Expected values

The **expected value** $E(X)$ or EX of a random variable X on a probability space Ω is a kind of *weighted average* of the values of X, with the weights being the probabilities of the different inputs/outputs. The precise definition is

$$\text{expected value of } X = E(X)$$

$$= \sum_{\omega \in \Omega} P(\omega) \cdot X(\omega)$$

We can *group* the inputs according to the output value produced, so this is also equal to

$$E(X) = \sum_{\text{values } x \text{ of } X} P(X = x) \cdot x$$

where (again) the notation $P(X = x)$ means the probability that X takes value x:

$$P(X = x) = P(\{\omega \in \Omega : X(\omega) = x\})$$

About notation

Yes, the notation and terminology for random variables is different from, and in conflict with, the notation used for functions and their values in calculus and differential equations.

First, and most importatly, yes, random *variables* are actually *functions*.

Yes, the random variable's name is often X, unlike the f or g in calculus.

Yes, usually the *input* to a function is called x, not the *output*, as in $X(\omega) = x$.

Examples of expected values

With X being the random variable counting `H`s in a single flip of a fair coin,

$$E(X) = \sum_{\text{values } x \text{ of } X} P(X = x) \cdot x$$

$$= P(X = 0) \cdot 0 + P(X = 1) \cdot 1$$

$$= \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$$

Note that we will never actually get 1/2 head in a flip of a fair coin.

But, as with many averages, the average or weighted average of integer values may be a non-integer.

That's ok.

With X being the random variable counting `H`s in 3 flips of a fair coin,

$$E(X) = \sum_{\text{values } x \text{ of } X} P(X = x) \cdot x$$

$$= P(X = 0) \cdot 0 + P(X = 1) \cdot 1$$

$$+ P(X = 2) \cdot 2 + P(X = 3) \cdot 3$$

$$= \binom{3}{0} 2^{-3} \cdot 0 + \binom{3}{1} 2^{-3} \cdot 1$$

$$+ \binom{3}{2} 2^{-3} \cdot 2 + \binom{3}{3} 2^{-3} \cdot 3$$

$$= \frac{0 + 3 \cdot 1 + 3 \cdot 2 + 1 \cdot 3}{8} = \frac{3}{2}$$

This may be an intuitively appealing answer, if we imagine that we get an *average* of $1/2$ head per flip in 3 flips.

But notice that the *definition* hands us an expression whose value is not obviously the answer what we expect, though it turns out to be so.

Sums and products of random variables

The **sum random variable** $X + Y$ made from two random variables X, Y defined on the *same* probability space Ω is defined, reasonably enough, to be the function whose values are the sum of the values of X and Y. That is, for $\omega \in \Omega$

$$(X + Y)(\omega) = X(\omega) + Y(\omega)$$

Similarly, the **product random variable** $X \cdot Y$ is

$$(X \cdot Y)(\omega) = X(\omega) \cdot Y(\omega)$$

The basic theorem on expected values

Our intuition about certain examples (like flipping a coin several times) is justified by the basic theorem about expected values:

Theorem: Let $X_1, \ldots, X_n$ be random variables on a common probability space Ω. Then

$$E(X_1 + \ldots + X_n) = E(X_1) + \ldots + E(X_n)$$

That is, the expected-value function E is **additive** (or *linear*).

Most functions do not have the additive property, though naive presumption of additivity (or linearity) is common. For example, despite many errors by novices, *generally*

$$\sin(a+b) \neq \sin a + \sin b$$

$$\sqrt{a+b} \neq \sqrt{a} + \sqrt{b}$$

$$(a+b)^2 \neq a^2 + b^2$$

For example, to compute the expected number of Hs in 10 flips of a fair coin, let X be the random variable on the probability space of all possible outcomes of 10 flips. The *definition* of expected value of X is what we want, namely

$$\text{expected no. Hs in 10 flips} = E(X)$$

$$= \sum_{k=0}^{10} P(X = k) \cdot k = \sum_{k=0}^{10} \binom{10}{k} \cdot 2^{-10} \cdot k$$

$$[1 \cdot 0 + 10 \cdot 1 + 45 \cdot 2 + 120 \cdot 3$$

$$+ 210 \cdot 4 + 252 \cdot 5 + 210 \cdot 6 + 120 \cdot 7$$

$$+ 45 \cdot 8 + 10 \cdot 9 + 1 \cdot 10] /1024$$

$$= (amazingly) 5$$

It is completely *not* obvious that this big computation will yield the intuitively suggested answer

$$10 \cdot \frac{1}{2} = 5 \ \text{ expected Hs in 10 flips}$$

Invocation of the *Theorem* allows us to legitimize our intuition here. Define random variables $X_1, \ldots, X_{10}$ by

$$X_i = \text{ no. } \texttt{H}\text{s on the } i^{\text{th}} \text{ flip of } 10$$

Note that these are all defined on the same probability space. Then

$$X = X_1 + \ldots + X_{10}$$

By the theorem,

$$E(X) = E(X_1) + \ldots + E(X_{10})$$

We evaluate each $E(X_i)$ via the definition

$$E(X_i) = \sum_{\text{values} k} P(X_i = k) \cdot k$$

$$= P(X_i = 0) \cdot 0 + P(X_i = 1) \cdot 1$$

Since the flips are independent and the coin is fair, for any index i the probability that H appears on the i^{th} flip is 1/2, so this is

$$E(X_i) = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$$

Then

$$E(X) = E(X_1) + \ldots + E(X_{10})$$

$$= \underbrace{\frac{1}{2} + \ldots + \frac{1}{2}}_{10} = 10 \cdot \frac{1}{2} = 5$$

It bears repeating that this is *not* the definition of expected value, and that our intuition is *not* obviously correct. Happily, it *is* intuitively correct *and* in the end our intuition (in this case) is vindicated by the Theorem.

Beware, though, that not all functions are additive or linear.

Evaluation by generating functions

But, even though it turns out that we do not need it in the above example, we might also want to be able to evaluate expressions such as

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} \cdot k$$

directly. This is possible, and the methodology has many applications.

Recall the **Binomial Theorem**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k \, y^{n-k}$$

Partial differentiation with respect to x gives

$$n(x+y)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k x^{k-1} \, y^{n-k}$$

Anticipating that we'll let $x = p$ and $y = 1 - p$ eventually, we see we're missing a factor of x on the right in that equality

$$n(x+y)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k x^{k-1} y^{n-k}$$

so multiply through by x:

$$nx(x+y)^{n-1} = \sum_{k=0}^{n} \binom{n}{k} k x^{k} y^{n-k}$$

Letting $x = p$ and $y = 1 - p$ gives

$$p \cdot n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k}\, k$$

For $p = 1/2$ we get the same conclusion as earlier via the Theorem, but use of the Theorem is **much better** because it is both simpler and more intuitive.

Example: How long we should expect to wait in flipping a fair coin until we get an H?

That is, let X be the random variable which counts the number of flips up to and including the first flip which gives a H. Then

$$E(X) = \sum_{k=0}^{\infty} P(X = k) \cdot k$$

$$= P(\mathtt{H}) \cdot 1 + P(\mathtt{TH}) \cdot 2 + P(\mathtt{TTH}) \cdot 3$$

$$+ P(\mathtt{TTTH}) \cdot 4 + P(\mathtt{TTTTH}) \cdot 5 + \dots$$

$$= P(\mathtt{H}) \cdot 1 + P(\mathtt{T})P(\mathtt{H}) \cdot 2 + P(\mathtt{T})^2 P(\mathtt{H}) \cdot 3$$

$$+ P(\mathtt{T})^3 P(\mathtt{H}) \cdot 4 + P(\mathtt{T})^4 P(\mathtt{H}) \cdot 5 + \dots$$

by *independence* of flips. Without even thinking about the fairness, let $P(\mathtt{H}) = p$ and $P(\mathtt{T}) = q$, where $p + q = 1$.

Then we're wanting to evaluate

$$\sum_{k=0}^{\infty} p\, q^{k-1} \cdot k$$

The infinite series we know how to evaluate is the **geometric series**

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$$

for $|q| < 1$. Differentiating both sides of this with respect to q gives

$$\sum_{k=0}^{\infty} q^{k-1} \cdot k = \frac{1}{(1-q)^2}$$

This is missing a factor of p, so multiply both sides by pq and using $p + q = 1$

$$\sum_{k=0}^{\infty} pq^{k-1} \cdot k = \frac{p}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p}$$

In the identity

$$\sum_{k=0}^{\infty} pq^{k-1} \cdot k = \frac{1}{p}$$

let $p = q = \frac{1}{2}$ to obtain

expected flips of fair coin to get a H

$$= \sum_{k=0}^{\infty} \frac{1}{2}\left(\frac{1}{2}\right)^{k-1} \cdot k = \frac{1}{1/2} = 2$$

This might suggest that we should *expect* to get a `H` on the second flip, so that we get a `T` on the first flip? But the same discussion would say that the expected number of flips to get a `T` is also 2.

No, it's just that *we should not expect to get the expected value*, since it's just an average.

Monty Hall Paradox [sic]

In case elementary probability seems too easy, here is a popular example that is less trivial.

In a game show *Let's Make a Deal* in which players were faced with 3 doors, behind one of which was a prize. The player chose a door, but the door was not opened. The host *Monty Hall* (who knew where the prize was) opened *another* door than the one guessed by the player, but *not* the one with the prize. The player was offered the chance to change their guess. *Should the player change their guess?*

Thus, the player was faced with one open door with no prize, and two closed doors, one of which was their original guess, and behind one of which is the prize.

The contestant should always change their guess.

This may be counter-intuitive.

One way to explain this in colloquial terms is to say that the probability of originally guessing the correct door is 1/3, and that does not change. Thus, the probability is

$$1 - \frac{1}{3} = \frac{2}{3}$$

that you're *wrong*, and should change your guess.

Among many *incorrect* arguments there is the one that says that, not knowing what else is going on, since there are two doors, the probability is 1/2. This approach, in which *ignorance of facts is interpreted as equal probability*, was already disdained by Laplace 300 years ago, and we should not use it now.

Conditional probability

The **conditional probability** that an event A will occur **given** that an event B occurs is *defined* to be

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

It is important to note that this is a *definition*, which turns out to have both practical and mathematical virtues. For one thing, we do not need to try to *intuit* the meaning of *probability that A will occur given that B occurs*.

Example: The conditional probability that at a fair coin comes up heads in at least 3 of 6 flips, given that the first two flips are tails, is

$$P(\text{at least 3 H's in 6}|\text{first two T's})$$

$$= \frac{P(\text{first two T's } \textit{and} \text{ at least 3 H's})}{P(\text{first two T's})}$$

Birthday paradox [sic]

It may seem strange that *in a set of at least 23 people the probability is* $\geq 1/2$ *that two have the same birthday*.

Not $365/2$, but more like $\sqrt{365}$.

For n things chosen at random with equal probabilities (and independently) from N things (with replacement), for

$$n > \sqrt{2 \ln 2} \cdot \sqrt{N} \sim \frac{17}{10} \cdot \sqrt{N}$$

the probability that two things are the same is $> \frac{1}{2}$.

Computation for birthday paradox

We compute the probability that *no two* outcomes are the same, and subtract this result from 1 to obtain the desired result.

After two trials, there is $1/N$ chance that the second outcome was equal to the first one, so the probability is $1 - \frac{1}{N}$ that the outcomes of two trials will be different.

After 3 trials, *given* that the first two outcomes are different, the conditional probability is $2/N$ that the third trial would give an outcome equal to *one* of the first two. Thus, given that the first two outcomes are different, the conditional probability that the third will differ from both is $1 - \frac{2}{N}$. Since the probability that the first two were different was $1 - \frac{1}{N}$, the formula above gives

$$P(\text{first 3 different}) = (1 - \frac{1}{N})\,(1 - \frac{2}{N})$$

After 4 trials, *given* that the first two outcomes are different, the *conditional probability* is $3/N$ that the third trial would give an outcome equal to *one* of the first two. Thus, given that the first two outcomes are different, the conditional probability that the third will differ from all of the first 3 is $1-\frac{3}{N}$. Using the previous step, and the formula above,

$$P(\text{first 4 different}) = (1 - \frac{1}{N})\,(1 - \frac{2}{N})\,(1 - \frac{3}{N})$$

Continuing, we get

$$P(n \text{ trials all different})$$

$$= (1 - \frac{1}{N})\,(1 - \frac{2}{N})\,(1 - \frac{3}{N})\ldots\,(1 - \frac{n-1}{N})$$

The logarithm of the probability that they're all different is

$$\ln(1 - \frac{1}{N}) + \ln(1 - \frac{2}{N}) + \cdots + \ln(1 - \frac{n-1}{N})$$

The first-order Taylor expansion for $\ln(1-x)$ for $|x| < 1$

$$\ln(1 - x) = -(x + \frac{x^2}{2} + \frac{x^3}{3} + \frac{x^4}{4} + \cdots)$$

In particular for $0 < x < 1$

$$\ln(1 - x) \leq -x$$

so

$$\ln(1 - \frac{1}{N}) + \ln(1 - \frac{2}{N}) + \cdots + \ln(1 - \frac{n-1}{N})$$

$$\leq -(\frac{1}{N} + \frac{2}{N} + \cdots + \frac{n-1}{N})$$

Recall

$$1 + 2 + 3 + 4 + \cdots + (k-1) + k = \frac{1}{2}k(k+1)$$

Then

$$\ln\left(P(n \text{ trials all different})\right) \leq \frac{-\frac{1}{2}(n-1)n}{N}$$

As n gets larger and larger, the expression $(n-1)n$ is for practical purposes n^2. Thus, we have an *approximate* formula

$$\ln\left(P(n \text{ trials all different})\right) \leq -\frac{n^2}{2N}$$

or

$$P(n \text{ trials all different}) \leq e^{-n^2/2N}$$

$$P(2 \text{ of } n \text{ trials the same}) \geq 1 - e^{-n^2/2N}$$

The probability that some two will be the same is therefore bigger than or equal 1/2 when the probability that no two are the same is *less* than 1/2. Thus, for given N we *solve* to find the smallest n so that

$$-\frac{n^2}{2N} < \ln\,\frac{1}{2}$$

which gives the formula

$$n \geq \sqrt{2 \cdot \ln 2} \cdot \sqrt{N} \sim \frac{17}{10} \cdot \sqrt{N}$$

for the size of n to assure that the probability is bigger than 1/2 that two choices are the same.
