Caltech Department of Mathematics

Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2021

Lecture 8: Expectation in Action

Relevant textbook passages:

Pitman [6]: Chapters 3 and 5; Section 6.4–6.5

Larsen–Marx [5]: Sections 3.7, 3.8, 3.9, 3.11, 3.12

8.1 The usefulness of linear operators

We have seen that **expectation is a positive linear operator** on the set $L_1(p)$ of random variables that finite expectation. This enables us to find expectations simply even when the formulas look formidable.

The next example is the basis for the Law of Large Numbers.

8.1.1 Example (Averages and sums) Let $X_1, \dots, X_n$ be random variables each with expectation (mean) μ, and let

$$S_n = X_1 + \cdots + X_n, \quad \text{and} \quad A_n = (X_1 + \cdots + X_n)/n.$$

Then since **expectation is a positive linear operator**,

$$\boldsymbol{E}\, S_n = n\mu \quad \text{and} \quad \boldsymbol{E}\, A_n = \mu.$$

If in addition the random variables are independent and each variance σ^2, then

$$\boldsymbol{Var}\, S_n = n\sigma^2 \quad \text{and} \quad \boldsymbol{Var}\, A_n = \frac{\sigma^2}{n}$$

□

8.1.2 Example (Binomial distribution) Then Binomial(n, p) distribution is the distribution of the number of success in n independent trials when the probability of success in each trial is p. It has the mass function

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, \dots, n.$$

If X is a random variable with this distribution, then

$$\boldsymbol{E}\, X = \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}.$$

This is a not very appealing formula, but there is a simpler way to compute the expectation. The number of successes is simply the sum of n independent Bernoulli(p) random variables $X_1, \dots, X_n$, where X_i is 1 if the i^{th} trial is a success and 0 otherwise. It is trivial to see that

$$\boldsymbol{E}\, X_i = 1p + 0(1-p)p.$$

Since **expectation is a positive linear operator**,

$$\boldsymbol{E}\, X = \boldsymbol{E}(X_1 + \cdots + X_n) = \boldsymbol{E}\, X_1 + \cdots + \boldsymbol{E}\, X_n = np.$$

This proves that

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

We know that for *independent* random variables, the variance of the sum is the sum of the variances, so since the variance of a Bernoulli trial is $(1-p)^2 p + (0-p)^2(1-p) = p(1-p)$, the variance of a Binomial random variable is just $np(1-p)$. By the way, this proves that

$$\sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} - \Big(\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k}\Big)^2 = np(1-p).$$

□

8.2 The "method of indicators"

Since the expectation of the indicator function $\mathbf{1}_E$ of the event E is just $P(E)$, we can use the linearity of expectation to simplify seemingly complicated expressions. Both Pitman [6, pp. 168–174] and Ash [2, § 3.5, pp. 122–124] refer to this practice as the **method of indicators**. Watch for it.

8.2.1 Proposition (The number of events that occur) *Let $E_1, \dots, E_n$ be an arbitrary collection of events, and let X be the number of these events that occur. See Figure 8.1. In terms of the sample space,*

$$X(\omega) = \#\{i : \omega \in E_i\}.$$

Then

$$\boldsymbol{E}\, X = P(E_1) + \cdots + P(E_n).$$

Proof: Note that

$$X = \mathbf{1}_{E_1} + \cdots + \mathbf{1}_{E_n}.$$

So since **expectation is a positive linear operator**,

$$\boldsymbol{E}\, X = \boldsymbol{E}(\mathbf{1}_{E_1} + \cdots + \mathbf{1}_{E_n}) = \boldsymbol{E}\, \mathbf{1}_{E_1} + \cdots + \boldsymbol{E}\, \mathbf{1}_{E_n} = P(E_1) + \cdots + P(E_n).$$

∎

The next example is a special case of the previous one.

8.2.2 Example (Balls in bins) There are n balls and n bins, numbered $1, \dots, n$. The balls are placed in the bins (one ball per bin) randomly (equally likely to put any ball in any bin). Let X be the number of balls placed in the corresponding-numbered bin. The distribution of X is a little complicated, but its expectation is simple. Let E_i be the event that ball i is placed in bin i. These events are not disjoint, and not independent. (For instance, it is impossible for exactly $n-1$ of these events to occur.) But observe that

$$X = \mathbf{1}_{E_1} + \cdots + \mathbf{1}_{E_n}.$$

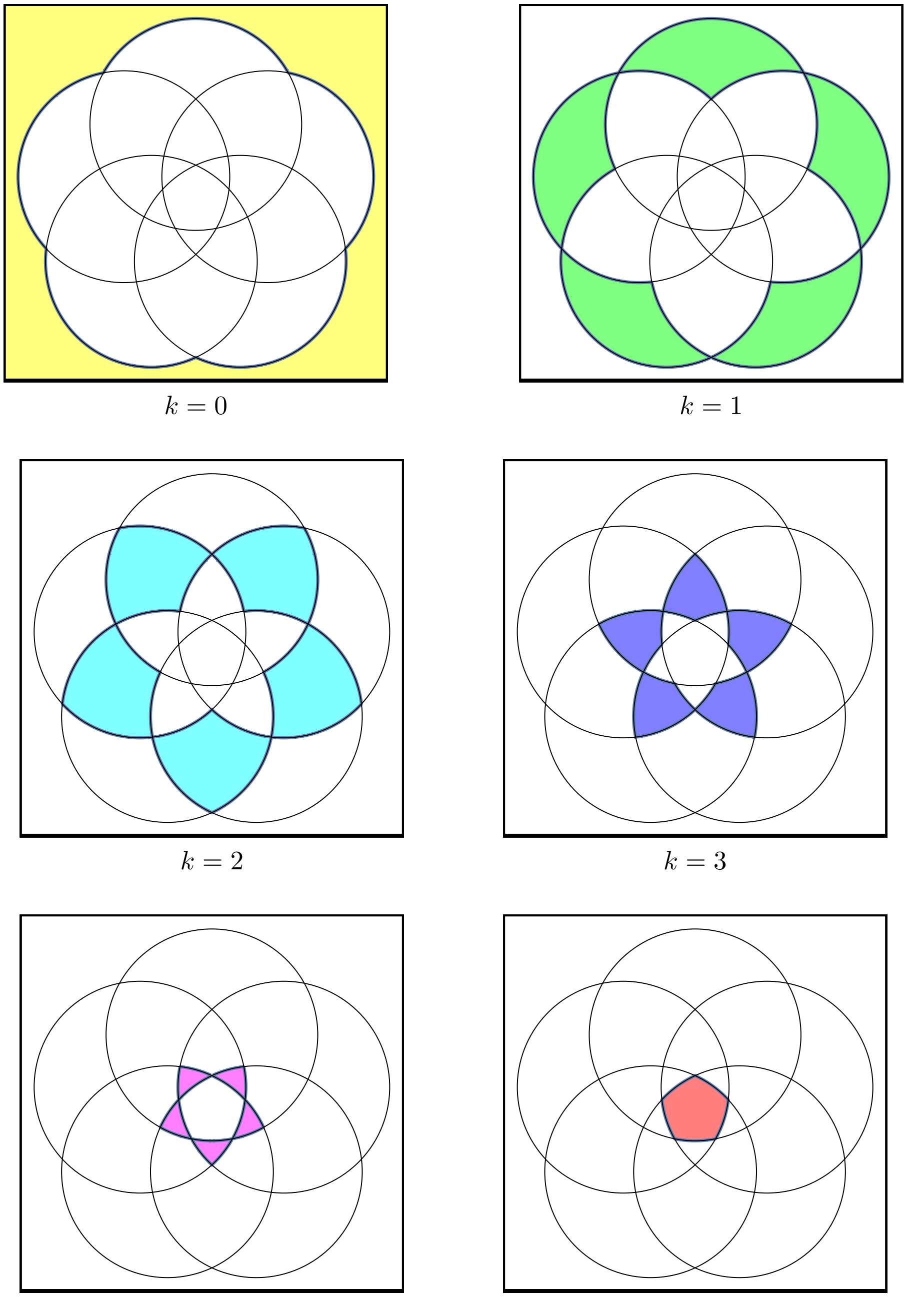
(If putting ball i in bin i is counted as a success, then X is the number of successes. Unlike the binomial case tough, the indicators of success are not independent random variables, but that doesn't matter.) Since **expectation is a positive linear operator**,

$$\boldsymbol{E}\, X = \boldsymbol{E}(\mathbf{1}_{E_1} + \cdots + \mathbf{1}_{E_n}) = P(E_1) + \cdots + P(E_n).$$

But what is the probability of E_i? Since ball i is equally likely to put in any bin, $P(E_i) = 1/n$, so

$$\boldsymbol{E}\, X = 1.$$

□



$k = 0$

$k = 1$

$k = 2$

$k = 3$

$k = 4$

$k = 5$

Figure 8.1. The five events $E_1, \ldots, E_5$ are represented by the five circles. Each figure depicts the set of points that belongs to exactly k of these events, $k = 0, \ldots, 5$.

8.2.3 Example (Binomial revisited) In a sequence of Bernoulli trials, let E_i be the event that the i^{th} trial is a success. Each E_i has probability p. Then the Binomial random variable X which is the number of success in n trials is just the count of the events E_i that have occurred. By Proposition 8.2.1, $\boldsymbol{E}\,X = \sum_{i=1}^{n} P(E_i) = np$. □

Pitman [6, p. 171] uses the above technique to prove the following.

8.2.4 Proposition (Tail probabilities) *Let X be a random variable that only takes on values in the set $\{0, 1, \dots, n\}$. Then*

$$\boldsymbol{E}\,X = \sum_{k=0}^{n} P(X > k).$$

Proof: For $k = 0, \dots, n$, let E_k be the event that $(X > k)$. These are nested and decreasing: $E_0 \supset E_1 \supset \cdots \supset E_n = \varnothing$. When $X = k$, the events $E_0, \dots, E_{k-1}$ occur (and there are k of these), but $E_k, E_{k+1}, \dots, E_n$ do not occur. This means the value of X is simply the number of the events $E_0, \dots, E_n$ that occur, so by Proposition 8.2.1,

$$\boldsymbol{E}\,X = \sum_{k=0}^{n} P(E_k) = \sum_{k=0}^{n} P(X > k),$$

where the second equality is just the definition of E_k. ∎

Compare this to Proposition 6.4.2.

8.3 The Inclusion/Exclusion Principle

I havea voided proving the Inclusion/Exclusion Principle so far, but it is a snap using the method of indicators.

8.3.1 A multinomial formula

Let us refer to the symbols $x_1, \dots, x_n$ as **letters**. We can use these letters to form symbolic sums and products (**polynomials** or **multinomials**) such x_1x_3, $(1 + x_1)$, $(1 + x_1)(1 + 3x_2)$, etc. Terms in a multinomial are scalar multiples of symbolic products of letters. The **degree** of a term is the sum of the exponents of letters in the term. By convention the product of zero letters is 1 and has degree zero.

The next identity is easy to prove by induction on n.

8.3.1 Proposition (A multinomial identity)

$$(1 + x_1)(1 + x_2)\cdots(1 + x_n) = 1 + \sum_{k=1}^{n} \sum_{i_1<\cdots<i_k} x_{i_1}\cdots x_{i_k} = \sum_{k=0}^{n} \sum_{i_1<\cdots<i_k} x_{i_1}\cdots x_{i_k}. \tag{1}$$

The somewhat unusual notation for the second sum means to sum the product $x_{i_1}\cdots x_{i_k}$ over all sorted lists of letters having length k. The sorting guarantees that we take each set of k distinct letters exactly once. There are $\binom{n}{k}$ such products. For example, for $n = 2$,

$$(1 + x_1)(1 + x_2) = \underbrace{1}_{k=0} + \underbrace{(x_1 + x_2)}_{k=1} + \underbrace{x_1x_2}_{k=2},$$

and there is $\binom{2}{0} = 1$ term of degree $k = 0$, $\binom{2}{1} = 2$ terms of degree $k = 1$, and $\binom{2}{2} = 1$ term of degree $k = n = 2$. Also, for $n = 3$,

$$(1+x_1)(1+x_2)(1+x_3) = \underbrace{1}_{k=0} + \underbrace{(x_1+x_2+x_3)}_{k=1} + \underbrace{(x_1x_2+x_1x_3+x_2x_3)}_{k=2} + \underbrace{x_1x_2x_3}_{k=3},$$

which has $\binom{3}{0} = 1$ term of degree $k = 0$, $\binom{3}{1} = 3$ terms of degree $k = 1$, and $\binom{3}{3} = 1$ term of degree $k = 2$, and $\binom{3}{3} = 1$ term of degree $k = n = 3$.

Replacing x_i by $-x_i$ we have the following.

8.3.2 Corollary (Another multinomial identity)

$$(1-x_1)(1-x_2)\cdots(1-x_n) = 1+\sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^k x_{i_1}\cdots x_{i_k} = \sum_{k=0}^{n}\sum_{i_1<\cdots<i_k}(-1)^k x_{i_1}\cdots x_{i_k}. \quad (2)$$

So, for instance,

$$(1-x_1)(1-x_2) = \underbrace{1}_{k=0} - \underbrace{(x_1+x_2)}_{k=1} + \underbrace{x_1x_2}_{k=2},$$

$$(1-x_1)(1-x_2)(1-x_3) = \underbrace{1}_{k=0} - \underbrace{(x_1+x_2+x_3)}_{k=1} + \underbrace{(x_1x_2+x_1x_3+x_2x_3)}_{k=2} - \underbrace{x_1x_2x_3}_{k=3}.$$

8.3.2 The Inclusion/Exclusion Principle

We now harness the expectation operator, and the important fact that **expectation is a positive linear operator**, to prove the mysterious Inclusion/Exclusion Principle (Pitman [6, p. 31]). Kaplansky [4] refers to it as Poincaré's formula, but it is often attributed to de Moivre.

8.3.3 Inclusion/Exclusion Principle *Let $E_1, \dots, E_n$ be an indexed family of events, not necessarily disjoint, nor even distinct apart from the indexing. Then*

$$\begin{aligned} P\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_{i=1}^{n} P(E_i) - \sum_{i_1<i_2} P(E_{i_1}E_{i_2}) + \cdots + (-1)^{n+1}P(E_1\cdots E_n) \\ &= \sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^{k+1}P(E_{i_1}E_{i_2}\cdots E_{i_k}). \end{aligned}$$

Proof: This is the proof outlined in Pitman [6, Exercise 21, p. 184]. Start by using de Morgan's laws to write

$$P\left(\bigcup_{i=1}^{n} E_i\right) = 1 - P\left(\left(\bigcup_{i=1}^{n} E_i\right)^c\right) = 1 - P\left(\cap_{i=1}^{n} E_i^c\right)$$

Recall that Proposition 5.2.1 asserts that

$$\mathbf{1}_{\cap_{i=1}^{n} E_i^c} = \prod_{i=1}^{n} \mathbf{1}_{\boldsymbol{E}_i^c} \quad \text{and} \quad \mathbf{1}_{\boldsymbol{E}_i^c} = 1 - \mathbf{1}_{\boldsymbol{E}_i}.$$

So

$$P\left(\bigcup_{i=1}^{n} E_i\right) = \boldsymbol{E}\,\mathbf{1}_{\bigcup_{i=1}^{n} E_i} = 1 - \boldsymbol{E}\,\mathbf{1}_{\cap_{i=1}^{n} E_i^c} = 1 - \boldsymbol{E}\prod_{i=1}^{n}(1-\mathbf{1}_{E_i}). \quad (3)$$

By the second multinomial identity (2),

$$\prod_{i=1}^{n}(1-\mathbf{1}_{E_i}) = \sum_{k=0}^{n}\sum_{i_1<\cdots<i_k}(-1)^k \mathbf{1}_{E_{i_1}}\cdots\mathbf{1}_{E_{i_k}} = \sum_{k=0}^{n}\sum_{i_1<\cdots<i_k}(-1)^k \mathbf{1}_{(E_{i_1}\cdots E_{i_k})}.$$

We now use the fact that **expectation is a positive linear operator** to conclude

$$\begin{aligned}\boldsymbol{E}\prod_{i=1}^{n}(1-\mathbf{1}_{E_i}) &= \sum_{k=0}^{n}\sum_{i_1<\cdots<i_k}(-1)^k\,\boldsymbol{E}\,\mathbf{1}_{(E_{i_1}\cdots E_{i_k})}\\ &= \sum_{k=0}^{n}\sum_{i_1<\cdots<i_k}(-1)^k P(E_{i_1}\cdots E_{i_k})\\ &= 1+\sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^k P(E_{i_1}\cdots E_{i_k})\end{aligned}$$

Substituting this back into (3) gives

$$\begin{aligned}P\left(\bigcup_{i=1}^{n}E_i\right) &= -\sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^k P(E_{i_1}\cdots E_{i_k})\\ &= \sum_{k=1}^{n}\sum_{i_1<\cdots<i_k}(-1)^{k+1} P(E_{i_1}\cdots E_{i_k})\end{aligned}$$

and the result is proven. ∎

8.4 Higher moments

Recall that in analogy to a balance beam, the expectation $\boldsymbol{E}\,X$ of X was also called the **first moment** of X.

8.4.1 Definition *The* $\boldsymbol{n}^{\mathbf{th}}$ ***moment of*** $\boldsymbol{X}$ *is*

$$\boldsymbol{E}(X^n),$$

the $\boldsymbol{n}^{\mathbf{th}}$ ***central moment of*** $\boldsymbol{X}$ *is*

$$\boldsymbol{E}\big((X-\boldsymbol{E}\,X)^n\big)$$

and the $\boldsymbol{n}^{\mathbf{th}}$ ***absolute moment of*** $\boldsymbol{X}$ *is*

$$\boldsymbol{E}\big(|X|^n\big).$$

(Usually the outer parentheses following $\boldsymbol{E}$ are omitted.) Note that the **variance** of X is the **second central moment**. Recall (Definition 6.13.1) that the set of random variables with finite p^{th} moment is denoted $L_p(P)$.

8.5 Finite higher moments imply finite lower moments

Jensen's Inequality 6.9.4 allows to prove the following useful result about the L_p spaces.

8.5.1 Proposition *If* $1 \leqslant p < q < \infty$*, then*

$$\|X\|_p \leqslant \|X\|_q$$

with equality only if X *is degenerate. Consequently, if the* q^{th} *moment is finite, then so is the* p^{th} *moment, and so*

$$Lp(P) \subset L_q(P).$$

Proof: Assume X is nondegenerate, let $q > p \geqslant 1$, and define $f \colon \boldsymbol{R}_+ \to \boldsymbol{R}_+$ via $f(x) = x^{q/p}$. Then f is strictly convex (its second derivative is strictly positive), and so

$$\|X\|_p^q = \big(\boldsymbol{E}\,|X|^p\big)^{q/p} = f\big(\boldsymbol{E}\,|X|^p\big) < \boldsymbol{E}\,f\big(|X|^p\big) = \boldsymbol{E}\,|X|^q = \|X\|_q^q,$$

where the strict inequality follows from Jensen's Inequality. Now raise each side to the $1/q$ power.

∎

8.6 The "moment problem"

The moments of a random variable depend only the distribution or density of the random variable. Suppose you know *all* of the moments of a random variable. Is this enough to pin down the entire distribution?

The answer to this question is one that economists love to give, namely, "It all depends." Let X be a random variable and let

$$\mu_k = \boldsymbol{E}\,X^k.$$

It is well beyond the scope of this course, but we do have the following theorem. For a proof, see Breiman [3, Proposition 8.49, p. 182].

8.6.1 Theorem *If*

$$\limsup_k \frac{|\mu_k|^{1/k}}{k} < \infty,$$

Then there is at most one distribution F satisfying

$$\mu_k = \boldsymbol{E}_F\,x^k, \quad (k = 1, 2, \dots)$$

What this amounts to is that if all the k^{th} moments are finite and if they do not grow too fast as $k \to \infty$, then the infinite sequence of the moments do pin down the distribution. However, there are examples of distinct distributions that have the same moment sequences. See, for instance, Section I.8 (p. 22) in Shohat and Tamarkin [7].

We shall return to this issue in Section 12.8⋆.

8.7 Moment generating functions

Larsen–Marx [5]: Section 3.12, pp. 207–216

For a random variable X, its **moment generating function** (mgf) M_X is defined by

$$M_X(t) = \boldsymbol{E}\,e^{tX},$$

provided the expectation is finite. For $t = 0$,

$$M_X(0) = 1,$$

which is finite, but the expectation may be infinite for $t \neq 0$. The mgf is most useful if there is an open interval containing 0 on which $M_X(t)$ is finite.

There are several uses for the mgf that appear in more advanced courses, but the name derives from the following fact. If $M(t)$ is finite on an open interval around zero, then

$$\boldsymbol{E}\,X^n = M_X^{(n)}(0),$$

provided the n^{th} moment exists. (Here $M_X^{(p)}(0)$ is the n^{th} derivative of the function M_X at the point 0.) In order for the derivative at 0 to exist we need that M_X is defined on a neighborhood of 0. I claim that if M_X is defined on a neighborhood of zero, then

$$\frac{d}{dt}M(t) = \boldsymbol{E}\,\frac{d}{dt}e^{tX} = \boldsymbol{E}\,Xe^{tX},$$

so for $t = 0$, we have

$$\frac{d}{dt}M(0) = \boldsymbol{E}\, Xe^{0X} = \boldsymbol{E}\, X.$$

Similarly

$$\frac{d^2}{dt^2}M(t) = \boldsymbol{E}\, \frac{d^2}{dt^2}e^{tX} = \boldsymbol{E}\, X^2 e^{tX},$$

so

$$\frac{d^2}{dt^2}M(0) = \boldsymbol{E}\, X^2 e^{0X} = \boldsymbol{E}\, X^2,$$

and so on. For a proof of the claim see Appendix 8.10.1.

Once you have the mgf, since differentiating is usually easier than integrating, it might be easier to find moments this way. And remember that $\boldsymbol{Var}\, X = \boldsymbol{E}(X^2) - (\boldsymbol{E}\, X)^2$, so you can find variances that way.

You can usually find the mgf for a distribution online on `wikipedia.com`, so I'm not going to have you compute a lot of them. But let's do a really simple case.

Let X be a Bernoulli(p) random variable. Then its mgf is

$$M(t) = \boldsymbol{E}\, e^{tX} = pe^{t1} + (1-p)e^{t0} = pe^t + 1 - p.$$

Then $M'(t) = pe^t$ and $M^{(n)}(t) = pe^t$ for all n, so $\boldsymbol{E}\, X^n = M^{(n)}(0) = p$ for all n. This is trivial to verify directly.

Here's another useful result.

8.7.1 Proposition *If X and Y are independent, then*

$$M_{X+Y}(t) = M_X(t)M_Y(t).$$

That is, the mgf of an independent sum is the product of the mgfs.

Proof: By definition

$$M_{X+Y}(t) = \boldsymbol{E}\, e^{t(X+Y)} = \boldsymbol{E}\, e^{tX}e^{tY},$$

and since X and Y are independent we have

$$\boldsymbol{E}\, e^{tX}e^{tY} = (\boldsymbol{E}\, e^{tX})\, \boldsymbol{E}(e^{tY}) = M_X(t)M_Y(t).$$

■

A consequence is that if $X_1, \dots, X_n$ are independent and identically distributed, with common mgf M, then the mgf M_n of their average is

$$M_n(t) = M(t/n)^n.$$

So the first derivative satisfies

$$M_n'(t) = nM(t/n)^{n-1}M'(t/n)/n = M(t/n)^{n-1}M'(t/n).$$

Since $M(0) = 1$, we have $M_n'(0) = M'(0)$, so the first moment of the average is just the first moment of each summand. But we already knew that. Still, it's reassuring when math works.

But here is the real reason that moment generating functions are useful:

8.7.2 Fact *If M_X and M_Y agree on an open interval containing 0, then X and Y have the same distribution.*

So if we can recover the distribution from the mgf, then we have a nice way to find the distribution of an independent sum. The proof of the fact is well beyond the scope of this course. It is stated without proof in Larsen–Marx [5, Theorem 3.12.2, p. 214].

Of course, there is the problem for some random variables $\boldsymbol{E}\, e^{tX}$ might be infinite for $t \neq 0$. That is why hard-core probability uses the **characteristic function** $\varphi(t) = \boldsymbol{E}\, e^{itX}$, where i is the complex square root of -1. This has many of the same features as the mgf: the characteristic function determines the distribution, and the cf of an independent sum is the product of the cfs. Characteristic functions are also well beyond the scope of this course, but you can learn a little about them in the forbidden Appendix 12.12★.

8.8 Skewness

If a random variable X has a pmf or pdf that is symmetric about zero, that is if

$$p(-x) = p(x), \quad \text{or} \quad f(-x) = f(x)$$

then its expectation is zero, provided it exists. In fact for any odd-numbered moment,

$$\boldsymbol{E}\, X^m = \int_{-\infty}^{\infty} x^m f(x)\, dx = 0, \quad m \text{ odd.}$$

If the distribution is not symmetric about zero, then we do not expect the odd moments to be zero. Thus the odd moments can perhaps be used to measure out severe the asymmetry is.

Pitman [6]: p. 198

8.8.1 Definition *The **skewness** of a random variable X (or its distribution) is third moment of its standardization, provided the third moment exists. That is,*

$$\operatorname{skewness} X = \boldsymbol{E}(X^*)^3 = \boldsymbol{E}\, \frac{(X-\mu)^3}{\sigma^3}.$$

Some pictures.

8.8.1 Skewness of an i.i.d. sum

8.8.2 Proposition *Let X and Y be independent random variables with $\boldsymbol{E}\, X = \boldsymbol{E}\, Y = 0$. Then*

$$\boldsymbol{E}(X+Y)^3 = \boldsymbol{E}\, X^3 + \boldsymbol{E}\, Y^3.$$

Proof: Simply use the Binomial Theorem to write

$$(X+Y)^3 = X^3 + 3X^2Y + 3XY^2 + Y^3,$$

and use the fact that **expectation is a positive linear operator** to write

$$\boldsymbol{E}(X+Y)^3 = \boldsymbol{E}\, X^3 + 3\,\boldsymbol{E}(X^2Y) + 3\,\boldsymbol{E}(XY^2) + \boldsymbol{E}\, Y^3,$$

use independence of X and Y to note that

$$\boldsymbol{E}(X+Y)^3 = \boldsymbol{E}\, X^3 + 3\,\boldsymbol{E}(X^2)\underbrace{(\boldsymbol{E}\, Y)}_{=0} + 3\underbrace{(\boldsymbol{E}\, X)}_{=0}(\boldsymbol{E}\, Y^2) + \boldsymbol{E}\, Y^3.$$

■

8.8.3 Corollary *If $X_1, \ldots, X_n$ are independent and identically distributed random variables with common mean μ and standard deviation σ and skewness γ, then*

$$\operatorname{skewness}(X_1 + \cdots + X_n) = \frac{\gamma}{\sqrt{n}}.$$

Proof: Let $S = X_1 + \cdots + X_n$. Then S has mean $n\mu$ and standard deviation $\sqrt{n}\sigma$. The standardization of S_n is then

$$S_n^* = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{X_i - \mu}{\sigma} = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^*.$$

Now

$$\text{skewness}\, S_n = \boldsymbol{E}(S_n^*)^3 = \boldsymbol{E}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i^*\right)^3 = \left(\frac{1}{\sqrt{n}}\right)^3 \sum_{i=1}^n \underbrace{\boldsymbol{E}(X_i^*)^3}_{=\gamma} = \gamma/\sqrt{n},$$

where the penultimate equality follows (by induction) from Proposition 8.8.2. ∎

Thus the skewness of an independent and identically distributed sum also vanishes at the rate $1/\sqrt{n}$. We may also now compute the skewness of a Binomial(n, p) random variable.

8.8.4 Fact *The skewness of a Binomial(n, p) random variable is the skewness of a Bernoulli(p) random variable divided by $\sqrt{n}$, which simple tedious calculation reveals to be $(1-2p)/\sqrt{np(1-p)}$.*

8.9 Conditional Expectation

A **conditional expectation** is simply the expectation of a random variable using a conditional probability.

8.9.1 Example Let X be the value of the roll of a die. Then $p(k) = P(X = k) = 1/6$, $k = 1, \dots, 6$, and

$$\boldsymbol{E}\, X = \sum_{k=1}^6 = kp(k) = \frac{21}{6} = 3\frac{1}{2}.$$

Let A be the event $(X \geqslant 3)$. If you know that A has occurred, you should update your probabilities on the value of X to the new conditional probabilities:

$$P\bigl(X = k \bigm| A\bigr) = \frac{P(X = k \ \&\ k \in A)}{P(A)} \begin{cases} 0 & k = 1, 2 \\ \frac{1/6}{2/3} = \frac{1}{4} & k = 3, 4, 5, 6. \end{cases}$$

The conditional expectation of X given the event A is just the expectation of X computed using the conditional probabilities:

$$\boldsymbol{E}(X \mid A) = \sum_{k=1}^6 kP\bigl(X = k \bigm| A\bigr) = 1 \cdot 0 + 2 \cdot 0 + 3 \cdot \frac{1}{4} + 4 \cdot \frac{1}{4} + 5 \cdot \frac{1}{4} + 6 \cdot \frac{1}{4} = \frac{18}{4} = 4\frac{1}{2}.$$

Consider now the complementary event $A^{\mathrm{c}} = (X < 3)$. Then

$$P\bigl(X = k \bigm| A^{\mathrm{c}}\bigr) = \frac{P(X = k \ \&\ k \in A^{\mathrm{c}})}{P(A^{\mathrm{c}})} \begin{cases} \frac{1/6}{1/3} = \frac{1}{2} & k = 3, 4, 5, 6 \\ 0 & k = 3, 4, 5, 6. \end{cases}$$

Also,

$$\boldsymbol{E}(X \mid A^{\mathrm{c}}) = \sum_{k=1}^6 kP\bigl(X = k \bigm| A^{\mathrm{c}}\bigr) = 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2} + 3 \cdot 0 + 4 \cdot 0 + 5 \cdot 0 + 6 \cdot 0 = \frac{3}{2} = 1\frac{1}{2}.$$

Thus in this case we see that

$$\boldsymbol{E}(X \mid A)P(A) + \boldsymbol{E}(X \mid A^{\mathrm{c}})P(A^{\mathrm{c}}) = 4\frac{1}{2} \cdot \frac{2}{3} + 1\frac{1}{2} \cdot \frac{1}{3} = 3\frac{1}{2} = \boldsymbol{E}\, X.$$

This is not just a lucky coincidence. □

8.9.2 Proposition *Let $B_1, \dots, B_n$ be a partition of the sample space Ω into events where $P(B_i) > 0$, for $i = 1, \dots, n$. Then for any random variable X,*

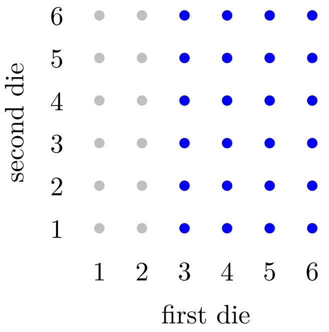
$$\boldsymbol{E}\,X = \sum_{i=1}^{n} \boldsymbol{E}\left(X \mid B_i\right) P(B_i).$$

Proof: To be added.
*********************** ∎

8.9.3 Example Here is a similar case. Roll two dice independently. Let X be the value of the first die, and Y the value of the second die. Let $S = X + Y$. Then

$$\boldsymbol{E}\,X = \boldsymbol{E}\,Y = 3\frac{1}{2}, \qquad \boldsymbol{E}\,S = \boldsymbol{E}\,X + \boldsymbol{E}\,Y = 7,$$

What if I now tell you that event $A = (X \geqslant 3)$ has occurred.? Now the sample space is reduced:



There are now 24 equally likely outcomes, and it is easy to see that the conditional probabilities are

$$P\big(S = 4 \mid A\big) = \frac{1}{24}, \qquad P\big(S = 5 \mid A\big) = \frac{2}{24},$$
$$P\big(S = 6 \mid A\big) = \frac{3}{24} \qquad P\big(S = 7 \mid A\big) = \frac{4}{24},$$
$$P\big(S = 8 \mid A\big) = \frac{4}{24}, \qquad P\big(S = 9 \mid A\big) = \frac{4}{24},$$
$$P\big(S = 10 \mid A\big) = \frac{3}{24}, \qquad P\big(S = 11 \mid A\big) = \frac{2}{24},$$
$$P\big(S = 12 \mid A\big) = \frac{1}{24}.$$

The expectation of S computed using the conditional probabilities is the expectation of$S$ conditional on A, and satisfies

$$\boldsymbol{E}(S \mid A) = \sum_{k=4}^{12} kP\big(S = k \mid A\big) = 8.$$

Another way to compute this is to note that

$$\boldsymbol{E}(S \mid A) = \boldsymbol{E}(X \mid A) + \boldsymbol{E}(Y \mid A) = \boldsymbol{E}(X \mid A) + \boldsymbol{E}\,Y = 4\frac{1}{2} + 3\frac{1}{2} = 8,$$

where we use the fact that Y is independent of X, and so of the event A, so $P\big(Y = y \mid A\big) = P(Y = y)$, which in turn implies $\boldsymbol{E}(Y \mid A) = \boldsymbol{E}\,Y$. □

8.10 Appendix: Parametrized expectations

Let X be a random variable on the probability space $(\Omega, \mathcal{F}, P)$ with range in $A \subset \boldsymbol{R}$ and let T be an open interval in $\boldsymbol{R}$. Let $g \colon A \times T \to \boldsymbol{R}$. We are interested in when the function $G \colon T \to \boldsymbol{R}$ defined by

$$G(t) = \boldsymbol{E}\, g\big(X, t\big)$$

is differentiable and satisfies

$$G'(t) = \boldsymbol{E}\, \frac{\partial g(X,t)}{\partial t}.$$

The following is clearly a necessary assumption.

8.10.1 Assumption *For each $t \in T$, the expectation $\boldsymbol{E}\, g\big(X,t\big)$ is finite For each $x \in A$ and $t \in T$, the partial derivative $\frac{\partial g(x,t)}{\partial t}$ exists and moreover for each $t \in T$, the expectation $\boldsymbol{E}\, \frac{\partial g(X,t)}{\partial t}$ is finite.*

But we need a bit more.

8.10.2 Assumption *There is a function $h \colon A \to \boldsymbol{R}_+$ such that*

1. *$\boldsymbol{E}\, h(X)$ is finite, and*
2. *for all $x \in A$, and for all $t \in T$,*

$$\left| \frac{\partial g(x,t)}{\partial t} \right| \leqslant h(x).$$

8.10.3 Theorem *Under Assumptions 8.10.1 and 8.10.2, the function $G \colon T \to \boldsymbol{R}$ defined by*

$$G(t) = \boldsymbol{E}\, g\big(X, t\big)$$

is differentiable and satisfies

$$G'(t) = \boldsymbol{E}\, \frac{\partial g(X,t)}{\partial t}.$$

This theorem is a translation of Aliprantis and Burkinshaw [1, Theorem 24.5, pp. 193–194] into the language of random variables. I will write out a proof at some point.

8.10.1 Application to moment generating functions

Recall that the moment generating function of the random variable X is defined to be

$$M(t) = \boldsymbol{E}\, e^{tX}.$$

Suppose that $M(t)$ is finite for all t in some interval $(-\alpha, \alpha)$ about zero. This entails that $\boldsymbol{E}\, e^{t\,|X|}$ is also finite. For an n, the n^{th}partial derivative of $g(x,t) = e^{tx}$ is $x^n e^{tx}$. Consider $0 < t < \alpha$ and choose $\varepsilon > 0$ so that $0 < t < t + \varepsilon < \alpha$. By assumption $\boldsymbol{E}\, e^{(t+\varepsilon)\,|X|} = \boldsymbol{E}\, e^{\varepsilon\,|X|} e^{t\,|X|}$ is finite. But for $|x|$ large enough,

$$e^{\varepsilon\,|x|} > |x|^n,$$

so

$$|x|^n\, e^{t\,|x|} < h(|x|),$$

where

$$h(x) = \max\{x^n, e^{\varepsilon x}\} e^{tx},$$

which satisfies $\boldsymbol{E}\, h(|X|) < \infty$, so the above theorem applies and

$$M'(0) = \boldsymbol{E}\, x^n e^{0x} = \boldsymbol{E}\, x^n.$$

Bibliography

[1] C. D. Aliprantis and O. Burkinshaw. 1998. *Principles of real analysis*, 3d. ed. San Diego: Academic Press.

[2] R. B. Ash. 2008. *Basic probability theory*. Mineola, New York: Dover. Reprint of the 1970 edition published by John Wiley and Sons.

[3] L. Breiman. 1968. *Probability*. Reading, Massachusetts: Addison Wesley.

[4] I. Kaplansky. 1944. Symbolic solution of certain problems in permutations. *Bulletin of the American Mathematical Society* 50(12):906–914.
http://projecteuclid.org/euclid.bams/1183506627

[5] R. J. Larsen and M. L. Marx. 2012. *An introduction to mathematical statistics and its applications*, fifth ed. Boston: Prentice Hall.

[6] J. Pitman. 1993. *Probability*. Springer Texts in Statistics. New York, Berlin, and Heidelberg: Springer.

[7] J. A. Shohat and J. D. Tamarkin. 1943. *The problem of moments*. New York: American Mathematical Society.