Probability distributions EBP038A05
Solutions for problems of BH

Nicky D. van Foreest, Ruben van Beesten
Joost Doornbos, Wietze Koops, Mikael Makonnen, Zexuan Yuan

August 15, 2022

CONTENTS

7 Chapter 7 5

8 Chapter 8 7

9 Chapter 9 9

10 Chapter 10 11

11 Hints 13

12 Solutions 19

CHAPTER 7

Ex 7.1 (BH.7.1)**.** We simulate this in one of the assignments.

Ex 7.9 (BH 7.9.)**.** We'll develop a simulation for this in the assignments.

Ex 7.10 (BH.7.10)**.** Recall that a conditional CDF given an event A is defined as $F(y|A) = \mathsf{P}\left\{Y \leq y|A\right\}$. Likewise, let us write here $F_T(t|x) = \mathsf{P}\{T \leq t|X = x\}$. Just use this in your derivation. However, there is one problem with the fact that the event $\{X = x\}$ has probablity zero. In the solution I'll discuss how to around this.

Don't forget to compare this exercise to BH.7.9, which is the same but for discrete memoryless rvs.

Ex 7.11 (BH.7.11)**.**

Ex 7.13. BH.7.13

Ex 7.15. BH7.15.

Ex 7.24. BH.7.24. In the assignments we'll develop a simulator.

Ex 7.29. BH.7.29

Ex 7.38. BH.7.38. Besides the solution of BH, read our solution.

Ex 7.53. BH.7.53. We simulate this in one of the assignments. The ideas of this exercise find much use in finance, physics, and actuarial sciences. In particular, the expected time it takes the drunken person—It's not only guys that sometimes consume too much alcohol—to hit some boundary is interesting. The notation of the book is a bit clumsy. Here is better notation. 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 $R_n^2 = S_n^2 + T_n^2$.

Ex 7.58. BH.7.58. This is a totally great exercise. First solve it yourself. In the solution, I'll explain why, in particular how to relate the concept of covariance to the determinant of a matrix.

Ex 7.59. BH.7.59. Read this exercise, then read (and do) BH.5.53 for some further background. You'll encounter these topics countless times in other courses! The final answer is really nice and intuitive.

Ex 7.71. BH.7.71.

Ex 7.86. BH.7.86. The concepts discussed here are a standard part of the education of GPs (i.e., medical doctors), and in data science in general.

CHAPTER 8

8

Ex 8.11. BH.8.11. With convolution we know how to add and subtract independent rvs. Now we make a start with division. You'll see that this operator is not as simple as you always thought.

Before solving the problem, let's take a step back. 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 rvs.

Here is an example to show you the relevance of this. 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 bags. 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 for example 500 kg of red paint. To make this, the operators follow a recipe, but dyes (in certain combinations) do not always give the same result. Therefore, the paint for each order is checked, and when it does not meet 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 rvs, such stock outs may be prevented, but this requires to be capable of determining the distribution of the something like Y/X.

Ex 8.15. BH.8.15. We'll use this exercise in a lecture to show how the normal distribution originates from astronomy (or dart throwing).

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

Ex 8.18. BH. 8.18. Here we deal with division of rvs.

Ex 8.23. BH.8.23. We already analyzed how to handle addition, subtraction and division. It remains to deal with multiplication.

Ex 8.31. BH.8.31

Ex 8.36. BH.8.36.

Ex 8.40. BH.8.40. A nice question on the exam could be to take another prior, e.g., p uniform on $[1/3, 2/3]$. How would that affect the solution?

Ex 8.52. BH.8.52. The concepts discussed here are useful to better understand how to generate exponential random numbers.

Ex 8.54. BH.8.54. We tackle this also with simulation in an assignment.

I find it easier to consider $Y = pX$, rather than pX/q. Note that since $q = 1-p \to 1$ as $p \to 0$, the factor $1/q$ is immaterial for the final result.

Read my solution too, as I develop some nice ideas in passing.

CHAPTER 9

9

Ex 9.1. BH.9.1. It is best to solve this problem with EVE's law.

Ex 9.25. BH.9.25. We tackle this problem also in an assignment with simulation. Check out `https://en.wikipedia.org/wiki/Kelly_criterion` you're interested.

Ex 9.28. BH.9.28.

Ex 9.32. BH.9.32. The results of this exercise are (or should be) used by nearly all software packages to control inventory levels of companies such as supermarkets and bol.com.

Ex 9.37. BH.9.37.

Bootstrapping is used in statistics to, for instance, construct confidence intervals. It is a much used and intuitive technique.

Extra exercise to help you recall some ideas of Ch 1. How many different bootstrap samples are possible?

I used some extra ideas to save some time. We say that the rvs $\{X_i\}$ are independent and distributed as the common rv X when $X_i \sim F_X$ where F_X is the CDF of the rv X. Then $\mathsf{E}[X_i] = \mathsf{E}[X]$, and so on. Next, I prefer to write $Y_j = X_j^*$, as this writes (and types) faster. Finally, it is easy to define $Y_j = \sum_{i=1}^n X_i I_{S_j=i}$, where $S_j \sim \text{DUnif}(\{1,\dots,n\})$ is the jth sample of the $\{X_i\}$.

Ex 9.39. BH.9.39. There are numerous examples of rvs with non-zero kurtosis, for instance, claim sizes of car accidents, the time patients spend in hospital beds, finance. This exercise helps to understand how a positive kurtosis may originate.

Ex 9.50. BH.9.50. We will also simulate this in an assignment.

Ex 9.52. BH.9.52

Ex 9.55. BH.9.55. Suppose first you draw just one number per day, what is then the recursion? Then suppose you draw 2 numbers per day.

An interesting variation is to find a recursion for the number of *draws* instead of *days* are needed until all numbers have been seen.

Ex 9.56. BH.9.56.

Ex 9.57. BH.9.57

Ex 9.58. BH.9.58. In part c. the prior is the uniform distribution. What would happen if you would take the prior of part b, i.e., a out of j wins?

CHAPTER 10

10

Ex 10.2. BH.10.2

Ex 10.3. BH.10.3 This is just a funny exercise, but I wonder whether it has a practical value.

Ex 10.6. BH.10.6

Ex 10.9. BH.10.9

Ex 10.23. BH.10.23.

Ex 10.26. BH.10.26.

Ex 10.28. BH.10.28. Note that standardized version of a rv X is $Y = (X - \mu)/\sigma$ where $\mathsf{E}[X] = \mu$ and $\mathsf{V}[X] = \sigma$.

Ex 10.30. BH.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.

Ex 10.36. BH.10.36.

Ex 10.39. BH.10.39.

11
HINTS

h.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 $\mathsf{P}\{|A-B| \leq 1/4\}$. (15 minutes is 1/4 hour.) Why is $f_{A,B}(x,y) = I_{x\in[0,1]}\, I_{y\in[0,1]}$? Now apply 2D-LOTUS to the function $g(x,y) = I_{|x-y|\leq 1/4}$.

h.7.9. a. $\mathsf{P}\{X=i, Y=j, N=n\} = \mathsf{P}\{X=i, Y=j\}\, I_{i+j=n}$.

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

h.7.11. a. First find $f_{Y|X}$ and $f Z|X$. 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 a realization of Y is really big. Since Y is dependent on X, X must be dependent on Y. But Z is in turn dependent on X. What are the consequences?

h.7.13. **??** contains all the explanations.

h.7.15. Make a drawing.

h.7.24. Check BH.7.1.24 and BH.7.1.25 First draw the area over which we have to integrate. Then use an indicator function over which to integrate. What is the joint PDF $f_{Y_1,Y-2}$?

h.7.53. Use the hint of the book and independence to see that $\mathsf{E}\left[S_n^2 T_n^2\right] = \mathsf{E}\left[S_n^2\right] \mathsf{E}\left[T_n^2\right]$. Then try to simplify.

b. It is immediate that $\mathsf{E}[S_n] = 0$. Hence, focus on $\mathsf{E}[S_n T_n]$. Expand the sums of $\mathsf{E}[S_n T_n]$, and consider the individual terms $\mathsf{E}\left[X_i Y_j\right]$. When $i \neq j$, are X_i and Y_j independent? What if $i = j$?

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.

h.7.58. 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

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

h.7.59. a. Use that expectation is linear.

b. Read the entire exercise in its entirety before trying to solve it. In this case trying to solve c. seems simpler because of the extra iid assumption. You might want to use this to formulate some simple guesses.

Thus first part c. It is given that the X_i and Y_j are iid. Then, 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. What does this idea imply for the weights?

Part b, continued. I always try to solve the problem myself without a hint. This lead to the following considerations, which gave me quite a bit of extra understanding beyond the problem itself. As a next piece of advice, before doing hard work, I prefer to look at some corner cases to acquire some intuitive understanding. I also use the rvs of Part c.

Suppose that $v_2 := \mathsf{V}[Y_j] = 0$, but $v_1 := \mathsf{V}[X_i] > 0$. (For instance, Y_j is the jth measurement of a perfect machine and X_j of an imperfect machine.) Then we know that the set $\{Y_j\}$ forms a set of perfect measurements. But then I am not interested in the $\{X_i\}$ measurements anymore; why should I as I have the 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.

How would you choose the weights such that this requirement is satisfied, but also the condition imposed by Part c.?

h.7.71. 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 hint based on recursion. Let S_n be the number of As in n individuals. We want to know $f_n(i) = \mathsf{P}\{S_n = i\}$. A simple recursive idea, i.e., one-step analysis by conditioning on the phenotype of the nth person, gives that

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

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. Derive this, and then use this to solve the problem.

d. It is easiest to work with $f(p) = \log \mathsf{P}\{X_1 = k, X_2 = l, X_3 = m\}$, where $\mathsf{P}\{X_1 = k, X_2 = l, X_3 = m\}$ follows from a., and then differentiate with respect to p.

e. Follow the same scheme as for d.

h.7.86. 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.

h.8.11. Start with the case $v = 0$. Use the proof of BH.8.1.1. Reason carefully; corner cases as simple to miss.

Then, make a graph of the two branches 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.

h.8.15. a. See BH.8.1.9.

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. What are the densities of X and Y when they are $N(0,1)$?

h.8.18. We can make a transform T, U such that $T = X/Y$ and $U = X$ to use a 2D transformation. Compute x and y as functions of t and u. Then the Jacobian.

h.8.23. You might want to follow the approach of BH.8.18.

h.8.31. Use the bank-post office Story 8.5.1 to see that T and W are independent.

h.8.36. a. See BH.8.5.1. The exponential is a special case of the gamma distribution. See also BH.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 BH.7.10 or BH.7.1.24.

c. First she has to wait for the first server to become free. This is the minimum of the two exponentials. 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.

h.8.40. Apply beta-binomial conjugacy.

h.8.52. a.

$$\mathsf{P}\left\{X_j \leq c\right\} = \mathsf{P}\left\{\log U_j \geq -c\right\} = \mathsf{P}\left\{U_j \geq e^{-c}\right\} = \mathsf{P}\left\{1 - U_j \leq 1 - e^{-c}\right\}.$$

What is the distribution of $1 - U_j$?

b. $\log \Pi_{j=1}^n U_j = \sum_{j=1}^n \log U_j = \sum_{j=1}^n (-X_j)$. But $-X_j \sim \text{Exp}(1)$, hence the sum is just a sum of iid Exp rvs. What is the distribution of this sum?

h.8.54. Use BH.4.3.9. Then, start with a geometic rv, then extend to a negative binomial rv.

h.9.25. Use Adam's law to express $\mathsf{E}[X_{n+1}]$ in terms of $\mathsf{E}[X_n]$, then use recursion.

h.9.32. a. Let Y be the amount purchased by the first customer that comes along, let P be the rv that is 1 if the customer does indeed purchase, and 0 otherwise, and let X be the size of the purchase. Why is $Y = XP$? What is $\mathsf{E}[P]$? What is $\mathsf{E}[Y\,|\,P]$? What is $\mathsf{E}\left[Y^2\,\middle|\,P\right]$. You might want to use BH.9.1.

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

h.9.50. a. $N|\lambda \sim \text{Pois}(\lambda)$.

b. Analogous to BH.9.6.1

c. and d. See BH.8.4.5.

h.9.56. Refresh your knowledge of the Beta distributions.

a. Since we include the win, the number of games $T|p$ (since we assume p given) must be $\sim \text{FS}(p)$. Hence, $\mathsf{E}\left[T|p\right] = 1/p$

To get $\mathsf{E}[T]$ use Adam's law. Realize that you have to take the integral with respect to p!

b. $1 + \mathsf{E}[G]$ is smaller than the expected time as computed in a. Why is this so?

c. The number of wins, conditional on p, out of n is $X|p \sim \text{Bin}(n, p)$. Then use Beta-Binomial conjugacy.

BTW, I find it easier to think about $f(p, X = k)$ instead of $f(p|X = k)$, since on the event $(p, X = k)$.

$$f(p, X = k) \propto p^{a-1} q^{b-1} \binom{n}{k} p^k q^{n-k} \propto p^{a-1+k} q^{b-1+n-k}.$$

Then, as $f(p|X = k) = f(p, X = k)/\mathsf{P}\{X = k\} \sim f(p, X = k)$ (because $\mathsf{P}\{X = k\}$ is just a constant) we get the same result op to a scaling factor. But we can use the reasoning of BH.8.3.3 to get the correct constant.

h.9.57. a. The prior of p is uniform on $[0, 1]$. But this is equal to Beta(1, 1). Now use Beta-Binomial conjugacy.

b. Write $S_n = \sum_{i=1}^{n} X_i$. What are $\mathsf{P}\left\{X_{n+1} = 1|p\right\}$ and $\mathsf{P}\left\{S_n = k|p\right\}$?

h.9.58. a. Recall that the uniform distribution on $[0, 1]$ is Beta(a, b) with $a = b = 1$. I prefer to write $S_n = \sum_{j=1}^{n} X_j$. First compute $\mathsf{E}\left[S_n\,\middle|\,p\right]$. Then compute $\mathsf{E}\left[\mathsf{E}\left[S_n|p\right]\right]$. Note that the outer expectation is an integral with respect to p and the density of Beta(1, 1).

For the variance, use Eve's law.

b. Use Beta-Binomial conjugacy. Or use the insights of BH.9.56 and BH.9.57.

c. Bayes Billiards.

h.10.3. 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?

h.10.6. Apple the idea of BH.10.1.3 to $W = (X-\mu)^2$.

h.10.9. a. Jensen's inequality, $g(x) = e^x$

b. Use symmetry: X and Y are iid.

c. Which set of events is larger?

d. Use Jensen's inequality and Cauchy-Schwarz.

e. Eve's law.

f. Use Markov's inequality and the triangle inequality

h.10.28. The idea is to prove that the MGF of X_n converges to the MGF of a $N(\mu, \sigma^2)$ rv as $n \to \infty$. Thus, read and follow the proof of the CTL, BH.10.3.1.

What are $\mathsf{E}[X_n]$ and $\mathsf{V}[X_n]$ if $X \sim \text{Pois}(n)$? Once you know that, explain that the MGF of the standardized version of X_n is equal to $\exp\{-n + s\sqrt{n} + ne^{-s/\sqrt{n}}\}$.

Perhaps you should do BH.10.27 first.

h.10.30. a. See BH.10.3.7. Try to convert the recursion for Y_n to a form as in that example.

b. Just substitute α in the relevant formula of part a.

12 SOLUTIONS

s.7.1. Use the hint. If you make a drawing, then you'll see that Alice and Bob will not meet on the triangles $\{(x,y) : x \in [0,3/4], y \in [1/4,1], y > x+1/4\}$ and $\{(x,y) : x \in [1/4,1], y \in [0,3/4], y < x-1/4\}$. The area of each triangle is $(3/4)^2/2$, hence, the combined area is $9/16$. Therefore the probability to meet is $1-9/16 = 7/16$.

We can also solve a 2D integral by first integrating along y, and then along x. Let's focus on the integral over y first.

$$\begin{aligned}
\int_0^1 I_{x<y+1/4}\, I_{y<x+1/4}\,\mathrm{d}y &= \int_0^1 I_{x-1/4<y<x+1/4}\,\mathrm{d}y \\
&= \int_0^1 I_{\max\{0,x-1/4\}<y<\min\{1,x+1/4\}}\,\mathrm{d}y \\
&= \min\{1, x+1/4\} - \max\{0, x-1/4\}
\end{aligned}$$

Now the integral over x:

$$\begin{aligned}
\int_0^1 (\min\{1,x+1/4\} - \max\{0,x-1/4\})\,\mathrm{d}x &= \int_0^1 \min\{1,x+1/4\}\,\mathrm{d}x - \int_0^1 \max\{0,x-1/4\}\,\mathrm{d}x \\
&= \int_0^{3/4} (x+1/4)\,\mathrm{d}x + \int_{3/4}^1 1\,\mathrm{d}x - \int_{1/4}^1 (x-1/4)\,\mathrm{d}x \\
&= 0.5x^2\Big|_0^{3/4} + 3/4\cdot 1/4 - 0.5x^2\Big|_{1/4}^1 + 3/4\cdot 1/4 \\
&= 1/2\cdot 9/16 + 3/16 - 1/2\cdot 15/16 + 3/16 = 7/16.
\end{aligned}$$

s.7.9. a. The hint is the solution.

b. Use the hint for a. Then, for $k = 0, 1, \ldots, n$,

$$\mathsf{P}\{X=k, N=n\} = \mathsf{P}\{X=k, Y=n-k\} = pq^k pq^{n-k} = p^2 q^n.$$

(Have we used independence somewhere?)

c. Observe that the right hand side does not depend on k. This implies that $\mathsf{P}\{X=k|N=n\}$ also does not depend on k. (Why?) But, since $\mathsf{P}\{X=k|N=n\}$ is a true PMF, is must be that $\sum_{k=0}^n \mathsf{P}\{X=k|N=n\}$ adds up to 1. These two ideas put together imply that $\mathsf{P}\{X=k|N=n\} = 1/(n+1)$.

With Bayes' expression, and using that $\mathsf{P}\{X=k|N=n\} = 1/(n+1)$,

$$\mathsf{P}\{X=k|N=n\} = \frac{\mathsf{P}\{X=k, N=n\}}{\mathsf{P}\{N=n\}},$$

it follows that

$$\mathsf{P}\{N=n\} = \frac{\mathsf{P}\{X=k,N=n\}}{\mathsf{P}\{X=k|N=n\}} = \frac{p^2q^n}{1/(n+1)} = (n+1)p^2q^n.$$

s.7.10. Just reasoning as if there is no problem, i.e., applying Bayes' rule in a naive way,

$$\begin{aligned} F_T(t|x) &= \mathsf{P}\{T \leq t | X = x\} = \mathsf{P}\{X + Y \leq t | X = x\} \\ &= \mathsf{P}\{Y \leq t - x, X = x\} / \mathsf{P}\{X = x\} = \mathsf{P}\{Y \leq t - x\}\, \mathsf{P}\{X = x\} / \mathsf{P}\{X = x\} \\ &= \mathsf{P}\{Y \leq t - x\}, 0 \leq x \leq t. \end{aligned}$$

where I use that Y and X are independent to split the probability.

The problem with this derivation is that we multiply and divide by 0 ($= \mathsf{P}\{X = x\}$) just as if all is ok. But hopefully, you know that when we multiply and divide by zero, we can get any answer we like. A better way is as follows. Note beforehand that I do not expect that you could have come up with such an answer, but you should definitely study it.

The first step is to realize that PDF $f_{T|X}(t|x) = f_{TX}(t,x)/f_X(x)$ *is* well defined; we don't divide by zero because $f_X(x) > 0$ on $x \geq 0$. By the proof of BH.8.2.1 we see that $f_{TX}(t,x) = f_X(x) f_Y(t-x)\, I_{0\leq x \leq t}$, where I include the indicator to ensure that we don't run out of the support of X and T. Thus,

$$f_{T|X}(t|x) = \frac{f_{TX}(t,x}{f_X(x)} = f_X(x) f_Y(t-x)\, I_{0\leq x\leq t} / f_X(x) = f_Y(t-x)\, I_{0\leq x\leq t}.$$

Now we know that a conditional PDF is a full-fledged PDF. So we can use idea that to *define* the conditional CDF as follows:

$$F_{T|X}(t|x) := \int_0^t f_{T|X}(v|x)\, I_{0\leq x\leq v}\, \mathrm{d}v = \int_x^t \lambda e^{-\lambda(v-x)}\, \mathrm{d}v = \int_0^{t-x} \lambda e^{-\lambda v}\, \mathrm{d}v = 1 - e^{-\lambda(t-x)}.$$

Isn't it a bit strange that we get the same answer? How to get out of this situation in a technically correct way is one of the hard parts of (mathematical) probability, and certainly not something we can deal with in this course. All books on elementary[1] probability, and lecturers similarly, struggle with this problem; this course is not an exception, nor am I.

b. See part a.

c. By the above,

$$\begin{aligned} &f_{X|T}(x|t) = f_{TX}(t,x)/f_T(t), \\ &f_{TX}(t,x) = f_X(x) f_Y(t-x)\, I_{0\leq x\leq t} \lambda^2 e^{-\lambda x} e^{-\lambda(t-x)}\, I_{0\leq x\leq t} = \lambda^2 e^{-\lambda t}\, I_{0\leq x\leq t}, \\ &\qquad \implies f_{X|T}(x|t) \propto \lambda^2 e^{-\lambda t}\, I_{0\leq x\leq t}, \end{aligned}$$

where the last follows because $f_T(t)$ is just a normalization constant. Now we use some real nice, but subtle, reasoning to avoid computing f_T by means of marginalizing out x from $f_{T,X}(t,x)$. Observe that $f_{TX}(t,x)$ is constant *as a function of* x on $0 \leq x \leq t$ (in other words, the RHS does not depend on x on this interval). But $f_{X|T}(x|t)$ is also a real PDF. This implies that the constant $f_T(t)$ (since it does not depend on x) must be such that $f_{X|T}(x|t)$ integrates to 1 on $0 \leq x \leq t$. The only possibility is that $f_{X|T}(x|t) = t^{-1}\, I_{0\leq x\leq t}$.

---

[1] When in mathematics someething is elementary, it doesn't necessarily mean that that thing is simple. In fact, it can be very difficult. Elementary means that we just dont use very advanced mathematical concepts.)

This reasoning gives some offspin. We can conclude that

$$f_T(t) = f_{TX}(t,x)/f_{T|X}(t,x) = \lambda^2 t e^{-\lambda t}.$$

This is more than a nice trick. Recall it, as it is not only used more often in the book, but also in more advanced courses on data science and machine learning.

s.7.11. a. Use the hint. Next, $f_{Y|X}(y|x) \propto e^{-(y-x)^2}$, and a similar expression holds for $f_{Z|X}(z|x)$. Now follow the steps of the hint.

b. Read the hint. When Y is really big, X must be big (with large probability), so Z must be big too.

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,

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

s.7.13. Read the material of **??** for many detailed explanations on the exponential. As we will not repeat that, here are just the results. $\mathsf{P}\{X < Y\} = 1/2$. Hence, $\mathsf{P}\{X \leq x | X < Y\} = 2\,\mathsf{P}\{X \leq x, X < Y\}$.

$$\begin{aligned} 2\,\mathsf{P}\{X \leq x, X < Y\} &= 2\lambda^2 \int_0^\infty \int_0^\infty I_{u\leq x} I_{u<v} e^{-\lambda u} e^{-\lambda v} \,\mathrm{d}v\,\mathrm{d}u \\ &= 2\lambda^2 \int_0^\infty I_{u\leq x} e^{-\lambda u} \int_0^\infty I_{u<v} e^{-\lambda v} \,\mathrm{d}v\,\mathrm{d}u \\ &= 2\lambda \int_0^\infty I_{u\leq x} e^{-\lambda u} e^{-\lambda u} \,\mathrm{d}u \\ 1 - e^{-2\lambda x}. \end{aligned}$$

b. If $X < Y$, then we know that $X = \min\{X, Y\}$. But,

$$\{\min\{X,Y\} \leq x\} = \{X \leq x, X < Y\} \cup \{Y \leq x, Y \leq X\},$$

and the two sets on the RHS are disjoint. Hence, $\mathsf{P}\{\min\{X,Y\} \leq x\}$ is the sum of the probabilities on the RHS. By symmetry, these are equal.

s.7.15. Use the hint, that is, really make the drawing of the rectangle mentioned in the exercise. (If you refuse to to this, then nothing can help you to understand the rest of the answer.) Then, in the drawing, note that $F(x,y)$ 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 exactly once. Realize that in the process, the square $(-\infty,a_1]\times(-\infty,b_1]$ is subtracted twice.

s.7.24. a. From the hint,

$$\begin{aligned}\mathsf{P}\{Y_1 < cY_2\} &= \int\int I_{x<cy}\lambda_1 e^{-\lambda_1 x}\lambda_2 e^{-\lambda_2 y}\,\mathrm{d}x\,\mathrm{d}y = \lambda_1\lambda_2\int_0^\infty e^{-\lambda_1 x}\int_{x/c}^\infty e^{-\lambda_2 y}\,\mathrm{d}y\,\mathrm{d}x\\ &= \lambda_1\int_0^\infty e^{-\lambda_1 x}e^{-\lambda_2 x/c}\,\mathrm{d}x = \frac{\lambda_1}{\lambda_1+\lambda_2/c}.\end{aligned}$$

Check the result for $c=0$ and $c=\infty$.

I prefer to use conditioning, like this:

$$\begin{aligned}\mathsf{P}\{Y_1 < cY_2\} &= \int \mathsf{P}\{Y_1 < cY_2 | Y_1 = x\}\,\lambda_1 e^{-\lambda_1 x}\,\mathrm{d}x = \int \mathsf{P}\{Y_2 > x/c | Y_1 = x\}\,\lambda_1 e^{-\lambda_1 x}\,\mathrm{d}x\\ &= \int e^{-\lambda x/c}\lambda_1 e^{-\lambda_1 x}\,\mathrm{d}x,\end{aligned}$$

and the rest goes as before. Actually, I tend to use conditioning as it helps to make the reasoning easier. In this case, suppose that I know that $Y_1 = x$, what can I say about $\mathsf{P}\{Y_2 > cx\}$?

BTW, conditioning does not always make things simpler. When rvs are dependent, then you have to watch out.

b. See the solutions of BH on the web.

s.7.29. All is covered in **??**.

s.7.38. First check **??**.

In general, I am always very careful with such 'shortcuts' such as $\max\{X,Y\}+\min\{X,Y\} = X+Y$. As a matter of fact, I try to avoid such arguments because it is easy to go wrong. Seemingly plausible arguments are often wrong due to overlooked dependency or non-linearity (effects of higher moments).

It is useful to write $\max\{x,y\} = x\,I_{x\geq y} + y\,I_{y>x}$, and something similar for the minimum. In the present case, $\mathsf{Cov}[X,Y] = \mathsf{E}[XY] - \mathsf{E}[X]\,\mathsf{E}[Y]$, and, similarly, $\mathsf{Cov}[M,L] = \mathsf{E}[ML] - \mathsf{E}[M]\,\mathsf{E}[L]$, where M is max, and L is min. With the above indicators, it is simple to show that $\mathsf{E}[ML] = \mathsf{E}[XY]$:

$$\begin{aligned}ML &= (X\,I_{X\geq Y} + Y\,I_{Y\geq X})((X\,I_{X<Y} + Y\,I_{Y<X})\\ &= XY\,I_{X\geq Y} + XY\,I_{Y<X} = XY\end{aligned}$$

since $I_{X\geq Y}\,I_{X<Y} = 0$.

However, take $X,Y\sim \mathrm{Exp}(\lambda)$. Then, $\mathsf{E}[M] = 3/(2\lambda)$ and $\mathsf{E}[L] = 1/(2\lambda)$, but $\mathsf{E}[X] = \mathsf{E}[Y] = 1/\lambda$.

s.7.53. a. In my notation, $X_i = 0 \implies Y_i \neq 0$ and $X_i \neq 0 \implies Y_i = 0$. The reason is that in step i, the drunkard makes a step left or right OR up or down. However, s/he cannot move to the right and up at the same time.

Here is 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, $\mathsf{E}\left[R_n^2\right] = n$.

s.7.58. b. Use the hint. Then, 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. This is the part of the exercise that explains 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 property i. of covariance 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 property ii., what happens to the area of rectangle if you stretch the sides? For property iii., realize that this is just a shift of a rectangle that leaves its area invariant. For property iv., what happens to the area if you put an extra rectangle on top or to the right?

BTW, property iii. follows directly from property 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.

Here are some final remarks.

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

1. It is bilinear, that is, 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. It is symmetric: $\mathsf{Cov}[X, Y] = \mathsf{Cov}[Y, X]$, 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 we do some geometry. Take three vectors $x, y, z \in \mathbb{R}^2$ (it's easy to generalize to $\mathbb{R}^n$). Then we know that the area $D(x, y)$ of the parallelogram spanned by vectors x and y satisfies the following properties.

1. Area is bilinear. 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 $a \in \mathbb{R}$. (Just make a drawing to convince you about this.) 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. $D(x, x) = 0$; there is no area between x and x.

3. $D((1, 0), (0, 1)) = 1$; the area of the square with side 1 is 1.

In fact, thex 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. In fact, this is the determinant of the matrix with as columns the vectors that span the parallellogram. Moreover, it can be shown that the second property can be replaced by the skew-symmetric property: $D(x, y) = -D(x, y)$. (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 $x = (a, b)$ and $y = (c, d)$ in 2D. Then

$$\begin{aligned} D(x, y) &= D((a, b), (c, d)) = D(a(1, 0)) + b(0, 1), c(1, 0) + d(0, 1)) \\ &= adD((1, 0), (0, 1)) + bcD((0, 1), (1, 0)) = ad - bc, \end{aligned}$$

where we use bilinearity in the first step, and skew-symmetry in the second and third. And this is indeed the determinant of the matrix with x and y as columns.

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.

s.7.59. a. Follows directly from the hint.

Check the hint!

c. If X_i and Y_j are iid, it must be that $w_1 = n/(n + m)$.

b. 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

$$\mathsf{V}[\hat{\theta}_1] = \mathsf{E}\left[\left(\frac{1}{n}\sum_i X_i - \theta\right)^2\right] = n^{-2}\,\mathsf{E}\left[\left(\sum_i (X_i - \theta)\right)^2\right] = n^{-2}\,\mathsf{V}\left[\sum_i X_i\right] = \mathsf{V}[X_1]/n = \sigma^2/n.$$

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.

Let us finally use the hint of BH 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'll get the above expression.

s.7.71. a. Multinomial.

b. With the hint we end up at $X_1 + X_2 \sim \mathrm{Bin}(n, p^2 + 2p(1-p))$.

c. Here is a short intermezzo on finding a recursion for the sum of a number of Bernouilli rvs. 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 recursion with $f_n(i)$ for the genes tox 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.

s.7.86. a. It is given that $\mathsf{P}\{T \le t \,|\, D = 1\} = G(t)$ and $\mathsf{P}\{T \le t \,|\, D = 0\} = H(t)$. From Theorem 5.3.1.i, we have that we can associate a rv. to a CDF F. Sometimes we say that the CDF F /induces/ a rv. X. So let us use this here to say that G induces the rv. 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:

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

s.8.11. From the hint, we first focus on a set $\{V \le 0\} = \{1/T \le 0\}$. Now, $1/T \le 0 \iff T \le 0$. And therefore $\mathsf{P}\{V \le 0\} = \mathsf{P}\{T \le 0\} = F_T(0)$.

If $v < 0$, then $1/T \le v \le 0 \iff 1/v \le T \le 0$. Therefore $F_V(v) = F_T(0) - F_T(1/v)$.

If $v > 0$, then $1/T \le v$ when $T < 0$ or $T \ge 1/v$. Hence, $F_V(v) = F_T(0) + 1 - F_T(1/v)$.

s.8.15. a. I remember this: $f_{X,Y}(x,y)\,\mathrm{d}x\,\mathrm{d}y = f_{R,\Theta}(r,\theta)\,\mathrm{d}r\,\mathrm{d}\theta$. From this,

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

Now, since $x = r\cos\theta$ and $y = r\sin\theta$,

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

which has determinant equal to r. It is given that $f_{X,Y}(x,y) = g(x^2 + y^2) = g(r^2)$. Hence,

$$f_{R,\Theta}(r,\theta) = f_{X,Y}(x,y)r = g(r^2)r,$$

with $r \geq 0, \theta \in [0, 2\pi$. The RHS does not depend on θ. Hence, $f_\Theta(\theta)$ must be a constant.

b. Use the hint. Since g is a constant, $f_{R,\Theta}(r,\theta) \propto r$. Thus,

$$\int_0^1 \int_0^{2\pi} r \, \mathrm{d}r \, \mathrm{d}\theta = 2\pi(1/2)r^2|_0^1 = \pi.$$

So, $1/\pi$ is the normalization constant.

c. $f_{X,Y}(x, y) = \exp - x^2/\sqrt{2\pi} \exp - y^2/\sqrt{2\pi} = \exp -(x^2 + y^2)/2\pi = \exp -r^2/2\pi$. Indeed, $f_{X,Y}(x, y)$ has the form $g(x^2 + y^2)$. The rest is as in part b.

s.8.18. I always start with this line: $f_{T,U}(t, u) \, \mathrm{d}t \, \mathrm{d}u = f_{X,Y}(x, y) \, \mathrm{d}x \, \mathrm{d}y$. Then, since $x = u$, $y = u/t$.

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

We don't need to take absolute signs in the last expression because X, Y are positive rvs. With this,

$$f_{T,U}(t, u) = f_{X,Y}(x, y) \left(\frac{\partial(t, u)}{\partial(x, y)} \right)^{-1} = f_{X,Y}(u, u/t) = f_X(u) f_Y(u/t) u/t^2.$$

b. Use part a.

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

Since f_X and f_Y are not given explicitly, we cannot make further progress.

All and all, division of rvs is not so simple.

s.8.23. a.

$$f_{X,T}(x, t) = f_{X,Y}(x, y) |\frac{\partial(x, y)}{\partial(x, t)}|,$$

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

$$\Longrightarrow$$

$$f_{X,T}(x, t) = f_{X,Y}(x, y)/x = f_{X,Y}(x, t/x)/x,$$

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.

s.8.36. a. I did not attempt any smart tricks. Take as transform $u = s/t$ and $v = s + t$, where I associate s to T_1 and t to T_2. Then,

$$\frac{\partial(u, v)}{\partial(s, t)} = \begin{pmatrix} 1/t & -s/t^2 \\ 1 & 1 \end{pmatrix} = \frac{1}{t} + \frac{s}{t^2} = \frac{t + s}{t^2} = \frac{v}{t^2}$$

With a bit of algebra: $s = uv/(u+1)$ and $t = v/(u+1)$. Therefore the Jacobian we need becomes

$$\frac{\partial(s,t)}{\partial(u,v)} = \frac{v}{t^2} = v\frac{v^2}{(u+1)^2} = \frac{(u+1)^2}{v}. \tag{12.0.1}$$

Next,

$$\begin{aligned}
f_{U,V}(u,v) &= f_{T_1,T_2}(s,t)\frac{\partial(s,t)}{\partial(u,v)} = f_{T_1}(uv/(u+1)f_{T_2}(v/(u+1))\frac{(u+1)^2}{v} \\
&= \lambda^2\exp(-\lambda uv/(u+1) - \lambda v/(u+1))\frac{(u+1)^2}{v} \\
&= \lambda^2\exp(-\lambda v)\frac{(u+1)^2}{v} \\
&= \lambda^2\exp(-\lambda v)/v\cdot(u+1)^2.
\end{aligned}$$

Clearly, this is a product of a function $f_V(f)$ that depends only v and another function $f_U(u)$ that depends only u. Hence, U and V are independent.

b. With the hint,

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

c. See the hint. Alice first has to wait for the first server to become free. The expected time in queue is $1/(\lambda_1+\lambda_2)$. If server 1 is the first, then Alice spends a time $1/\lambda_1$ in service. Thus, the total time is

$$\frac{1}{\lambda_1+\lambda_2} + \frac{\lambda_1}{\lambda_1+\lambda_2}\frac{1}{\lambda_1} + \frac{\lambda_2}{\lambda_1+\lambda_2}\frac{1}{\lambda_2} = \frac{3}{\lambda_1+\lambda_2}.$$

s.8.40. Let us solve the question from first principles. At the end, I'll give the short solution based on Beta-Binomial conjugacy.

Let $f(p)$ be our prior density (In the exercise it is taken to be uniform). Then

$$\mathsf{P}\{p > r\} = \int I_{p>r} f(p)\,\mathrm{d}p = \int_r^1 f(p)\,\mathrm{d}p$$

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 $\mathsf{P}\{p > r\} \geq c$?

We can start with one trial, i.e., $n = 1$. Then we analyze the case for $n = 2$, and so on, and hope to see a pattern. Here are the standard steps of Bayesian reasoning.

1. I want to know the density $f_1(p|N=1)$, i.e, the density of p after having seen one successful test. (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. It is clear that $f_1(N=1|p) = p$, because we know that an item passes a test with probability p, when p is given.

4. Perhaps I don't need $\mathsf{P}\{N=1\}$ if I can guess it (though see below), but here it is just for completeness' sake.

$$\mathsf{P}\{N=1\} = \int f(N=1|p) f(p)\,\mathrm{d}p = \int_0^1 p\,\mathrm{d}p = 1/2,$$

   because the prior $f(p) = I_{p\in[0,1]}$, i.e., uniform on $[0,1]$, i.e, it is Beta$(1,1)$.

5. With this, $f_1(p|N=1) = \frac{p}{1/2} I_{p\in[0,1]} = 2p\, I_{p\in[0,1]}$.

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

Sometimes we are lucky and we don't have to compute the denominator in Bayes' formula. We did this earlier, but let's show again how this works.

$$f_1(p|N=1) = \frac{f_1(p, N=1)}{\mathsf{P}\{N=1\}} \propto f_1(N=1|p) f(p) = p\, I_{0\le p\le 1}.$$

Now $f_1(p|N=1)$ is a PDF, hence must integrate to 1. Thus, $\int_0^1 p\,\mathrm{d}p = 1/2$, must be the normalization constant by which we have to divide to turn f_1 into a real PDf. In this case we don't save any work, but sometimes this really helps, in particular when dealing with integrals with Beta distributed rvs.

Now generalize to larger n, compute $f_2(p|N=2)$, then for $n=3$, and so on, until you see the pattern.

We can also directly use the ideas of the book. Starting with a prior Beta$(1,1)$, after n 'wins', the distribution becomes Beta$(1+n,1)$. Then,

$$\mathsf{P}\{p > r\} = \frac{\Gamma(n+2)}{\Gamma(n+1)\Gamma(1)} \int_r^1 p^n\,\mathrm{d}p = 1 - r^{n+1}. = (n+1)p^{n+1}|_r^1 = 1 - r^{n+1}.$$

s.8.52. a. By the hint and the fact that U_j is uniform on $[0,1]$, so that $1-U_j$ is also uniform, the last equality of the hint implies that $\mathsf{P}\{1-U_j \le 1-e^{-c}\} = \mathsf{P}\{U_j \le 1-e^{-c}\} = 1-e^{-c}$. But then, $X_j \sim \text{Exp}(1)$.

b. The sum of n iid exponentials is Gamma(n,λ). And so, if $S_n = \sum_i^n X_i$, then $\mathsf{P}\{S_n \le x\} = \int_0^x f(y)\,\mathrm{d}y$, with $f(y)$ the gamma density with n and $\lambda = 1$.

Just to test my skills, I used MGFs, because I know that the MGF of a sum of iid rvs is the product of the MGF of one them. Since $e^{\log u} = u$,

$$\mathsf{E}\left[e^{-s\log U}\right] = \int_0^1 e^{-s\log u}\,\mathrm{d}u = \int_0^1 u^{-s}\,\mathrm{d}u.$$

If $s \geq 1$ this does not converge (convince yourself that you understand this). With $s < 1$,

$$\mathsf{E}\left[e^{-s\log U}\right] = \frac{1}{-s+1} u^{-s+1}|_0^1 = \frac{1}{1-s}.$$

Therefore,

$$\begin{aligned}\mathsf{E}\left[e^{-sS_n}\right] &= \mathsf{E}\left[e^{-s\log U_1 - s\log U_2 - \cdots - s\log U_n}\right] = \left(\mathsf{E}\left[e^{-s\log U}\right]\right)^n \\ &= \left(\frac{1}{1-s}\right)^n,\end{aligned}$$

and this is the MGF of a Gamma$(n, \lambda = 1)$ rvs.

s.8.54.

$$M_Y(s) = \mathsf{E}\left[\exp sY\right] = p\sum_{i=0}^{\infty} e^{spi} q^i = p/(1-qe^{sp}).$$

Now, use that $e^{sp)} \approx 1 + sp$ for $p \ll 1$. (This is easier than using l' Hopital's rule as BH do in their solution). Hence, the denominator becomes $\approx 1-(1-p)(1+sp) = p(1-s) - sp^2 \approx p(1-s)$ when $p \ll 1$ Hence,

$$M_Y(s) \approx p/(p(1-s) = 1/(1-s).$$

In the limit $p \to 0$ the LHS converges to the RHS, which is the MGF of an exponential rv. For the rest, follow the solution of BH.

Here is another line of attack. Let us first use probability theory to find out what is $\sum_{i=0}^{\infty} q^1$ for some $|q| < 1$. Take $X \sim \text{Geo}(p)$, so that X corresponds to the number of failures (tails say) until we see a success (heads say). So, X corresponds to the number of tails until we see a heads. Now if we keep on throwing, then we know that eventually a heads will appear. Therefore $p + pq + pq^2 + \cdots = 1$, that, is $p\sum_i^{\infty} q^i = 1$. But this implies that $\sum_{i=0}^{\infty} q^i = 1/p = 1/(1-q)$.

By similar reasoning, if we keep on throwing the coin until we see r heads then we know that $p^r \sum_{i=0}^{\infty} \binom{r+i-1}{r} q^i = 1$. Therefore,

$$\sum_{i=0}^{\infty}\binom{r+i-1}{r} q^i = \frac{1}{p^2} = \frac{1}{(1-q)^r}.$$

With this insight, for $X \sim \text{NBin}(p, n)$

$$\begin{aligned}M_X(s) &= p^r \sum_{i=0}^{\infty}\binom{r+i-1}{r} q^i e^{si} = p^r \sum_{i=0}^{\infty}\binom{r+i-1}{r} (e^s q)^i \\ &= \frac{p^r}{(1-qe^s)^r} \approx \left(\frac{p}{p(1-s)}\right)^r,\end{aligned}$$

where we use again Taylor's expansion for $p \ll 1$.x

s.9.1. a.

$$\begin{aligned}
\mathsf{E}\left[T \,\middle|\, R=j\right] &= \mu_j \\
\mathsf{E}[T \mid R] &= \mu_R = \sum_j \mu_j I_{R=j}, \\
\mathsf{E}[T] = \mathsf{E}[\mathsf{E}[T \mid R]] &= \mathsf{E}\left[\sum_j \mu_j I_{R=j}\right] = \sum_j \mu_j p_j \\
&= (\mu_1+\mu_2+\mu_3)/3.
\end{aligned}$$

For b., use a. And then,

$$\begin{aligned}
\mathsf{V}\left[T|R=j\right] &= \sigma_j^2 \\
\mathsf{V}[T|R] &= \sum_j \sigma_j^2 I_{R=j} \\
\mathsf{E}[VT|R] &= (\sigma_1^2+\sigma_2^2+\sigma_3^3)/3, \\
\mathsf{V}[\mathsf{E}[T \mid R]] &= \mathsf{E}\left[(\mathsf{E}[T \mid R])^2\right] - (\mathsf{E}[\mathsf{E}[T|R]])^2 \\
(\mathsf{E}[T \mid R])^2 &= \left(\sum_j \mu_j I_{R=j}\right)^2 = \sum_j \mu_j^2 I_{R=j} + 2\mu_1\mu_2 I_{R=1} I_{R=2} + \cdot \\
&= \sum_j \mu_j^2 I_{R=j}. \\
\mathsf{E}\left[(\mathsf{E}[T \mid R])^2\right] &= \sum_j \mu_j^2 p_j = (\mu_1^2+\mu_2^2+\mu_3^2)/3 \\
\mathsf{V}[\mathsf{E}[T \mid R]] &= \mathsf{E}\left[(\mathsf{E}[T \mid R])^2\right] - (\mathsf{E}[\mathsf{E}[T|R]])^2 \\
&= (\mu_1^2+\mu_2^2+\mu_3^2)/3 - (\mu_1+\mu_2+\mu_3)^2/9. \\
\mathsf{V}[T] &= \mathsf{V}[\mathsf{E}[T|R]] + \mathsf{E}[\mathsf{V}[T|R]] \\
&= (\mu_1^2+\mu_2^2+\mu_3^2)/3 - (\mu_1+\mu_2+\mu_3)^2/9 + (\sigma_1^2+\sigma_2^2+\sigma_3^3)/3.
\end{aligned}$$

s.9.25.

$$\begin{aligned}
\mathsf{E}[X_{n+1}|X_n=100] &= 100 + pf100 - (1-p)f100 = 100(1-f+2pf) \\
\mathsf{E}[X_{n+1}|X_n] &= X_n(1-f+2pf) \\
\mathsf{E}[X_{n+1}] &= (1-f+2pf)\,\mathsf{E}[X_n] \\
\mathsf{E}[X_{n+1}] &= (1-f+2pf)^2\,\mathsf{E}[X_{n-1}] = (1-f+2pf)^{n+1} X_0.
\end{aligned}$$

s.9.32. Use that $P^2 = P$ (indicator funtion), Adam and Eve, and that $N \sim \text{Pois}(8\lambda)$,

$$\begin{aligned}
\mathsf{E}[Y|P] &= \mathsf{E}[PX|P] = \mathsf{E}[X]\,\mathsf{E}[P|P] = \mu P, & \mathsf{V}[Y|P] &= \mathsf{V}[XP|P] = P^2\,\mathsf{V}[X|P] = P\sigma^2 \\
\mathsf{E}[Y] &= \mu p, & \mathsf{V}[Y] &= \mathsf{E}[\mathsf{V}[Y|P]] + \mathsf{V}[\mathsf{E}[Y|P]] = \sigma^2 p + \mu^2 p(1-p), \\
\mathsf{E}[S|N] &= N\,\mathsf{E}[Y], & \mathsf{V}[S|N] &= N\,\mathsf{V}[Y] \\
\mathsf{E}[N] &= 8\lambda, & \mathsf{V}[N] &= 8\lambda.
\end{aligned}$$

Now use BH.9.6.1. It's just a matter of filling in.

s.9.37. a. Here you should assume that the X_i are not yet known. Thus, the expectation over X_i is taken with respect to the CDF F_X. Using the independence of X_j and S_j, $I_{S_j=i} I_{S_j=k} = 0$ if $i \neq k$, and that $\mathsf{E}\left[I_{S_j=k} \right] = 1/n$,

$$\begin{aligned}
\mathsf{E}[Y_j] &= \sum_i \mathsf{E}[X_i]\, \mathsf{E}\left[I_{S_j=i} \right] = \mu, \\
\mathsf{E}\left[Y_j^2 \right] &= \mathsf{E}\left[\sum_k \sum_l X_k X_l I_{S_j=k} I_{S_j=l} \right] = \mathsf{E}\left[\sum_k X_k^2 I_{S_j=k} \right] = \sum_k \mathsf{E}\left[X^2\right] n^{-1} = \mathsf{E}\left[X^2\right], \\
\mathsf{V}[Y_j] &= \mathsf{E}\left[Y_j^2 \right] - (\mathsf{E}[Y_j])^2 = \sigma^2.
\end{aligned}$$

b. Now we are given the outcomes (samples) $X_i = x_i$ of n experiments. I prefer to write $D = X_1, \ldots, X_n$ as it is shorter. Noting that S_j and D are independent, and that $\mathsf{E}[X_k | D] = X_k$,

$$\mathsf{E}[Y_j | D] = \sum_k X_k \mathsf{E}\left[I_{S_j=k} | D \right] = \frac{1}{n} \sum_k X_k := \bar{X}$$

Observe that this average need not be the same as μ!

The conditional variance. Since S_j and S_k are independent when $j \neq k$, it must be that $Y_j|D$ and $Y_k|D$ are also conditionally independent. Moreover, $\{Y_j|D\}$ are conditionally iid. Therefore,

$$\begin{aligned}
\mathsf{E}\left[Y_j^2 | D \right] &= \mathsf{E}\left[\sum_k \sum_l X_k X_l I_{S_j=k} I_{S_j=l} | D \right] \\
&= \mathsf{E}\left[\sum_k X_k^2 I_{S_j=k} | D \right] = \sum_k X_k^2 \mathsf{E}\left[I_{S_j=k} | D \right] \\
&= \frac{1}{n} \sum_k X_k^2, \\
\mathsf{V}[Y_j | D] &= \frac{1}{n} \sum_k X_k^2 - (\bar{X})^2 = \frac{1}{n} \sum_k (X_k - \bar{X})^2 = \frac{n-1}{n} \sigma^2, \\
\mathsf{V}[\bar{Y} | D] &= \mathsf{V}\left[\frac{1}{n} \sum_j Y_j | D \right] = \frac{1}{n^2} \sum_j \mathsf{V}[Y_j | D] = \frac{1}{n} \mathsf{V}[Y_1 | D].
\end{aligned}$$

c. For $\mathsf{E}[\bar{Y}]$ use linearity and Adam's law:

$$\mathsf{E}[\bar{Y}] = \mathsf{E}[\mathsf{E}[\bar{Y} | D]] = \frac{1}{n} \sum_k \mathsf{E}[X_k] = \mathsf{E}[X] = \mu.$$

Here are the details for $\mathsf{V}[\bar{Y}]$. Using BH.6.3.3 and BH.6.3.4,

$$\begin{aligned}
\mathsf{E}[\mathsf{V}[\bar{Y} | D]] &= \frac{1}{n} \mathsf{E}[\mathsf{V}[Y_1 | D]] = \frac{1}{n^2} \mathsf{E}\left[\sum_{i=1}^n (X_i - \bar{X})^2 \right] \\
&= \frac{n-1}{n^2} \mathsf{E}\left[\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right] = \frac{n-1}{n^2} \mathsf{E}\left[S_n^2\right] = \frac{(n-1)\sigma^2}{n^2} \\
\mathsf{V}[\mathsf{E}[\bar{Y} | D]] &= \mathsf{V}[\bar{X}] = \frac{1}{n^2} \sum_i \mathsf{V}[X_i] = \frac{1}{n} \sigma^2.
\end{aligned}$$

Now use Eve's law to add both terms to get $V\bar{Y}$.

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

The extra exercise: 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?

s.9.50. a. Using the hint gives us $\mathsf{E}[N\,|\,\lambda] = \lambda$ and $\mathsf{V}[N\,|\,\lambda] = \lambda$.

Now use Adam and Eve.

b. Just copy the formulas of BH.9.6.1

c. With the hint, observe that $\text{Exp}(()1) = \Gamma 1, 1$. In the relevant formula of BH.8.4.5 ($\mathsf{P}\{Y=y\}$), take $t = r_0 = b_0 = 1$ and conclude that $\mathsf{P}\{N=n\} = 2^{-n-1}$. Hence, $N \sim \text{Geo}(1/2)$.

d. Same story. The relevant formula is $f_1(\lambda|y)$.

s.9.56. a. From the hint,

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

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.

b. Take $Y = 1+G$, then Y has the first success distribution since G is geometric. Hence, $\mathsf{E}[Y] = (a+b)/a = 1+b/a$. Clearly, this is smaller than $1+b/(a-1) = \mathsf{E}[T]$.

But why is this so?

I must miss something here. The prior is $\text{Beta}(a,b)$. 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. Immediate from BH.8.3.3: $p|X=7 \sim \text{Beta}(a+7, b+3)$.

s.9.57. a. By the hint,

$$f(p|X_1=x_1) \propto f(p, X_1=x_1) \propto p^{a-1}q^{b-1}p^{x_1}q^{1-x_1} \propto p^{a+x_1-1}q^{b+(1-x_1)-1}.$$

Hence, $p|X_1=x_1 \sim \text{Beta}(a+x_1, b+(1-x_1))$. We can now use this as prior to see that $p|X_1 = x_a, X_2 = x_2 \sim \text{Beta}(1+x_1+x_2, 1+(1-x_1)+(1-x_2))$, and so on. Hence, $p|X_1,\dots X_n \sim \text{Beta}(1+S_k, 1+n-S_k)$.

b. With the hint, $\mathsf{P}\{X_{n+1}=1|p\}=p$ and $\mathsf{P}\{S_n=k|p\}=\binom{n}{k}p^kq^{n-k}\propto p^kq^{n-k}$. Also $X_{n+1}|p$ and $S_n|p$ are conditionally independent. Therefore,

$$\mathsf{P}\{X_{n+1}=1,S_n=k|p\}\propto pp^kq^{n-k}=p^{k+1}q^{n-k},$$

which in turn implies that

$$\mathsf{P}\{X_{n+1}=1|S_n=k,p\}\propto=p^{k+1}q^{n-k}.$$

Hence, $X_{n+1}|S_n=k,p\sim\mathrm{Beta}(k+2,n-k+1)$. Now, $X_{n+1}\in\{0,1\}$, so that $\mathsf{P}\{X_{n+1}=1|S_n=k,p\}=\mathsf{E}\left[X_{n+1}|S_n=k,p\right]=(k+2)/(n+3)$, since $X_{n+1}|S_n=k,p\sim\mathrm{Beta}(k+2,n-k+1)$.

The last step is to realize that $\mathsf{E}[X_{n+1}|S_n=k]=\mathsf{E}\left[\mathsf{E}\left[X_{n+1}|S_n=k,p\right]|S_{n=k}\right]$.

Here is another way to get the same result.

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

Now simplify.

s.9.58.

$$\begin{aligned}
\mathsf{E}\left[S_n|p\right]&=np\\
\mathsf{E}\left[p\right]&=\frac{1}{\beta(a,b)}\int_0^1 pp^{a-1}q^{b-1}\,\mathrm{d}p=\frac{\beta(a+1,b)}{\beta(a,b)}=\frac{a}{a+b}=1/2\\
\mathsf{E}\left[\mathsf{E}\left[S_n|p\right]\right]&=n\mathsf{E}\left[p\right]=n/2.\\
\mathsf{V}\left[S_n|p\right]&=npq\\
\mathsf{E}\left[\mathsf{V}\left[S_n|p\right]\right]&=n\mathsf{E}\left[pq\right]=n\mathsf{E}\left[p\right]-n\mathsf{E}\left[p^2\right]=n/2-n\mathsf{E}\left[p^2\right]\\
\mathsf{E}\left[p^2\right]&=\frac{1}{\beta(a,b)}\int_0^1 p^2p^{a-1}q^{b-1}\,\mathrm{d}p=\frac{\beta(a+2,b)}{\beta(a,b)}=\frac{a(a+1)}{(a+b)(a+b+1)}=\frac{2}{2\cdot 3}=1/3\\
\mathsf{V}\left[\mathsf{E}\left[S_n|p\right]\right]&=\mathsf{V}\left[np\right]=n^2\mathsf{V}\left[p\right]=n^2/12.
\end{aligned}$$

The rest of Eve's law is now trivial.

b. We start with a $\mathrm{Beta}(1,1)$ prior on p. After the first win, the prior gets updated to $\mathrm{Beta}(1+1,1)$, after a loss to $\mathrm{Beta}(1,1+1)$. Reasoning like this, after a wins and $j-a$ losses, the distribution for a win becomes $\mathrm{Beta}(1+a,a+j-1)$. Therefore, by using the hint in the book, $\mathsf{E}\left[p|S_j=a\right]=(a+1)/(j+2)$.

c. 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 BH.8.3.2. Hence, $\mathsf{P}\{X=k\}=1/(n+1)$. Observe that we use the prior $p \sim$ Beta$(1,1)$.

When the prior is Beta$(a, j-a)$, we should get the negative hypergeometric distribution, see the remark in BH.8.3.3.

d. Shanille scores the first and missed the second. Hence, there are 98 shots left, out which she has to score 49. Thus, we ask for $\mathsf{P}\left\{S_{98}=49|p\right\}$, where $p \sim$ Beta$(a=1, b=1)$ is the prior since she hit $a=1$ out of $a+b=2$ shots. This places us in the situation of part c above, with $n=98$. Hence, $\mathsf{P}\left\{S_{98}=49|p\right\}=1/99$.

s.10.6. Take W as in the hint and $Z=1$. By the inequality of Cauchy-Schwarz, $(\mathsf{E}[W])^2 \geq \mathsf{E}\left[W^2\right]$. The LHS is σ^4, the RHS is $\mathsf{E}\left[(X-\mu)^4\right]$. The rest follows right away from the definition of kurtosis.

s.10.9. a. $\leq$ Immediate from the hint.

b. =: immediate from the hint

c.

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

Both terms on the RHS are non-negative.

d. Use the hint. $(\mathsf{E}[XY])^2 \leq \mathsf{E}\left[X^2\right]\mathsf{E}\left[Y^2\right]=(\mathsf{E}\left[X^2\right])^2 \leq \mathsf{E}\left[X^4\right]$, where we use that X and Y are iid, so that $\mathsf{E}\left[X^2\right]$ and $\mathsf{E}\left[Y^2\right]$ are equal.

e. =: since X and Y are independent, $\mathsf{V}[Y|X]=\mathsf{V}[Y]$.

f. From the hint, $\mathsf{P}\{|X+Y|>3\} \leq \mathsf{E}[|X+Y|]/3 \leq \mathsf{E}[|X|]/3+\mathsf{E}[|Y|]/3=2\mathsf{E}[|X|]/3 \leq \mathsf{E}[|X|]$. Why is there not an < in the last step?

s.10.26. a. I did things a bit differently than in the book. Take $S_n=\sum_{i=1}^n X_i$ with $X_i \sim$ Bern(p). Then I know this:

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

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 Pois$(\lambda) \sim N(n,n)$.

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.

s.10.28. Since $X_n \sim \text{Pois}(n)$, $\mathsf{E}[X_n] = n$, $\mathsf{V}[X_n] = n$. Using the hints, with Y_n the standardized version of X_n:

$$\begin{aligned} M_{Y_n}(s) &= \sum_{i=0}^{\infty} e^{-n} n^i / i! \cdot e^{s(i-n)/\sqrt{n}} = e^{-n} e^{s\sqrt{n}} \sum_{i=0}^{\infty} (n e^{s/\sqrt{n}})^i / i! \\ &= \exp\{-n + s\sqrt{n} + n e^{-s/\sqrt{n}}\}. \end{aligned}$$

With Taylor's expansion for e^x to second order,

$$-n + s\sqrt{n} + n e^{-s/\sqrt{n}} \approx -n + s\sqrt{n} + n\left(1 - s/\sqrt{n} + s^2/2n\right) = s^2/2.$$

Now follow the proof of the CTL, BH.10.3.1.

s.10.30. a. Define I_n as the success indicator: it is 1 if I win, and 0 if I loose. For round 1, suppose I win, then $Y_1 = Y_0/2 + 1.7Y_0/2 = 1.35Y_0$. If I lose, $Y_1 = Y_0/2 + 0.5Y_0/2 = 0.75Y_0$. Therefore,

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

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

b. Use the hint.

$$\begin{aligned} Y_n &= Y_{n-1}(1+0.7\alpha)^{I_n}(1-0.5\alpha)^{1-I_n} \implies \\ \log Y_n &= \log Y_{n-1} + I_n \log(1+0.7\alpha) + (1-I_n)\log(1-0.5\alpha) \\ &= \log Y_0 + \log(1+0.7\alpha)\sum_{i=1}^{n} I_i + \log(1-0.5\alpha).\sum_{i=1}^{n}(1-I_i) \end{aligned}$$

By the strong law, $\sum I_i/n \to 1/2$ and $\sum(1-I_i)/n \to 1/2$. Therefore

$$n^{-1}\log Y_n \to 0.5\log(1+0.7\alpha) + 0.5\log(1-0.5\alpha) = 0.5\log((1+0.7\alpha)(1-0.5\alpha)) = g(\alpha)$$

For the maximum, take the derivative with respect to α. This gives $\alpha = 2/7$.

s.10.36. See the solution manual of BH.

s.10.39. 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 have the first success distribution, and $N-1$ be geometric.

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 $\sim \text{Exp}(\lambda)$, $S_n \sim \text{Pois}(\lambda t)$.

c. The sum of n iid Exp(1) rvs is Gamma$(n,1)$. Since $\bar{X}_n$ has mean 1, $X_n \sim \text{Gamma}(n,n)$. Then $\mathsf{V}[X_n] = 1/n$ (I just looked it up in the back of the book). By the CLT, $\bar{X}_n$ is approximated well by a Norm(μ, σ^2) rv with $\mu = 1, \sigma^2 = 1/n$.