Examples & Complements

Chapter 4

1 Linearity

When considering more than one RV, we use the notation for intersection of corresponding events

$$P\left[X=x,Y=y,\ldots\right]$$

(See the file on independence on line). The book climbs mirrors to avoid considering more than one RV at a time in this and the following chapter, but it actually uses the facts we are discussing here, so we might as well make them explicit. This is not difficult, since we've been dealing with intersection of events all along.

Consider now

$$E\left[X+Y\right]=\sum_{k,j}\left(x_k+y_j\right)P\left[X=x_k,Y=y_j\right]=$$

$$\sum_k x_k\sum_j P\left[X=x_k,Y=y_j\right]+\sum_j y_j\sum_k P\left[X=x_k,Y=y_j\right]$$

Now, using conditional probabilities, as well as the total probabilities formula (or just thinking how the partition $\{Y=y_j\}$ covers the event $\{X=x_k\}$, and reciprocally for the second term)

$$\sum_j P\left[X=x_k,Y=y_j\right]=\sum_j P\left[X=x_k\left|Y=y_j\right.\right]P\left[Y=y_j\right]=P\left[X=x_k\right]$$

Similarly for the second term, so that the expression is equal to

$$\sum_k x_kP\left[X=x_k\right]+\sum_j y_jP\left[Y=y_j\right]=E\left[X\right]+E\left[Y\right]$$

Combining this result with the fact that $E\left[aX+b\right]=aE\left[X\right]+b$, we conclude that *expectation is a linear operator*:

$$E\left[\sum_i a_iX_i\right]=\sum_i a_iE\left[X_i\right]$$

1.1 Variance of a sum

Having defined independence, we compute, in general

$$E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} EX_i\right)^2\right] = E\left[\left(\sum_{i=1}^{n} X_i\right)^2\right] - \left(\sum_{i=1}^{n} EX_i\right)^2 =$$

$$= E\left[\sum_{i=1}^{n} X_i^2 + 2\sum_{i<j} X_i X_j\right] - \sum_{i=1}^{n} \left(EX_i\right)^2 - 2\sum_{i<j} EX_i EX_j =$$

$$= \sum_{i=1}^{n} \left(E\left[X_i^2\right] - \left(EX_i\right)^2\right) + 2\sum_{i<j} \left(E\left[X_i X_j\right] - EX_i EX_j\right) =$$

$$\sum_{i=1}^{j} Var\left(X_i\right) + 2Cov\left(X_i X_j\right)$$

$Cov\left(X_i X_j\right) = E\left[X_i X_j\right] - EX_i EX_j = E\left[\left(X_i - EX_i\right)\left(X_j - EX_j\right)\right]$ is the "covariance". Note that if the X_i's are independent,

$$E\left[X_i X_j\right] = \sum_{k,l} x_k^i x_l^j P\left[X_i = x_k^i, X_j = x_l^j\right] = \sum_{k,l} x_k^i x_l^j P\left[X_i = x_k^i\right] P\left[X_j = x_l^j\right] = EX_i EX_j$$

so that $cov\left(X_i X_j\right) = 0$. The reverse is false! Take X, such that $EX = EX^3 = 0$, and $Y = X^2$. Now $E\left(XY\right) = E\left[X^3\right] = 0 = EX \cdot EX^2$.

Remark: The *distribution* of a sum requires more work. For discrete random variables, we can argue as follows. Consider two random variables X and Y with respective probability mass functions p_X and p_Y. Let $Z = X + Y$:

$$P\left[Z = z\right] = \sum_k P\left[X + Y = z \left|Y = y_k\right.\right] P\left[Y = y_k\right] =$$

$$\sum_k P\left[X = z - Y \left|Y = y_k\right.\right] p_Y\left(y_k\right) = \sum_k P\left[X = z - y_k \left|Y = y_k\right.\right] p_Y\left(y_k\right)$$

The last step is justified by the fact that the conditional probability reduces us to the event $Y = y_k$. We can't go any further in general, as we need information on the *conditional distribution of* X*, given* $Y = y_k$*, for all* y_k. The simplest case is, as usual, when the two variables are independent. If so, $P\left[X = z - y_k \left|Y = y_k\right.\right] = P\left[X = z - y_k\right] = p_X\left(z - y_k\right)$. In this case,

$$P\left[X + Y = z\right] = \sum_k p_X\left(z - y_k\right) p_Y\left(y_k\right)$$

The expression on the right is called the *convolution product* of the two probability mass functions, in our discrete case.

2 The Law of Rare Events

2.1 Limit of Binomials

Assume n very large, p very small, with $np = \lambda$. Now $p = \frac{\lambda}{n}$, and consider random variables X_n, with distribution $bin(n,p)$. Then,

$$
\begin{aligned}
P\left[X_n = k\right] = \binom{n}{k} p^k \left(1-p\right)^{n-k} &= \frac{n!}{k!\,(n-k)!}\frac{\lambda^k}{n^k}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} = \\
&= \frac{\lambda^k}{k!}\frac{n(n-1)\cdots(n-k+1)}{n^k}\frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} = \\
&\frac{\lambda^k}{k!}\cdot\frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\left(1-\frac{\lambda}{n}\right)^n\frac{1}{\left(1-\frac{\lambda}{n}\right)^k} \qquad (1)
\end{aligned}
$$

As $n \to \infty$, $\frac{n-c}{n} \to 1$, $\left(1-\frac{\lambda}{n}\right)^n \to e^{-\lambda}$, $\left(1-\frac{\lambda}{n}\right)^k \to 1$, that is the limiting distribution is a Poisson distribution.

Remark: This is an example of *convergence in distribution*. This definitely says nothing about the behavior of random variables, thought of as functions on a sample space. That is, we cannot say anything about possible limits of the sequence of *functions* X_n – for that matter, they may even be defined on different sample spaces, so that it would make no sense at all to talk about "$\lim_n X_n$".

2.2 The Poisson Process

Now, consider the number of independent arrivals over $[0,t]$, N_t. Divide time in units of $\frac{1}{n}$, assume probability of one arrival in each slot is proportional to $\frac{1}{n}$, $\frac{\lambda}{n}$, and of two, by independence, is proportional to $\frac{1}{n^2} \ll \frac{1}{n}$, so we will neglect it (one can account for this more precisely). In time t we have nt time slots (if that's not an integer, we can adjust that as well, so we'll ignore this issue)[1]. The sequence of time slots, with 1 if an arrival occurs, and 0 if it does not, form a sequence of independent Bernoulli trials, each with parameter $\frac{\lambda}{n}$.

Then

$$
\begin{aligned}
P\left[N_t = k\right] &= \frac{(nt)!}{k!\,(nt-k)!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^{nt}\left(1-\frac{\lambda}{n}\right)^{-k} = \\
&= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{nt}\frac{nt\cdot(nt-1)\cdots(nt-k+1)}{n^k}\left(1-\frac{\lambda}{n}\right)^{-k}
\end{aligned}
$$

[1] The book has a precise construction of a Poisson process, but it may leave the deep connection with Bernoulli trials less obvious.

The limit is as before, except $\frac{nt-c}{n} \to t$, and $\left(1-\frac{\lambda}{n}\right)^{nt} \to e^{-\lambda t}$. Thus

$$P\left[N_t = k\right] \to \frac{(\lambda t)^k}{k!} e^{-\lambda t}$$

which a Poisson distribution, with parameter λt.

We can now observe, intuitively, that $N_t = N_s + (N_t - N_s)$, and that the two random variables, N_s and $N_t - N_s$ should be independent, as they come from separate Binomial experiments[2]. Additionally, since the original Bernoulli trials are independent, if we started from time s instead of time 0, and repeated the construction up to time t, we would end up with N_{t-s}, with Poisson distribution with parameter $\lambda(t-s)$. But that should also be the same as $N_t - N_s = N_{t-s}$. The family N_t is called a ***Poisson Process***, and describes things like requests for service from independent sources[3], traffic flow, arrivals of customers at a counter, etc.

2.3 Using Generating Functions

The law of rare events concerns the limit of a family of distributions. As such it can also be proved by using a powerful result, whose proof lies beyond our scope.

Recall the definitions:

- Moment Generating Function. For finite range RV's knowing enough moments determines the distribution. In general, even having all moments (assuming they exist) may not determine the distribution. A variation that does (when it is defined) is the MGF (aka Laplace Transform)

  $$M(t) = E\left[e^{tX}\right]$$

  In general, it won't be defined for all t, since we need a series to converge (e.g., for an integer-valued RV, we need $\sum_{k=0}^{\infty} e^{kt} p(k) < \infty$, so the $p(k)$ must go to zero faster than e^{-k} for this to be defined for $t > 0$. If X takes values over *all* integers, then $p(k)$ needs to go down more than exponentially fast at both ends. Obviously, $M(0) = 1$. Expanding the exponential shows why the name.

- Characteristic Function. $C(t) = E\left[e^{itX}\right]$. We define $e^{ix} = \cos x + i \sin x$, so that $\left|e^{ix}\right| = 1$. Now $\left|Ee^{itX}\right| \le E\left|e^{itX}\right| = 1$, so this is aways defined (aka Fourier Transform). Both the MGF (when defined) and the CF determine the distribution. Recovering the distribution from the MGF or the CF is, in general, an advanced problem (inversion of transforms). We'll mention other important benefits later.

[2] More generally, given times, $s \le t \le u \le v$, it is reasonable to expect that $N_v - N_u$ will be independent of $N_t - N_s$ – as it turns out to be the case. Note that our construction assumes $N_0 = 0$, so that $N_t = N_t - N_0$.

[3] In the past, the standard example of our construction was the activity of a switchboard handling telephone calls.

- Generating Function. If X takes only non negative integer values, we can write
$$E\left[e^{itX}\right] = E\left[\left(e^{it}\right)^X\right] = E\left[z^X\right] = \sum_{k=0}^{\infty} p\left(k\right) z^k$$
$z = e^{it}$. The series converges for $z = 1$, so it converges for all $z : |z| < 1$.

For all of these functions (when they are defined) the following useful facts hold

1. If $X_1, X_2, \ldots, X_n$ are *independent* random variables with moment generating/characteristic/generating functions $M_1, M_2, \ldots, M_n$, the sum $\sum_{k=1}^{n} X_k$ has corresponding function equal to $\prod_{k=1}^{n} M_k$
2. Given a sequence of random variables with cumulative distribution functions $F_k\left(x\right) = P\left[X_k \leq x\right]$ and moment generating/characteristic/generating functions M_k, if $\lim_{k\to\infty} M_k = M$, where M is a moment generating/characteristic/generating function (there are theorems that give conditions for this to be true – for example, it is clearly necessary that $M(0) = 1$ for the first two), the there is a distribution function F, corresponding to M, and $\lim_{k\to\infty} F_k = F$.

It is easy to check that, for example, the moment generating function corresponding to (2) is given by

$$\sum_{k=0}^{\infty} e^{tk} \frac{n!}{k!\left(n-k\right)!} \frac{\lambda^k}{n^k} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k} = \sum_{k=0}^{\infty} \frac{n!}{k!\left(n-k\right)!} \frac{\left(e^t\lambda\right)^k}{n^k} \frac{\left(1-\frac{\lambda}{n}\right)^n}{\left(1-\frac{\lambda}{n}\right)^k}$$

and that this function has a limit equal to $e^{\lambda\left(e^t-1\right)}$, as $n \to \infty$. It is also easy to see that this moment generating function corresponds to the Poisson distribution of parameter λ.

3 Complements

3.1 Some Interesting Properties of Expectation

3.1.1 Monotonicity

From the definition, you can see right away that

$$X \leq Y \Rightarrow E\left[X\right] \leq E\left[Y\right]$$

This is the *monotonicity* property of expectation.

3.1.2 A Useful Formula

Suppose X takes values $0, 1, 2, \ldots$. Then,

$$E\left[X\right] = \sum_{n=1}^{\infty} P\left[X \geq n\right]$$

The proof goes like this

$$P\left[X\geq n\right]-P\left[X\geq n+1\right]=P\left[X=n\right]$$

and

$$E\left[X\right]=\sum_{n=0}^{\infty}nP\left[X=n\right]=\sum_{n=0}^{\infty}n\left(P\left[X\geq n\right]-\sum_{n=0}^{\infty}P\left[X\geq n+1\right]\right)=$$

$$=\sum_{n=0}^{\infty}\left\{nP\left[X\geq n\right]-\left(n+1\right)P\left[X\geq n+1\right]\right\}+\sum_{n=0}^{\infty}P\left[X\geq n+1\right]$$

The first sum is a telescoping sum, whose only surviving term is $0\cdot P\left[X\geq 0\right]=0$, so we are left with

$$\sum_{n=0}^{\infty}P\left[X\geq n+1\right]=\sum_{n=1}^{\infty}P\left[X\geq n\right]$$

3.2 Markov and Chebyshev Inequalities

One applications of moments is straightforward. Compute (assuming it exists)

$$E\left[|X|\right]=\sum_{k}|x_k|\,p\left(x_k\right)$$

Pick a number c and split the sum as

$$\sum_{|x_k|<c}|x_k|\,p\left(x_k\right)+\sum_{|x_k|\geq c}|x_k|\,p\left(x_k\right)\geq\sum_{|x_k|\geq c}|x_k|\,p\left(x_k\right)\geq c\sum_{|x_k|\geq c}p\left(x_k\right)=cP\left[|X|\geq c\right]$$

The first term was just thrown brutally away (it's non negative), while in the second we minorize with $c\leq|x_k|$. Thus

$$P\left[|X|\geq c\right]\leq\frac{E\left[|X|\right]}{c}\tag{2}$$

If we have more moments

$$E\left[|X|^m\right]\geq c^m\sum_{|x_k|\geq c}p\left(x_k\right)=c^mP\left[|X|\geq c\right]$$

Thus

$$P\left[|X|\geq c\right]\leq\frac{E\left[|X|^m\right]}{c^m}$$

Now, apply this to $Y=X-E\left[X\right]$, $m=2$:

$$P\left[|X-E\left[X\right]|\geq c\right]\leq\frac{Var\left(X\right)}{c^2}\tag{3}$$

which is known as Chebyshev's Inequality. Note that the only assumptions here are existence of moments – results apply to any distribution at all. Also, it turns out that Chebyshev is *sharp*: you can exhibit an example where the equality holds.

3.2.1 The (Weak) Law of Large Numbers

Equation (3) has a significant consequence. Consider n random variables $X_1, X_2, \ldots, X_n$, that have the same mean μ, the same variance σ^2, and are *uncorrelated* (of course, the usual situation will be that they are independent). Then it is easy to check that

$$E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}\cdot n\mu = \mu$$

$$Var\left(\frac{1}{n}\sum_{i=1}^{m} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var\left(X_i\right) = \frac{1}{n^2}\cdot n\sigma^2 = \frac{\sigma^2}{n}$$

Now, looking at a countable sequence of such variables, we have that, for any n, using (3)

$$P\left[\left|\frac{1}{n}\sum_{i=1}^{m} X_i - \mu\right| \geq \varepsilon\right] \leq \frac{\sigma^2}{n\varepsilon^2} \tag{4}$$

which vanishes as $n \to 0$, for any fixed value of ε. In words, the arithmetic average of independent variables with the same mean and variance has a vanishingly small probability of deviating from the mean, provided we are averaging a sufficiently large numbers of variables. This is one of the cornerstones of statistical analysis, as it promises an arbitrarily close approximation to the (generally unknown) mean of a distribution, if we observe a large enough number of "copies" of its observation. Note that (4) also allows us to calculate n so that the estimate of μ will be close to a given approximation, with a predetermined high probability. In most cases, this is a pessimistic estimate of n, as, if some additional assumptions hold, we can refer to the other basic limit theorem (the so-called Central Limit Theorem) to get a sharper estimate for n – the nice thing about the estimate from (4), though, is that it is valid under the minimal condition that the variables be uncorrelated (they don't even need to be independent), and that mean and variance are finite (well defined).

3.3 Side Comment: Expectation of Absolute Value is (Apparently) Stronger

Assuming that $E\left[|X|\right]$ exists can be thought as being stronger than assuming that EX exists. It's very much the distinction between simple convergence and absolute convergence. As an artificial example, consider a RV taking positive odd integer values and negative even integer values with

$$P\left[X = 2k-1\right] = \frac{6}{\pi^2}\frac{1}{2k-1}$$

$$P\left[X = -2k\right] = \frac{6}{\pi^2}\frac{1}{2k}$$

Thus

$$EX = \frac{6}{\pi^2}\sum_{k=1}^{\infty}\left(-1\right)^{k-1}\frac{1}{k} \tag{5}$$

which is convergent (see the following subsection 3.3.1 for a reminder of the proof), while the corresponding series for $|X|$ is a divergent harmonic series.

However, it is not a good idea to make this distinction. More precisely, if a convergent series does not converge absolutely, it converges *conditionally*, which means that by reordering terms, we can force it to converge to some different limit, or not converge at all. This is clearly the case for our example: if we first sum all positive terms we have a divergent series, and similarly if we sun the negative terms. By cleverly connecting the ways these two sub-series diverge, we can force convergence to any limit we want. While this might not seem a big problem, it does make such "contingent" expectations a lot less useful.

The reasonable way to handle this issue is the following. We'll state it, more generally, for the expectation of $f\left(X\right)$. We first define the *positive* and *negative* parts of f:

$$f^+ := \begin{cases} f & \text{wherever } f > 0 \\ 0 & \text{elsewhere} \end{cases}$$

$$f^- = \begin{cases} -f & \text{wherever } f < 0 \\ 0 & \text{elsewhere} \end{cases}$$

so that $f = f^+ - f^-$ and $|f| = f^+ + f^-$. We then consider $E\left[f^+\right]$ and $E\left[f^-\right]$ separately. If they are both finite, we define $E\left[f\right] = E\left[f^+\right] - E\left[f^-\right]$. If one diverges, and the other does not, we may define, if we wish, $E\left[f\right]$ to be positive or negative infinity, depending on which the divergent part is. If both diverge, the expectation is undefined, even if we allow infinity as a valid expectation. Our example matches the last case. With this convention f and $|f|$ have or don't have an expectation (finite or infinite) simultaneously. This definition allows the extension of the notion of expectation to much more general cases (even more general than the ones we will consider next).

3.3.1 Leibniz Criterion

Convergence of the series (5) can follows from a result named after Leibniz.

Theorem *If* $a_0 \geq a_1 \geq a_2 \geq \ldots > 0$, *and* $a_k \to 0$. *Then the series* $\sum_{k=0}^{\infty}(-1)^k a_k$ *converges.*

A proof can go like this. Consider the partial sums from 0 to $2n+1$ $S_{2n+1} = \sum_{k=0}^{2n+1}(-1)^k a_k$. Then,

$$S_{2n+1} = a_0 - (a_1 - a_2) - (a_3 - a_4) - \ldots - a_{2n+1} \leq a_0$$

since all succeeding terms are negative. Hence, the sequence is bounded. Also

$$S_{2n+1} = S_{2n-1} + (a_{2n} - a_{2n+1}) \geq S_{2n-1}$$

so that the sequence is increasing. A bounded increasing sequence has a limit, so that $S_{2n+1} \to s$ for some number s. Looking now at the sequence of sums from 0 to $2n$, we have

$$S_{2n} = S_{2n-1} + a_{2n}$$

Taking limits of both sides,

$$\lim_{n\to\infty} S_{2n} = \lim_{n\to\infty} S_{2n-1} + \lim_{n\to\infty} a_{2n} = s + 0 = s$$

All in all, $S_n \to s$.

In our specific case, the proof above can be made slightly more explicit. Consider the sums up to $2n$. Then

$$\frac{\pi^2}{6}\left(S_{2n} - S_{2n-2}\right) = \frac{1}{2n-1} - \frac{1}{2n-2} = \frac{2n-2+1-2n}{(2n-1)(2n-2)} = -\frac{1}{4n^2-5n+2}$$

Hence, $\sum_{n=1}^{2m}\left(S_{2n} - S_{2n-2}\right) = S_{2m}$ is convergent. A similar argument shows that $S_{2n-1}$is convergent, and since $S_{2n+1} - S_{2n} = \frac{6}{\pi^2}\frac{1}{2n+1} \to 0$, the limits are equal.