CSE 103: Probability and statistics **Winter 2010**

Topic 3 — Random variables, expectation, and variance

3.1 Random variables

A *random variable* (r.v.) is defined on a probability space $(\Omega, \Pr)$ and is a mapping from Ω to $\mathbb{R}$.

The value of the random variable is fully determined by the outcome $\omega \in \Omega$. Thus the underlying probability space (probabilities $\Pr(\omega)$) induces a probability distribution over the random variable. Let's look at some examples.

Suppose you roll a fair die. The sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, all outcomes being equally likely. On this space we can then define a random variable

$$X = \begin{cases} 1 & \text{if die is} \geq 3 \\ 0 & \text{otherwise} \end{cases}$$

In other words, the outcomes $\omega = 1, 2$ map to $X = 0$, while the outcomes $\omega = 3, 4, 5, 6$ map to $X = 1$. The r.v. X takes on values $\{0, 1\}$, with probabilities $\Pr(X = 0) = 2/3$ and $\Pr(X = 1) = 1/3$.

Or say you roll this same die n times, so that the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}^n$. Examples of random variables on this larger space are

$$\begin{aligned} X &= \text{the number of 6's rolled,} \\ Y &= \text{the number of 1's seen before the first 6.} \end{aligned}$$

The sample point $\omega = (1, 1, 1, 1, \ldots, 1, 6)$, for instance, would map to $X = 1, Y = n - 1$. The variable X takes values in $\{0, 1, 2, \ldots, n\}$, with

$$\Pr(X = k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}$$

(do you see why?).

As a third example, suppose you throw a dart at a dartboard of radius 1, and that it lands at a random location on the board. Define random variable X to be the distance of the dart from the center of the board. Now X takes values in $[0, 1]$, and for any x in this range, $\Pr(X \leq x) = x^2$.

Henceforth, we'll follow the convention of using capital letters for r.v.'s.

3.2 The mean, or expected value

For a random variable X that takes on a finite set of possible values, the *mean*, or *expected value*, is

$$\mathbb{E}(X) = \sum_x x \Pr(X = x)$$

(where the summation is over all the possible values x that X can have). This is a direct generalization of the notion of *average* (which is typically defined in situations where the outcomes are equally likely). If X

is a continuous random variable, then this summation needs to be replaced by an equivalent integral; but we'll get to that later in the course.

Here are some examples.

1. *Coin with bias (heads probability) p.*

   Define X to be 1 if the outcome is heads, or 0 if it is tails. Then

   $$\mathbb{E}(X) \;=\; 0 \cdot \Pr(X=0) + 1 \cdot \Pr(X=1) \;=\; 0 \cdot (1-p) + 1 \cdot p \;=\; p.$$

   Another random variable on this space is X^2, which also takes on values in $\{0,1\}$. Notice that $X^2 = X$, and in fact $X^k = X$ for all $k = 1, 2, 3, \ldots$! Thus, $\mathbb{E}(X^2) = p$ as well. This simple case shows that in general, $\mathbb{E}(X^2) \neq \mathbb{E}(X)^2$.

2. *Fair die.*

   Define X to be the outcome of the roll, so $X \in \{1,2,3,4,5,6\}$. Then

   $$\mathbb{E}(X) \;=\; 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} \;=\; 3.5.$$

3. *Two dice.*

   Let X be their sum, so that $X \in \{2, 3, 4, \ldots, 12\}$. We can calculate the probabilities of each possible value of X and tabulate them as follows:

x	2	3	4	5	6	7	8	9	10	11	12
$\Pr(X=x)$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

   This gives $\mathbb{E}(X) = 7$.

4. *Roll n die; how many sixes appear?*

   Let X be the number of 6's. We've already analyzed the distribution of X, so

   $$E(X) \;=\; \sum_{k=0}^{n} k \Pr(X=k) \;=\; \sum_{k=0}^{n} k \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k} \;=\; \frac{n}{6}.$$

   The last step is somewhat mysterious; just take our word for it, and we'll get back to it later!

5. *Toss a fair coin forever; how many tosses to the first heads?*

   Let $X \in \{1, 2, \ldots\}$ be the number of tosses until you first see heads. Then

   $$\Pr(X=k) \;=\; \Pr((T,T,T,\ldots,T,H)) \;=\; \frac{1}{2^k}.$$

   It follows that

   $$\mathbb{E}(X) \;=\; \sum_{k=1}^{\infty} \frac{k}{2^k} \;=\; 2.$$

   We saw in class how to do this summation. The technique was based on the formula for the sum of a geometric series: if $|r| < 1$, then

   $$a + ar + ar^2 + \cdots \;=\; \frac{a}{1-r}.$$

6. *Toss a coin with bias p forever; how many tosses to the first heads?*

   Once again, $X \in \{1, 2, \ldots\}$, but this time the distribution is different:

   $$\Pr(X = k) \;=\; \Pr((T, T, T, \ldots, T, H)) \;=\; (1 - p)^{k-1} p.$$

   Using the same technique as before, we get $\mathbb{E}(X) = 1/p$.

   There's another way to derive this expectation. We always need at least one coin toss. If we're lucky (with probability p), we're done; otherwise (with probability $1 - p$), we start again from scratch. Therefore $\mathbb{E}(X) = 1 + (1 - p)\mathbb{E}(X)$, so that $\mathbb{E}(X) = 1/p$.

7. *Pascal's wager: does God exist?*

   Here was Pascal's take on the issue of God's existence: if you believe there is some chance $p > 0$ (no matter how small) that God exists, then you should behave as if God exists.

   Why? Well, let the random variable X denote your amount of suffering.

   Suppose you behave as if God exists (that is, you are good). This behavior incurs a significant but finite amount of suffering (you are not able to do some of the things you would like to). Say $X = 10$.

   On the other hand, suppose you behave as if God doesn't exist – that is, you do all the things you want to do. If God really doesn't exist, you're fine, and your suffering is $X = 0$. But if God exists, then you go straight to hell and your suffering is $X = \infty$. Thus your *expected* suffering if you behave badly is $\mathbb{E}(X) = 0 \cdot (1 - p) + \infty \cdot p = \infty$.

   So: to minimize your expected suffering, behave as if God exists!

3.3 Linearity of expectation

If you double each value of X, then you also double its average; that is, $\mathbb{E}(2X) = 2\mathbb{E}(X)$. Likewise, if you raise each of its values by 1, you will also increase the average by 1; that is, $\mathbb{E}(X + 1) = \mathbb{E}(X) + 1$. More generally, for any constants a, b,

$$\mathbb{E}(aX + b) \;=\; a\mathbb{E}(X) + b.$$

Another exceptionally useful formula says that the mean value of the sum of variables is simply the sum of their individual means. Formally, for any random variables X, Y,

$$\mathbb{E}(X + Y) \;=\; \mathbb{E}(X) + \mathbb{E}(Y).$$

For example, recall our earlier example about two rolls of a die, in which we let X be the sum of the rolls and derived $\mathbb{E}(X)$ by first computing $\Pr(X = x)$ for all $x \in \{2, 3, \ldots, 12\}$. Well, now we can do it much more easily: simply write X_1 for the first roll and X_2 for the second roll, so that $X = X_1 + X_2$. We already know $\mathbb{E}(X_i) = 3.5$, so $\mathbb{E}(X) = 7$.

More generally, for any random variables $X_1, X_2, \ldots, X_n$,

$$\mathbb{E}(X_1 + \cdots + X_n) \;=\; \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_n).$$

Some quick examples:

1. Roll n dice and let X be the number of sixes. What is $\mathbb{E}(X)$?

   This time, let X_i be 1 if the ith roll is a six, and 0 otherwise. Thus $\mathbb{E}(X_i) = 1/6$, so $\mathbb{E}(X) = n/6$.

2. Toss n coins of bias p and let X be the number of heads. What is $\mathbb{E}(X)$?

   Let X_i be 1 if the ith coin turns up heads, and 0 if it turns up tails. Then $\mathbb{E}(X_i) = p$ and since $X = X_1 + \cdots + X_n$, we have $\mathbb{E}(X) = np$.

3. Toss n coins of bias p; what is the expected number of times HTH appears in the resulting sequence?

   Let X_i be 1 if there is an occurrence of HTH starting at position i (so $1 \leq i \leq n-2$). The total number of such occurrences is $X = X_1 + X_2 + \cdots + X_{n-2}$. Since $\mathbb{E}(X_i) = p^2(1-p)$, we have $\mathbb{E}(X) = (n-2)p^2(1-p)$.

3.3.1 Fixed points of a permutation

The *fixed points* of a permutation are the numbers that remain in their original position. For instance, in the permutation

$$(1,2,3,4,5,6) \to (6,2,5,4,1,3)$$

the fixed points are 2 and 4. Let X be the number of fixed points in a random permutation of $(1,2,\ldots,n)$; what is $\mathbb{E}(X)$?

Linearity is very helpful here. Define the random variable X_i to be 1 if i is a fixed point, and 0 otherwise. Then $\mathbb{E}(X_i) = 1/n$. Therefore

$$\mathbb{E}(X) \;=\; \mathbb{E}(X_1 + \cdots + X_n) \;=\; 1.$$

The expected number of fixed points is 1, regardless of n.

3.3.2 Coupon collector, again

Recall the setting: each cereal box holds one of k action figures (chosen at random), and you want to collect all the figures. What is the expected number of cereal boxes you need to buy?

Suppose you keep buying boxes until you get all the figures. Let X_i be the number of boxes you buy to get from $i-1$ distinct figures to i distinct figures. Therefore $X = X_1 + X_2 + \cdots + X_k$, and of course $X_1 = 1$.

What is $\mathbb{E}(X_i)$? Well, you already have $i-1$ of the figures, so the chance of getting a new figure in a cereal box is $(k-(i-1))/k$. Call this p. Therefore, the expected amount of time you have to wait to get a new figure is $1/p$: just like waiting for a coin with bias p to turn up heads. That is,

$$\mathbb{E}(X_i) \;=\; \frac{k}{k-i+1}.$$

Invoking linearity of expectation,

$$\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X_1) + \cdots + \mathbb{E}(X_k) \\
&= \frac{k}{k} + \frac{k}{k-1} + \frac{k}{k-2} + \cdots + \frac{k}{1} \\
&= k\left(1 + \frac{1}{2} + \cdots + \frac{1}{k}\right) \\
&\approx k \ln k.
\end{aligned}$$

This confirms our earlier observations about the coupon collector problem: you need to buy about $k \ln k$ boxes.

3.3.3 Balls in bins, again

Toss m balls in n bins; what is the expected number of *collisions*? Let's make this more precise. For any $1 \leq i < j \leq m$, define the random variable X_{ij} to be 1 if balls i and j land in the same bin, and 0 otherwise. Then the number of collisions is defined to be

$$X = \sum_{1 \leq i < j \leq m} X_{ij}.$$

Since $\mathbb{E}(X_{ij}) = 1/n$ (do you see why?), it follows that the expected number of collisions is

$$\mathbb{E}(X) = \binom{m}{2} \frac{1}{n} = \frac{m(m-1)}{2n}.$$

So if $m < \sqrt{2n}$, the expected number of collisions is < 1, which means every ball goes into a different bin.

This relates back to the birthday paradox, where m is close to the threshold $\sqrt{2n}$.

3.4 Independent random variables

Random variables X and Y are *independent* if

$$\Pr(X = x, Y = y) = \Pr(X = x)\Pr(Y = y)$$

for all x, y. In words, the joint distribution of (X, Y) factors into the product of the individual distributions. This also implies, for instance, that

$$\Pr(X = x | Y = y) = \Pr(X = x).$$

Which of the following pairs (X, Y) are independent?

1. Pick a random card out of a standard deck. Define X to be 1 if it is a heart; and 0 otherwise. Define Y to be 1 if it is a jack, queen, or king; and 0 otherwise.

2. Toss a fair coin n times, and define X to be the number of heads, and Y to be 1 if the last toss is heads (and 0 otherwise).

3. X and Y take values in $\{-1, 0, 1\}$, and their joint distribution is given by the following table of probabilities.

		Y		
		−1	0	1
	−1	0.4	0.16	0.24
X	0	0.05	0.02	0.03
	1	0.05	0.02	0.03

If X, Y are independent, they satisfy the following useful *product rule*:

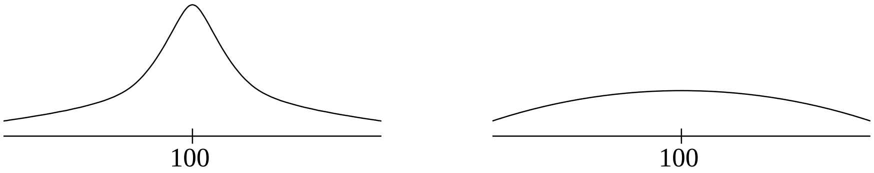
$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

Another useful fact is that $f(X)$ and $g(Y)$ must also be independent, for any functions f and g.

3.5 Variance

If you need to summarize a probability distribution by a single number, then the mean is a reasonable choice – although often the *median* is better advised (more on this later). But neither the mean nor median captures how *spread out* the distribution is.

Look at the following two distributions:



They both have the same expectation, 100, but one is concentrated near the middle while the other is pretty flat. To distinguish between them, we are interested not just in the mean $\mu = \mathbb{E}(X)$, but also in the typical distance from the mean, $\mathbb{E}(|X-\mu|)$. It turns out to be mathematically convenient to work with the square instead: the *variance* of X is defined to be

$$\mathrm{var}(X) \;=\; \mathbb{E}((X-\mu)^2) \;=\; \mathbb{E}((X-E(X))^2).$$

In the above example, the distribution on the right has a higher variance that the one on the left.

3.5.1 Properties of the variance

In what follows, take μ to be $\mathbb{E}(X)$.

1. The variance cannot be negative.

   Since each individual value $(X-\mu)^2$ is ≥ 0 (since its squared), the average value $\mathbb{E}((X-\mu)^2)$ must be ≥ 0 as well.

2. $\mathrm{var}(X) = \mathbb{E}(X^2) - \mu^2$.

   This is because

$$\begin{aligned}\mathrm{var}(X) &= \mathbb{E}((X-\mu)^2)\\ &= \mathbb{E}(X^2+\mu^2-2\mu X)\\ &= \mathbb{E}(X^2)+\mathbb{E}(\mu^2)+\mathbb{E}(-2\mu X) \quad \text{(linearity)}\\ &= \mathbb{E}(X^2)+\mu^2-2\mu\mathbb{E}(X)\\ &= \mathbb{E}(X^2)+\mu^2-2\mu^2 \;=\; \mathbb{E}(X^2)-\mu^2.\end{aligned}$$

3. For any random variable X, it must be the case that $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$.

   This is simply because $\mathrm{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 \geq 0$.

4. $\mathbb{E}(|X-\mu|) \leq \sqrt{\mathrm{var}(X)}$.

   If you apply the previous property to the random variable $|X-\mu|$ instead of X, you get $\mathbb{E}(|X-\mu|^2) \geq (\mathbb{E}(|X-\mu|))^2$. Therefore, $\mathbb{E}(|X-\mu|) \leq \sqrt{\mathbb{E}(|X-\mu|^2)} = \sqrt{\mathrm{var}(X)}$.

The last property tells us that $\sqrt{\mathrm{var}(X)}$ is a good measure of the typical spread of X: how far it typically lies from its mean. We call this the *standard deviation* of X.

3.5.2 Examples

1. Suppose you toss a coin with bias p, and let X be 1 if the outcome is heads, or 0 if the outcome is tails. Let's look at the distribution of X and of X^2.

Prob	X	X^2
p	1	1
$1-p$	0	0

From this table, $\mathbb{E}(X) = p$ and $\mathbb{E}(X^2) = p$. Thus the variance is $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = p(1-p)$.

2. Roll a 4-sided die (a tetrahedron) in which each face is equally likely to come up, and let the outcome be $X \in \{1, 2, 3, 4\}$.

We have two formulas for the variance:

$$\begin{aligned}\text{var}(X) &= \mathbb{E}\left((X-\mu)^2\right) \\ \text{var}(X) &= \mathbb{E}(X^2) - \mu^2\end{aligned}$$

where $\mu = \mathbb{E}(X)$. Let's try both and make sure we get the same answer. First of all, $\mu = \mathbb{E}(X) = (1+2+3+4)/4 = 2.5$. Now, let's tabulate the distribution of X^2 and $(X-\mu)^2$.

Prob	X	X^2	$(X-\mu)^2$
1/4	1	1	2.25
1/4	2	4	0.25
1/4	3	9	0.25
1/4	4	16	2.25

Reading from this table,

$$\begin{aligned}\mathbb{E}(X^2) &= \frac{1}{4}(1+4+9+16) &= 7.5 \\ \mathbb{E}(X-\mu)^2 &= \frac{1}{4}(2.25+0.25+0.25+2.25) &= 1.25\end{aligned}$$

The first formula for variance gives $\text{var}(X) = \mathbb{E}(X-\mu)^2 = 1.25$. The second formula gives $\text{var}(X) = \mathbb{E}(X^2) - \mu^2 = 7.5 - (2.5)^2 = 1.25$, the same thing.

3. Roll a k-sided die in which each face is equally likely to come up. The outcome is $X \in \{1, 2, \ldots, k\}$.

The expected outcome is

$$\mathbb{E}(X) = \frac{1+2+\cdots+k}{k} = \frac{\frac{1}{2}k(k+1)}{k} = \frac{k+1}{2},$$

using a special formula for the sum of the first k integers. There's another for the sum of the first k squares, from which

$$\mathbb{E}(X^2) = \frac{1^2+2^2+\cdots+k^2}{k} = \frac{\frac{1}{6}k(k+1)(2k+1)}{k} = \frac{(k+1)(2k+1)}{6}.$$

Then

$$\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2 = \frac{(k+1)(2k+1)}{6} - \frac{(k+1)^2}{4} = \frac{k^2-1}{12}.$$

The standard deviation is thus approximately $k/\sqrt{12}$.

4. X is the number of fixed points of a random permutation of $(1, 2, \ldots, n)$.

   Proceeding as before, let X_i be 1 if i is a fixed point of the permutation, and 0 otherwise. Then $\mathbb{E}(X_i) = 1/n$. For $i \neq j$, the product $X_i X_j$ is 1 only if both i and j are fixed points, which occurs with probability $1/n(n-1)$ (why?). Thus $\mathbb{E}(X_i X_j) = 1/n(n-1)$.

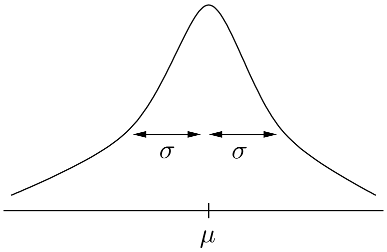
   Since X is the sum of the individual X_i, we have $\mathbb{E}(X) = 1$ and

$$\begin{aligned}
\mathbb{E}(X^2) &= \mathbb{E}((X_1 + \cdots + X_n)^2) \\
&= \mathbb{E}\left(\sum_{i=1}^n X_i^2 + \sum_{i \neq j} X_i X_j\right) \\
&= \sum_i \mathbb{E}(X_i^2) + \sum_{i \neq j} \mathbb{E}(X_i X_j) \\
&= n \cdot \frac{1}{n} + n(n-1) \cdot \frac{1}{n(n-1)} \;=\; 2.
\end{aligned}$$

   Thus $\text{var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X)^2) = 1$. This means that the number of fixed points has mean 1 and variance 1: in short, it is quite unlikely to be very much larger than 1.

3.5.3 Another property of the variance

Here's a cartoon picture of a well-behaved distribution with mean μ and standard deviation σ (that is, $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{var}(X)$).



The standard deviation quantifies the *spread* of the distribution whereas the mean specifies its *location*. If you increase all values of X by 10, then the distribution will shift to the right and the mean will increase by 10. But the spread of the distribution – and thus the standard deviation – will remain unchanged.

On the other hand, if you double all values of X, then its distribution becomes twice as wide, and thus its standard deviation σ is doubled. Which means that its variance, which is the square of the standard deviation, gets multiplied by 4.

In summary, for any constants a, b:

$$\text{var}(aX + b) = a^2 \text{var}(X).$$

Contrast this with the mean: $\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$.