Expected value and higher moments of discrete variables [1]

1 Definition of Expected Value

If you have an X with $P(X=-1)=P(X=1)=1/2$ - which is a 50-50 bet with possible gains $-1,1$ - how much do you expect to gain? Zero. Why? Because $-1*1/2+1*1/2=0$. We multiply values times probabilities and then add up. That is how we get the expected value of a random variable. Of course if $P(X=-1)=2/3$ and $P(X=1)=1/3$ we expect to lose money: in fact, by the same kind of computation, $-1*2/3+1*1/3<0$. The general definition reflects this idea.

Definition. For $X\colon\Omega\to\mathbb{R}$ discrete

$$E(X)=\sum\nolimits_i x_i\cdot P(X=x_i).$$

Multiply values times probabilities and then add up. has We usually write $p(x_i)=P(X=x_i)$, so that the above reads $EX=\sum_i x_i p(x_i)$. To avoid subscripts we may also write $EX=\sum_{x\in X(\Omega)} xp(x)$. Simplest cases: if $X(\omega)=c$ with probability 1 then $EX=c$; and $E\left(1_E\right)=P(E)$.

Remark. When X takes infinitely many values the sum $\sum_i x_i p(x_i)$ is a series, and we require that it converges absolutely to ensure that it converges, so that EX is well defined. The assumption that expectation is well defined is maintained throughout in what follows and we will not mention it any more, but one has to know that it is a real assumption, it is not *pour parler*. We show how it may fail in a famous example.

Example (The St. Petersburg paradox). Suppose someone offers you the following bet: they toss a fair coin repeatedly, and you win 2^k Euros if the first Heads occurs at the k-th toss, for $k\geq 1$. As we know this event has probability 2^{-k} (and $\sum_{k\geq 1}2^{-k}=1$), so we are talking of the random variable X on $\Omega=\{H,T\}^{\mathbb{N}}$ with $X(\Omega)=\left\{2^k\colon k\geq 1\right\}$ and distribution $P(X=2^k)=2^{-k}$. Observe that with probability close to 95% Heads arrives within the first 4 tosses and you get at most 16 Euros. Still, if we multiply values times probabilities and then add up we get

$$\sum\nolimits_{k\geq 1}2^k\cdot\frac{1}{2^k}=2\cdot\frac{1}{2}+2^2\cdot\frac{1}{2^2}+\cdots=\infty.$$

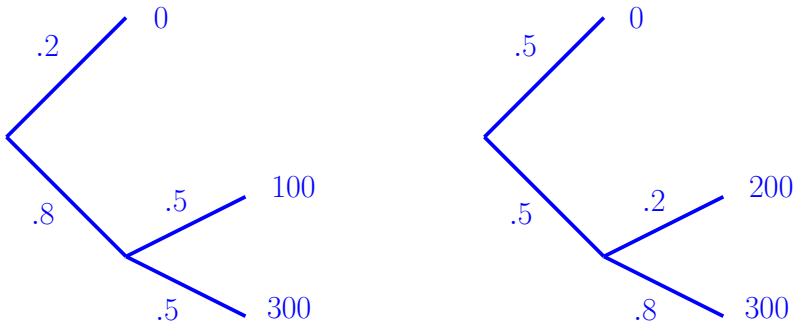
Of course notwithstanding this infinite value you would not give away all you have to enter such a bet, and this is the "paradox". But that issue belongs to a Decision Theory course.

We go back (for good) to finite sums.

Example (Choosing a problem). This is exactly the example we had in the first lecture; now we can confirm that the calculations there are correct. There are two problems: an easier one which

[1] S Modica 2026

you solve correctly with probability .8 and a harder one which which you solve with probability .5. Guessing the easy one yields 100, guessing the hard one gives 200. You can choose which one to take first; if you guess you proceed to the other one; if you fail you get zero. The yield add up. So the two choices can be visualized as in the picture below, where choosing the easy problem first is on the left. Which is it best to take first, in expected value terms?



Computing expectations (and using $P(AB) = P(A)P(B \mid A)$) we get that the left choice has an expected value of $.2 * 0 + .8 * .5 * 100 + .8 * .5 * 300 = .4 * 400 = 160$, while the right one gives $.1 * 200 + .4 * 300 = 140$. Therefore it is better to try the easy problem first.

2 $E(g(X))$

The results in this section are mostly a rewriting of those in Baldi section 2.6. Let $X\colon \Omega \to \mathbb{R}^n$ be an n-dimensional discrete random variable with density $p(x)$, $x \in X(\Omega)$; and consider a function $g : \mathbb{R}^n \to \mathbb{R}$. Then $g(X)$ is the discrete variable taking value $g(X(\omega))$ at ω.

Proposition. $Eg(X) = \sum_{x\in X(\Omega)} g(x) \cdot p(x)$.

Proof. Letting $Y = g(X)$ we have by definition $EY = \sum_{y\in Y(\Omega)} yP(Y = y)$. Now $\{Y = y\} = \cup_{x\colon g(x)=y} \{X = x\}$, union disjoint. Therefore

$$Eg(X) = \sum\nolimits_{y\in Y(\Omega)} y \sum\nolimits_{x\colon g(x)=y} p(x) = \sum\nolimits_{y\in Y(\Omega)} \sum\nolimits_{x\colon g(x)=y} g(x)p(x).$$

But since any $x \in X(\Omega)$ must end up in some $y \in Y(\Omega)$ the above double sum covers $X(\Omega)$, and the result follows. □

Corollary. *For real random variables X, Y one has*

(i) $E(cX) = cE(X)$ *if* $c \in \mathbb{R}$

(ii) $E(X + Y) = EX + EY$.

Proof. (i) follows from the above with $g(x) = cx$. For (ii) consider $(X, Y)\colon \Omega \to \mathbb{R}^2$ with joint density $p(x, y)$ and marginals $p_X(x) = \sum_y p(x, y)$ and $p_Y(y) = \sum_x p(x, y)$, and take $g(x, y) = x+y$.

Then, summing over the image of (X, Y) we get

$$E(X+Y) = \sum_{x,y}(x+y)\cdot p(x,y) = \sum_{x,y} xp(x,y) + \sum_{x,y} yp(x,y)$$
$$= \sum_x xp_X(x) + \sum_y yp_Y(y) = EX + EY. \qquad \square$$

As a direct consequence we get *linearity of expectation*: $E\left(\sum_{i=1}^n c_i X_i\right) = \sum_{i=1}^n c_i E(X_i)$. *It is important to observe that this holds whether or not the variables are independent.*

Example (trivial). $E(a+X) = a + EX$ for any $a \in \mathbb{R}$.

Example. Given random variables X and Y let $M^- = \max\{X, Y\}$ and $M_- = \min\{X, Y\}$. Since $M^- + M_- = X + Y$ we have $EM_- = E(X+Y) - EM^-$; similarly for EM^-.

Example (from Bertsekas, *Introduction to Probability*). n people throw their hats in a box; then the hats are extracted and assigned at random to the people. What is the expected number of people who end up with their own hat?

Let Ω consist of the $n!$ permutations π of the integers from 1 to n, and on Ω let X be the number of people getting their own hat; letting $X_i = 1$ if $\pi(i) = i$ and zero otherwise for $i = 1, \ldots, n$ we have $X = \sum X_i$. We have $P(X_i = 1) = (n-1)!/n! = 1/n$, therefore $EX = \sum EX_i = n \cdot 1/n = 1$.

Example (Expectation of our main distributions).

(i) *Binomial.* Recall that $X \sim B_{n,p}$ if $X = \sum_i \omega_i$ on $\Omega = \{0,1\}^n$ - number of successes in a Bernoulli scheme. That the probability of success in each repetition is p means that for the coordinate variable $X_i(\omega) = \omega_i$ we have $P(X_i = 1) = p$. Of course $E(X_i) = p$; but $X = \sum_i X_i$, so $EX = E\left(\sum_i X_i\right) = \sum_i E(X_i) = np$.

(ii) *Geometric.* Recall that $T \sim \text{Geom}_p$ if $P(T = k) = (1-p)^{k-1}p$. One can compute $ET = \sum_{k\geq 0} k \cdot (1-p)^{k-1}p$, but we shall find that $ET = 1/p$ in a much more clever way in Section 3. Notice that the expected time of first success is inversely proportional to the probability of success in the single trial.

(iii) *Poisson.* Recall that $X \sim \text{Poisson}_\lambda$ if $P(X = k) = e^{-\lambda}\lambda^k/k!$ for $k \geq 0$. Here you can compute directly, as an exercise, that

$$E(X) = \sum_{k=0}^{\infty} k \cdot e^{-\lambda}\frac{\lambda^k}{k!} = \lambda.$$

(iv) *Hypergeometric with two types.* This case is here for completeness, but frankly the effort it takes to go through it is too much. The message is this: in an urn with two types of balls, labeled zero and one, if you extract n of them the average number of ones you draw is the same whether you do the drawings with or without replacement. Formally, there are $N = N_0 + N_1$ balls, N_0 labeled 0

and N_1 labeled 1. You extract n balls without replacement. Letting X_i the indicator of the event "the i-th ball is a 1" we want $E\left(\sum_{I=1}^{n} X_i\right) = \sum_{I=1}^{n} E\left(X_i\right)$ so the whole task is to find $E\left(X_i\right)$ that is $P\left(X_i = 1\right)$, the probability that the i-th ball is a 1. Consider extracting all of the N balls. The probability that a given, fixed ball of type 1 is in the i-th place is $(N-1)!/N! = 1/N$; but the events that the first or, second or ... N_1-th ball is in the i-th place are disjoint, so the probability we look for is just N_1/N. Therefore $EX = n \cdot (N_1/N)$. The point is that if you do the drawings *with* replacement you get a Bernoulli process with success probability N_1/N - therefore as we just saw the expected number of 1's is again $n \cdot (N_1/N)$.

Example (Sum of random number of terms). We can do it from Example 2.40 in Baldi.

Example (Expected distance). Consider X and Y independent random variables with uniform distribution on the set $\{0, 1/n, 2/n, \ldots, 1\}$, that is $P(X = i/n) = P(Y = i/n) = 1/(n+1)$. We want to compute $E\left|X - Y\right|$, the average distance between the coordinates of the points uniformly distributed on (a grid of) the unit square. We shall find that $E\left|X - Y\right| = \frac{n+2}{3(n+1)}$, and it is interesting to note that it goes to $1/3$ when $n \to \infty$ - so that the grid becomes finer and finer and in the limit we may think of the variables as distributed on the whole of the unit interval. We shall return to it.

To find $E\left|X - Y\right|$, letting $M^- = \max\{X, Y\}$ and $M_- = \min\{X, Y\}$ we shall use the identity $|X - Y| = M^- - M_-$. We already know that $EM_- = E(X + Y) - EM^-$, so we need only EX and EM^-; but $EX = \frac{1}{n+1}(\frac{1}{n} + \ldots + \frac{n}{n}) = \frac{1}{2}$ so $E(X + Y) = 1$. Given that we obtain $E\left|X - Y\right| = EM^- - EM_- = 2EM^- - 1$. For M^- observe that its values are $\{0, 1/n, \ldots, 1\}$; letting F denote its cdf we can derive $P(M^- = k/n)$ as $F(k/n) - F((k-1)/n)$. Since X and Y are independent we have $F(k/n) = P(X \leq k/n)P(Y \leq k/n) = (k+1)^2/(n+1)^2$ (note carefully the denominator). Therefore

$$
\begin{aligned}
&P(M^- = k/n) = (k+1)^2/(n+1)^2 - k^2/(n+1)^2 = (2k+1)/(n+1)^2, \quad k = 0, \ldots, n \\
&EM^- = \sum_{k=0}^{n} \frac{k}{n} \frac{2k+1}{(n+1)^2} = \frac{1}{n(n+1)^2} \sum_{k=1}^{n} \left(2k^2 + k\right) = \frac{1}{n(n+1)^2} \left[\frac{n(n+1)(2n+1)}{3} + \frac{n(n+1)}{2}\right] \\
&\quad = \frac{n(n+1)}{n(n+1)^2} \left[\frac{(2n+1)}{3} + \frac{1}{2}\right] = \frac{4n+5}{6(n+1)},
\end{aligned}
$$

and finally

$$
E\left|X - Y\right| = \frac{4n+5}{3(n+1)} - 1 = \frac{n+2}{3(n+1)}.
$$

2.1 Monotonicity of E

Monotonicity is the other fundamental property of the expectation operator (a version of it is Baldi Proposition 2.36). The result is that if $P(X \geq Y) = 1$ then: (i) $EX \geq EY$, and (ii) equality can occur only if $X = Y$ with probability 1. Knowing that $EX - EY = E(X - Y)$, letting $Z = X - Y$ the above can be more concisely restated as follows:

Proposition. *Let $Z\colon \Omega \to \mathbb{R}$, discrete. Suppose $P(Z \geq 0) = 1$. Then*
(i) $EZ \geq 0$
(ii) $EZ = 0 \iff P(Z = 0) = 1$.

Proof. (i) $z \in Z(\Omega)$ always understood, we know that $EZ = \sum_z zP(Z = z)$. By hypothesis $P(Z = z) = 0\ \forall z < 0$, therefore

$$EZ = \sum\nolimits_{z\geq 0} zP(Z = z) \geq 0. \tag{1}$$

As to (ii), from (1)

$$\begin{aligned} EZ = \sum\nolimits_{z\geq 0} zP(Z = z) = 0 &\iff \forall z \geq 0\ zP(Z = z) = 0 \\ &\iff \forall z > 0\ P(Z = z) = 0 \iff P(Z = 0) = 1 \text{ (since } P(Z \geq 0) = 1). \end{aligned}$$ □

Remark. In the general case $EZ = \int_\Omega Z(\omega)\, dP(\omega)$, integral in the sense of Lebesgue, and this important result remains true (same statement). We cannot prove it in this course, and this is why it is instructive to have a complete proof in the discrete case.

Corollary. *Whenever EX exists we have $|EX| \leq E\,|X|$.*

Proof. We have $-\,|X| \leq X \leq |X|$ so from monotonicity and linearity $-E\,|X| \leq EX \leq E\,|X|$ which implies the above inequality since $E\,|X| \geq 0$ (again from monotonicity). □

2.2 $E(XY)$

Recall the notion of independence of two discrete real variables X and Y defined on a space $(\Omega, \mathcal{F})$. As above let $p(x, y)$ be their joint density and p_X, p_Y the marginals. Then X and Y are *independent* if for all x, y we have $p(x, y) = p_X(x) \cdot p_Y(y)$.

Proposition. *If X and Y are independent then $E(XY) = E(X)E(Y)$.*

Proof. Taking $g(x, y) = xy$ we get

$$\begin{aligned} E(XY) &= \sum\nolimits_{x,y} xy \cdot p(x, y) = \sum\nolimits_{x,y} xy \cdot p_X(x) \cdot p_Y(y) \\ &= \sum\nolimits_x xp_X(x) \cdot \sum\nolimits_y yp_Y(y) = EX \cdot EY. \end{aligned}$$ □

3 Random variables with nonnegative integer values

Lemma. *The sum of a convergent non-negative series is invariant to rearrangements.*

Proof. If $a_n \geq 0$ for all n then

$$\sum_{n=1}^{\infty} a_n \ \ (= \lim_{N\to\infty} \sum_{n=1}^{N} a_n) = \sup_{I \text{ finite} \subset \mathbb{N}} \sum_{n\in I} a_n$$

and the right hand side shows that the sum of the series does not change under a rearrangement. □

A quite useful result about random variables taking nonnegative integer values is this:

Proposition. *Let X be a r.v. with non-negative integer values. Then*

$$EX = \sum_{k\geq 1} P(X \geq k) = \sum_{k\geq 0} P(X > k).$$

Proof. In this case $E[X] = \sum_{n=0}^{\infty} nP(X = n)$, so we are adding up $P(X = 0)$ zero times, $P(X = 1)$ once, $P(X = 2)$ twice, etc. This can be represented in array form, where we're adding column by column:

$$\begin{array}{ccccccc} P(X=1) & P(X=2) & P(X=3) & P(X=4) & P(X=5) & \dots \\ & P(X=2) & P(X=3) & P(X=4) & P(X=5) & \dots \\ & & P(X=3) & P(X=4) & P(X=5) & \dots. \\ & & & P(X=4) & P(X=5) & \dots \\ & & & & P(X=5) & \dots \end{array}$$

Since (from the Lemma above) we can rearrange positive series arbitrarily we can add up these numbers row by row and get the same result. The first row has everything but $P(X = 0)$ and so its total is $P(X > 0)$. The second row has everything but $P(X = 0)$ and $P(X = 1)$ so it has $P(X > 1)$; in general, the sum of row $n + 1$ is $P(X > n)$, and so adding row by row gives $E[X] = \sum_{n=0}^{\infty} P(X > n)$.[2] □

Example (The geometric distribution). If $T \sim \text{Geom}(p)$ - that is $P(T = k) = (1 - p)^{k-1}p$ - we have $P(T > k) = (1 - p)^k$ so we immediately find $ET = \sum_{k\geq 0}(1 - p)^k = 1/p$.

Example (How long it takes to Tombola). This is from G. Letta, *Probabilità Elementare*. Tombola is a kind of Bingo. There are 90 balls numbered from 1 to 90 and are extracted one by one, without replacement. You are the only player and have a "card" with 15 numbers. When all of them come out you make Tombola. What is the expected number of draws required to complete

[2] A more "professional" argument is the following. Here $X(\omega) = \sum_{k\geq 1} \mathbf{1}_{\{X\geq k\}}(\omega)$ - as it is easily verified; and w know that $E\mathbf{1}_{\{X\geq k\}} = P(X \geq k)$; what we do *not* know, and have to accept here, is that in this case we can bring the expectation operator inside the sum. Then we obtain

$$EX = E\sum_{k\geq 1} \mathbf{1}_{\{X\geq k\}} = \sum_{k\geq 1} E\mathbf{1}_{\{X\geq k\}} = \sum_{k\geq 1} P(X \geq k).$$

A different (somewhat pedantic) point is the following: if X can take the value $+\infty$ but $P(X = +\infty) = 0$ (as the time of first success in the Bernoulli scheme) then since $\lim_{n\to\infty} nP(X = n) = 0$ we take $+\infty \cdot P(X = +\infty) = 0$.

the card? Letting T denote the number of draws required to complete the card (which can take integer values from 15 to 90) we want to find $ET = \sum_{k=0}^{89} P(T > k)$.[3]

We work with $\{T \leq k\}$. For this to occur your 15 numbers must be included in the first k draws. From the hypergeometric with two types - your numbers and the other ones - we get

$$P(T \leq k) = \binom{15}{15}\binom{75}{k-15} \Big/ \binom{90}{k} = \binom{k}{15} \Big/ \binom{90}{15}$$

(you can check the last equality). Therefore

$$\begin{aligned} ET &= \sum\nolimits_{k=0}^{89} P(T > k) = \sum\nolimits_{k=0}^{89} [1 - P(T \leq k)] \\ &= 90 - \sum\nolimits_{k=15}^{89} P(T \leq k) = 90 - \binom{90}{15}^{-1} \sum\nolimits_{k=15}^{89} \binom{k}{15} = 90 - \binom{90}{15}^{-1} \binom{90}{16}. \end{aligned}$$

where the last equality is obtained by repeatedly used $\binom{k}{h} = \binom{k+1}{h+1} - \binom{k}{h+1}$.[4] You can check that $ET \approx 85.3$, a surprisingly high number.

4 Conditional expectation

This concept is important and useful. Again we consider discrete variables. The useful result is that we can "disintegrate" expectations as well as probabilities.

4.1 Conditioning with respect to a random variable

Remember that X and Y the conditional density of Y given X is

$$p(y|x) = \frac{p(x,y)}{p_X(x)} \equiv \frac{P(X = x, Y = y)}{P(X = x)}.$$

[3] For the curious here is a formal $(\Omega, \mathcal{F}, P)$ model. Take Ω to be the set of permutations of the integers from 1 to 90 so $\omega = (\omega_1, \ldots, \omega_{90})$. Since Ω is finite, $\mathcal{F}$ can be the power set of Ω; and let P be the uniform probability measure on Ω. Define $Y_k(\omega) = \{\omega_i \colon 1 \leq i \leq k\}$, representing the first k numbers drawn. Letting C denote the set of numbers on your card, the number of elements of C drawn in the first k draws is $Z_k(\omega) = \#\{C \cap Y_k(\omega)\}$. The number of draws required to complete the card is $T(\omega) = \min\{k \colon Z_k(\omega) = 15\}$.

[4] You should know that $\binom{n+1}{k+1} = \binom{n}{k} + \binom{n}{k+1}$: the combinations $\binom{n+1}{k+1}$ can be divided into those that contain a certain element - which are $\binom{n}{k}$ - and those that do not - which are $\binom{n}{k+1}$. This implies $\binom{n}{k} = \binom{n+1}{k+1} - \binom{n}{k+1}$. Therefore

$$\begin{aligned} \sum\nolimits_{k=15}^{89} \binom{k}{15} &= 1 + \left[\binom{17}{16} - \binom{16}{16}\right] + \left[\binom{18}{16} - \binom{17}{16}\right] + \cdots + \left[\binom{90}{16} - \binom{89}{16}\right] \\ &= 1 - \binom{16}{16} + \binom{90}{16} = \binom{90}{16}. \end{aligned}$$

The conditional expectation of Y given $X = x$ is simply the expectation of Y computed using the conditional density:

$$E(Y|X = x) = \sum\nolimits_{y\in Y(\Omega)} y\, p(y|x).$$

Clearly if X and Y are independent then $E(Y|X = x) = EY$ for all x.

It is then natural to define the random variable $E(Y \mid X)$ on the image of X as taking the value $E(Y|X = x)$ at $x \in X(\Omega)$.

Proposition (Iterated Expectations). $E\left[E(Y \mid X)\right] = EY$.

Proof. The statement is $\sum_{x\in X(\Omega)} E(Y|X = x)P(X = x) = EY$. But

$$\begin{aligned}\sum\nolimits_{x\in X(\Omega)} E(Y|X = x)P(X = x) &= \sum\nolimits_{x\in X(\Omega)} \left[\sum\nolimits_{y\in Y(\Omega)} y\,\frac{p(x,y)}{p_X(x)}\right] p_X(x)\\ &= \sum\nolimits_{y\in Y(\Omega)} y \sum\nolimits_{x\in X(\Omega)} p(x,y) = \sum\nolimits_{y\in Y(\Omega)} y\, p_Y(y) = EY.\end{aligned}$$

□

Example. A miner is trapped in a mine containing 3 doors. The first door leads to a tunnel that will take him to safety after 3 hours of travel. The second and third doors lead to tunnels that will return him to the starting point, after after 5 and 7 hours of travel respectively. Assume that the miner is equally likely to choose any one of the doors *at any stage* (for instance because he does not remember past choices). We wish to find the expected length of time until he reaches safety.[5]
To compute it let Y denote the amount of time (in hours) until the miner reaches safety, and let X denote the door he initially chooses. Therefore $P(X = i) = 1/3$, $i = 1, 2, 3$ and

$$\begin{aligned}E[Y] &= E[Y \mid X = 1]P\left(X = 1\right) + E[Y \mid X = 2]P\left(X = 2\right) + E[Y \mid X = 3]P\left(X = 3\right)\\ &= \frac{1}{3}\big(E[Y \mid X = 1] + E[Y \mid X = 2] + E[Y \mid X = 3]\big).\end{aligned}$$

On the other hand $E[Y \mid X = 1] = 3$, $E[Y \mid X = 2] = 5 + E[Y]$ and $E[Y \mid X = 3] = 7 + E[Y]$. For instance $E[Y \mid X = 2] = 5 + E[Y]$ because if he takes the second door he walks 5 hours and then the situation is the same as the initial one. Hence $E[Y] = \big(3+5+E[Y]+7+E[Y]\big)/3$, which gives $E[Y] = 15$.
Of course it is crucial to assume that the miner keeps choosing at random when at the starting point. The result would be different if we assumed that he remembers his past choices and wants to get out as soon as he can.

[5] Taken from Ross *First Course*

4.2 Conditioning with respect to a partition

For an event $A \subseteq \Omega$ (of positive probability) define

$$E(Y \mid A) = \sum\nolimits_{y \in Y(\Omega)} y \cdot P(Y = y \mid A) = \frac{1}{P(A)} \sum\nolimits_{y \in Y(\Omega)} y \cdot P\left(\{Y = y\} \cap A\right).$$

Then it follows directly that if $\{A_i\}$ is a partition of Ω then

$$\begin{aligned}\sum\nolimits_i P(A_i) E(Y \mid A_i) &= \sum\nolimits_i \sum\nolimits_y y P\left(\{Y = y\} \cap A_i\right) = \sum\nolimits_y y \sum\nolimits_i P\left(\{Y = y\} \cap A_i\right) \\ &= \sum\nolimits_y y P\left(Y = y\right) = EY.\end{aligned}$$

Example (How long does it take to the first HH). We toss a coin repeatedly, where probability of H is p. Let Y be the number of tosses up to the first HH. What is EY. We know that here $\Omega = \{H, T\}^{\mathbb{N}}$ and $Y(\omega) = \inf\{n \colon (\omega_{n-1}, \omega_n) = (H, H)\}$. Partition Ω into the following 3 sets: $T \equiv \{\omega \colon \omega_1 = T\}$, $HT \equiv \{\omega \colon (\omega_1, \omega_2) = (H, T)\}$ and $HH \equiv \{\omega \colon (\omega_1, \omega_2) = (H, H)\}$, with respective probabilities $1 - p, p(1 - p)$ and p^2 (which sum to 1). Then by conditioning we find $EY = (1 + p)/p^2$, since

$$\begin{aligned}EY &= (1 - p) E(Y \mid T) + p(1 - p) E(Y \mid HT) + p^2 E(Y \mid HH) \\ &= (1 - p)(1 + EY) + p(1 - p)(2 + EY) + p^2 * 2 = (1 - p^2) EY + 1 + p.\end{aligned}$$

5 Moments of order higher than 1

The definitions and results of this section are actually valid for *general* random variables. Since we have defined expectations only for discrete random variables we continue to suppose the variables are discrete; but be aware that with a well defined notion of expectation in place all of this section applies to the more general setting.

Assuming as usual that the series involved converge absolutely, letting $\mu = EX$ take $g(x) = x^k$ and $g(x) = (x - EX)^k$; we obtain

$$E\left(X^k\right) = \sum\nolimits_i x_i^k p(x_i) \quad \text{and} \quad E\left((X - \mu)^k\right) = \sum\nolimits_i (x_i - \mu)^k \, p(x_i).$$

These are called *moments* and *central moments*. The most relevant is the central moment of order 2 which is called the *variance* of X:

$$\mathrm{Var}(X) := E\left((X - \mu)^2\right) = E\left(X^2\right) - \left(E\left(X\right)\right)^2.$$

The standard deviation of X is the square root of the variance: $\mathrm{std}(X) = \sqrt{\mathrm{Var}(X)}$.

Example (Don't keep doubling against the House). An event has a probability of success

equal to p, and the House accepts bets on independent repetitions; as usual $q = 1 - p$. You play as follows: on the first toss you bet 1 Euro; if you win you stop; if you lose you double the bet, continuing in this manner until you either win or run out of money. What is the expected gain G of this strategy?

So the game unfolds as follows: at the first shot you and the House put down 1 Euro; if you win you take the 2 Euros on the table and gain 1 Euro (net gain of course). If you lose, at the second shot you put down 2 more Euros; if you win you take the 4 on the table - and again your gain is the only 1 Euro the House put down at the beginning. If you lose you repeat, and at whatever shot you win you gain that same 1 Euro and no more.

What is your cumulative loss? At the first shot you put $1 = 2^0$, at the second $2 = 2^1$, and at the k-th shot you put 2^{k-1}; therefore if you lose n times your total loss is $\sum_{k=1}^{n} 2^{k-1} = (2^n - 1)/(1-2) = 2^n - 1$. So if you have $2^n - 1$ Euros you can afford to play at most n times. If success occurs after the n-th shot - probability q^n - you lose everything, meaning net gain $-(2^n - 1)$.

Let G denote your net gain. What is EG? Easy: since if you win you gain 1 (no matter at which shot)

$$EG = 1 * (1 - q^n) - (2^n - 1) * q^n = 1 - (2q)^2 .$$

If $p < 1/2$ (as is the rule in every casino) then $q > 1/2$ so $EG < 0$. But suppose $p = 1/2$; then $EG = 0$. Observe that it is the same expected gain you get if you play just once. In that case you just bet 1 Euro on a fair coin, with expected gain $1 * 1/2 + (-1) * 1/2 = 0$.

Are they the same bets? Not at all. In the latter case the variance is clearly 1. Since $EG = 0$ the variance in the doubling strategy is easy to compute; for $p = 1/2$ we have

$$\mathrm{Var}(G) = 1 * (1 - q^n) + (2^n - 1)^2 * q^n = 1 + q^n \left[2^{2n} - 2^{n+1}\right] = 1 + 2^n - 2 = 2^n - 1.$$

This is exactly the value of your initial capital: the higher the stakes the riskier the bet.

5.1 Moments of order k

The following result has practical importance:

Proposition. *(i) If X has finite moment of order k the same holds for any $r \leq k$. (ii) If X and Y have finite moment of order k also $X + Y$ does. (iii) If X has finite moment of order k it also has finite* central *moment of that order.*

Proof. For any $x > 0$ we have $x^r \leq 1 + x^k$ - because $x \leq 1 \Rightarrow x^r \leq 1$ and $x > 1 \Rightarrow x^r \leq x^k$. From this (i) follows. For (ii), we shall show in a moment that $|x + y|^k \leq 2^{k-1} \left(|x|^k + |y|^k\right)$. Therefore $E\left(|X + Y|^k\right) \leq 2^{k-1} \left[E\left(|X|^k\right) + E\left(|Y|^k\right)\right]$ (by linearity of E), which is finite by assumption. To prove the stated inequality observe first that for $x \geq 0$ and $k \geq 1$ the function $f(x) = x^k$ is convex, hence $f\left(\frac{1}{2}|x| + \frac{1}{2}|y|\right) \leq \frac{1}{2} f(|x|) + \frac{1}{2} f(|y|)$ which reads $\left(\frac{1}{2}|x| + \frac{1}{2}|y|\right)^k \leq \frac{1}{2}|x|^k + \frac{1}{2}|y|^k$; multiplying

by 2^k and recalling that $|x+y| \leq |x| + |y|$ yields the result. For (iii) just take $X + Y = X + (-\mu)$ in (ii). □

5.2 Basics on variance and covariance

Talking about variance, first of all why this power of 2 in $E(X-\mu)^2$, and why then $\text{std}(X) = \sqrt{\text{Var}(X)}$? Of course we $E(X-\mu) = 0$ so that gives no information. A good alternative would be $E|X-\mu|$ - but the problem there is that the absolute value function is not everywhere differentiable. So we move to the next exponent and consider $E(X-\mu)^2$ which is manageable. Now observe that if X is $-1, 1$ with equal probabilities then $\text{Var}(X) = 1$, but if X is $-10, 10$ with same probabilities then $\text{Var}(X) = 100$ - of course, since we are squaring everything. On the other hand the standard deviation scales well: $\text{std}(X) = 10$ - because by taking square root we go back to exponent 1 as it were. This is why the standard deviation is so widely used.

We next consider the *covariance* between X and Y, which is defined as

$$\text{Cov}(X,Y) := E((X-EX)(Y-EY)) = E(XY) - E(X)E(Y).$$

It is positive if on average the two variables are high together and low together. Observe that *independence implies zero covariance*. The converse is not true (because zero covariance only implies linear independence), as in the following classical example.

Example (Nonlinear dependence.). Let X have uniform density on $\{-1,0,1\}$: $p_X(i) = 1/3$, $i = -1,0,1$; and let $Y = X^2$, so that $p_Y(0) = 1/3, p_Y(1) = 2/3$. Obviously X and Y are not independent - for example, letting p be their joint density, $p(0,0) = p_X(0) = 1/3$ but $p_X(0)p_Y(0) = 1/9$.[6] On the other hand $XY = i \iff X = i$ so that $E(XY) = 0 = EXEY$.

The following results *are often used* (you should verify them as exercise):

Proposition. *For random variables, not necessarily distinct:*

(i) $\text{Var}(aX) = a^2 \text{Var}(X)$

(ii) $\text{Var}(a+X) = \text{Var}(X)$

(iii) $\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$*, and more generally*

$$\text{Var}\left(\sum_{i=1}^n X_i\right) = \sum_i \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j)$$

(iv) Covariance operator is linear: $\text{Cov}(X, aY + bZ) = a\,\text{Cov}(X,Y) + b\,\text{Cov}(X,Z)$

Example. $\text{Cov}(X, aX + Z) = a\,\text{Var}(X) + \text{Cov}(X,Z)$.

[6] You may complete the specification of the joint density as an exercise.

Example (Variance of Binomial and Poisson). Just to have them in sight:

(i) $B_{n,p}$. If $X \sim B_{n,p}$ it is the sum of n independent Bernoulli$_p$ variables. By direct computation $\text{Var}(\text{Bernoulli}_p) = p(1-p)$ so by independence we get $\text{Var}(B_{n,p}) = np(1-p)$. Remember this one.

(ii) Poisson$_\lambda$. We apply $\text{Var}(X) = EX^2 - (EX)^2$. If $X \sim \text{Poisson}_\lambda$ then

$$EX^2 = e^{-\lambda} \sum\nolimits_{k\geq 0} k^2 \frac{\lambda^k}{k!} = e^{-\lambda}\lambda \sum\nolimits_{k\geq 1} k \frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda \sum\nolimits_{n\geq 0} (n+1)\frac{\lambda^n}{n!} = \lambda E(X+1)$$

so $EX^2 = \lambda(\lambda+1)$ and from this $\text{Var}\left(\text{Poisson}_\lambda\right) = \lambda$. Observe that in this case $E(X) = \text{Var}(X)$: the Poisson is on average higher and more volatile the higher its parameter λ.

5.3 The correlation coefficient

The *correlation coefficient* between the variables X and Y is defined as

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\text{std}(X)\cdot\text{std}(Y)}.$$

It is the covariance standardized, in the sense that $|\rho| \leq 1$. More precisely, the correlation coefficient measures the degree of *linear* dependence between the two variables.

Proposition. *(i)* $|\text{Corr}(X,Y)| \leq 1$. *(ii)* $|\text{Corr}(X,Y)| = 1$ *iff with probability 1 it is* $Y = aX + b$, *with*

$$a = \frac{\text{Cov}(X,Y)}{\text{Var}(X)},\ b = EY - aEX.$$

Proof. (i) Consider the function

$$\begin{aligned} h(t) &= E\left(t(X-EX) + (Y-EY)\right)^2 \\ &= t^2\,\text{Var}(X) + 2t\,\text{Cov}(X,Y) + \text{Var}\left(Y\right). \end{aligned}$$

This is always positive so the discriminant is at most zero, that is

$$[\text{Cov}(X,Y)]^2 \leq \text{Var}(X)\cdot\text{Var}(Y).$$

which implies that $|\text{Corr}(X,Y)| \leq 1$.[7] (ii) In addition, $|\text{Corr}(X,Y)| = 1$ iff the above discriminant is zero, so that $h(t)$ has unique root

$$t_0 = -\frac{\text{Cov}(X,Y)}{\text{Var}(X)}$$

and one has $t_0(X-EX) + (Y-EY) = 0$ *with probability 1.* In other words with probability 1 one

[7] The displayed inequality is a simple version of the *Cauchy-Schwarz inequality.*

has

$$Y = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \cdot X + (EY + t_0 EX) \equiv aX + b$$

where, given $|\mathrm{Corr}(X,Y)| = 1$, we have $a > 0 \iff \mathrm{Cov}(X,Y) > 0 \iff \rho = 1$. □

Example. If $X + Y = c$ with probability 1 then $\mathrm{Corr}(X,Y) = -1$.

Example (Hedging in financial markets). This is of course a toy model - financial markets are far more complex, but still we can get the idea of balancing portfolios. A number from 1 to 8 will be drawn at random (each has probability 1/8 of being selected). You can bet 1 (Euro) on any set A of four numbers, so that your net gain is $X = 2\mathbf{1}_A - 1$ (1 on A and -1 on A^c). Since $\#A = 4$ we have $P(A) = 1/2$ and $EX = 0$; notice also that $\mathrm{std}(X) = 1$.

Now suppose you may also bet 1 Euro on a second four-element set B; your gain on the bet is $Y = 2\mathbf{1}_B - 1$, also with $EY = 0$. The total gain is then

$$S = X + Y = 2(\mathbf{1}_A + \mathbf{1}_B - 1).$$

Whatever the choices of A and B you get $ES = 0$ - but the variability of S, as expressed by $\mathrm{Var}(S)$, does depend on the choice. We have

$$\mathrm{Var}(S) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y) = 2\,[1 + \mathrm{Corr}(X,Y)]$$

where the last equality follows from $\mathrm{std}(X) = \mathrm{std}(Y) = 1$. You can easily check that $\mathrm{Corr}(X,Y) = \mathrm{Corr}(\mathbf{1}_A, \mathbf{1}_B) = \mathrm{Cov}(\mathbf{1}_A, \mathbf{1}_B)/(1/4)$ since $\mathrm{std}(\mathbf{1}_A) = \mathrm{std}(\mathbf{1}_B) = 1/2$ from $P(A) = P(B) = 1/2$; now $\mathbf{1}_A \cdot \mathbf{1}_B = \mathbf{1}_{AB}$, so $E(\mathbf{1}_A \cdot \mathbf{1}_B) = P(AB) = \#(AB)/8$, whence $\mathrm{Cov}(\mathbf{1}_A, \mathbf{1}_B) = \#(AB)/8 - 1/4 = [\#(AB) - 2]/8$; in conclusion

$$\mathrm{Var}(S) = 2\,[1 + \mathrm{Corr}(X,Y)] = 2\,[\#(AB)/2] = \#(AB)$$

which says that the variance of the gain depend on the size of the overlap between A and B. The two extreme cases are:

Complete Hedging: $\#(AB) = 0$. Here $B = A^c$ and $S = 0$ with probability 1: whenever you win on A you lose on B and vice versa.

Double your bet: $\#(AB) = 4$. Here $B = A$ so you are effectively betting 2 Euros on A, and $\mathrm{std}(S) = 2$ - twice as much as in the 1-Euro bet. This is the case of "all eggs in one basket".

Independence of bets: $\#(AB) = 2$. This is an intermediate case. The $\mathrm{std}(S) = \sqrt{2}$. Here $P(AB) = P(A)P(B)$ so that A and B are independent, which implies that the two bets X and Y are independent too. It is as if you bet on A in two independent games/trials with probability of success 1/2. The number of successes is $\mathbf{1}_A + \mathbf{1}_B \sim \mathrm{Binom}(2, 1/2)$.

5.3.1 Correlation and conditioning

The idea is simple. Consider height and reading ability in children. It is reasonable to assume that they are independent, but if you measure the two variables in a sample of children in elementary school (5 to 10 years of age) you surely find a positive correlation. The reason is that both are positively correlated with age: higher children are older hence read better. To find reliable results you have to condition on age, and measure the two variables in a sample of children in the same age group. In this case the solution is simple, but you can imagine that this problem comes up in a great varieties of contexts, and the "confounding" variables are usually not at all easy to spot. The same clearly occurs if you take different age groups in different moments and then pool all data and compute correlation, as in the following example.[8]

To set the stage consider a Bernoulli scheme with n repetitions and probability p of success on the basic space: $\Omega = \{0,1\}^n$ with $\text{Prob}(1) = p$; $q \equiv 1-p$. Let $X = \sum_i \omega_i$ and $Y = n - X$ the number of successes and failures. Since $X + Y = n$ we have $\text{Corr}(X,Y) = -1$.

Now suppose we repeat the above Bernoulli scheme but let the number of repetitions be a random variable N, taking values $n \in \mathbb{N}$. Given $N = n$ both $X \sim B_{n,p}$ and $Y \sim B_{n,q}$, so

$$E(X \mid N=n) = np, \quad E(Y \mid N=n) = nq \quad \text{and} \quad \text{Var}(X \mid N=n) = \text{Var}(Y \mid N=n) = npq.$$

For the covariance, using iterated expectations we have

$$\text{Cov}(X,Y) = E(XY) - EXEY = E\left[E(XY \mid N)\right] - E\left[E(X \mid N)\right] E\left[E(Y \mid N)\right]$$

and we need $E(XY \mid N)$. To compute it *fix* $N = n$ and drop the conditioning notation. We have $X + Y = n$ so $\text{Var}(X+Y) = 0$, and from $0 = \text{Var}(X) + \text{Var}(Y) + 2\,\text{Cov}(X,Y)$ we obtain $\text{Cov}(X,Y) = -npq$; then from $-npq = E(XY) - EXEY = E(XY) - n^2pq$ we obtain - reinstating the proper notation - $E(XY \mid N=n) = \left(n^2 - n\right)pq$ or $E(XY \mid N) = \left(N^2 - N\right)pq$. From this we arrive at

$$\text{Cov}(X,Y) = pq\left[E\left(N^2\right) - E(N) - \left(E(N)\right)^2\right] = pq\left[\text{Var}(N) - E(N)\right].$$

In conclusion, inserting a variable as N with which both X and Y are correlated the covariance between X and Y becomes positive if $\text{Var}(N) > E(N)$.

Incidentally, $\text{Cov}(X,N) = p\,\text{Var}(N)$ and $\text{Cov}(Y,N) = q\,\text{Var}(N)$; details in footnote.[9]

[8] This section is based on Pitman, *Probability,* chapter 6.

[9] $\text{Cov}(X,N) = E(XN) - EXEN$. First, $EX = E(E(X \mid N)) = E(Np) = pEN$. Next $E(XN \mid N)$; here for each n we have $E(XN \mid N=n) = E(Xn \mid N=n) = nE(X \mid N=n)$, in other words $E(XN \mid N) = NE(X \mid N) = N^2p$. We have obtained $\text{Cov}(X,N) = E(pN^2) - p\left(EN\right)^2 = p\,\text{Var}(N)$. The other is analogous.

5.4 Conditional variance

We define *conditional variance* using conditional mean as the variable

$$\text{Var}(Y \mid X) := E[Y^2 \mid X] - [E(Y|X)]^2 .$$

Then $\text{Var}(Y)$ may be decomposed as follows:

Proposition ("Analysis of Variance"). $\text{Var}(Y) = \text{Var}(E(Y \mid X)) + E(\text{Var}(Y \mid X))$.

Proof. The two terms on the right are

$$\text{Var}(E(Y \mid X)) = E\left\{[E(Y|X)]^2\right\} - \{E\left[E(Y \mid X)\right]\}^2 = E\left\{[E(Y|X)]^2\right\} - (EY)^2$$
$$E(\text{Var}(Y \mid X)) = E\left\{E[Y^2 \mid X] - [E(Y|X)]^2\right\} = E(Y^2) - E\left\{[E(Y|X)]^2\right\}.$$

Adding up we are left with $E(Y^2) - (EY)^2 = \text{Var}(Y)$. □

A concrete relevant example this: you want to explain income Y as a function of education level X. How much of the variability of Y - that is of $\text{Var}(Y)$ - is explained by X? What moves with education is $E(Y \mid X)$ so the explained part is the variance of this variable: $V(E(Y \mid X))$. For each x the variance $V(Y \mid X = x)$ is due to something else - not, X which is fixed here. On average then the unexplained variance is $E(V(Y \mid X))$.

6 Examples and exercises

All those in Baldi Chapter 2, except sections 2.8-2.10. We study sections 2.3-2.5 directly from the book.