Ma 3/103 KC Border
Introduction to Probability and Statistics Winter 2017

# 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 of random variables that finite expectation). This enables 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) $\mu$, 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 $\sigma^2$, 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 $i^{\text{th}}$ 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 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

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

□

**8.1.3 Example (Bins and Balls)** 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

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

□

**8.1.4 Example (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 the events that occur. Then

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

so since expectation is a positive linear operator,

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

See Figure 8.1. □

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

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

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

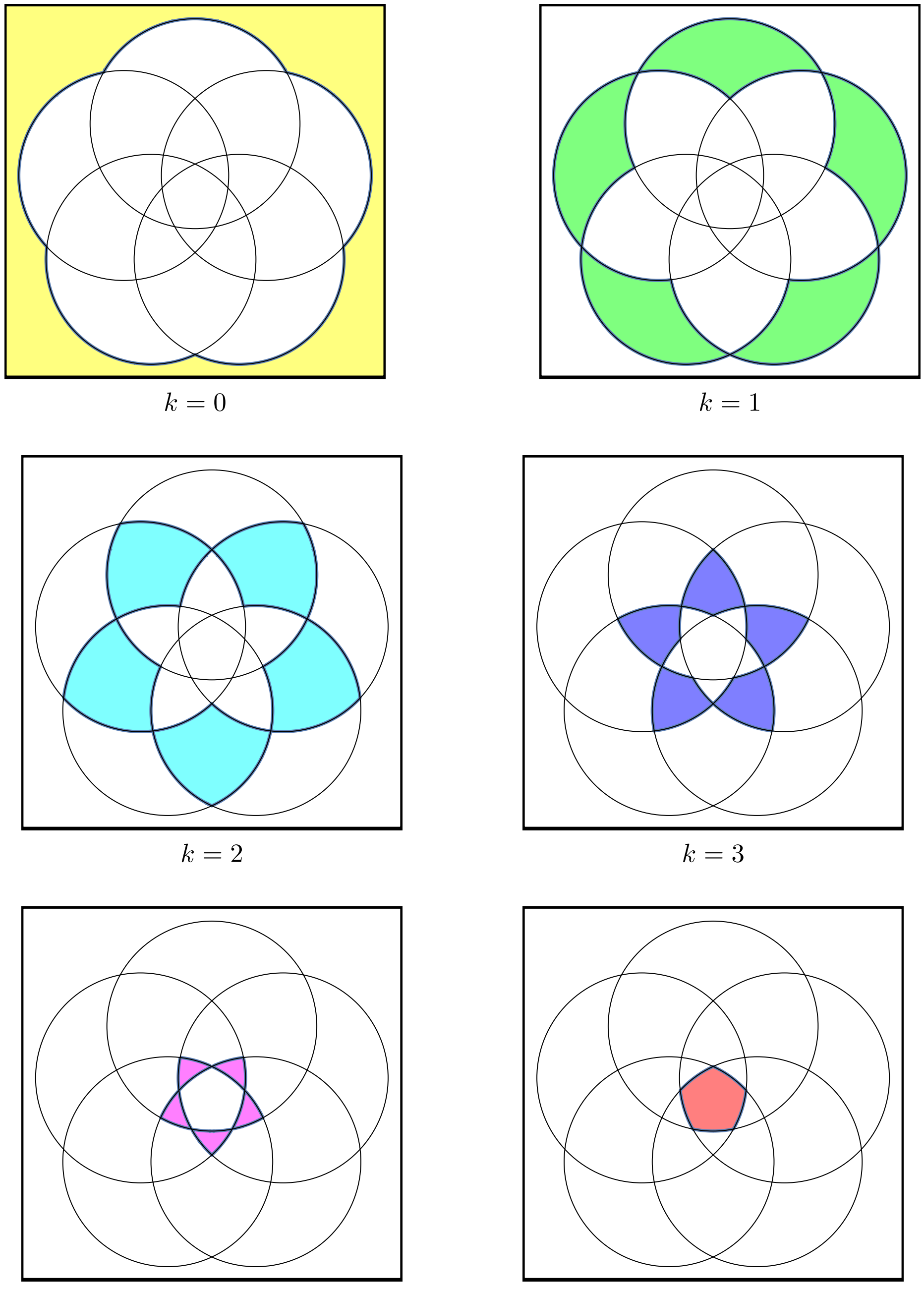



Figure 8.1. The five events $E_1, \ldots, E_5$ are represented by the five circles. Each figure depicts the set of point that belongs to exactly $k$ of these events, $k = 0, \ldots, 5$.

*Proof*: To see this, let $E_k$ be the event that $X > k$, for $k = 0, \ldots, n$. When $X = x$, the events $E_0, \ldots, E_{x-1}$ occur (and there are $x$ of these), but $E_x, E_{x+1}, \ldots, E_n$ do not occur. (Remember, $x$ must be an integer.) This means the value of $X$ is simply the number of the events $E_0, \ldots, E_n$ that occur, so

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

∎

There is a generalization of this formula, but the proof is different. Note that in the above proposition, if $F$ is the cumulative distribution function of $X$, then $P\left(X > k\right) = 1 - F(k)$.

**8.1.6 Proposition (Tail probabilities II)** *Let $X$ be a nonnegative random variable with a density $f$ on the interval $[0, b]$, and cumulative distribution function $F$, so $f = F'$. Then*

$$\boldsymbol{E}\, X = \int_0^b 1 - F(x)\, dx.$$

*Proof*: By definition,

$$\boldsymbol{E}\, X = \int_0^b x f(x)\, dx.$$

Integrating by parts gives

$$\int_0^b x f(x)\, dx = bF(b) - 0F(0) - \int_0^b F(x)\, dx.$$

But $bF(b) = b$, and $\int_0^b 1\, dx = b$, so

$$\boldsymbol{E}\, X = \int_0^b 1 - F(x)\, dx.$$

∎

## 8.2 The Inclusion/Exclusion Principle

### 8.2.1 A multinomial formula

Let us refer to the symbols $x_1, \ldots, x_n$ as **letters**. We can use these letters to form symbolic sums and products (**polynomials** or **multinomials**) such $x_1 x_3$, $(1 + x_1)$, $(1 + x_1)(1 + 3x_2)$, 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 $n$.

**8.2.1 Proposition (A multinomial identity)**

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

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. The case $k = 0$ corresponds to taking no letters, and the product of zero letters is the symbol 1. For example,

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

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

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

**8.2.2 Corollary (Another multinomial identity)**

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

So, for instance,

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

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

### 8.2.2 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.

**8.2.3 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*

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

*Proof*: This is the proof outlined in Pitman [6, Exercise 21, p. 184]. Start by observing that

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

Now use Proposition 5.2.1 to rewrite this in the language of indicator functions.

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

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

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

Substituting this back into (3) proves the result. ∎

**8.2.4 Example (An application)** In many engineering projects redundancy is desirable. If say you want to make sure a safety devise has power, you may give it multiple separate power supplies. As long as any power supply is connected, the device will be operable. If there are $n$ power supplies, and supply $i$ has probability $1-p_i$ of failing, and if the failures are stochastically independent, then what is the probability that the device will be operable?

If $E_i$ is the even that power supply $i$ is operable, then we want $P\left(\bigcup_{i=1}^{n} E_i\right)$. In order to use the Inclusion/Exclusion Principle we need the probabilities of the various intersections, but by independence, these are just the products of the probabilities, so the probability the device will be operable is

$$\sum_{i=1}^{n} p_i - \sum_{i_1<i_2} p_{i_1}p_{i_2} + \cdots + (-1)^{n+1}p_1\cdots p_n).$$

By the way, you might want to examine the assumption of independence. For instance, it might be that the power supplies are lithium ion batteries. When one fails it often catches fire. If the batteries are all in the same location, the fire might spread and destroy all the batteries. In this case, you will have seriously miscalculated your safety margin. □

## 8.3 Higher moments

**8.3.1 Definition** *The* ***$n^{\text{th}}$ moment of $X$*** *is*

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

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

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

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

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

*The set of random variables with finite $n^{\text{th}}$ moment is denoted $L_n(P)$,.*

(Usually the outer parentheses are omitted.)

## 8.4 Higher moments imply lower moments

There is a theorem called **Hölder's Inequality** that has as one of its consequences the following result. This result appears in any good analysis text, but out of laziness I'll just cite Aliprantis–Border [1, Corollary 13.3, p. 463].

**8.4.1 Fact** *If $1 \leqslant p \leqslant q$, if $\boldsymbol{E}|X^q|$ is finite, then $\boldsymbol{E}|X^p|$ is finite.*

## 8.5 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].

**8.5.1 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 $k^{\text{th}}$ 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 11.8$\star$.

## 8.6 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 formula:

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

provided the $n^{\text{th}}$ moment exists. (Here $M_X^{(p)}(0)$ is the $n^{\text{th}}$ 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. To fully understand this result, you have to know when you can differentiate an expectation. I have notes on this on the course web site. But trust me and write

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

so for $t = 0$, we have

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

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.

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 it 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. If $X$ and $Y$ are independent, then the mgf of their sum satisfies

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

and since $X$ and $Y$ are independent we have

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

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

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.

Another really useful result is that if $M_X$ and $M_Y$ agree on an open interval containing 0, then $X$ and $Y$ have the same distribution. If we know how to recover the distribution from the mgf, then we have a nice way to fins the distribution of a independent sum.

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.

## 8.7 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

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

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

Some pictures.

### 8.7.1 Skewness of an i.i.d. sum

**8.7.2 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

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

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

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

∎

**8.7.3 Corollary** *If $X_1, \dots, X_n$ are independent and identically distributed random variables with common mean $\mu$ and standard deviation $\sigma$ and skewness $\gamma$, then*

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

*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 quality follows (by induction) from Proposition 8.7.2. ∎

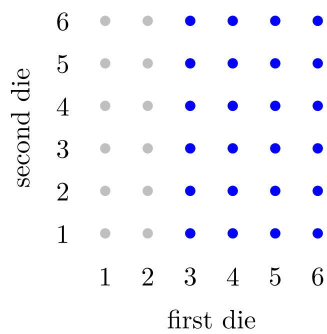
Thus the skewness of an independent and identically distributed sum also vanishes at the rate $1/\sqrt{n}$. We may also now compute the skewness of a Binomial$(n,p)$ random variable.

**8.7.4 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 $(1-2p)/\sqrt{np(1-p)}$.*

## 8.8 Conditional Expectation

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

Here is a very simple case. Roll two die. The expected value of the sum $S$ is 7. But what if I now you that the first die is $\geqslant 3$. Now the sample space is reduced:




There are now only 24 equally likely outcomes, and it is easy to see that

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

Thus

$$\boldsymbol{E}\left(S \,\middle|\, \text{first die} \geqslant 3\right)=8.$$

(You could just multiply each value of $S$ by its probability and sum, but smarter idea is to note that the expected value of the first die given that it is 3 or more is just $4\frac{1}{2}$ and add the expected value of the second die, which is $3\frac{1}{2}$.)

Later we will learn about more sophisticated ways to compute conditional expectations.

## 8.9 Density of a function of a random variable

Find the appropriate textbook passages

If $X$ is a random variable with cumulative distribution function $F$ and density $f=F'$, then $g(X)$ is a random variable.[1] What is the density of $g$? Well if $g$ is strictly increasing, then $g$ has an inverse, and

$$g(X)\leqslant y \iff X\leqslant g^{-1}(y),$$

so

$$F_g(y)=P\left(g(X)\leqslant y\right)=P\left(X\leqslant g^{-1}(y)\right)=F\big(g^{-1}(y)\big)$$

is the cumulative distribution function of $g(X)$. The density $f_g$ of $g(X)$ is found by differentiating this.

If $g$ is strictly decreasing,

$$g(X)\leqslant y \iff X\geqslant g^{-1}(y),$$

and if $f$ has density this is just $1-F\big(g^{-1}(y)\big)$, and we may differentiate that (with respect to $y$, to get the density.

When $g$ is a differentiable strictly increasing function, with $g'>0$, then $g$ has an inverse, and in light of the Inverse Function Theorem [2, Theorem 6.7, p. 252], then

$$f_g(y)=\frac{d}{dy}F\big(g^{-1}(y)\big)=F'\big(g^{-1}(y)\big)\frac{d}{dy}g^{-1}(y)=f\big(g^{-1}(y)\big)\frac{1}{g'\big(g^{-1}(y)\big)}.$$

Or letting $y=g(x)$, $f_g\big(g(x)\big)=f(x)/g'(x)$.

[1] Provided $g$ is a Borel function.

**8.9.1 Example** If $X \sim F_X$ with density $f_X$ and $Y = aX + b$ where $a > 0$, then

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right),$$

so

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

□

**8.9.2 Example** Let $X \sim U[0,1]$, (so $f(x) = 1$ for $0 \leqslant x \leqslant 1$). Let $g(x) = x^a$, so $g'(x) = ax^{a-1}$, $g^{-1}(y) = y^{1/a}$ and $(g^{-1})'(y)(1/a)y^{(1-a)/a}$. Then the density of $g(X) = X^a$ is given by

$$f_g(y) = \frac{1}{ax^{a-1}} = \frac{1}{a} y^{(a-1)/a} \qquad (0 \leqslant y \leqslant 1).$$

□

## Bibliography

[1] C. D. Aliprantis and K. C. Border. 2006. *Infinite dimensional analysis: A hitchhiker's guide*, 3d. ed. Berlin: Springer–Verlag.

[2] T. M. Apostol. 1967. *Calculus*, 2d. ed., volume 1. Waltham, Massachusetts: Blaisdell.

[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.