf shows that the probability distribution is uniform or rectangular, a continuous counterpart to the single die-roll example discussed above. Such distributions are rarely if ever found in nature. The die and spinner are both examples of human contrivances, devices which constrain random processes to operate in a particularly simple and "orderly" manner. Two aspects of this are noteworthy. First, each device allows only a single random process to influence the outcome: the toppling of a near-perfect cube or the rotary motion of the spinner on its bearing. Second, the outcomes are constrained to a specific range, via the integer dot patterns on the die or the (arbitrary) 0 to 12 range of the clock face. Consider by contrast (for example) the heights or weights of a population of animals or people. There will be numerous "random" influences on the height of any individual, stemming from both genetic endowment and environment, and there is no fixed range. We can be pretty sure we'll never see an adult human less than 1 foot tall or greater than 12 feet tall, but there's no fixed limit, no guarantee that somebody won't come along tomorrow and beat the current "tallest person" entry in the Guinness Book of World Records.

A proof—even a precise mathematical statement—of the Central Limit Theorem is beyond the scope of this class, but the general idea is roughly as follows. If a random variable X represents the summation of numerous independent random factors then, *regardless of the specific distribution of the individual factors*, X will tend to follow the normal or Gaussian distribution, the familiar symmetrical "bell curve" of statistics (Figure 6).
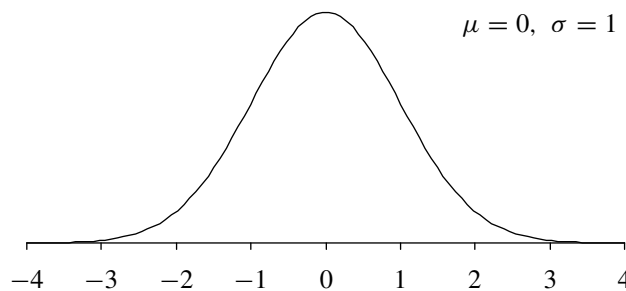

Figure 6: Normal or Gaussian pdf

The general formula for the normal pdf is:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad -\infty < x < \infty \tag{15}$$

where μ denotes the mean of the distribution and σ its standard deviation. The "standard normal" distribution is obtained by setting $\mu = 0$ and $\sigma = 1$; its pdf is then

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad -\infty < x < \infty \tag{16}$$

The probability of x falling into any given range can be found by integrating the above pdf from the lower to the upper limit of the range. A couple of results to commit to memory are $P(\mu - 2\sigma < x < \mu + 2\sigma) \approx 0.95$ and $P(\mu - 3\sigma < x < \mu + 3\sigma) \approx 0.997$. Other values can be looked up in a normal distribution table if they are needed. The Gaussian pdf does not rule out extreme values, but it assigns them a very low probability: as you can see from the diagram there is little chance of a normal random variable being found more than three standard deviations from the mean of the distribution.

A compact notation for saying that x is distributed normally with mean μ and variance σ^2 is $x \sim N(\mu, \sigma^2)$.

11   Practice problems

Problems 1 to 4 are taken from Jeffrey Wooldridge.

1. Let X be a random variable distributed as $N(5, 4)$. Find the following probabilities:

   (a) $P(X \leq 6)$

   (b) $P(X > 4)$

   (c) $P(|X - 5| > 1)$

   If you don't have a standard normal table to hand, see `http://en.wikipedia.org/wiki/Standard_normal_table`.

2. Much is made of the fact that certain mutual funds outperform the market year after year. Consider a 10-year period and let the population be the 4,170 mutual funds reported in *The Wall Street Journal* on January 1, 1995. By saying that performance relative to the market is random, we mean that each fund has a 50:50 chance of beating the market in any given year, and performance is independent from year to year. (That is, whether a given fund beats the market in a given year is like a coin-flip.) In answering the following, suppose that performance relative to the market is indeed random.

   (a) What is the probability that any particular fund beats the market in all 10 years?

   (b) Find the probability that *at least one* out of 4,170 funds beats the market in all 10 years. What do you make of your answer?

3. For a randomly selected county in the US, let X denote the proportion of adults over the age of 65 who are employed, the elderly employment rate ($0 \leq X \leq 1$). Let the cdf of X be $F(x) = 3x^2 - 2x^3$ for $0 \leq x \leq 1$. Find the probability that the elderly employment rate is at least 0.6 or 60 percent.

4. Just prior to jury selection for O. J. Simpson's murder trial in 1995, a poll found that 20 percent of the adult population believed Simpson was innocent. Suppose for the sake of argument that this figure accurately gauged the proportion of US adults believing in Simpson's innocence. And assume that the 12 jurors were selected randomly from the adult population (although this turned out not to be the case).

   (a) Find the probability that the jury had at least one member who believed in Simpson's innocence prior to selection.

   (b) Find the probability that the jury had at least two members who believed in Simpson's innocence. (Hint: "at least two" means not 0 and not 1).

5. The police department of a city studies the safety of cyclists at night. They find that 60 percent of cyclists involved in accidents at night are wearing light-colored clothing.

   (a) Should we conclude that wearing light-colored clothing is dangerous? Why, or why not?

   (b) Express in terms of conditional probabilities the information you would need in order to judge whether or not light-colored clothing is helpful in avoiding accidents.