Random variables and their properties

As we have discussed in class, when observing a random process, we are faced with uncertainty. It is easier to study that uncertainty if we make things numerical. We define a **random variable** as follows: it is a rule that assigns a number to each of the outcomes of a random process.

ex: roll a die — assign the numbers $1, 2, 3, 4, 5, 6$ as the random variable D

ex: students attending class on a given day — assign the number of people who actually attend as the random variable C

ex: a person walking in the door — assign 0 if the person is male, 1 if the person is female as the random variable G

Once we have defined a random variable, we can examine the random process through its properties. The pattern of probabilities that are assigned to the values of the random variable is called the **probability distribution** of the random variable.

ex: roll a die — $P(D = 1) = 1/6; P(D = 2) = 1/6$; etc.

ex: students attending class on a given day — $P(C = 0) = \ldots$

ex: a person walking in the door — $P(G = 0) = .7; P(G = 1) = .3$

Why is there a benefit to examining a random process through a random variable? The benefit comes from the ability to manipulate the numerical values in useful and interesting ways. One example of how we can do this is in the construction of summary values for random variables (technically called the *moments* of random variables).

We often hear phrases like "the expected sales for this product over the next six months is 2 million units," or "the life expectancy for males in this country is 72 years." What do such phrases mean? They do not refer to sample means for observed data, since the event hasn't occurred yet; rather, they refer to properties of a random variable.

Consider as an example the New York State Daily Numbers lottery game. The simplest version of the game works as follows: each day a three digit number between 000 and 999, inclusive, is chosen. You pay $1 to bet on a particular number in a game. If your number comes up, you get back $500 (for a net profit of $499), and if your number doesn't come up, you get nothing (for a net loss of $1).

Consider the random process that corresponds to one play of the game, and define the random variable W to be the net winnings from a play. The following table summarizes the properties of the random variable as they relate to the random process:

Outcome of process	*Probability*	*Value of W*
Your number comes up	$p = 1/1000$	$W = 499$
Your number doesn't come up	$p = 999/1000$	$W = -1$

In the long run, we expect to win 1 time out of every 1000 plays, where we'd win \$499, and we expect to lose 999 out of every 1000 plays, where we'd lose \$1 (this is just the frequency theory definition of probabilities). That is, our rate of winnings per play, in the long run, would be \$499, .001 of the time, and −\$1, .999 of the time, or

$$(499)(.001) + (-1)(.999) = -.5.$$

In the long run, we lose 50¢ each time we play. Note that on any one play, we *never* lose 50¢ (we either win \$499 or lose \$1); rather, this is saying that if you play the game 10000 times, you can expect to be roughly \$5000 down at the end. An even better way to look at it is that if 10 million people play the game every day, the state can expect to only have to give back about \$5 million, a daily profit of a cool \$5 million (this is why states run lotteries!).

In general, the **expected value** of a random variable (also called the **mean**) is the sum of products of the value of the random variable for each outcome of the random process times the probability of that outcome occurring; that is, for a random variable X,

$$\begin{aligned} E(X) \equiv \mu_X &\equiv \sum_{\text{all outcomes}} (\text{Value of } X \text{ for that outcome})(\text{Probability of that outcome}) \\ &= \sum_i x_i p_i \end{aligned}$$

Just as was true when examining data, we are often interested in variability, as well as location. That is, the expected value gives a sense of a typical value, but is the random variable concentrated tightly around that value, or does it vary widely? We can define the **variance** of a random variable as the sum of squared differences from the mean, weighted by the probability of occurrence,

$$V(X) \equiv \sigma_X^2 = \sum_i (x_i - \mu_X)^2 p_i.$$

The variance is in squared units, so we usually think in terms of the **standard deviation** $\sigma_X = \sqrt{\sigma_X^2}$.

So, for the Daily Numbers lottery game,

$$\begin{aligned}\sigma^2 &= [499 - (-.5)]^2 \times .001 + [-1 - (-.5)]^2 \times .999 \\ &= (499.5)^2(.001) + (-.5)^2(.999) \\ &= 249.75,\end{aligned}$$

so the standard deviation is $\sigma = \$15.8$. Note that the standard deviation is much larger than (the absolute value of) the mean. The ratio of these numbers,

$$\frac{\sigma}{|\mu|},$$

is called the **coefficient of variation** (the inverse of this value, $|\mu|/\sigma$, is commonly used in finance as a measure of the reward/risk tradeoff for equities, and is called the *Sharpe ratio*). For the lottery data, the coefficient of variation equals 31.6. The coefficient of variation gives a sense of how variable the random variable is, relative to a typical value. Values of the coefficient variation over 1 are indicative of highly variable processes. The value of 31.6 here is *very* large, which makes sense; a random variable with payoffs of either 499 or -1 *is* highly variable. Another example of a highly variable random variable is stock returns; for example, the coefficient of variation for the daily return of the New York Stock Exchange Composite Index is typically around 15 or so.

The following shortcut formula is equivalent to the one above, but is sometimes easier to use:

$$\sigma_X^2 = \left[\sum_i x_i^2 p_i\right] - \mu_X^2.$$

So, the variance of the Numbers game payoff, using the shortcut formula, is

$$\begin{aligned}\sigma^2 &= (499^2)(.001) + (-1)^2(.999) - (-.5)^2 \\ &= 250 - .25 = 249.75.\end{aligned}$$

As measures of location and scale, respectively, the mean and standard deviation satisfy certain intuitive characteristics. For example, the expected value of a linearly transformed random variable is just the linear transform of the expected value. That sounds like a mouthful, but all that it says is that

$$E(aX + b) = aE(X) + b.$$

This result can be easily verified using the definition of the expected value, and an appendix gives the details. The result is also completely intuitive, as it says, for example, that simple rescalings of random variables do not change the expected behavior. So, for example, it doesn't matter if sales are recorded in dollars or Euros — the expected sales translate from one to the other in the exact way that individual values do.

Scale measures like the standard deviation also operate in the expected way under linear transformation. Shifting all of the values of a random variable up or down does not affect the variability of the random variable, so it does not affect the standard deviation. On the other hand, multiplying all of the values by a constant makes those values farther away from the mean by that amount, and thus multiplies the standard deviation by that amount. That is,

$$SD(aX + b) = |a|\, SD(X).$$

(Note also the relationship

$$V(aX + b) = a^2 V(X),$$

which is implied by the first.) The derivations of these relationships are also given in an appendix.

The standard deviation of a random variable is often useful as a measure of **volatility**. Say, for example, that the random variable S is the price of a stock on a given day, while C is the change in price from the previous trading day. In a time of a flat market, we might expect that $E(C) = 0$, while $SD(C)$ measures the tendency for the stock price to vary (that is, its volatility). Since most investments provide proportional returns on investments, we wouldn't be surprised to see $SD(C)$ be directly related to the level of the stock; that is, S. So, we might envision a class of stocks that have a standard deviation of 1% of the stock price; if the stock price S_1 of one stock is 10, that's a standard deviation of the change in price C_1 of .1, and if the stock price S_2 of another stock is 50, that's a standard deviation of the change in price C_2 of .5. This would imply that in terms of volatility, there is no difference between buying five shares of stock 1 or one share of stock 2, since in either case, the standard deviation of the change in price of the total purchase is the same (if C_1 is the change in price for one share of stock 1, $5C_1$ is the change in price of the total purchase of five shares of stock 1, and $SD(5C_1) = 5SD(C_1) = (5)(.1) = .5 = SD(C_2)$).

We have, of course, seen the names mean, variance and standard deviation before, in the context of analyzing data. The use of the same terms is not accidental. The sample mean $\overline{X}$ and sample standard deviation s are estimates of the population mean μ and standard deviation σ, respectively. We will talk about the properties of such estimates in a little while.

If we use μ_X to represent the mean of the random variable X, we see that the variance is also an expected value: if we center the random variable by subtracting the mean, and then square it, the variance is the expected value of the result. That is,

$$V(X) = E[(X - \mu_X)^2].$$

Of course, the standard deviation satisfies $SD(X) = \sqrt{V(X)}$. It turns out that the relationship between two random variables X and Y can also be summarized using an expected value. The *covariance* of two random variables X and Y is defined as

$$Cov(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

The covariance is a useful concept for several reasons.

(1) Consider a new random variable that is the sum of X and Y. The variance of this sum is a simple function of the variances of X and Y and their covariance:

$$V(X + Y) = V(X) + V(Y) + 2Cov(X, Y).$$

Consider, for example, two investments, whose returns are the random variables X and Y, respectively. This formula tells us that the variance of the return of a portfolio that consists of the sum of these two investments is the sum of the variances of the returns, plus twice the covariance of the returns.

(2) If two random variables X and Y are independent, then $Cov(X, Y) = 0$. Note that this implies that if two random variables X and Y are independent, then

$$V(X + Y) = V(X) + V(Y).$$

The converse of this statement is *not* true; that is, two random variables having zero covariance does not guarantee that they are statistically independent.

(3) The *correlation coefficient*, ρ, is a scaled version of the covariance, with

$$\rho = \frac{Cov(X, Y)}{SD(X)SD(Y)}.$$

The correlation coefficient satisfies $-1 \le \rho \le 1$, with $\rho = \pm 1$ representing a perfect straight line relationship between X and Y (either direct for positive ρ, or inverse for negative ρ), and $\rho = 0$ representing the absence of any linear relationship between X and Y. In the latter case we say that the two random variables are *uncorrelated*.

(4) This generalizes to weighted sums in a straightforward way; the variance of a weighted sum of two random variables is just

$$V(aX + bY) = a^2V(X) + b^2V(Y) + 2ab\rho SD(X)SD(Y).$$

Note that if X and Y represent the returns of two investments, and a and b are positive numbers that sum to 1, this corresponds to the variance of the return of a portfolio consisting of the two investments with weights (a, b).

The sample correlation coefficient r that I mentioned in the discussion of basic data analysis is a sample–based estimate of this population parameter. The formulas given above apply to the sample–based versions in the same way that they do to the population–based versions. So, for example,

$$s^2_{x+y} = s^2_x + s^2_y + 2r_{xy}s_xs_y,$$

where s^2_x, s^2_y, and s^2_{x+y} are the sample variances of a set of data values $\{x_i\}$, $\{y_i\}$, and $\{x_i + y_i\}$, respectively, and r_{xy} is the sample correlation between the $\{x_i\}$ and $\{y_i\}$ values.

Appendix — the shortcut formula for the variance

The variance of a random variable is defined as

$$V(X) = \sum_i (x_i - \mu_X)^2 p_i.$$

Expanding the square term on the right side of the equation gives

$$\begin{aligned} V(X) &= \sum_i (x_i^2 - 2x_i\mu_X + \mu_X^2)p_i \\ &= \sum_i x_i^2 p_i - 2\mu_X \sum_i x_i p_i + \mu_X^2 \sum_i p_i. \end{aligned}$$

Since $\mu_X = \sum_i x_i p_i$ and $\sum_i p_i = 1$, this yields

$$V(X) = \sum_i x_i^2 p_i - 2\mu_X^2 + \mu_X^2 = \sum_i x_i^2 p_i - \mu_X^2,$$

which is the shortcut formula.

Appendix — linear transformations of random variables and the mean and standard deviation

Consider first the expected value of the linear transform of a random variable. By definition

$$E(X) = \sum_i x_i p_i,$$

so

$$\begin{aligned} E(aX + b) &= \sum_i (ax_i + b)p_i \\ &= \sum_i ax_i p_i + \sum_i bp_i \\ &= a\sum_i x_i p_i + b\sum_i p_i. \end{aligned}$$

Since $E(X) = \sum_i x_i p_i$ and $\sum_i p_i = 1$, this yields

$$E(aX + b) = a\,E(X) + b.$$

Now consider the variance. By definition

$$V(X) = \sum_i [x_i - E(X)]^2 p_i,$$

so

$$\begin{aligned} V(aX+b) &= \sum_i [ax_i + b - E(aX+b)]^2 p_i \\ &= \sum_i \{ax_i + b - [aE(X)+b]\}^2 p_i \\ &= \sum_i [ax_i - aE(X)]^2 p_i \\ &= a^2 \sum_i [x_i - E(X)]^2 p_i \\ &= a^2 V(X). \end{aligned}$$

Taking the square root of both sides then yields

$$SD(aX+b) = a\,SD(X).$$

Appendix — fun with random variables

Expected number of children

Here is an example of the use of the expected value of a random variable that isn't related to gambling games. It is well–known that in the People's Republic of China, there are more male children than female children (while the worldwide rate of male births is .515, in China the reported value is .539). Various explanations have been given for this fact, including infanticide. A Letter to the Editor of the *Washington Post* on May 1, 1993, suggested that the reason for this is that when a Chinese family has a male child, they stop having children, and this leads to more male children, on average.

Does this possibility explain the observed pattern of more male children? A Letter to the Editor of the *Washington Post* by Paul Glewwe on May 11, 1993, pointed out that it does not. Here's the way to prove it. Say $P(\text{male child}) = .5$; if a couple has a male child, they stop having children, or, they stop after three children. Is the expected number of male children larger than the expected number of female children? Here are the numbers:

Result	*Probability*	*# males*	*# females*
M	.5	1	0
FM	.25	1	1
FFM	.125	1	2
FFF	.125	0	3

Now,

$$E(M) = (1)(.5) + (1)(.25) + (1)(.125) + (0)(.125) = .875$$

and

$$E(F) = (0)(.5) + (1)(.25) + (2)(.125) + (3)(.125) = .875.$$

Thus, even if Chinese couples are using this system, the expected number of male children still equals the expected number of female children. As Paul Glewwe noted, this does not account for the observed pattern more male children. Note, by the way, that the choice of three children above is arbitrary; this pattern holds true for any number of children.

One possible explanation, by the way, for the smaller number of female children is that Chinese families don't report female children to the authorities, given the strong disincentives from the government to having more than one child. The paper "Five Decades of Missing Females in China" by A.J. Coale and J. Danister in the August 1994 issue of *Demography* attributes this historical pattern to high rates of female infanticide in the 1930's and 1940's, differential death rates from a famine in 1959–1960, and gender–selective abortion starting in the 1980's. Recently, it has been shown that families in which the husband is significantly older than the wife tend to produce more boys (families in which the wife is significantly older than the husband tend to produce more girls); perhaps this is a contributing factor as well.

Size bias

Here's another example to consider that is related to *size bias*. Size bias occurs when the method of observing a random process results in observation of too many (or too few) observations that are larger (or smaller). An example of this is as follows. Say you want to know how fast people actually drive on a highway. You might imagine finding this out by standing at a particular spot on a highway, putting a radar gun on all the cars that come by in a three hour period (say), and averaging the resultant speeds. This is *not* the right thing to do (at least not without adjusting your results). Why? Because you are more likely to observe people who drive faster, simply because they drive faster, biasing the speed figures upwards. If you're having trouble seeing this, think of it this way. Consider a large geographic region with automobiles spread evenly over the region. Say half of all cars drive 50 miles per hour, and half drive 75 miles per hour, and you're going to sample at a certain spot from 1:00-4:00 PM. The autos that drive 50 miles per hour that can possibly be in your sample are the ones within a 150 mile radius of your position (of course, not all of these cars will pass you, but they are the only ones that can). On the other hand, the autos that drive 75 miles per hour that can possibly be seen are those within a 225 mile radius of your position, a region that is more than twice as large. In general, faster cars are more likely to be seen, simply because they're faster.

Say a certain course at Stern has four sections, which will have 25, 40, 50, and 85 students, respectively. Four professors will be assigned to the four sections at random. What is the expected class size for one of them, Professor A? This is a fairly straightforward problem. Each class has probability .25 of being Professor A's, so the expected class size is

$$(.25)(25) + (.25)(40) + (.25)(50) + (.25)(85) = 50.$$

Now, assume that the 200 students taking the course will be assigned randomly to each section according to the distribution given above. Consider a particular student, student B. What is the expected class size for student B? This is a different problem from Professor A's question. The probability that student B is put in a given section equals the proportion of the 200 students that fall in that section; so, for example, the probability of being put in the section with 25 students is $25/200 = .125$. Thus, the expected class size from student B's point of view is

$$(.125)(25) + (.2)(40) + (.25)(50) + (.425)(85) = 59.75.$$

How is this related to size bias? If you try to answer one of the questions using the other approach, you're going to have size bias. If you're interested in the average number of classmates a particular (randomly chosen) student has, you have to look at students; looking at classes (by examining a set of rosters, for example) will give too low a result. Similarly, if you're interested in the average number of students a professor has in classes, look at classes; asking students will give too high a result.

"The statistician's attitude to variation is like that of the evangelist to sin; he sees it everywhere to a greater or lesser extent."

— W. Spendley

"It is important to note that: many beers and wines are stronger than average."

— Drinking and driving campaign leaflet, British Department of Transport, 1996

"If you confront a statistician with a man with one foot in a bucket of boiling water and the other foot in a bucket of ice–cold water, he will say that, on average, the subject is comfortable."

— Anonymous

"Variance is what any two statisticians are at."

— Anonymous

"The overwhelming majority of people have more than the average number of legs."

— E. Grebenik

"Variation itself is nature's only irreducible essence. Variation is the hard reality, not a set of imperfect measures for a central tendency. Means and medians are the abstractions."

— Stephen Jay Gould

"Only 25% of households consist of the classic couple with 2.4 children."

— Matthew Fort, *The Observer*, 10 November 1996.

"We look forward to the day when everyone will receive more than the average wage."

— Australian Minister of Labor (1973)

"Thorstein the Learned says that there was a settlement on the Island of Hising which alternately belonged to Norway and to Gautland. So the kings agreed between them to draw lots and throw dice for this possession. And he was to have it who threw the highest. Then the

Swedish king threw two sixes and said that it was no use for King Olaf to throw. He replied, while shaking the dice in his hand, 'There are two sixes still on the dice, and it is a trifling matter for God, my Lord, to have them turn up.' He threw them, and two sixes turned up. Thereupon Olaf, the king of Norway, cast the dice, and one six showed on one of them, but the other split in two, so that six and one turned up; and so he took possession of the settlement."

— 13th century Norse saga of Saint Olaf