

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 — $P(D = 1) = 1/6; P(D = 2) = 1/6; \text{etc.}$

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

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:

<i>Outcome of process</i>	<i>Probability</i>	<i>Value of W</i>
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

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

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

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 \bar{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.

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 $-1 \leq \rho \leq 1$, 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,

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2r_{xy}s_x s_y,$$

where s_x^2 , s_y^2 , and s_{x+y}^2 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.

“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