Probability distributions: Notes, hints and some solutions

EBP038A05

Nicky D. van Foreest

2021-03-13

Chapter 7, notes

General background

Here is one example to see why joint distributions are important. Suppose we throw a coin twice, the probability to see a head twice is $\mathsf{P}[X_1 = H, X_2 = H]$. But this is precisely a joint PMF! More generally, in any experiment that involves a sequence of measurements, such as multiple throws of a coin, or the weighing of a bunch of chimpanzees, we have to deal with joint CDFs and PMFs.

7.1.17

This is a subtle point, and we partly deal with it in Chapter 9. For general background, you have to do a course on probability based on measure-theory.

Page 318, half way.

The authors write 'marginal distribution', but this must be the marginal density.

Example 7.1.24

This works best with an indicator function and the fundamental bridge (recall, $\mathsf{P}[A] = \mathsf{E}[I_A]$ for an event A). It is easy to use, and prevents many mistakes. Consider this example.

$$\begin{aligned}\mathsf{P}[T_1 < T_2] = \mathsf{E}\left[I_{T_1<T_2}\right] = &= \int_0^\infty \int_0^\infty I_{t_1<t_2}\, \mathsf{P}[T_1 \in \mathrm{d}t_1, T_2 \in \mathrm{d}t_2] \\ &= \int_0^\infty \int_0^{t_2} \mathsf{P}[T_1 \in \mathrm{d}t_1]\, \mathsf{P}[T_2 \in \mathrm{d}t_2] \\ &= \int_0^\infty \int_0^{t_2} \lambda_1 e^{-\lambda_1 t_1}\, \mathrm{d}t_1\, \lambda_2 e^{-\lambda_2 t_2}\, \mathrm{d}t_2.\end{aligned}$$

Theorem 7.5.7

This theorem is fundamentally important. You'll save yourself many hours for the rest of your study if you remember this.

BH.7.1 and BH.7.2

`Memoryless excursions.pdf` illustrates many technical skills that are useful for the exercises of BH.7

Hints and remarks on exercises

7.1

Check BH 7.2.2. Bigger hint: Let A the arrival time of Alice, and B the time of Bob. Then we want to compute $\mathsf{P}\left[|A-B| \leq 1/4\right]$. (15 minutes is 1/4 hour.) Why is $f_{A,B}(a,b) = I_{a\in[0,1]}\, I_{b\in[0,1]}$? Now apply 2D-LOTUS to the function $g(a,b) = I_{|a-b|\leq 1/4}$.

7.9

a. $\mathsf{P}\left[X=i, Y=j, N=n\right] = \mathsf{P}\left[X=i, Y=j\right]\, I_{i+j=n}$.

c. $\mathsf{P}\left[X=i \,|\, N=n\right] = 1/(n+1)$. Why is this uniform?

7.10

This is good exercise to straighten things out about conditional probability. In particular, it is important to see how to use conditional PDFs. Don't forget to relate this exercise to Exercise 7.9; it's the same for continuous r.v.s.

a. Perhaps I missed it, but the conditional CDF seems not to be formally defined in the book.

Spoiler alert: here is the answer, in two different ways. But before answering the problem, I want to discuss the problem. Note beforehand that I do not expect that you could have come up with these answers, but you should definitely study them.

The problem. Let's tackle the problem first from a purely intuitive point of view. I am interested in $\mathsf{P}\left[T \leq t \,|\, X=x\right]$. Then, applying Bayes' rule in a naive way:

$$\begin{aligned}\mathsf{P}\left[T \leq t \,|\, X=x\right] &= \frac{\mathsf{P}\left[T\leq t, X=x\right]}{\mathsf{P}\left[X=x\right]} = \frac{\mathsf{P}\left[X+Y\leq t, X=x\right]}{\mathsf{P}\left[X=x\right]} = \frac{\mathsf{P}\left[Y\leq t-x, X=x\right]}{\mathsf{P}\left[X=x\right]} \\ &= \mathsf{P}\left[Y\leq t-x\right],\end{aligned}$$

where in the last equality I use that Y and X are independent to split the probability. But I have committed a cardinal sin here: I divided by 0 because $\mathsf{P}\left[X=x\right]=0$. However, as we will see below, the end result is correct.

How to get out of this situation in a technically correct way is one of the hard part of (mathematical) probability, and certainly not something we can solve in this course. All books on elementary probability, and lecturers similarly, struggle with this problem; this course is not an exception, nor am I. BTW, 'elementary' is not the same as 'simple'. An elementary book can still be hard; it simply does not use very advanced mathematical concepts.

Now that you understand the problem we have to avoid—dividing by 0—I discuss two methods to work around the problem.

Solution Method. Let me show I apply a general approach to find a conditional density, here $f_{T|X}(t|x)$.

- I make a *guess* of the conditional density
- I try to *verify* the guess by using formal tools. For this, I follow these steps:
  - Compute the CDF $F_{T,X}(t,x)$
  - Compute the PDF $f_{T,X}(t,x) = \partial_t \partial_x F_{T,X}(t,x)$.
  - Compute $f_X(x)$, either I know it, or I marginalize out t by solving $f_X(x) = \int f_{T,X}(t,x)\,\mathrm{d}t$.
  - Use the definition: $f_{T|X}(t|x) = f_{T,X}(t,x)/f_X(x)$.

My *guess* for the conditional CDF is this. If $X = x$, then $T \le t \iff Y \le t - x$. Hence, $\mathsf{P}[T \le t \mid X = x] = \mathsf{P}[Y \le t - x] = 1 - e^{-\lambda(t-x)}$. BTW, this is the same as my earlier answer.

Now the next step, *verification*. It is reasonable to *define* the conditional CDF in terms of the conditional PDF, and the above approach gives me a recipe to compute the latter. So, let's see how far I can get with this idea. Define the *conditional CDF* like this:

$$\mathsf{P}[T \le t \mid X = x] = \int I_{0 \le v \le t} f_{T|X}(v|x)\,\mathrm{d}v.$$

Clearly, I need $f_{T|X}(t|x)$. To get this, I *guess first*, and *then I check*. My guess is this. If $X = x$, then $Y = t - x$ so that $T = t + x$. Thus, $f_{T|X}(t|x) = f_Y(t - x) = \lambda e^{-\lambda(t-s)}\, I_{t \ge x}$, because $t \ge x$. (Note that this is tricky business; I treat densities as real probabilities, and this is not always correct. However, my guess serves as a goal, and as an intuitive check at the end.)

For the verification of $f_{T|X}$, I can use the *definition* of conditional PDFs:

$$f_{T|X}(t|x) = \frac{f_{T,X}(t,x)}{f_X(x)}.$$

I know that $f_X(x) = \lambda e^{-\lambda x}$, but what is $f_{T,X}$? For this, I again guess first, and then I check. When $X = x$ and $T = t$, then it must be that $Y = T - X = t - x$. Thus the guess is this:

$$f_{T,X}(t,x) = f_Y(t-x) f_X(x)\, I_{t \ge x} = \lambda e^{-\lambda(t-x)} \lambda e^{-\lambda x} = \lambda^2 e^{-\lambda t},$$

where I use the independence of X and Y. Here is the verification. The joint distribution of T and X can be computed without any formal problem as an integral over an indicator:

$$\mathsf{P}[T \le t, X \le x] = \mathsf{E}[\, I_{T \le t, X \le x}] = \mathsf{E}[\, I_{X+Y \le t, X \le x}]\,.$$

To compute this, *you* should apply the fundamental bridge on the indicator $I_{u+v \le t, u \le x}$ (Once you read BH.7.2, you'll see that this is the same as 2D LOTUS applied to the function $g(u,v) = I_{u+v \le t, u \le x}$). Specifically, *you* should show that

$$\mathsf{P}[T \le t, X \le x] = \iint I_{u+v \le t, u \le x} f_{X,Y}(u,v)\,\mathrm{d}u\,\mathrm{d}v = 1 - e^{-\lambda \min\{t,x\}} - \lambda \min\{t,x\} e^{-\lambda t};$$

recall, $f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)}$. (If you find this hard, check `memoryless_excursions.pdf`.) Then,

$$f_{T,X}(t,x) = \partial_t \partial_x F_{T,X}(t,x) = \lambda^2 e^{-\lambda t},$$

precisely what I found earlier.

With this,

$$f_{T|X}(t|x) = \frac{f_{T,X}(t,x)}{f_X(x)} = \frac{\lambda^2 e^{-\lambda t}}{\lambda e^{-\lambda x}} = \lambda e^{-\lambda(t-x)}\, I_{t \ge x};$$

this checks with my earlier guess. And now I can do the integral in the conditional CDF:

$$\begin{aligned}\mathsf{P}[T \le t \mid X = x] &= \int I_{x \le v \le t} f_{T|X}(v|x)\,\mathrm{d}v = \int_x^t \lambda e^{-\lambda(v-x)}\,\mathrm{d}v = e^{\lambda x}(-e^{-\lambda t} + e^{-\lambda x}) \\ &= 1 - e^{-\lambda(t-x)}.\end{aligned}$$

I can also define the *conditional CDF* as

$$\mathsf{P}[T \le t \mid X = x] = \frac{\partial_x F_{T,X}(t,x)}{f_X(x)};$$

note that there is just the partial derivative ∂_x. With this definition,

$$\mathsf{P}[T \le t \mid X = x] = (\lambda e^{-\lambda x} - \lambda e^{-\lambda t}) \lambda e^{-\lambda x} = 1 - e^{-\lambda(t-x)}.$$

It's the same. Neat!

Discussion But why do the guesses and the 'method' by which I divided by $\mathsf{P}[X = x]$ (sort of) work? The reason is that a density (a PDF) is the value of a function at some point, for example, $f_X(x)$. And when this function is positive, we can divide by it without problem. There is also a simple, intuitive, reasoning underlying this, cf., BH.7.1.17, namely, $\mathsf{P}[X \in [x, x+\epsilon]] \approx \epsilon f_X(x)$. So, it seems OK to define $f_X(x) = \lim_{\epsilon \downarrow 0} \mathsf{P}[X \in [x, x+\epsilon]]/\epsilon$.

However, this is not the end of the matter; in fact, it is the start of many problems. If you like to know more of this, read about the Borel-Kolomogorov paradox.

I admit that all this is subtle and quite difficult. Probability theory for continuous r.v.s is, simply put, difficult and tricky.

I see two ways to come to grips with all this business. One way is to do a few good courses on real analysis, topology, integration and measure theory, and Fourier theory. With this, you can understand the Radon-Nikodym theorem, weak convergence, and the proof of the central limit law (the main topic of Ch 10). The other—which I came to like more and more over the years, and only after having invested (wasted?) a considerable amount of time on the first way—is to dispense with infinite sets altogether. Once this is done, lots of the problems are cleared up. But, mathematicians are (obsessively?) fond of infinity (and use it in the same way as Aladdin uses his magic lamp to solve all his problems). However, let's not enter a philosophical discussion on the foundations of mathematics here. We stick to the book!

b. Use method 2 of part a. to get $f_{T,X}$, then use the conditional PDF definition.

c. The answer is $1/t$ (I leave it to you to get the notation right.). This is interesting: Given that the second arrival was at time t, the arrival time of the first customer is uniform on t.

d. Marginalize X out of $f_{T,X}$.

7.11

a. First find $f_{Y|X}$ and $fZ|X$. Then, given X, Z and Y are iid. Hence $f_{X,Y,Z} = f_{Y,Z|X} f_X$. Use independence to split $f_{Y,Z|X}$ into a product.

b. Suppose that Y is really big. Since Y is dependent on X, X must be dependent on Y. But Z is in turn dependent on X. And therefore...

c. Here is the answer. The ideas are important, you'll need them during nearly any course in statistics, given the importance of the normal distribution.

$$f_{Y,Z}(y,z) = \int \frac{1}{2\pi} e^{-(y-x)^2/2} e^{-(z-x)^2/2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, \mathrm{d}x.$$

It remains to simplify $(y-x)^2 + (z-x)^2 + x^2$. With a bit of work, it follows that this can be written as

$$3(x - (y+z)/3)^2 - (y+z)^2/3 + y^2 + z^2.$$

When plugging this in the integral, the last two terms appear in front of the integral. The term $(y+z)/2$ is just a shift, hence can be neglected in the integration over x. The 3 has to be absorbed in the standard deviation $\sigma = 1/\sqrt{3}$. And therefore,

$$f_{Y,Z}(y,z) = \frac{1}{2\pi} \frac{1}{\sqrt{3}} e^{-y^2/2 - z^2/2 + (y+z)^2/6}.$$

7.13

See 'Memoryless excursions'.

7.15

Make a drawing. $F(x,y)$ is the area of an (infinite) square lying south west of the point (x,y). Add and subtract such (infinite) squares until the square $[a_1,a_2]\times[b_1,b_2]$ is covered. Realize that in the process, the square $(-\infty,a_1]\times(-\infty,b_1]$ is subtracted twice.

7.20

a. Use independence, and

$$\begin{aligned}\mathsf{P}[M\le m] &= \mathsf{P}[U_1\le m, U_2\le m, U_3\le m] = \ldots\\ \mathsf{P}[M\le m, L\le l] &= \mathsf{P}[M\le m] - \mathsf{P}[M\le m, L> l]\\ \mathsf{P}[M\le m, L> l] &= \mathsf{P}[l< U_1\le m, l< U_2\le m, l< U_3\le m] = \ldots,\end{aligned}$$

to determine the joint CDF. Take partial derivatives $\partial_l\partial_m$ to find the joint PDF.

b. Find the marginal PDF of L and fill in the definition of the conditional PDF.

7.23

a. What is the volume of $[-1,1]^n$? Then divide the relevant volumes.

c. $X_n=\sum_i I_{|U_i|>c}$. Hence, $\mathsf{E}[X]=n\,\mathsf{E}\left[I_{|U_1|>c}\right]x=\ldots$ But, now realize that $I_{|U_1|}$ is Bernoulli distributed. Express the parameter p in terms of c. Then realize that X_n is the sum of a set of Bernoulli r.v.s.

d. Just fill in, in the result of c. The insight is relevant.

7.29

See 'Memoryless excursions'.

7.32

Write $M=\max\{X,Y\}$ and $L=\min\{X,Y\}$. Then $M-L=|X-Y|$ and $M+L=X+Y$.

7.35

We need a. as preparation for Ch 8.

a. Take $g(x,y)=\sqrt{x^2+y^2}=r$. Then convert the integral over $\mathsf{E}[R]=\mathsf{E}[g(X,Y)]$ over the unit circle to polar coordinates. Don't forget the r of the Jacobian, cf., Appendix A.7.2.

b. The idea is similar to Exercise 7.23. The area of a circle of radius r is πr^2. So, $F_{R^2}(x)=\mathsf{P}\left[R^2\le x\right]=?$. With this, $F_R(t)=\mathsf{P}[R\le t]=\mathsf{P}\left[R^2\le t^2\right]=F_{R^2}(t^2)$.

You should check your expression for F_R with the expression of Theorem 5.3.8. This should give the result of part a.

7.36

I give the answer here, because of the point I want to make after the proof. Take $g(x,y)=x+y$. I also assume that X and Y are continuous; for the proof it has no consequences.

As an aside, proving that $X+Y$ is a random variable when X and Y are random variables is not entirely trivial.

a.

$$\begin{aligned}\mathsf{E}\left[g(X,Y)\right] &= \iint g(x,y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint (x+y) f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint x f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y + \iint y f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \int x f_X(x)\,\mathrm{d}x + \int y f_Y(x,y)\,\mathrm{d}y = \mathsf{E}\left[X\right] + \mathsf{E}\left[Y\right].\end{aligned}$$

It is hard to overestimate the importance of this insight: Taking the expectation is, plain and simple, just integration (or summation, or a combination of the two). The linearity of the expectation is due to the fact that the integral is linear. (If you are interested in more detail, take a course on measure theory. There is also a way to define the integral by starting to require that it has to be linear. This is called a Daniell integral, cf., Wikipedia).

7.38

It is useful to write $\max\{x,y\} = x\, I_{x\geq y} + y\, I_{y>x}$, and something similar for the minimum. Then add both up. BTW, memorize this trick with indicators. It is often useful in probability, statistics, queueing, inventory theory, optimization, and so on.

For the covariance, I am always very careful with such 'shortcuts' as given in the exercise; as a matter of fact, I try to avoid such arguments because it is so easy to go wrong. Seemingly plausible arguments are often wrong due to overlooked dependency or non-linearity (effects of higher moments). So, I just fill in the above expressions for the max and the min (in terms of indicators). Then I use bilinearity (check the list of key properties of covariance) of the covariance to see what I get. Hopefully the result is easy, otherwise I need some algebra to simplify.

In the present case, $\mathsf{Cov}\left[X,Y\right] = \mathsf{E}\left[XY\right] - \mathsf{E}\left[X\right]\mathsf{E}\left[Y\right]$, and, similarly, $\mathsf{Cov}\left[M,L\right] = \mathsf{E}\left[ML\right] - \mathsf{E}\left[M\right]\mathsf{E}\left[L\right]$, with M is max, and L is min. With the above indicators, it is simple to show that $\mathsf{E}\left[ML\right] = \mathsf{E}\left[XY\right]$. However, take $X, Y \sim \mathrm{Exp}(\lambda)$. Then, $\mathsf{E}\left[M\right] = 3/(2\lambda)$ and $\mathsf{E}\left[L\right] = 1/(2\lambda)$, see the 'Memoryless excursions'. But $\mathsf{E}\left[X\right] = \mathsf{E}\left[Y\right] = 1/\lambda$.

7.42

Use example 4.4.5.

7.46

Take $I_j = 1$ if person j takes the slip with its own name. What is $\mathsf{E}\left[I_j\right]$? Then use BH.7.3.8 to see how to tackle the variance. Write, $\mathsf{Cov}\left[I_1, I_2\right] = \mathsf{E}\left[I_1 I_2\right] - \mathsf{E}\left[I_1\right]\mathsf{E}\left[I_2\right]$. But, $\mathsf{E}\left[I_1 I_2\right]$ is the probability that persons 1 and 2 pick their own slip.

7.48

Reread Example 5.7.4. Bigger hint: Write $I_j = I_{X_j = \max\{X_1,\ldots,X_j\}}$. Then, what is $\mathsf{E}\left[I_j\right] = \mathsf{P}\left[I_j = 1\right]$? What is the meaning of $T = \sum_{j=1}^{n} I_j$? Why is $\mathsf{Cov}\left[I_i, I_j\right] = 0$ when $i \neq j$? What is $\mathsf{V}\left[T\right]$?

Without knowing the distribution of the X_i it is not possible to compute $\mathsf{E}\left[X_i^* - X_{i-1}^*\right]$ where X_i^* and X_{i-1}^* are the ith and $i-1$th records. In sports statistics, people are interested in estimating the distribution of the differences between subsequent maxima. (Too) large differences can be due to technical innovation, but also due to cheating. See M.E. Robinson and J.A. Tawn, Statistics for exceptional athletics records, Applied Statistics 44(4), 1995 if you find this topic interesting.

7.51

Use the hint of the book and the independence to see that $\mathsf{E}\left[X^2Y^2\right] = \mathsf{E}\left[X^2\right]\mathsf{E}\left[Y^2\right]$. Then play with the formulas. For instance, we want to convert $\mathsf{E}\left[X^2\right]\mathsf{E}\left[Y^2\right]$ to $\mathsf{V}\left[X\right]\mathsf{V}\left[Y\right]$. One thing we can try to add and subtract $(\mathsf{E}\left[X\right])^2\mathsf{E}\left[Y^2\right]$. And so on.

7.53

The ideas of this exercise find much use in finance, physics, and actuarial sciences. In particular, the expected time it takes the drunken guy to hit some boundary is interesting. For those of you that are interested in physics, the theory of random walk helps to understand potential theory (the theory of gravitation and electro-statics).

The notation of the book is a bit clumsy. Let X_i be the movement along the x-axis at step i, and Y_i along the \$y\$-axis. Then $S_n = \sum_{i=1}^n X_i$ and $T_n = \sum_{j=1}^n Y_j$, and $R_n^2 = S_n^2 + T_n^2$. For your convenience, I'll stick to the notation of the book.

a. If you don't how to prove such a thing, it is always good to think about whether you can find a counterexample. Bigger hint: Is $\mathsf{P}[X_n = n, Y_n = n] > 0$ or must it be 0? Is $\mathsf{P}[X_n = n] > 0$?

b. It is immediate that $\mathsf{E}[X_n] = 0$. Hence, focus on $\mathsf{E}[X_n Y_n]$. Now the bad notation breaks up a simple analysis. So, I want to write $X_i \in \{-1, 0, 1\}$ to denote the direction along the x-axis of the i step, and likewise for Y_i. Then, $S_n = \sum_{i=1}^n X_i$, and T_n is the sum over the Y_i. In our new notation, $\mathsf{E}[S_n] = 0$. Then, multiply all terms in the sums of $\mathsf{E}[S_n T_n]$, and focus on the individual terms $\mathsf{E}\left[X_i Y_j\right]$. When $i \neq j$, then X_i and Y_j are independent (or not?); and when $i = j$?

c. It is clear that $R_n^2 = S_n^2 + T_n^2$. Now use linearity to split $\mathsf{E}\left[R_n^2\right]$. Finally, realize that $\mathsf{E}[S_n] = 0$, hence $\mathsf{E}\left[S_n^2\right] = \mathsf{V}[S_n]$. But then we can use the formula of the variance of a sum to split it up into a sum of variances plus covariances.

And now an argument based on recursion. (By now I hope you see that I like this method in particular).

$$\mathsf{E}\left[R_n^2\right] = \mathsf{E}\left[(R_{n-1} + X_n + Y_n)^2\right],$$

but R_{n-1} and $X_n + Y_n$ are independent, and $\mathsf{E}\left[(X_n + Y_n)^2\right] = 1$. Using the recursion, $\mathsf{E}\left[R_n^2\right] = n$.

7.58

This is a totally great exercise. At the end, I'll explain why. With the insights obtained we can relate the covariance to the determinant of a matrix, which we will use in Chapter 8.

a. Expand the brackets in the expression for the sample variance r to see that

$$r = 1/n \sum_i x_i y_i - \bar{x}\bar{y}.$$

Next, we choose with probability $1/n$ one the points (x_i, y_i). Under this probability, $\mathsf{E}[XY] = 1/n \sum_i x_i y_i$, $\mathsf{E}[X] = \bar{x}$, $\mathsf{E}[Y] = \bar{y}$. So, how do $\mathsf{Cov}[X, Y]$ and r relate?

b. Expand the brackets and use iid and linearity properties to show that the expected area spanned by two random points (X, Y) and $(\tilde{X}, \tilde{Y})$ satisfies

$$\mathsf{E}\left[(X - \tilde{X})(Y - \tilde{Y})\right] = 2\,\mathsf{Cov}[X, Y].$$

If we choose two points at random from the sample, then $(x_i - x_j)(y_i - y_j)$ is the area spanned by these two points. More generally, I have n choices for my first point, and also n choices for the second point (if both points are the same, the area of the rectangle is 0, so we don't have to exclude such choices). Hence, the expected area of the rectangle spanned by the two random points (X, Y) and $(\tilde{X}, \tilde{Y})$ is

$$\frac{1}{n^2} \sum_{i,j} (x_i - x_j)(y_i - y_j).$$

Simplify this to show that

$$2\frac{1}{n}\sum_i x_i y_i - 2\bar{x}\bar{y} = 2r$$

Hence, by part a., the expected area is twice the covariance.

Why is $\mathsf{Cov}[X, a] = 0$ for a a constant? Because the 'area' of rectangles, all with the same y-coordinate, is zero, i.e., they lie on a line.

c. And this is what the above is all about. Since there is a direct relation between covariance and area, we can use geometric arguments to derive (and memorize!) all properties of covariance! Write i. as $\mathsf{Cov}[X, Y] = \mathsf{Cov}[Y, X]$. Suppose I flip the x and y-axis, does the area of a rectangle change? For ii., What happens to the area of rectangle if you stretch the sides? For iii., realize that this is just a shift of a rectangle that leaves its area invariant. For iv., what happens to the area if you put an extra rectangle on top or to the right?

BTW, iii. follows directly from iv. In iv., take W_3 equal to a constant a_2, in other words $\mathsf{P}[W_3 = a_2] = 1$. We know that $\mathsf{Cov}[X, a] = 0$ for a constant a.

Final remarks Let's put all the above in a yet more general frame. The covariance has a number of interesting properties:

1. *bilinearity*: The covariance is linear in both arguments. The linearity in the first argument means that $\mathsf{Cov}[X + Y, Z] = \mathsf{Cov}[X, Z] + \mathsf{Cov}[Y, Z]$ and $\mathsf{Cov}[aX, Z] = a\,\mathsf{Cov}[X, Z]$ for $a \in \mathbb{R}$. The linearity in the second argument means that $\mathsf{Cov}[X, Y + Z] = \mathsf{Cov}[X, Y] + \mathsf{Cov}[X, Z]$ and $\mathsf{Cov}[X, aZ] = a\,\mathsf{Cov}[X, Z]$ for $a \in \mathbb{R}$.
2. *symmetry*: $\mathsf{Cov}[X, Y] = \mathsf{Cov}[Y, X]$, from which we define $\mathsf{V}[X] = \mathsf{Cov}[X, X]$.
3. $\mathsf{Cov}[X, a] = 0$ for all $a \in R$.

If you memorize the first two properties of covariance, all the rest follows.

Now, take three vectors $x, y, z \in \mathbb{R}^2$ (it's easy to generalize to $\mathbb{R}^n$). Then we want the area $D(x, y)$ of the parallelogram spanned by vectors x and y to satisfy:

1. *bilinearity*: The determinant is linear in both arguments. The linearity in the first argument means that $D(x + y, z) = D(x, z) + D(y, z)$ and $D(ax, z) = aD(x, z)$ for $a \in \mathbb{R}$. The linearity in the second argument means that $D(x, y + z) = D(x, y) + D(x, z)$ and $D(x, az) = aD(x, z)$ for $a \in \mathbb{R}$.
2. $D(x, x) = 0$
3. $D((1, 0), (0, 1)) = 1$.

These properties are entirely natural. The first property means that stretching vectors and stacking parallelograms result in stretching and adding areas. The second says that the area of a parallelogram spanned by two parallel vectors is zero. The third specifies that the area of the unit square is 1. Now it can be proven that there exists just one function D that satisfies these properties, and that is the determinant. Moreover, it can be shown that the second property can be replaced by

- *skew-symmetry*: $D(x, y) = -D(x, y)$.

(Note that $D(x, x) = -D(x, x) \implies 2D(x, x) = 0 \implies D(x, x) = 0$.) Let us use the properties to compute the area of a parallelogram spanned by the vectors (a, b) and (c, d). Then

$$D((a, b), (c, d)) = D(a(1, 0)) + b(0, 1), c(1, 0) + d(0, 1)) = adD((1, 0), (0, 1)) + bcD((0, 1), (1, 0)) = ad - bc,$$

where we use bilinearity in the first step, and skew-symmetry in the second and third.

So, all in all, this is what I remembered throughout the years: the covariance and the determinant are bi-linear forms, the first is symmetric, the second skew- (or anti-)symmetric.

Finally, I don't see why the areas of the rectangles have to have a sign in this problem. Interestingly, for the determinant, the areas of the parallelograms do have to have a sign to make the concept useful for physics.

7.59

a. Use that expectation is linear.

c. I think that with a bit of reasoning, it is possible to do this part before part b. In fact, I use the assumption of part c to find a guess for the answer of b.

Spoiler alert, here is the answer for c.

Suppose that the X_i and Y_j are iid. Now if I could improve the estimator $\hat{\theta}$ by splitting the measurements into two sets X_i and Y_j, then I would certainly do that. And not only I would do that; anybody in his right mind would do that. But, I never heard of this idea, and I am sure you have neither, so this must be impossible (because if it would, people would have been using this trick for ages.) Hence, we can place this in the context of the maxim: 'we cannot obtain information for free'. For this case, this must imply that splitting iid measurements into smaller sets cannot help with improving the estimator. Hence, if X_i and Y_j are iid, it must be that $w_1 = n/(n+m)$.

b. I always try to solve the problem without a hint; my answer here requires a bit more work, but is also more direct. As an aside, the result of Exercise 5.53 is really important; that I don't use the hint, doesn't mean that I find it unimportant.

First read part c, because I'll use that the notation.

Before doing the hard work, I prefer to look at some corner cases to acquire some intuitive understanding. Suppose that $v_2 := \mathsf{V}\left[Y_j\right] = 0$, but $v_1 := \mathsf{V}\left[X_i\right] > 0$. Then we know that the Y_j form a set of perfect measurements. But then I am not interested in the X_i measurements anymore; why should I, I have perfect measurements Y_j at my disposal. So, then I put $w_1 = 0$, because I don't want the X_i measurements to pollute my estimator. In other words, the final result should be such that $v_2 = 0 \implies w_1 = 0$, and vice versa.

More generally, I learned from this corner case that I want this for the final result: when $v_2 < v_1 \implies w_1 < w_2$, and vice versa.

Can we make some further progress, just by keeping a clear mind? Well, in fact we can by using our insights of part c. If we have $n+m$ iid measurements of which we call n measurements of type X_i, and m of type Y_j, then

$$\mathsf{V}\left[\hat{\theta}_1\right] = \mathsf{E}\left[\left(\sum_i X_i/n - \theta\right)^2\right] = n^{-2}\,\mathsf{E}\left[\left(\sum_i (X_i - \theta)\right)^2\right] = n^{-2}\,\mathsf{V}\left[\sum_i X_i\right] = \mathsf{V}\left[X_1\right]/n = \sigma^2/n.$$

So, $n = \sigma^2/\mathsf{V}\left[\hat{\theta}_1\right]$, and likewise $m = \sigma^2/\mathsf{V}\left[\hat{\theta}_2\right]$. Finally, plug this into our earlier expression for w_1 to get

$$w_1 = \frac{n}{n+m} = \frac{\sigma^2/\mathsf{V}\left[\hat{\theta}_1\right]}{\sigma^2/\mathsf{V}\left[\hat{\theta}_1\right] + \sigma^2/\mathsf{V}\left[\theta_2\right]} = \frac{\mathsf{V}\left[\hat{\theta}_2\right]}{\mathsf{V}\left[\hat{\theta}_1\right] + \mathsf{V}\left[\theta_2\right]}.$$

If we check our earlier insight, then we see that if $\mathsf{V}\left[Y_j\right] = 0$, then $\mathsf{V}\left[\theta_2\right] = 0$, hence $w_1 = 0$ in that case. This is precisely what we wanted.

So, here is the hint to check that the above expression for w_1 is correct.

$$\mathsf{E}\left[(\hat{\theta} - \theta)^2\right] = \mathsf{E}\left[(w_1(\hat{\theta}_1 - \theta) + w_2(\hat{\theta}_2 - \theta))^2\right] = \mathsf{V}\left[w_1\hat{\theta}_1\right] + \mathsf{V}\left[w_2\hat{\theta}_2\right],$$

by independence. Take the w's out of the variances, then write $w_2 = 1 - w_1$, take ∂_{w_1} of the expression, set the result to 0, and solve for w_1. You should get the above expression.

7.64

I found it a bit hard to understand the problem, so I pushed a bit harder. You should read well, and think hard about making a good model. Recall, it's not just mathematics we are dealing with; modeling is also important (and hard).

a. What confused me was what would happen if $X < 4$? But then I realized that the question is about expectations, so I just have to focus on the cases in which $X \geq 4$.

Use LOTUS to calculate $\mathsf{E}\left[\binom{X}{4}\right]$. You'll get an elegant answer for the expected number of different choices

b. Read well. We are spreading the courses over mornings and afternoons. Hence, we are thinning the stream of courses. The morning receives half of the courses, the afternoon the other half. So each has $\lambda/2$. Bigger hint: chicken and egg story.

c. Analogous to part b.. Bigger hint, the expected number of courses in slot 1 is $\mathsf{E}[X_1] = \lambda/4$. Of these I can choose 1. So, for four blocks, I have $(\lambda/4)^4$ choices in total. Divide this by the number obtained in part a.

7.65

See the solutions on the web.

7.68

The modeling here is hard! The probability itself not directly.

a. A multinomial distribution.

b. I found it difficult to understand what is meant with 'decisive'. So, after reading the exercise a couple of times, it suddenly occurred to me that the authors write 'if exactly 2 of the 3 choices…', so this means that for a decision, one choice (Scissors say) does not appear, i.e., all choices should belong to just two groups (only Rocks and Paper). But what if all players make the same choice? Then there cannot be a winner. In that case, Scissors and Paper, say, have not been chosen. All in all, then, for a decision, it is necessary that only one choice is 'empty'. This we can compute from part a.

c. Just generalize parts a. and b.

7.71

With genetics we can find out numerous interesting insights into history. I just read a nice book on this: L'odyssée des gènes by Évelyne Heyer.

a. Multinomial.

b. The people in the sample of size n with an A is $X_1 + X_2$. But this is the same as $n - X_3$. Hence, what is $\mathsf{P}[X_3 = n - i]$?

c. I found this a hard problem. Here is my solution, but it is a bit of a 'hack'. Let S_n be the number of A in n individuals. We want to know $\mathsf{P}[S_n = i]$. I like to use recursions, and this time it helped again. Let Y_n be the phenotype of person n. For ease, let $f_n(i) = \mathsf{P}[S_n = i]$. The conditioning on the nth person:

$$f_n(i) = f_{n-1}(i-2)p^2 + f_{n-1}(i-1)2pq + f_{n-1}(i)q^2,$$

with $q = 1 - p$ as always. Now I was a bit stuck, but just to try to see whether I could see some structure, I tried a simpler case, namely, a recursion for the binomial distribution.

So, short intermezzo. Let S_n be the number of successes in the binomial, and write $g_n(i) = \mathsf{P}[S_n = i]$ for this case. Then,

$$\begin{aligned} g_n(i) &= g_{n-1}(i-1)p + g_{n-1}(i)q \\ &= (g_{n-2}(i-2)p + g_{n-2}(i-1)q)p + (g_{n-2}(i-1)p + g_{n-2}(i)q)q \\ &= g_{n-2}(i-2)p^2 + g_{n-2}(i-1)2pq + g_{n-2}(i)q^2. \end{aligned}$$

I also know that $g_n(i) = \binom{n}{i}p^i q^{n-i}$. End of intermezzo.

Now compare the expression for the genes with the expression for the binomial. They are nearly the same, except that in the genes case, the 'n' seems to run twice as fast. I then tried the guess $f_n(i) = \binom{2n}{i}p^i q^{2n-i}$. For you, plug it in, and show that it works.

So, what was my overall approach? I used recursion, but got stuck. Then I used recursion for a simpler case whose solution I know by heart. I compared the recursions for both cases to see whether I could recognize a pattern. This lead me to a guess, which I verified by plugging it in. Using recursion is not guaranteed to work, of course, but often it's worth a try.

Now, looking back, I realize that it is as if individual n adds the outcome of two coin flips (with values in \$AA, *Aa* or *aa*) to the sum S_n of \$A'\$s. For you to solve: what is the distribution of two coin flips? Next, S_n is just the sum of n individual 'double coin flips'. Hence, what must the distribution of S_n be?

d. It is easiest to work with $f(p) = \log \mathsf{P}[X_1 = k, X_2 = l, X_3 = m]$. With part a. this can be written as

$$f(p) = C + (2k + l)\log p + (l + 2m)\log(1 - p),$$

where C is a constant (the log of the normalization constant). (BTW, with this you can check your answer for part a.) Compute $\mathrm{d}f(p)/\,\mathrm{d}p = 0$, because at this p, $\log f$, hence f itself, is maximal. Observe that C drops out of the computation, because when differentiating, it disappears.

e. Now we like to know what p maximizes $\mathsf{P}[X_3 = n - i]$. Take $g(q) = \log \mathsf{P}[X_3 = n - i]$, then

$$g(q) = C + i\log(1 - q^2) + 2(n - i)\log q.$$

(With this, check your answer of part b.) Again, take the derivative (with respect to q), and solve for q.

7.73

Standard, just check the definitions.

7.77

Transforms to remove the correlation, hence dependency, between normal r.v.s is common in statistics. Just follow the hint; it's easy. Bigger hint. Use Definition 7.5.1, Theorem 7.5.7, and Example 7.5.10.

7.80

As soon as I read exercises like this, I think about *thinned Poison processes*. Such processes abound in many practical settings. Take a hospital; the number of patients that arrive at the first aid is mostly Poisson dependent (not directly in the morning, because people who had a a small accident, will not come during the night.), but for the rest the Poisson distribution is reasonable. A fraction p is severely injured, hence as to stay at the hospital, for surgery say. The arrivals at the surgery must then form a Poisson process with rate λp. Another fraction q of patients has to go to the medical imaging department, for X ray. In shops we can make similar selections.

If it helps you, memorize all such examples as 'Chicken-egg' stories. I just think about *thinning*; that works for me.

a. It is perhaps helpful to understand how the voting process works in the US. The people that have the right to vote have to register first. (Depending to the state, adults in jail are not allowed to vote, but since the US prefers to have about $1\backslash$

Hence, the initial arrival process of voters is thinned twice, first with p, then with s.

b. See Section 4.8. Why is it binomial? What is the success probability, and what is the trial size?

c. Why is it binomial? What is now the success probability, and what is the trial size?

d. See Section 3.4. We select n balls from an urn containing d white and r black balls.

7.82

Perhaps only in my book the assumption that X and Y are independent is missing. Anyway, make this assumption.

Take Y continuous (for the discrete case the reasoning is analogous), and write

$$\mathsf{P}[Y < X] = \int I_{y<x} f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y.$$

Use independence to split $f_{X,Y}$ into the product f_Y and f_X. Then use that $X \sim \mathrm{Exp}(\lambda)$ to find f_X. Integrate over x. Then, write down the *definition* of $M_Y(s)$ as an integral, and compare the two results.

7.86

The concepts discussed here are a standard part of the education of GPs (i.e., medical doctors). The challenge for you is to try to understand the mathematics behind these concepts. Read the exercise a number of times. I found it quite difficult to capture the concepts in formulas. (I solved it once. After two weeks, I tried to solve it again, and found it just as hard as the first time.) Once you have the model, the technical part itself is simple.

a. Kind of spoiler alert. Here is, up to some computational details, my answer.

It is given that $\mathsf{P}[T \leq t \mid D = 1] = G(t)$ and $\mathsf{P}[T \leq t \mid D = 0] = H(t)$. From Theorem 5.3.1.i, we have that we can associate a r.v. to a CDF F. Sometimes we say that the CDF F *induces* a r.v. X. So let us use this here to say that G induces the r.v. T_1 and H induces T_0. So the *sensitivity* is $\mathsf{P}[T_1 > t_0] = 1 - G(t_0)$ and the *specificity* is $\mathsf{P}[T_1 < t_0] = H(t_0)$.

To make the ROC plot, I first made two plots, one of the sensitivity and the other for 1 minus the specificity, i.e., $1 - H(t_0)$. Then, in the ROC plot, we put a specificity of s on the x-axis, then we search for a t such that $1 - H(t) = s$, and then we plug this t into $1 - G(t)$ to get the sensitivity. To help you understand this better, check that $s = 0 \implies t = b \implies 1 - G(t) = 0$. Moreover,

check that $s = 1 \implies t = a \implies 1 - G(t) = 1$. Hence, the ROC curve starts in the origin and stops at the point $(1, 1)$.

With this insight, the area under the ROC curve can be written as

$$\int_0^1 (1 - G(H^{-1}(1-s)))\,\mathrm{d}s = 1 - \int_0^1 G(H^{-1}(1-s))\,\mathrm{d}s = 1 - \int_a^b G(t)h(t)\,\mathrm{d}t,$$

where, in the last step, we use the 1D change of variable $H(t) = 1 - s \implies h(t)\,\mathrm{d}t = -\mathrm{d}s$. It remains to interpret the integral, so let's plug in the definitions:

$$\int_a^b G(t)h(t)\,\mathrm{d}t = \int_a^b \mathsf{P}\left[T_1 \leq t\right] f_{T_0}(t)\,\mathrm{d}t = \int_a^b \mathsf{P}\left[T_1 \leq T_0 \,|\, T_0 = t\right] f_{T_0}(t)\,\mathrm{d}t = \mathsf{P}\left[T_{1 \leq T_0}\right].$$

7.87

a. Just use the definition, $\mathsf{E}[J] = \sum_j j\,\mathsf{P}[J = j] = 1/n \sum_j j$ and use the appendix.

For $\mathsf{V}[J]$, use that $\mathsf{E}\left[J^2\right] = 1/n \sum_{j=1}^n j^2$ and use the appendix to simplify the sum. Combine with $(\mathsf{E}[J])^2$ and simplify.

c. Read well $I_j = 1$ when $X_n > X_j$. Why is $\mathsf{E}\left[I_j\right] = 1/2$? Use linearity to get a result for $\mathsf{E}[R_n]$. Don't forget: the sum in R_n runs up to $n - 1$, not to n! For $\mathsf{V}[R_n]$, split this into a sum over variances and covariances. Now realize that $\mathsf{Cov}[I_1, I_2] \neq 0$. To see why, explain that

$$\mathsf{E}[I_1 I_2] = \mathsf{P}[X_1 < X_n, X_2 < X_n] = 1/3,$$

but $\mathsf{E}[I_1]\,\mathsf{E}[I_2] = 1/4$. Now find a closed-form expression for $\mathsf{V}[R_n]$.

d. Why is $\mathsf{E}[J] = \mathsf{E}[R_n]$? And once you see this, realize that $\mathsf{E}\left[J^2\right] = \mathsf{E}\left[R_n^2\right] = \mathsf{V}[R_n] + (\mathsf{E}[R_n])^2$. Use part c. to express this in terms of n.

Chapter 8, notes

Once you finished this chapter you can have a look at this paper. They apply the beta distribution and Bayes' formula to modeling the risks of flooding in any region of England. You have all tools to understand the entire paper.

8.1 Change of variables

The problem

In probability theory we often want to consider functions of r.v.s, as explained in BH.8.1. For instance, starting with X, we like to express $\mathsf{P}[Y \leq y]$ where $Y = g(X)$ for some function g. When g is strictly increasing this is easy:

$$F_Y(y) = \mathsf{P}[Y \leq y] = \mathsf{P}[g(X) \leq y] = \mathsf{P}\left[X \leq g^{-1}(y)\right] = F_X(g^{-1}(y)). \tag{1}$$

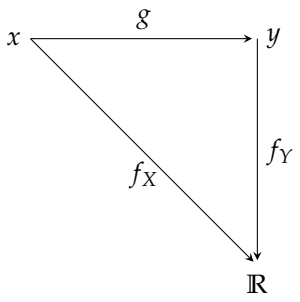
However, when g is not increasing everywhere, finding F_Y in terms of F_X is not so simple anymore because there can be no or multiple points x such that $g(x) = y$. Moreover, when we want to generalize to 2D cases (or yet more dimensions), working with CDFs becomes messy. Of course, we can use the 'manager's' or 'politician's' strategy to deal with complications: just avoid, or ignore, the complications. But this strategy does not work for probability theory, not even for elementary probability problems. Here are some examples.

In 1D, take $g(x) = x^2$. We know that $\mathsf{V}[X] = \mathsf{E}[X^2] - (\mathsf{E}[X])^2$; so have to deal with functions like g, or yet higher moments. But $\mathsf{E}[X^2] = \mathsf{E}[g(X)]$, and g is not monotone increasing. Hence, even to define the variance, we already have to deal with a function that is not one-to-one everywhere on $\mathbb{R}$. In 2D, consider the maximum of a number of r.v.s, which is interesting for people that like to bet on horse races or 100 m sprint events. Then, $g(X, Y) = (\min\{X, Y\}, \max\{X, Y\})$ is also not one-to-one. More generally, 2D function can have saddle points, i.e., points in which the function increases in one direction and decreases in another. Then finding the set of points x such that[1] $g(x, y) \leq (u, v)$ (which we need if we want to express $\mathsf{P}[g(X, Y) \leq (u, v)]$ in terms of the distribution $F_{X,Y}$) is not a particularly attractive task, to say the least.

The easiest way seems to be the avoid using the CDFs, but characterize $Y = g(X)$ in terms of the PDFs f_Y and f_X instead. Then we only have to require *locally* that g is one-to-one, and we don't have to work with inequalities, but can focus on *points* at which $g(x) = y$. Thus, below we will concentrate on the transformation of densities. I like to warn you right away: the reasoning below may seem simple, but as soon as you'll try to explain it to yourself with the book *closed*, you'll see that it is quite hard to get the details right. It's easy to mess things up; it happens to me regularly, hence I *check*.

Change of variables in 1D

To organize the relations between the r.v.s and g, I always draw the figure below.



We are given a density f_X that maps a point x to a real number $f_X(x)$, and we are given a function g that maps x to y, i.e, $y = g(x)$. The problem is to express the density f_Y, that maps y to $\mathbb{R}$, in terms of g and f_X.

To proceed, we can take the derivative at the LHS and RHS of eq. (1). This gives with the chain rule

$$f_Y(y) = F'_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} F_X(g^{-1}(y)) = f_X(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = f_X(x) \frac{\mathrm{d}x}{\mathrm{d}y},$$

where we write $x = g^{-1}(y)$ in the last step. To get the derivative of g^{-1}, consider the next equation. Taking derivatives with respect to y at both sides, and then applying the chain rule,

$$g(g^{-1}(y)) = y \implies \frac{\mathrm{d}}{\mathrm{d}y} g(g^{-1}(y)) = 1 \iff g'(g^{-1}(y)) \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = 1 \implies \frac{\mathrm{d}}{\mathrm{d}y} g^{-1}(y) = 1/g'(x),$$

where we use that $g^{-1}(y) = x$. And so we get

$$f_Y(y) = f_X(x) \frac{1}{g'(x)}.$$

[1] We say that $x < y$ on $\mathbb{R}^2$ when both $x_1 < y_1$ and $x_2 < y_2$.

There is one subtlety, when g is decreasing at some point x, then $f_Y(y) = -F_Y'(y)$. But since at such a point we also have $g'(x) < 0$, the signs cancel. Hence, in general we have

$$f_Y(y) = f_X(x)\frac{1}{|g'(x)|}.$$

There is a simple way to memorize this. We have that $g(x) = y$ and we want the probability to be conserved under the transformation g. From this,

$$f_Y(y)\,\mathrm{d}y = f_X(x)\,\mathrm{d}x \implies f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y}.$$

How to handle the case $g(x) = x^2$? Well, we can split the line into disjoint intervals in which g is either strictly increasing or decreasing, and then we apply the above rule in each of the intervals. When $g(x) = x^2$, we have that $g'(x) = 2x$, hence,

$$f_Y(y) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}.$$

It is simple to generalize when there are yet more points x such that $g(x) = y$.

The general *1D change of variables formula* is therefore like this:

$$f_Y(y) = \sum_{x_i : g(x_i) = y} f_X(x_i)\frac{1}{|g'(x_i)|}.$$

Example Take $X \sim \mathrm{Exp}(1)$, what is the density of $Y = g(X) = \lambda X$, and $\mathsf{E}[Y]$?

Solution: Let's first use the change-of-variables theorem. Take $y = g(x) = \lambda x$. Then,

$$f_Y(y) = f_X(x)\frac{\mathrm{d}x}{\mathrm{d}y} = f_X(x)\frac{1}{g'(x)} = e^{-y/\lambda}\frac{1}{\lambda}.$$

With this,

$$\mathsf{E}[Y] = \int_0^\infty y f_Y(y)\,\mathrm{d}y = \int_0^\infty y e^{-y/\lambda}\frac{1}{\lambda}\,\mathrm{d}y.$$

To solve this integral, I recognize y/λ in the exponent, and I want to get rid of the $1/\lambda$ factor. Hence, I write $u = y/\lambda$, and use this to see that

$$u = y/\lambda \implies \mathrm{d}u = \mathrm{d}y/\lambda \implies \mathrm{d}y = \lambda\,\mathrm{d}u.$$

Then, including a and b for the boundaries to show explicitly what is going on,

$$\int_a^b y/\lambda e^{-y/\lambda}\,\mathrm{d}y = \int_{a/\lambda}^{b/\lambda} u e^{-u}\lambda\,\mathrm{d}u = \lambda\int_{a/\lambda}^{b/\lambda} u e^{-u}\,\mathrm{d}u.$$

Since $0/\lambda = 0$ and $\infty/\lambda = \infty$,

$$\mathsf{E}[Y] = \lambda\int_0^\infty u e^{-u}\,\mathrm{d}u = \lambda\,\mathsf{E}[X].$$

Thus, to solve the integral over y, I used the *reverse* of g in the substitution. Indeed, to transform from Y to X, I have to divide by λ.

Example Show that the 1D change-of-variables formula relates directly to the substitution rule to solve 1D integrals.

Solution: when we have the density f_Y and the function g, then the substitution rule says that,

$$\int_a^b f_Y(g(x))g'(x)\,\mathrm{d}x = \int_{g(a)}^{g(b)} f_Y(y)\,\mathrm{d}y.$$

We also want that the transformation from X to Y does not affect the probability of the set (event) $A = [a, b]$, hence,

$$\int_{g(a)}^{g(b)} f_Y(y)\,\mathrm{d}y = \int_a^b f_X(x)\,\mathrm{d}x.$$

Combining the above two equations gives that

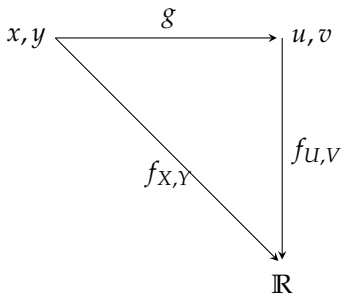
$$\int_a^b f_Y(g(x)))g'(x)\,\mathrm{d}x = \int_a^b f_X(x)\,\mathrm{d}x.$$

Since this holds for any a and b, it follows that

$$f_Y(g(x))g'(x) = f_X(x).$$

Change of variables in 2D, and higher.

Here is the figure to show how the r.v.s and g relate.



We are given a (joint) density $f_{X,Y}$ that maps a point (x, y) to a real number, and we are given a transformation g that maps (x, y) to (u, v). We want to express the density $f_{U,V}$, that maps (u, v) to $\mathbb{R}$, in terms of g and $f_{X,Y}$.

For the 2D (and higher dimensional) case, we need the following theorem of analysis, which is not entirely straightforward to prove, hence we skip the proof. Let A be the set of points (x, y) such that $g(x, y) \in B$; in other words, A is the inverse of B, i.e., $A = \{(x, y) : g(x, y) = (u, v) \in B\}$. Then we have the *change-of-variables* theorem for integration:

$$\iint_B f_{U,V}(u, v)\,\mathrm{d}u\,\mathrm{d}v = \iint_A f_{U,V}(g(x, y))\frac{\partial(u, v)}{\partial(x, y)}\,\mathrm{d}x\,\mathrm{d}y,$$

where $\partial(u, v)/\partial(x, y)$ is the absolute value of the Jacobian of g computed at (x, y). (Here my notation deviates from BH, they don't include the absolute values in the definition of $\partial(u, v)/\partial(x, y)$, but this is inconsistent with most of the mathematics books, and it leads to a extra pair of bars, which is ugly.) The functions g and $f_{U,V}$ and the set B should be 'reasonable'; the functions and sets you'll meet in practice are always reasonable. For a bit more detail about what is needed in general, see the appendix of BH, for yet more detail check the web, and if you really like such stuff, read the following books of Bressoud (in that order): Second Year Calculus: From Celestial Mechanics to Special Relativity, A Radical Approach to Real Analysis, and A Radical Approach

to Lebesgue's Theory of Integration. And once your done with that, read Capinski, Zastawniak, Jerzy: Probability Through Problems.

Now, to relate Y to X, remark that $\mathsf{P}[X, Y \in A] = \mathsf{P}[U, V \in B]$, i.e.,

$$\iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_B f_{U,V}(u,v)\,\mathrm{d}u\,\mathrm{d}v.$$

But the integral at the RHS over B can be rewritten as an integral over A by the above theorem,

$$\iint_A f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = \iint_A f_{U,V}(g(x,y))\frac{\partial(u,v)}{\partial(x,y)}\,\mathrm{d}x\,\mathrm{d}y.$$

As this holds for any (reasonable) set A,

$$f_{X,Y}(x,y) = f_{U,V}(g(x,y))\frac{\partial(u,v)}{\partial(x,y)}.$$

Finally, since $(x,y) = g^{-1}(u,v)$,

$$f_{U,V}(u,v)) = f_{X,Y}(x,y)\frac{\partial(x,y)}{\partial(u,v)}.$$

Analogous to the 1D case, this can be extended when there are several points x_i such that $g(x_i) = y$.

To memorize things, recall that $g(x,y) = (u,v)$ and we want the probability to be conserved under the transformation g, hence

$$f_{U,V}(u,v)\,\mathrm{d}u\,\mathrm{d}v = f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y \implies f_{,V}(u,v) = f_{X,Y}(x,y)\frac{\partial(x,y)}{\partial(u,v)}.$$

Example Let $U = X + Y$ and $V = X - Y$, and $X, Y \sim U[0,1]$ and independent. Find $f_{U,V}$.

Solution: $(u,v) = g(x,y) = (x+y, x-y)$. Hence $x = (u+v)/2$, $y = (u-v)/2$, and

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = |-2| = 2.$$

Hence,

$$f_{U,V}(u,v) = f_{X,Y}(x,y)\frac{\partial(x,y)}{\partial(u,v)} = f_{X,Y}(x,y)/2 = f_X(x)f_Y(y)/2 = \frac{1}{2}.$$

The difficulty is in the domain, however. Note that x and y satisfy $0 \le x \le 1$, $0 \le y \le 1$. So $0 \le (u+v)/2 \le 1$ and $0 \le (u-v)/2 \le 1$, which simplifies to $-v \le u \le 2 - v$ and $v \le u \le 2 + v$, which can also be written as $|v| \le u \le 2 - |v|$. Hence,

$$f_{U,V}(u,v) = \frac{1}{2}\, I_{|v| \le u \le 2-|v|}.$$

Example Let X and Y be independent, and take $g(x,y) = (\min\{x,y\}, \max\{x,y\})$. Determine $f_{U,V}$, with $(U,V) = g(X,Y)$. Simplify for the case X, Y iid.

Solution: Now $g^{-1}(u,v) = \{(u,v),(v,u)\}$, i.e., a set of two points. For the Jacobian:

$$\frac{\partial(u,v)}{\partial(x,y)} = \begin{bmatrix} I_{x \le y} & I_{y<x} \\ I_{y<x} & I_{x \le y} \end{bmatrix} = |\, I_{x \le y} - \, I_{x>y}| = 1.$$

Hence,

$$f_{U,V}(u,v) = f_X(u)f_Y(v) + f_X(v)f_Y(u),$$

where I use that X and Y are independent. If X, Y iid, this becomes

$$f_{U,V}(u,v) = 2f_X(u)f_Y(v).$$

Example 8.1.6

This is very important. Location-scale transformations are used in python (scipy) and R to generate random numbers of many standard distributions.

Section 8.6

You can skip this. But there are interesting examples. During WWII, order statistics were used to estimate the number of tanks the Nazi's were capable of producing per month. See the web, if you find this interesting. For instance here.

Hints and remarks on exercises

8.1

Take $g(x) = e^{-x}$. Since g is decreasing, so is g^{-1}. Then, $Y = g(X)$, so

$$\mathsf{P}[Y \leq y] = \mathsf{P}[g(X) \leq y] = \mathsf{P}\left[X \geq g^{-1}(y)\right] = 1 - \mathsf{P}\left[X \leq g^{-1}(y)\right] = 1 - (1 - exp - g^{-1}(y)) = exp - g^{-1}(y).$$

What is $g^{-1}(y)$? Is the condition on g of the change of variables theorem satisfied? Note that I follow the proof, which I find that easy to memorize, of this theorem.

8.5

Write $R = |Z|$. Observe that $\mathsf{P}[R \leq x] = \mathsf{P}[-x \leq Z \leq x]$. Rewrite this in terms of $\Phi(x)$ and $\Phi(-x)$, and then take the derivative. Don't forget, $x \geq 0$.

8.10

This is just a funny exercise. Relate it to Exercise 7.89.b.

Write $a = \mathsf{P}[Y = 0]$ and $b = \mathsf{P}[Y = 1]$. Clearly, $a + b = 1$. From the hint in the book, write $a - b$ as a series to find that $a - b = e^{-2\lambda}$. Now solve for a and b.

8.11

Background. Such exercises are interesting if you take a step back a bit. You learned arithmetic at primary school. In all those problems, the numbers you had to add, subtract, etc. where supposed to be known precisely. At secondary school, you learned how to arithmetic with symbols. And now, at university, your next step is learn how to do arithmetic with r.v.s. The relevance of this is clear to you, hopefully; but to ensure, here is an example. In a paint factory at which a couple of my students did their master's thesis, the inventory level of dyes and other raw materials is often not known exactly. There are plenty of simple explanations for this. Raw materials are kept in big bags, and personnel uses shovels to take it out of the bag. Of course, occasionally, there is some spillage on the floor, and this extra 'demand' is not reported.

The demand side is also not exact. A customer orders say 500 kg of red paint. To make this, the operators follow a recipe, but dyes (in certain combinations) do not always give the same end result. Therefore, the paint for each order is checked, and when it does not the quality level, the batch has to be adjusted by adding a bit more of certain dyes or solvents, or other chemical products.

When the planner has to make a decision on when to reorder a certain raw material, s/he divides the total amount of raw material by the average demand size. And this leads to occasional stock outs. When the stock level and the demands are treated as a r.v.s, such stock outs may be prevented, but this requires to be capable of determining the distribution of the something like Y/X.

Of course, for paint dyes it is not always necessary to enter such complications—dyes are often cheap, so planners keep enormous inventory levels—but for making medicine or other critical materials, it is important to find out how uncertainty propagates under arithmetic operations. You'll see that this is not entirely trivial.

Hint. Make a graph of the two branched of the hyperbola's $1/t$, one branch for $t > 0$, the other for $t < 0$. Then draw a horizontal line to indicate the level $V = v$; this shows with part(s) of the hyperbola's lie below v. Then compute the probability for each branch. This will give the answer of the book immediately.

Extra Determine $f_V(v)$. You'll need this anyway for Exercise 8.12

8.12

a. Follow the hint, and take the derivative of F_V, with $V = 1/T$, of the previous exercise. Then use that $1/v = t$.

b. Recall, $T = X/Y$, to $1/T = Y/X$. But what are Y and X...?

8.14

a. Make the pullback figures for this case, and just follow the logic. Then all is easy. Extra hint: Compute

$$\frac{\partial(t,w)}{\partial(x,y)}, \quad \text{or} \quad \frac{\partial(x,y)}{\partial(w,t)},$$

whatever is easier, and compute the determinant. Write the result in terms of t and w by expressing $x^2 + y^2$ in t and w. Even more hints: the Jacobian is $ad - bc$. Any idea why the exercise makes the assumption that this should not be equal to 0?

b. Observe that this result implies that T and W are independent.

8.15

The notation is a bit clumsy for the angle coordinate. Write Θ for the r.v., and θ for its value.

a. I remember this: $f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = f_{R,\Theta}(r,\theta)\,\mathrm{d}r\,\mathrm{d}\theta$. Then,

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x,y)\left|\frac{\partial(x,y)}{\partial(r,\theta)}\right|.$$

To memorize the right sequence, I write $\mathrm{d}x\,\mathrm{d}y/\,\mathrm{d}r\,\mathrm{d}\theta$ in a fancier way. Finally,

$$\frac{\partial(x,y)}{\partial(r,\theta)} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix},$$

which has determinant equal to r. What is $f_{X,Y}$ as given in the exercise? Then note that there is no dependence on θ.

b If (X, Y) are uniform on the disk, then the function $g(x,y)$ must be constant on this disk. Use an indicator to ensure that $X^2 + Y^2 \leq 1$. Finally, normalize.

c. Similar, realize that $x^2 + y^2 = r^2$.

8.16

See Exercise 8.14.

8.18

As motivation I explained above that the distribution of the division of r.v.s is important.

a. Use Pullbacks, and note that,

$$\frac{\partial(x,t)}{\partial(x,y)} = \begin{pmatrix} 1 & 0 \\ 1/y & -x/y^2 \end{pmatrix}.$$

Take the absolute value of the determinant of this. Then, use this, or its inverse (in my first attempt I did it the wrong way around).

b. Here is the answer.

$$f_T = \frac{1}{t^2}\int_0^\infty x f_X(x) f_Y(x/t)\,\mathrm{d}x.$$

Weird that probabilistic arithmetic is so hard, isn't it?

8.23

This is the last operation we use in elementary arithmetic.

a. You can also follow the approach of Exercise 8.18.

$$\begin{aligned} f_{X,T}(x,t) &= f_{X,Y}(x,y)|\frac{\partial(x,y)}{\partial(x,t)}|, \\ |\frac{\partial(x,t)}{\partial(x,y)}| &= |\begin{pmatrix} 1 & 0 \\ y & x \end{pmatrix}| = x. \\ &\Longrightarrow \\ f_{X,T}(x,t) &= f_{X,Y}(x,y)/x = f_{X,Y}(x,t/x)/x, \end{aligned}$$

since $y = t/x$. Finally, for f_T, marginalize x out by integration.

b. Just do the algebra. With part a. you have the answer, so you can check.
I did not try, but would 8.18 also work with taking logarithms?

8.27

This is simple, but comes down to good bookkeeping. I found the answer in a second with Googling: 'Sum of three uniform random variables'. There is no point in doing it here.

8.31

Use the bank-post office Story 8.5.1 to see that T and W are independent. Bigger hint: This implies that T and any (reasonable) function of W are also independent. Find a suitable function to apply to W to get X/Y. Biggest hint: $X/(X+Y) = 1/(1+Y/X)$. Take $f(x) = (1-x)/x$, then $f(1/(1+Y/X)) = Y/X$. Hence, $Y/X = f(X/(X+Y))$ is independent of T. How to get to X/Y?

8.33

I don't like to say that the arrival rate Λ is a random variable. In my opinion, it is better to say that, as Λ is unknown, we express *our belief* about its value by means of a probability distribution. For the mathematics this does not change, but for the interpretation of the results it does.

See also my discussion of Exercise 8.40.

Suppose I know λ. Then, $X \sim \text{Poi}(2\lambda)$ (Why the 2?), so that $\mathsf{P}[X = 2 \mid \lambda] = e^{-2\lambda}(2\lambda)^2/2!$. With Bayes' rule, $f(\lambda|X = 2) = \mathsf{P}[X = 2 \mid \lambda] f_\Lambda(\lambda)/\mathsf{P}[X = 2]$. The prior is given to be $\mathsf{P}[\Lambda \leq \lambda] = 1 - e^{-3\lambda}$, hence $f_\Lambda(\lambda) = 3e^{-3\lambda}$. Finally, compute the following integral,

$$\mathsf{P}[X = 2] = \int \mathsf{P}[X = 2 \mid \lambda] f_\Lambda(\lambda)\, \mathrm{d}\lambda = \int_0^\infty e^{-2\lambda}(2\lambda)^2/2! \cdot 3e^{-3\lambda}\, \mathrm{d}\lambda,$$

and simplify all.

8.34

Relate this to Example 8.5.2.

a. The simplest I can think of is to take $X \equiv 1$, i.e., $\mathsf{P}[X = 1] = 1$, and $Y \equiv 1$. But this is not a counter example. What about taking $X = 1$ or $X = 2$, both with probability $1/2$? Does this work? If not, try the same idea but now for Y. Mind that as X and Y have to be positive, the dumbest possible choice $X = Y \equiv 0$ is not allowed.

As I mentioned earlier, trying dumb choices is often remarkably effective, not only in finding counter examples, but also in understanding the structure of a problem.

b. Use symmetry. Bigger hint: for the LHS of the equation in the problem, use,

$$1 = \mathsf{E}\left[\frac{X+Y}{X+Y}\right] = \mathsf{E}\left[\frac{X}{X+Y}\right] + \mathsf{E}\left[\frac{Y}{X+Y}\right]$$
$$\mathsf{E}\left[\frac{Y}{X+Y}\right] = \mathsf{E}\left[\frac{X}{X+Y}\right].$$

Observe that we now have two equations, and two unknowns. For the RHS, note that $\mathsf{E}[X] = \mathsf{E}[Y]$.

c. I was a bit confused about the meaning of X^c. Sometimes authors write A^c for the complement of the set A. So I was wondering what the complement of a r.v. could be, in fact, whether that could be sensibly defined. For instance, take I_A, that is an r.v.. And then I_{A^c} is also a r.v.. But, what if we would define I_A^c as I_{A^c}?

After some further musing, I suddenly realized that it is just X to the power c. I think that if the authors would have written X^r I would not have been confused.

Once I understood this, I also understood what the problem is about. In part a. we should see that splitting the expectations is generally not OK. However, when X and Y are iid, splitting is allowed. Now in this part, i.e., part c, X and Y are not iid, but does splitting hold for the Gamma distribution case? Finally, by taking powers, we cannot use the straightforward linearity of expectations.

To solve it, read Example 8.5.2 critically. Realize that by story 8.5.1, $X + Y$ and $X/(X + Y)$ are independent. Hence, functions of these r.v.s. are also independent. In this particular case take $f(x) = x^c$ as function. Finally, what is

$$\mathsf{E}\left[(X+Y)^c \frac{X^c}{(X+Y)^c}\right]?$$

8.36

a. See Story 8.5.1

The exponential is a special case of the gamma distribution. See also Exercise 8.34.c. T_1/T_2 is a function of $T_1/(T_1+T_2)$.

b. This can be solved with a joint distribution function and integration over the event $\{T_1 < T_2\}$. However, we can use Exercise 7.10 (or the tools of Ch 9), or Example 7.1.24. Bigger hint,

$$\begin{aligned}\mathsf{P}[T_1 < T_2] &= \int_0^\infty \mathsf{P}[T_1 < T_2 \mid T_1 = s]\, f_{T_1}(s)\,\mathrm{d}s \\ &= \int_0^\infty \mathsf{P}[s < T_2 \mid T_1 = s]\, \lambda_1 e^{-\lambda_1 s}\,\mathrm{d}s \\ &= \lambda_1 \int_0^\infty e^{-\lambda_2 s} e^{-\lambda_1 s}\,\mathrm{d}s = \frac{\lambda_1}{\lambda_1+\lambda_2}.\end{aligned}$$

c. First she has to wait for the first server to become free. Then she is assigned to the first free server. With $\mathsf{P}[T_1 < T_2]$ server 1 is the first. What is the probability that the other server is empty first? Then, once she is at a server, what is her expected service time? The total time in the system is the time in queue plus the service time.

With the tools of Ch 9 this exercise is really easy. You should definitely do this exercise again!

8.40

Let $f(p)$ be our prior density (In the exercise it is taken to be uniform). Then $\mathsf{P}[p > r] = \int I_{p>r} f(p)\,\mathrm{d}p$ is our belief that $p > r$. For this exercise, we are interested in the relation $\mathsf{P}[p > r] \geq c$. For instance, suppose we take $c = 0.95$, then we like to know which value for r achieves that $\mathsf{P}[p > r] \geq c$?

Suppose we do $n = 1$ trial (I don't know of a simpler case.) I follow the standard steps of Bayesian reasoning.

1. I want to know the density $f_1(p|N=1)$, i.e, the density of p when we do $n = 1$ trial indicated by f_1, and that $N = 1$ trials are successful. (Note here that I am careful about notation. We do $n = 1$ trials, and then the number of successes is given by the random variable N.)

2. Now I use Bayes' rule:

$$f_1(p|N=1) = \frac{f_1(p, N=1)}{\mathsf{P}[N=1]} = \frac{f_1(N=1|p)}{\mathsf{P}[N=1]} f(p).$$

   Here $f(p)$ acts as the prior density on p.

3. I can use $f(N=1|p)f(p)$ to compute $\mathsf{P}[N=1]$:

$$\mathsf{P}[N=1] = \int f(N=1|p) f(p)\,\mathrm{d}p = \int p\, I_{p\in[0,1]}\,\mathrm{d}p = 1/2,$$

   because $f(p) = I_{p\in[0,1]}$.

4. And then, $f_1(p|N=1) = \frac{p}{1/2} I_{p\in[0,1]} = 2p\, I_{p\in[0,1]}$.

5. Thus, $\mathsf{P}[p > r \mid N=1] = \int_0^1 I_{p>r} f_1(p|N=1)\,\mathrm{d}p = \int_r^1 2p\,\mathrm{d}p = 1 - r^2$.

Now generalize to larger n, compute $f_2(p|N=2)$, then for $n=3$, and so on, until you see the pattern. (In my first attempt, I messed up, by making the dumb error to set $f(p|N=1) = f(N=1|p)$. We all make mistakes..., but some check better than others.)

There are three important remarks to make about Bayesian analysis.

1. Not all people like the idea of using a prior. Where does it come from, and what if your prior differs from mine? To me, for instance, Bayesian analysis makes a lot of sense, and I don't have objections against the fact that people don't agree on the prior. The entire stock market is based on those differences!

2. In particular I like the idea of a prior because animals (such as dogs, ants, people) have an astonishingly efficient capacity to learn priors. (I have no clue how a baby learns from seeing a cat a couple of times that that animal is a cat. But a computer has to see a billion photo's with a cat before it 'knows' what is a cat, and even then it can go wrong easily.) So, humans can make help making good models by formulating priors.

   But there is an interesting, and very scary, counterargument to using human-made priors for machine learning. When computers learn from massive amounts of data, they can learn things we humans cannot learn. For instance, we can train computers to use the sound of the voice on the phone to make a prediction about the state of health. What happens if insurance companies use this to update (real time) the price of an insurance product you would like to buy?

3. It can be very hard (impossible) to carry out step 3 above. In particular when the integral is over many dimensions, it is a research problem how to do this.

Suppose you would take another prior, e.g., p uniform on $[1/3, 2/3]$. How would that affect the solution?

8.52

The concepts discussed here are useful to better understand random number generators.

a.

$$\mathsf{P}\left[X_j \leq c\right] = \mathsf{P}\left[\log U_j \geq -c\right] = \mathsf{P}\left[U_j \geq e^{-c}\right].$$

But U_j is uniform on $[0, 1]$, so what is the above probability?

So, why is this useful? You should know that much research has been spent on finding extremely efficient algorithms to generate uniform *random deviates*. Random deviates are random numbers but generated by an algorithm; they are not really random, but they 'behave' as random numbers. Hence the different name. As far as I know, these random deviates pass all statistical tests to try to distinguish them from 'real' random numbers (for instance generated by physical noise). For this reason, they can be treated in simulation for instance as if they are really random. You should also know that humans are *extremely* bad at generating random patters. Never think you are an exception, turn to the computer right away.

Since these algorithms are so fast, it is possible with the trick of the exercise to generate exponentially distributed r.v.s, and with similar ideas r.v.s. with many other different distributions.

8.54

Use Theorem 4.3.9.

In more detail, write X_i for the number of throws to go from the $i-1$th success until the ith success. Then $X_i \sim \text{Geo}(p)$. Then,

$$M_{Xp/(1-p)}(s) = \mathsf{E}\left[\exp sp/(1-p) \sum_{i=1}^{r} X_i\right] = (MX_1 p/(1-p))^r,$$

where the RHS is the moment-generation function of $X_1 \sim \text{Geo}(p)$. Compute $MX_1 p/(1-p)$, fill this in the above, use that $e^{sp/(1-p)} \approx 1 + sp/(1-p)$ for $p \ll 1$, and take the limit $p \to 0$. You should get a Gamma distribution.

Chapter 9, notes

Example 9.1.6

You can skip this. I am not a particular fan of such type of paradoxes; at least not for students new to probability. To start with, why is this a probability problem at all? For this, we should start with making a sample space, but what is it here? We also need probabilities (formally, a measure), but are those? As a challenge, if you like, try turning this 'paradox' into a well-formulated probability problem.

Example 9.1.7

I like this example as it is a simple start to understand how to set prices for insurance, and, more generally, it is an example of finding optimal decisions under uncertainty.

This problem is more interesting as follows. If $b < \alpha V$, for some $\alpha \in (0, 1)$ the bid is rejected. By the solution, if $\alpha = 2/3$, you should not enter this game. But is there a smaller α at which entering the game is interesting?

Spoiler alert: I solved the problem in a different way. (And I also find it more natural, but that is personal bias I suppose). Suppose I offer b and the value is V. Then my payoff is $(V - b)\, I_{b \geq \alpha V}$. Then my expected gain is

$$\mathsf{E}[W] = \int_0^1 (V - b)\, I_{b \geq \alpha V}\, \mathrm{d}V = \int_0^{\min\{1, b/\alpha\}} 1(V - b)\, \mathrm{d}V.$$

Now solve this for $b < \alpha$ and $b > \alpha$.

An interesting generalization: Suppose that V is the size of a claim with a certain CDF F (not uniform as in the example) what price b should you ask to insure this product?

Example 9.1.8

What happens if you take X to be the number of *throws*, rather than Heads, i.e., $X \sim \mathrm{FS}(p)$? In that case, $\mathsf{E}[X] = 1 + q\,\mathsf{E}[X]$, because we have to throw at least once, and with probability q, we start again. Hence, $\mathsf{E}[X] = 1/(1 - q) = 1/p$.

With this idea, it is also easy to find $N_r := \mathsf{E}[X]$ when $X \sim \mathrm{NBin}(r, p)$. Suppose the first throw is a success, then we need $r - 1$ more successes, if the first throw is a failure, we are back at 'hole one'. Thus, $N_r = pN_{r-1} + q(1 + N_r)$. Simplifying (and using that $p/(1 - q) = 1$) gives $N_r = N_{r-1} + q/p$, which implies $N_r = rq/p$.

Example 9.1.9

I reason slightly differently here. Write N_r for the number of throws required to reach r heads in row. Then I need N_{r-1} throws in expectation to reach the state in which there are $r - 1$ heads in row. Suppose now that we are in state $r - 1$. Then, if I throw heads, with probability p, I reach the state with r heads in row, and I am done. However, if I throw tails, with probability q, I have to start all over again. Therefore,

$$N_r = N_{r-1} + p \cdot 1 + q(1 + N_r) \implies N_r = N_{r-1}/p + 1/p \implies N_r = \sum_{i=1}^{r} 1/p^i. \tag{2}$$

Here is a fun variation. We have a mouse that sits on one corner of a cube whose edges are made of wire, and at some other point (for instance diagonally opposite the mouse) there is some cheese. The mouse chooses at random (uniform) any edge and moves to the next corner. How long does it take, in expectation, for the mouse to reach the cheese, if traversing an edge takes one minute (and the cheese does not move). A really interesting extension is to think of the cheese also moving.

Here is the solution. The mouse starts in a position in which it has at minimum three edges to walk. Let the expected time to hit the cheese be given by T_3; this is not an r.v.! Now the mouse takes one edge, and then it has at least two edges to travel, which takes expected time T_2. T_1 is defined similarly. A tiny bit of though shows that the times should satisfy the relations:

$$\begin{aligned} T_3 &= 1 + T_2 \\ T_2 &= 1 + T_1 \cdot 2/3 + T_3 \cdot 1/3 \\ T_1 &= 1 + T_2 \cdot 2/3. \end{aligned}$$

Solving gives $T_1 = 7, T_2 = 9, T_3 = 10$.

The general pattern is like this. Identity a set of states that capture what you want to know. (Like in the example of the mouse, I am not interested in which specific corner the mouse visits, I only want to know how many edges it still has to travel.) Then find out how to move from one state to another with which probability. And then do the algebra (which can be hard technically, but conceptually simple). Sometimes you can find closed-form solutions, like here; otherwise you can use numerical tools.

Example 9.1.10

Skip the b part. The problem is not well-specified. Suppose the drunk hits b and makes a step to the right. Then, with quite a bit of mathematics (harder than we do in this book), it is possible to prove that the expected time to hit b again is ∞, but that the probability to hit b is 1 (weird, not?). To repair for this, suppose that the drunk can only go to the left after hitting b, so that the drunk is confined to stay on the numbers $0, \ldots, b$. Then we first have to prove that the hitting time T has finite expectation, which is necessary for the answer of the book to work.

For 9.1.10.b, a more useful problem is to compute $\mathsf{E}[T]$, i.e., the expected time to hit b for the first time, assuming the drunk stays on $0, \ldots, b$, starts at some level a, and $\mathsf{E}[T] < \infty$.

Definition 9.2.1

This definition is subtle, and it takes time to understand. Here is a slightly different explanation; perhaps it's useful for you. Take some random variable X, say. Then, as in Chapter 7, we can be interested in $\mathsf{E}[g(X)]$, i.e., the expectation of the *random variable* $g(X)$. As another example, in Example 9.1.7, I computed $\mathsf{E}[g(V)]$ with $g(v) = (v - b)\, I_{\text{¤}v \leq b}$, and V as r.v.

When Y is continuous we can compute $\mathsf{E}[Y \mid X = x]$ with the conditional CDF

$$\mathsf{E}[Y \mid X = x] = \int f_{Y|X}(y|x)\, \mathrm{d}y.$$

(For discrete r.v., replace the integral by the PMF.) Observe that this is *just a function of* x; define this function as $g(x) = \int f_{Y|X}(y|x)\, \mathrm{d}y$. And now, as before, we consider the random variable $g(X)$, and give this r.v. a name: the conditional expectation of Y given X.

It is true that X plays some sort of double role—first we use it in the conditioning in the definition of the function g, and then we plug it into g again—and this is perhaps confusing. But I finally 'got it', when I understood that g can be interpreted as just some function. And then we compute $\mathsf{E}[g(X)]$, and so on.

Warning 9.2.2

What type is $\mathsf{E}[Y \mid I_A]$? Is it a number or a r.v.?

Example 9.2.5

Nice exercise, but try to prove it with the tools of Section 7.1; that is much more useful than the answer in the book. For me, the answer is, to some extent, a bit of a trick that works just for

this example. For instance, I didn't know this property before doing this example, and only in hindsight, I find it clear that $M - L$ is independent of L. But, what if we take $X, Y \sim \text{Geo}(p)$, or $X \sim \text{Exp}(\lambda)$ and $Y \sim \text{Exp}(\mu)$?

Adam's law

This is mostly know as the *Tower Rule* or *Tower Property*.

Figure 9.6

It is important to remember the idea that $\mathsf{E}[Y \mid X]$ is the best estimator of Y given X. For me Figure 9.6 captures this idea well.

Example 9.5.4.

For the variance, you can also compute $\mathsf{E}\left[W^2 \,\middle|\, Z\right] = \mathsf{E}\left[W^2 \,\middle|\, X\right]$. Then use the expression for W in terms of X and Y. Finally, realize that since Y and X are independent, $\mathsf{E}[f(Y) \mid X] = \mathsf{E}[f(Y)]$, for any reasonable function f.

Hints and remarks on exercises

9.1

Let R be the route taken. Then, $\mathsf{E}[T \mid R] = \sum_j \mathsf{E}\left[R_j\right] I_{R=j}$, where R_j is the time of route j. Hence, $\mathsf{E}\left[R_j\right] = \mu_j$. Next, realize that $\mathsf{V}[T] = \mathsf{E}\left[T^2\right] - (\mathsf{E}[T])^2$. For $\mathsf{E}\left[T^2|R\right]$, realize that $\mathsf{E}\left[R_j^2\right] = \mu_j^2 + \sigma_j^2$, because $\mathsf{V}\left[R_j\right] = \mathsf{E}\left[R_j^2\right] - (\mathsf{E}\left[R_j\right])^2$. Finally, with these ideas, check that

$$\begin{aligned}\mathsf{V}[\mathsf{E}[T \mid R]] &= \mathsf{E}\left[(\mathsf{E}[T \mid R])^2\right] - (\mathsf{E}[\mathsf{E}[T \mid R]])^2 \\ &= (\mu_1^2 + \mu_2^2 + \mu_3^2)/3 + (\mu_1 + \mu_2 + \mu_3)^2/9.\end{aligned}$$

9.3

The value of p is not relevant for the answer, but it is important that it is the same for women and men. If one person would be affected, the probability that a woman has the disease is $21/35$. Now generalize to $N = 5$ persons.

A generalization: what if the probability of having the disease is different between women and men?

9.7

See my remark above. As a variation, what if the bid is lost if it is rejected? So you pay b always, and only get the prize if $b > \alpha V$.

9.12

Such problems find applications in medicine testing and DNA sequencing. Let X_1 the outcome of the first throw. Then, given X_1, the run length $R_1 \sim \text{FS}(p)$ or $\sim \text{FS}(q)$, depending on X_1.

9.15

Check Example 9.3.6. Bigger hint: Use linearity of conditional expectation, then symmetry to conclude that $\mathsf{E}[X_i \mid \bar{X}] = \bar{X}$.

9.16

Use linearity

9.18

Expand the square. Take out what is known, in other words, realize that $\mathsf{E}[Y \mid X]$ is a function of X, hence, $\mathsf{E}[\mathsf{E}[Y \mid X] \mid X] = \mathsf{E}[Y \mid X]$.

9.19

Check Example 9.5.4. The ideas of this exercise are important for your statistics courses.

a. My guess was that in the relation $X = dY + W$, $d = 1/a$. Why is this wrong; that is the aim of this exercise.

b. Take $V = Y - cX$. We are given that (X, Y) has the bivariate normal distribution, so (X, V) has the bivariate normal distribution. Therefore, when $\mathsf{Cov}[X, V] = 0$, X and V are independent.

9.21

Using indicators like this lies at the heart of more advanced probability theory and measure theory. For instance, the proof of LOTUS is based on indicators.

a. Interpret $\mathsf{E}[Y \mid I_A]$ as a function of I_A. Then, what is $g(1) = \mathsf{E}[Y \mid I_A = 1]$?

b. The PMF of X is the joint PMF of $Y\, I_A$. Hence, $\mathsf{P}[X = i] = \mathsf{P}[Y = i,\, I_A = 1]$, when $i \neq 0$. What is $\mathsf{P}[X = 0]$? By using the ideas of Ch 7 the rest follows.

c. Of course, $1 = I_A + I_{A^c}$. Hence, $Y = Y \cdot 1 = \ldots$? Use linearity of expectations to split $\mathsf{E}[Y]$ into two parts.

9.24

An interesting variation is like this. After n throws, you observe k tails. What can you say about the probability of having picked the coin with p_1?

9.25

Follow the hint, and use recursion. Kelly betting strategies are used at the stock markets. Check out Wikipedia if you're interested. (It must be possible to make a very nice assignment based on this exercise and Kelly strategies.) The hint is an example of the important concept of *martingales*.

9.28

a. The marginal second moment of θ is $\mathsf{E}\left[\theta^2\right]$. Recall, by the assumptions, we are given the joint distribution of X and θ. Thus, we can obtain (the moments of) θ (or X) by marginalizing out X (or θ). For the rest, follow the hint of the book. Expand the square, use linearity of conditional expectation, and use unbiasedness.

b. Use the hint and follow the same steps of a. Do the algebra, it may seem that you'll get the same answer of part a., but it is slightly different.

c. Note that $\mathsf{E}\left[(\hat{\theta}-\theta)^2\right] \geq 0$. By a., if $\hat{\theta}$ is unbiased, then $\mathsf{E}\left[\hat{\theta}^2\right] \geq \mathsf{E}\left[\theta^2\right]$, while by b., if $\hat{\theta}$ is the Bayes procedure, $\mathsf{E}\left[\hat{\theta}^2\right] \leq \mathsf{E}\left[\theta^2\right]$. What are the consequences of these inequalities?

9.32

The results of this exercise are (or should be) used by nearly all software packages to control inventory levels of companies. For sure, your supermarket uses this, Bol.com, and so.

a. Let X be the amount purchased by the first customer that comes along. Let P be the r.v. that is 1 if the customer does purchase, and 0 otherwise. What is $\mathsf{E}[P]$? What is $\mathsf{E}[X \mid P]$? What is $\mathsf{E}\left[X^2 \mid P\right]$; for this, use the insights of Exercise 9.1.

b. Let $N \sim \text{Poi}(8\lambda)$ be the number of customers that pass by. Given $N = n$, what is $\mathsf{E}[S \mid N]$, where $S = \sum_{i=1}^{N} X_i$ is the total sales. Now use the law of total expectation. What is $\mathsf{V}[S \mid N]$? Use Eve's law to compute $\mathsf{V}[S]$. Bigger hint, read Example 9.6.1.

9.35

To keep the analysis simple, the authors assume that the winning probability p is constant during Judit's career. But, is this a realistic assumption?

a See Exercise 9.32.

b. Write $X_i = 1$ for a win of tournament i, and $X_i = 0$ otherwise. Then, $T = \sum_{i=1}^{N} X_i$. Then, use independence to simplify $\mathsf{E}\left[e^{sT} \mid N = n\right]$ to $\left(\mathsf{E}\left[e^{sX_1}\right]\right)^n$. Then, $f(s) = \mathsf{E}\left[e^{sX_1}\right] = pe^s + 1 - p$. Use the tower property to get rid of the condition on N, and use the geometric series to simplify. Finally, writing $M(s) = \mathsf{E}\left[e^{sT}\right]$, realize that $\mathsf{E}[T] = M'(0)$, and $\mathsf{E}\left[T^2\right] = M''(0)$. With this, check the results of a.

9.37

Bootstrapping is used in statistics to, for instance, construct confidence intervals. It is a much used and intuitive technique. The extra exercise below helps you to recall some ideas of Ch 1.

I prefer to write $Y_j = X_j^*$, as this writes (and types) easier.

Here are the solution; for the moment, I don't know a set of good hints to coach to the answer. If you do, tell me. It took me quite a bit of effort to get the details right.

a. Suppose $S_j \sim \text{DUnif}(\{1, \ldots, n\})$ is the jth sample. Then, $Y_j = \sum_{i=1}^{n} X_i \, I_{S_j=i}$. And then, using the independence of X_j and S_j, and $I_{S_j=i} \, I_{S_j=k} = 0$ if $i \neq k$,

$$\begin{aligned}
\mathsf{E}\left[Y_j\right] &= \sum_i \mathsf{E}\left[X_i\right] \mathsf{E}\left[\, I_{S_j=i}\right] = \mu, \\
\mathsf{V}\left[Y_j\right] &= \mathsf{E}\left[(\sum_i X_i \, I_{S_j=i})^2\right] - (\,\mathsf{E}\left[X_j\right])^2 \\
&= \mathsf{E}\left[\sum_i X_i^2 \, I_{S_j=1} + \sum_i \sum_{k \neq i} X_i X_k \, I_{S_j=i} \, I_{S_j=k}\right] - \mu^2 \\
&= \sum_i \mathsf{E}\left[X_i^2\right] \mathsf{E}\left[\, I_{S_j=1}\right] - \mu^2 \\
&= \frac{1}{n} \sum_i \mathsf{E}\left[X_j^2\right] - \mu^2 = \sigma^2.
\end{aligned}$$

b. Now we are given the outcomes (samples) $X_i = x_i$ of n experiments. I prefer to write $\mathsf{E}_D[Y] = \mathsf{E}[Y \mid X_1, \ldots, X_n]$. It is shorter, and by labeling the expectation with the data $D = \{X_1, \ldots, X_n\}$, I am reminded that the expectation is taken only with respect to the random variables S_j (the data D is considered fixed).

$$\begin{aligned}\mathsf{E}_D[\bar{Y}] &= \mathsf{E}_D\left[\frac{1}{n}\sum_j Y_j\right] = \frac{1}{n}\,\mathsf{E}_D\left[\sum_j\sum_i X_i\, I_{S_j=i}\right] \\ &= \frac{1}{n}\sum_i X_i \sum_j \mathsf{E}_D\left[\, I_{S_j=i}\right] = \frac{1}{n}\sum_i X_i := \bar{X}.\end{aligned}$$

Given D, X_i is just a number, hence can be taken out of expectation. This is not the same as μ!. The conditional variance:

$$\begin{aligned}\mathsf{V}_D[\bar{Y}] &= \mathsf{V}_D\left[\frac{1}{n}\sum_j Y_j\right] \\ &= \frac{1}{n^2}\,\mathsf{V}_D\left[\sum_j Y_j\right] \\ &= \frac{1}{n^2}\sum_j \mathsf{V}_D[Y_j] && \text{by the hint} \\ &= \frac{1}{n}\,\mathsf{V}_D[Y_1] && Y_j \text{ are iid.}\end{aligned}$$

Let's guess what $\mathsf{V}_D[Y_1]$ is. We select arbitrarily an element of the data D. The variance must then be $\frac{1}{n}\sum_j (X_j - \bar{X})^2$. We must be able to get this from our definition $Y_1 = \sum_i X_i\, I_{S_1=i}$. Again using that $I_{S_1=i}\, I_{S_1=j} = 0$ if $i \neq j$,

$$\begin{aligned}\mathsf{V}_D[Y_1] &= \mathsf{V}_D\left[\sum_i X_i\, I_{S_1=i}\right] \\ &= \sum_i \left(\mathsf{V}_D\left[X_i\, I_{S_1=i}\right] + \sum_{j\neq i} \mathsf{Cov}\left[X_i\, I_{S_1=i}, X_j\, I_{S_1=j}\right]\right) \\ &= \sum_i \left(X_i^2\,\mathsf{V}_D\left[\, I_{S_1=i}\right] + \sum_{j\neq i} X_i X_j\,\mathsf{Cov}\left[\, I_{S_1=i},\, I_{S_1=j}\right]\right) \\ &= \sum_i \left(X_i^2 \frac{n-1}{n^2} - \sum_{j\neq i} X_i X_j\,\mathsf{E}_D\left[\, I_{S_1=i}\right]\,\mathsf{E}_D\left[\, I_{S_1=j}\right]\right) \\ &= \frac{n-1}{n^2}\sum_i X_i^2 - \frac{1}{n^2}\sum_i\sum_{j\neq i} X_i X_j \\ &= \frac{1}{n}\sum_i X_i^2 - \frac{1}{n^2}\sum_i X_i^2 - \frac{1}{n^2}\sum_i\sum_{j\neq i} X_i X_j \\ &= \frac{1}{n}\sum_i X_i^2 - (\bar{X})^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2.\end{aligned}$$

c. $\mathsf{E}[\bar{Y}]$ is simple, use linearity and Adam's law (i.e, the tower property). For the variance, use Eve's law.

$$\mathsf{E}[\bar{Y}] = \mathsf{E}[\,\mathsf{E}_D[\bar{Y}]] = \mathsf{E}[\bar{X}] = \frac{1}{n}\sum_i \mathsf{E}[X_i] = \mathsf{E}[X] = \mu.$$

Here are the details for $\mathsf{V}[\bar{Y}]$.

$$\mathsf{V}[\mathsf{E}_D[\bar{Y}]] = \mathsf{V}[\bar{X}] = \frac{1}{n^2}\sum_i \mathsf{V}[X_i] = \frac{1}{n}\sigma^2,$$
$$\mathsf{E}[\mathsf{V}_D[\bar{Y}]] = \mathsf{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n^2}\mathsf{E}\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2\right] = \frac{n-1}{n^2}\mathsf{E}\left[S_n^2\right],$$

where we use BH.Def 6.3.3 and BH.Th 6.3.4. Note that S_n^2 is now the sample variance, so we do not use the notation from part a. and b. anymore. The rest is simple.

d. We add randomness twice, first we draw samples to get D, and then we draw randomly from D.

Extra exercise. How many different bootstrap samples are possible?

Spoiler alert: here is the answer. Immediate from Example 1.4.22. We are not interested in the sequence of the bootstrap sample. BTW, the story that goes for me with this example is the 'balls and bars story'. I have n balls to distribute over k boxes. Hence, there are $k-1$ bars to separate the boxes. For the bootstrap sample, I have to distribute n bootstrap samples (the X_i^*) over n boxes (the initial sample X_i.)

If n is small, say $n = 4$. Does it make sense to take more than 1000 bootstrap samples?

9.39

There are numerous examples of r.v.s with non-zero kurtosis, for instance, the claim sizes of car accidents, the time patients spend in hospital beds, finance. This exercise helps to understand how a positive kurtosis may originate.

a Why is $\mathsf{E}[X \mid I_1] = \mu_1 I_1 + \mu_2(1 - I_1)$? if $\mathsf{E}[X_1] = \mu_1$ and $\mathsf{E}[X_2] = \mu_2$? Next, to simplify, use that $\mu_1 = \mu_2 = 0$. Then, show that $\mathsf{V}[X \mid I_1 = 0] = \sigma_1^2$. Finally, put everything together and use Eve's law. For the covariance, note that $\mathsf{E}[I_1 X_1 I_2 X_2] = 0$; explain why.

b. Use that $\mu = 0$. Then expand the brackets in $\mathsf{E}\left[(X-\mu)^4\right]$ and realize that all terms with a μ are zero. We also know that $K(X_1) = 0$, hence $\mathsf{E}\left[X^4 \mid I_1 = 1\right] = \mathsf{E}\left[X_1^4\right] = 3\sigma_1^4$. The computation of $\mathsf{E}\left[X^4\right]$ now follows from the LOTE. Finally, from part a, we have the σ for $\mathsf{E}[X]$. It remains to assemble all in one nice formula.

9.42

Use Theorem 9.3.8. Bigger hint: take $Z = S$ in the formula of that theorem.

Intuitively, why does this hold? To see that, I introduce some new notation; it is slightly abstract, but very useful. Let us write $\mathcal{F}$ for all information we can learn from the sample $\mathbb{X}$. Then, in fact, $S = \mathsf{E}[Z \mid \mathcal{F}]$. But then it is automatic that $\mathsf{E}[Z \mid S, \mathcal{F}] = \mathsf{E}[Z \mid S]$, because if we are told S (i.e, we are provided with the expectation of Z based on all information $\mathcal{F}$ about $\mathbb{X}$ that is available to us), then telling us $\mathcal{F}$ once more does not add any extra information.

9.50

a. I prefer to write $f_\Lambda(\lambda) = e^{-\lambda}$ for the density of the prior of Λ. If we are told that $\Lambda = \lambda$, then $N|\Lambda = \lambda \sim \text{Poi}(\lambda)$. This idea gives us $\mathsf{E}[N \mid \Lambda = \lambda]$ and $\mathsf{V}[N \mid \Lambda = \lambda]$. By replacing λ by Λ we get the *r.v.s* $\mathsf{E}[N \mid \Lambda]$ and $\mathsf{V}[N \mid \Lambda]$.

Now use Adam and Eve.

Perhaps it is a good idea to try the generalization $\Lambda \sim \text{Exp}(\alpha)$, and see how explicitly how all works (the 1 in $\text{Exp}(1)$ hides a bit of what is going on). Then, $f_\Lambda(\lambda) = \alpha e^{-\alpha\lambda}$, and we have to integrate over λ!

b. Use Example 9.6.1

c. and d. I had to do all the work myself, because I read the book in a piece-meal way. In particular, I did not read Story 8.4.4 before I started working on some of the problems . So, avoid the mistakes I made, read this story first.

9.52

a. Adam and Eve. $\mathsf{E}[T|W=i] = 1/\lambda_i$, hence $\mathsf{E}[T|W] = \sum_{i=1}^{2} 1/\lambda_i \, I_{W=i}$.

b. Write S_i for the service time at server i. Then $X = \min\{S_1, S_2\}$. Since S_1 and S_2 are exponential, X is also exponential. But then, X is memoryless. So, whatever amount of time has been invested (here 24 hours), the expected time to finish remains the same. Kind of weird, not? The exponential distribution is natural, but some times strange too.

9.55

When I don't know how to proceed, I always try to make some simplifications. Sometimes I am lucky, and see how to generalize; sometimes it doesn't help, but I still had a nice time playing with a simpler problem. So, I always gain :-)

Here, the simplest I can think of is to assume that there is one number drawn per day. In that case,

$$a_j = 1 + \frac{j}{n} a_{j-1} + \frac{n-j}{n} a_j,$$

if there are n different numbers; here $n = 35$. Now I recognize this structure as the expected time it takes from having j numbers to go to having $j-1$ to go, plus a_{j-1}. In other words, $a_j - a_{j-1} \sim \text{FS}(j/n)$.

How about drawing two numbers per day? I can draw two new numbers and then a time a_{j-2} remains, I can draw one new number and one known number and then a_{j-1} remains, or I can draw two known numbers and then a_j remains (since I did not make any progress). Assign probabilities to each event and add.

Now that I know how to do 2, the structure is clear.

9.56

This is good exercise to rehearse your knowledge on Beta distributions.

I prefer to say that I model my uncertainty about the value of p by means of a prior.

a. Since we include then win, the number of games $T|p$ (since we assume p given) must be $\sim \text{FS}(p)$.

I use recursions, because I like using recursions.

$$\mathsf{E}[T \mid p] = 1 + q\,\mathsf{E}[T \mid p] \implies \mathsf{E}[T \mid p] = 1/p.$$

And thus, $\mathsf{E}[T \mid p] \sim \text{FS}(p)$. To get $\mathsf{E}[T]$ use the tower property. For this, write the integral with the Beta distribution, Definition 8.3.1. Take the integral with respect to p!

Spoiler alert. Here is the rest.

$$\begin{aligned}\mathsf{E}\,[1/p] &= \frac{1}{\beta(a,b)}\int_0^1 \frac{1}{p}p^{a-1}(1-p)^{b-1}\,\mathrm{d}p = \frac{1}{\beta(a,b)}\int_0^1 p^{a-2}(1-p)^{b-1}\,\mathrm{d}p = \frac{\beta(a-1,b)}{\beta(a,b)}\\ &= \frac{a+b-1}{a-1}.\end{aligned}$$

To get the last equation, use the definition of $\beta(a,b)$ in terms of factorials (see the Bayes' billiards story) to simplify. This is easy, many terms cancel.

BTW, with Bayesian analysis, it is simple to mix up the 'thing' against with one has to integrate, x, or p, or something else. I make mistakes with this, I suspect you will too. Hence, be warned.

b. We should compare the expectation $\mathsf{E}\,[1/p]$ we computed in part a to $1+\mathsf{E}\,[G]$, where $G\sim$ Geo(x), with $x = a/(a+b)$. Clear, $1+\mathsf{E}\,[G] = 1+(1-x)/x = 1/x = (a+b)/a$, and this is smaller than $\mathsf{E}\,[T]$ of the part a.

I must miss something here. The prior is Beta(a,b). Then Beta-Binomial conjugacy story, we assume that Vishy won $a-1$ games, and lost $b-1$ games. My guess for Vishy winning the next game would be $(a-1)/(a+b-2)$, not $a/(a+b)$. But I make an error here. Check the BH problem 9.57. You'll see that we should indeed use $a/(a+b)$! Tricky!

c. Use Story 8.3.3

9.57

This has simple solution.

$$\begin{aligned}\mathsf{P}\,[X_{n+1}=1\,|\,S_n=k] &= \mathsf{P}\,[X_{n+1}=1, S_n=k]\,/\,\mathsf{P}\,[S_n=k]\\ \mathsf{P}\,[S_n=k] &= \frac{1}{n+1},\ \text{Bayes' billiard,}\\ \mathsf{P}\,[X_{n+1}=1,S_n=k] &= \int_0^1 \mathsf{P}\,[X_{n+1}=1,S_n=k\,|\,p]\,f(p)\,\mathrm{d}p = \int_0^1 p\binom{n}{k}p^k(1-p)^{n-k}f(p)\,\mathrm{d}p\\ &= \frac{k+1}{n+1}\int_0^1\binom{n+1}{k+1}p^{k+1}(1-p)^{n-k}f(p)\,\mathrm{d}p\\ &= \frac{k+1}{n+1}\frac{1}{n+2},\ \text{Bayes' billiard.}\end{aligned}$$

Now simplify.

9.58

Recall that the uniform distribution on $[0,1]$ is Beta(a,b) with $a=b=1$. I prefer to write $S_n = \sum_{j=1}^n X_j$.

a. First compute $\mathsf{E}\,[S_n\,|\,p]$. Then compute $\mathsf{E}\,[\,\mathsf{E}\,[S_n|p]]$. Note that the outer expectation is an integral with respect to p and the density Beta$(1,1)$.

For the variance, use EVE's law. Spoiler alert. Here is the VE part of it.

$$\mathsf{V}\,[\,\mathsf{E}\,[S_n|p]] = \mathsf{E}\left[(\,\mathsf{E}\,[S_n|p])^2\right] - (\,\mathsf{E}\,[\,\mathsf{E}\,[S_n|p]])^2 = \int_0^1 n^2p^2\,\mathrm{d}p - (\int_0^1 np\,\mathrm{d}p)^2 = n^2/12.$$

With this, you can check the rest.

b. Use Beta-Binomial conjugacy. BTW, I find it a really elegant relation!

c. This time I don't use integrals, so I deviate from my normal approach (i.e., use standard tools to get a model, then do integrals or other straightforward, technical(?), work.) When somebody doesn't give me any information about what team can win, then any outcome must be equally likely. What else can it be? This is also my way to understand the expression in Story 8.3.2.

There is an interesting method to get probability distribution based on information and constraints. The idea is based on the *maximization of entropy*, see Ch 10. In a bit more detail. Suppose somebody tells you that the outcome of some experiment can be any non-negative number, and tells you also that the mean is $\alpha > 0$. What distribution assumes the least amount of information (i.e., entropy) beyond what you are given? That turns out to be $\text{Exp}(\alpha)$. Nice. Suppose that you are told that the outcome can be any positive or negative number, and the mean in μ and the std is σ. Then $\text{Norm}(\mu, \sigma^2)$ is the distribution that maximizes the entropy. Nicer yet!. If you find this interesting, search the web and the literature on 'Entropy maximization and probability'.

Chapter 10, notes

Example 10.1.3

Perhaps you find the following easier:

$$(\mathsf{E}[X])^2 = (\mathsf{E}[X\, I_{X>0}])^2 \leq \mathsf{E}\left[X^2\right]\, \mathsf{E}[\, I_{X>0}] = \mathsf{E}\left[X^2\right]\, \mathsf{P}[X > 0]\,,$$

because $1^2 = 1$.

Theorem 10.1.5

The proof is important. Convexity is often used in optimization, and the tangent line you see in Figure 10.1 is a convenient lower bound of the function g. Hence, memorize Figure 10.1.

I often forget the direction in Jensen's inequality. To check, the following reasoning works for me: I know that $\mathsf{V}[X] \geq 0$, but $\mathsf{V}[X] = \mathsf{E}\left[X^2\right] - (\mathsf{E}[X])^2 = \mathsf{E}[g(X)] - g(\mathsf{E}[X])$, if $g(x) = x^2$. From the graph of the parabola, i.e., the graph of g, I know that g is convex.

Examples 10.1.7, 10.1.8, 10.1.9

You can skip these. However, if you do physics, mathematics or computer science, you'll come across these examples anyway, during statistical physics, convergence of measures, or information theory.

Theorem 10.1.12

There is a body of interesting (advanced) probability around Chernoff's inequality, called *large deviations*. In fact, this inequality relates classical mechanics (formulated in terms of energy and the Euler-Lagrange equation) and probability theory. This is one of those unexpected relations that makes mathematics and physics fun.

Example 10.3.7

Note that $\mathsf{E}[\log Y_n] \sim -0.081n$, i.e., has has negative drift, but $\log \mathsf{E}[Y_n] \sim n \log 1.1$. Check that this is not in conflict with Jensen's inequality.

Example 10.4.3

Why is $n\bar{Z}_n^2 \sim \chi_1^2$? Here is the reason. $\bar{Z}_n$ is the sum of n normal r.v.s Z_j, hence normal itself. As each of these Z_j is standard normal, $\mathsf{E}[\bar{Z}_n] = 0$, and $\mathsf{V}[\bar{Z}_n] = n^{-2}\sum_j \mathsf{V}[Z_j] = 1/n$, by independence. Therefore, $\sqrt{n}\bar{Z}_n \sim N(0,1) \implies (\sqrt{n}\bar{Z}_n)^2 \sim \chi_1^2$, where we use Definition 10.4.1 and Theorem 10.4.2 in the last step.

Hints and remarks on exercises

10.2

Write $X_j \in \{0,1\}$ for whether person j in the sample supports the policy or not. Then $\hat{p} = \sum_{j=1}^n X_j/n$ is a r.v. Apply Chebyshev's inequality to $\mathsf{P}[|\hat{p} - p| > 2\sigma]$. Here $\sigma^2 = \mathsf{V}[X_j]$. Then, why is $\mathsf{E}[\hat{p}] = p$? With this, realize that the RHS of Chebyshev's inequality becomes $\mathsf{V}[\hat{p}]/(4\sigma^2)$. Use iid to simplify $\mathsf{V}[\hat{p}]$. This proves the inequality for $c = 1$.

10.3

This is just a funny exercise. I wonder whether it has a practical value. First check the assumption that $Y \neq aX$, for some $a > 0$; why is it there? Then, take a suitable g in Jensen's inequality. Bigger hint: $g(x) = 1/x$.

In the solution guide, the authors do not explain the $>$, while in Jensen's inequality there is a $\leq$. To see why the $>$ is allowed here, rethink the assumption in the exercise, and reread Theorem 10.1.5.

Finally, at what p is $p(1-p)$ maximal?

10.6

Apple the idea of Example 10.1.3 to the r.v. $W = (X - \mu)^2$. Bigger hint: $Z = I_{W>0}$. Hence $\mathsf{E}[W^2] \geq (\mathsf{E}[W])^2/\mathsf{P}[W > 0]$. I now need the assumption that $\sigma > 0 \implies \mathsf{P}[W > 0] > 0$. Relate $\mathsf{E}[W]$ to σ. When, on the other hand, $\sigma = 0$, then $\mathsf{P}[W = 0] = 1$ (why?). But then the inequality is trivially true.

The claim of kurtosis follows right from the definition.

10.9

a. Jensen's inequality, $g(x) = e^x$

b. Use symmetry: X and Y are iid.

c.

$$\mathsf{P}[X > Y - 3] = \mathsf{P}[X > Y + 3] + \mathsf{P}[Y - 3 \leq X \leq Y + 3].$$

d. Use Jensen's inequality. Since X and Y are iid, how do $\mathsf{E}[X^2]$ and $\mathsf{E}[Y^2]$ compare?

e. Use EVE's law. What does this imply about the direction of the inequality (recall, variances cannot be negative)? Since X and Y are iid, $\mathsf{E}[Y|X] = \mathsf{E}[Y]$. But then, what is $\mathsf{V}[\mathsf{E}[Y|X]]$?

f. Use Markov's inequality. Then use the triangle inequality to split $|x + y|$ into two terms. Then use that X and Y are iid.

10.26

a. I did things a bit differently then in the book. Take $S_n = \sum_{i=1}^n X_i$ with $X_i \sim \text{Bern}(p)$. Then I know this:

$$\mathsf{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k} \to e^{-\lambda}\lambda^k/k! = \mathsf{P}[N = k], \quad \text{if } N \sim \text{Poi}(\lambda),$$

for $n \to \infty$, $p \to 0$ but such that $pn = \lambda$. I also know from the CTL that $S_n \sim N(np, np(1-p))$ if n becomes large. But, $N(np, np(1-p)) \to N(\lambda, \lambda)$ in the above limit. Now take $\lambda = n$ to see that $\text{Poi}(\lambda) \sim N(n, n)$.

b. Check the solution manual. Then, with $\mu = \sigma = \lambda = n$, and $n \gg 1$,

$$\begin{aligned}
\Phi(n+1/2) - \Phi(n-1/2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \int_{n-1/2}^{n+1/2} e^{-(x-\mu)/2\sigma^2}\, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi n}} \int_{-1/2}^{1/2} e^{-x^2/2n}\, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi n}} \int_{-1/2}^{1/2} (1 - x^2/2n)\, \mathrm{d}x \\
&= \frac{1}{\sqrt{2\pi n}} (1 - 1/(24n)).
\end{aligned}$$

So, we found another term to approximate $n!$ yet better.

10.27

If you don't want to check the on-line solution guide right away, here is a hint.

a.

$$M_n(t) = \mathsf{E}\left[e^{s\bar{X}_n}\right] = \mathsf{E}\left[e^{s/n\sum_{i=1}^n X_i}\right] = \left(\mathsf{E}\left[e^{s/nX}\right]\right)^n = (M_X(s/n))^n,$$

where $X \sim \text{Poi}(\lambda)$.

b. Check Exercise 8.54 for some inspiration.
Suppose $Y \equiv \lambda$, then $M_Y(s) = e^{s\lambda}\,\mathsf{P}[Y = \lambda] = e^{s\lambda}$. Compare the limit of part a. to M_Y.

10.28

First do 10.27.

I could not find the definition of the *standardized version* of a r.v. X in the book, but it is $Y = (X - \mu)/\sigma$, if $\mathsf{E}[X] = \mu$ and $\mathsf{V}[X] = \sigma$. I hope you realize how important this concept is, in view of the CLT.

Look up the moment generating function for an $N(\mu, \sigma^2)$ r.v.

Now, what is $\mathsf{E}[X_n]$ and $\mathsf{V}[X_n]$ if $X \sim \text{Poi}(n)$; what is Y_n and $\mathsf{E}[Y_n]$ and $\mathsf{V}[Y_n]$? To find $M_{Y_n}(s)$, realize that X_n is the sum of n independent $\text{Poi}(1)$ r.v.s. Take $X \sim \text{Poi}(1)$. Then, explain that

$$\begin{aligned}
M_{Y_n}(s) &= \left(M_X(s/\sqrt{n})\right)^n e^{-s\sqrt{n}} \\
&= e^{n(e^{s/\sqrt{n}}-1)} e^{-s\sqrt{n}}.
\end{aligned}$$

Use Taylor's expansion for e^x to second order. Finally, what is $M_N(s)$ for a $N(0,1)$ r.v.?

10.29

Apply the strong law to the r.v defined in the hint.

The problem of the statistician is of course to find functions L and R, such that $\mathsf{P}\left[\theta \in [L(X), R(X)]\right] \geq 0.95$.

10.30

The problem demonstrates a simple investment strategy. If you plan to work as a quant in finance or as an actuary, or if you play poker, or some similar game, such strategies should interest you naturally.

a. Define I_n as the success indicator: it is 1 if I win, and 0 if I loose. Use recursion to see that

$$Y_n = Y_{n-1}/2 + (1.7)^{I_n}(0.5)^{1-I_n}Y_{n-1}/2.$$

Bigger hint: After one round I still have the money I put aside, and I invested the rest.

Simplify this to get:

$$Y_n = \frac{Y_{n-1}}{2}(1 + 0.5(3.4)^{I_n}) = Y_0 2^{-n}\Pi_{i=1}^{n}(1 + 0.5(3.4)^{I_i}).$$

Let me first digress, and show you a useful dead-end. I stared at the above for a while, wondering how to make progress with this. The form is different from the one in Example 10.3.7, so it appears that I am stuck. But suddenly I recalled that I once saw (in a book on complex function theory) the inequality

$$1 + \sum_{n=1}^{N} p_n \leq \prod_{n=1}^{N}(1 + p_n) \leq \exp\left(\sum_{n=1}^{N} p_n\right),$$

when $p_n \geq 0$. This is also useful for you as it relates compounding interests (e.g. monthly interest and other economic and finance ideas) to sums and exponentials. Perhaps I can use the left-most inequality. If this goes to ∞ as $N \to \infty$, then I can conclude that $Y_n \to \infty$.

So, I want to find a p_n such that

$$1 + p_n = (1 + 0.5(3.4)^{I_n})/2.$$

A bit of algebra gives

$$p_n = ((3.7)^{I_n} - 2)/4.$$

Hmm, this is disappointing: it is not true that $p_n \geq 0$ always, when $I_n = 0$, $p_n < 0$. I cannot use the inequalities.

What else can I think of? As a next step, let's try to convert my recursion for Y_n to a form as in the example. If $I_n = 1$, $Y_n = Y_{n-1}2.7/2$, while if $I_n = 0$, $Y_n = Y_{n-1}1.5/2$. Hence, I can also write:

$$Y_n = Y_{n-1}(1.35)^{I_n}(0.75)^{1-I_n}.$$

With this expression, the rest is simple, just follow Example 10.3.7. It turns out that $Y_n \to \infty$ as $n \to \infty$.

b. Just substitute α in the relevant formula of part a. For the maximum, take the derivative with respect to α.

10.35

Just keep your cool. It seems like a hard control problem, but it is very simple. The problem can be generalized to multi-armed bandits, which are used to improve web sites, pricing plans, medicine. There is an entire book on it.

a. Consider some corner cases. If $p = 1$, what would you guess for the next throw of the coin? Or if $p = 0$? In general, if the coin is loaded (biased), and you the bias, what do you do?

b. By the strong law, the fraction of heads (1's) will converge to p. So, if $p > 1/2$, what will you predict? And if $p < 1/2$, then what do you predict?

c. If somebody knows your strategy, and this person can control the outcome of the game, would you play this game if you have to pay for every round?

10.39

This is an easy exercise (finally).

a. $\mathsf{P}[N = n] = \mathsf{P}[X_1 < 1, X_2 < 1, \cdots X_{n-1} < 1, X_n > 1]$. But, then N must be related to the first success distribution.

b. Let X_i be the inter-arrival time between jobs $i - 1$ and i. Then $S_n = \sum_i^n X_i$ is the arrival time of job n. We want that $S_{M-1} < 10 \leq S_M$. Since the X_i are $\sim \text{Exp}(\lambda)$, $S_n \sim \text{Poi}(\lambda t)$.

c. Use the CLT.