Problem Sheet 7 [1]

Based on Chapters 5.1-5.4 of the course book.

Optional Revision Problems

Exercise 1. Consider the following simplified scenario based on Who Wants to Be a Millionaire?, a game show in which the contestant answers multiple-choice questions that have 4 choices per question. The contestant (Fred) has answered 9 questions correctly already, and is now being shown the 10th question. He has no idea what the right answers are to the 10th or 11th questions are. He has one "lifeline" available, which he can apply on any question, and which narrows the number of choices from 4 down to 2. Fred has the following options available.

- Walk away with \$16,000.
- Apply his lifeline to the 10th question, and then answer it. If he gets it wrong, he will leave with \$1,000. If he gets it right, he moves on to the 11th question. He then leaves with \$32,000 if he gets the 11th question wrong, and \$64,000 if he gets the 11th question right.
- Same as the previous option, except not using his lifeline on the 10th question, and instead applying it to the 11th question (if he gets the 10th question right).

Find the expected value of each of these options. Which option has the highest expected value? Which option has the lowest variance?

Hint: First derive the PMF-s, i.e. the probabilities of winning the different amounts, following each of the strategies respectively, outlined above.

Solution 1. Denote the amount of winnings following the strategies above with W_A, W_B and W_C respectively. Also, denote the event of getting the 10th question right with Q_{10}. To calculate the mean and the variance for these random variables (r.v.-s), first, we have to derive the respective PMF-s.

- W_A: $P(W_A = 16,000) = 1$, and $P(W_A = w) = 0$ for all $w \neq 16000$ as this strategy has no uncertainty: If you choose this, you are guaranteed to walk away with a fix amount of money.
- W_B: Here you have three options, winning \$1,000, \$32,000 or \$64,000, so $P(W_B = w) = 0$ for $w \notin \{1000, 32000, 64000\}$.

  We will use $P(Q_{10}) = 1/2$ with a slight abuse of notation, importantly, in this scenario everything is under the setting that we used the lifeline in round 10. (In the next part $P(Q_{10} = 1/4)$ will hold, as the lifeline is reserved for round 11. Technically we should either

[1] Exercises are based on the coursebook Statistics 110: Probability by Joe Blitzstein

condition on the choice of strategy or use subscripts for the probability measures P, but to avoid the notational burden, this is swept under the rug).

$$P(W_B = 1000) = P(Q_{10}^c) = 1/2,$$

as after using the lifeline we only have two options left, so with probability $1/2$ we choose the wrong answer and walk away with \$1,000.

Using the law of total probability we have

$$P(W_B = 32000) = P(W_B = 32000|Q_{10})P(Q_{10}) + P(W_B = 32000|Q_{10}^c)P(Q_{10}^c),$$

and similarly

$$P(W_B = 64000) = P(W_B = 64000|Q_{10})P(Q_{10}) + P(W_B = 64000|Q_{10}^c)P(Q_{10}^c).$$

After getting the 10th question wrong, we cannot progress to the 11th round, hence both $P(W_B = 32000|Q_{10}^c)$ and $P(W_B = 64000|Q_{10}^c)$ are equal to 0.

Since in the 11th round we have no lifeline we get the question right with probability $1/4$ and wrong with probability $3/4$, thus $P(W_B = 64000|Q_{10}) = 1/4$ and $P(W_B = 32000|Q_{10}) = 3/4$. Substituting back everything we have

$$P(W_B = 32000) = 3/4 \cdot 1/2 + 0 = 3/8,$$
$$P(W_B = 64000) = 1/4 \cdot 1/2 + 0 = 1/8.$$

- W_C: The same winning options are available as in the previous part, thus $P(W_B = w) = 0$ for $w \notin \{1000, 32000, 64000\}$ still holds. Similarly, all the probabilities of winning the different amounts can be rewritten as previously,

$$P(W_C = 1000) = P(Q_{10}^c),$$
$$P(W_C = 32000) = P(W_C = 32000|Q_{10})P(Q_{10}) + P(W_C = 32000|Q_{10}^c)P(Q_{10}^c),$$
$$P(W_C = 64000) = P(W_C = 64000|Q_{10})P(Q_{10}) + P(W_C = 64000|Q_{10}^c)P(Q_{10}^c),$$

but now the probabilities on the right-hand side (and consequently on the left-hand side) are different. In particular, as no lifeline is used in the 10th round, i.e. we have to choose the right answer from 4 options, $P(Q_{10}) = 1/4$ and $P(Q_{10}^c) = 3/4$. Because the lifeline is used in the 11th round if progressed that far, i.e. there are only two possible answers for the 11th question, $P(W_C = 32000|Q_{10}) = 1/2$ and $P(W_C = 64000|Q_{10}) = 1/2$.

By plugging back the numbers we have

$$P(W_C = 1000) = 3/4,$$
$$P(W_C = 32000) = 1/4 \cdot 1/2 + 0 = 1/8,$$
$$P(W_C = 64000) = 1/4 \cdot 1/2 + 0 = 1/8.$$

Advice: It's always useful and sometimes effortless sanity check to look at the derived PMF whether it really sums up to 1 (as it should, see Definition 3.2.2 in the book). In these three scenarios, it checks as $1 = 1$, $1/2 + 3/8 + 1/8 = 1$, and $3/4 + 1/8 + 1/8 = 1$.

Now that we have the PMF-s the expectations are the variances follow from the definitions.

Expectation:

- $E(W_A) = P(W_A = 16000) \cdot 16000 = 1 \cdot 16000 = \mathbf{16000}$
- $E(W_B) = P(W_B = 1000) \cdot 1000 + P(W_B = 32000) \cdot 32000 + P(W_B = 64000) \cdot 64000$
  $= 1/2 \cdot 1000 + 3/8 \cdot 32000 + 1/8 \cdot 64000 = \mathbf{20500}$
- $E(W_C) = P(W_C = 1000) \cdot 1000 + P(W_C = 32000) \cdot 32000 + P(W_C = 64000) \cdot 64000$
  $= 3/4 \cdot 1000 + 1/8 \cdot 32000 + 1/8 \cdot 64000 = \mathbf{12750}$

Variance: For calculating the variance we use the formula $Var(X) = E(X^2) - (E(X))^2$, and for calculating the first part of the expression we use the law of the unconscious statistician:

- $E(W_A^2) = P(W_A = 16000) \cdot 16000^2 = 1 \cdot 16000 = \mathbf{16000^2} (= 256000000 = 2.56 \cdot 10^8)$
- $E(W_B^2) = P(W_B = 1000) \cdot 1000^2 + P(W_B = 32000) \cdot 32000^2 + P(W_B = 64000) \cdot 64000^2$
  $= 1/2 \cdot 1000^2 + 3/8 \cdot 32000^2 + 1/8 \cdot 64000^2 = \mathbf{8.965 \cdot 10^8} (= 896500000)$
- $E(W_C^2) = P(W_C = 1000) \cdot 1000^2 + P(W_C = 32000) \cdot 32000^2 + P(W_C = 64000) \cdot 64000^2$
  $= 3/4 \cdot 1000^2 + 1/8 \cdot 32000^2 + 1/8 \cdot 64000^2 = \mathbf{6.4075 \cdot 10^8} (= 640750000)$,

therefore the variances are

- $Var(W_A) = E(W_A^2) - (E(W_A))^2 = 16000^2 - (16000)^2 = \mathbf{0}$
- $Var(W_B) = E(W_B^2) - (E(W_B))^2 = 8.965 \cdot 10^8 - (2.05 \cdot 10^4)^2 = (8.965 - 4.2025) \cdot 10^8 = \mathbf{4.7625 \cdot 10^8}$
- $Var(W_C) = E(W_C^2) - (E(W_C))^2 = 6.4075 \cdot 10^8 - (1.275 \cdot 10^4)^2 = (6.4075 - 1.625625) \cdot 10^8 = \mathbf{4.781875 \cdot 10^8}$.

Therefore the second strategy (W_B) has the highest expectation, and the first strategy (W_A) has the lowest variance, that is in fact 0, as there is no randomness involved.

If you only care about the expectation and the variance of a strategy, then the third strategy is strictly outperformed by both of the others. So you should not wait to use your lifeline, either walk away with the money or use it already in the 10th round.

Note: I was trying to be very descriptive and write a justification for each of the steps, so in the end, the solution got a bit long, but in an exam, your solution can be more succinct than this.

Exercise 2. For $X \sim Pois(\lambda)$, find $E(2^X)$, if it is finite.

Solution 2. By the law of the unconscious statistician, we have

$$E(2^X) = \sum_{x=0}^{\infty} 2^x \cdot P(X = x).$$

As X is $Poisson(\lambda)$ distributed $P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}$, so

$$\begin{aligned} E(2^X) &= \sum_{x=0}^{\infty} 2^x \frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda} \sum_{x=0}^{\infty} \frac{(2\lambda)^x}{x!} \\ &= e^{-\lambda} \cdot e^{2\lambda} = e^{\lambda}, \end{aligned}$$

where in the second line we took out $e^{-\lambda}$ from the sum and grouped to one expression 2^x and λ^x, and in the third line we used that the sum is just the Taylor series expansion for $e^{2\lambda}$ (see Appendix A.8.3 in the book). The Taylor series expansion of e^x holds for any x, so this expectation is always finite.

Week 7 Exercises

Exercise 3. Let F be the CDF of a continuous r.v., and $f = F'$ be the PDF.

1. Show that g defined by $g(x) = 2F(x)f(x)$ is also a valid PDF.
2. Show that h defined by $h(x) = \frac{1}{2}f(-x) + \frac{1}{2}f(x)$ is also a valid PDF.

Solution 3. To show that they are valid PDF-s we have to show two things: That they are non-negative and that they integrate to 1.

Since by definition F and f, are non-negative functions, if we take the product of them multiplied by a positive integer, or the sum of them after rescaling by a positive number, we must get a non-negative function, so the first part holds for both g and h.

In the following we will show that they integrate to 1.

1. $$\int_{-\infty}^{\infty} g(x)dx = \int_{-\infty}^{\infty} 2F(x)f(x)\,dx$$

   Using integration by substitution and the defintion of the PDF, in particular that $f(x) = F'(x)$, we can define $u = F(x)$ and then $du = f(x)dx$, thus

   $$\int_{-\infty}^{\infty} 2F(x)f(x)\,dx = 2\int_{u=F(-\infty)}^{F(\infty)} u\,du = 2\left[\frac{u^2}{2}\right]_{u=0}^{1} = 2\frac{1}{2} = 1,$$

   therefore $g(x)$ is a valid PDF.

   Alternatively, you can show this with integration by parts with $u = F(x)$ and $v' = f(x)$.

2. $$\int_{-\infty}^{\infty} h(x)\,dx = \int_{-\infty}^{\infty} \frac{1}{2}f(-x) + \frac{1}{2}f(x)\,dx.$$

   We can apply integration by substitution again for $u = -x$ and then $du = -1\,dx$:

   $$\int_{-\infty}^{\infty} \frac{1}{2}f(-x) + \frac{1}{2}f(x)\,dx = \frac{1}{2}\int_{u=\infty}^{-\infty} f(u)\cdot(-1)\,du + \frac{1}{2}\int_{x=-\infty}^{\infty} f(x)\,dx$$
   $$= \frac{1}{2}\left[-F(u)\right]_{u=\infty}^{-\infty} + \frac{1}{2}\left[F(x)\right]_{x=-\infty}^{\infty} = \frac{1}{2}(0-(-1)) + \frac{1}{2}(1-0) = 1,$$

   therefore $h(x)$ is a valid PDF as well.

Exercise 4. Let U be a Uniform r.v. on the interval $(-1, 1)$ (be careful about minus signs).

1. Compute $E(U)$, $Var(U)$, and $E(U^4)$.

2. Find the CDF and PDF of U^2. Is the distribution of U^2 Uniform on $(0,1)$?

Solution 4. 1. We have $E(U) = 0$ since the distribution is symmetric about 0. By the law of the unconscious statistician,

$$E(U^2) = \frac{1}{2}\int_{-1}^{1} u^2\,du = \frac{1}{3}.$$

So, $Var(U) = E(U^2) - (E(U))^2 = E(U^2) = \frac{1}{3}$. Again by LOTUS,

$$E(U^4) = \frac{1}{2}\int_{-1}^{1} u^4\,du = \frac{1}{5}.$$

2. Let $G(t)$ be the CDF of U^2. Clearly $G(t) = 0$ for $t \leq 0$ and $G(t) = 1$ for $t \geq 1$, because $0 \leq U^2 \leq 1$. For $0 < t < 1$,

$$G(t) = P(U^2 \leq t) = P(-\sqrt{t} \leq U \leq \sqrt{t}) = \frac{1}{2}\cdot(\sqrt{t} - (-\sqrt{t})) = \sqrt{t},$$

since the probability of U being in an interval in $(-1,1)$ is proportional to its length. The PDF is $G'(t) = \frac{1}{2}t^{-1/2}$ for $0 < t < 1$ (and 0 otherwise). The distribution of U^2 is *not* Uniform on $(0,1)$, as the PDF is not a constant on this interval (it is an example of a *Beta distribution*, an important distribution that is introduced in the following chapters of the course book).

A shorter solution would be, that a $Uniform(0,1)$ random variable has an expectation 0.5. However, we showed it in part 1. that $E(U^2) = 1/3$, hence U^2 cannot be uniformly distributed on $(0,1)$.

Exercise 5. A circle with a random radius $R \sim Unif(0,1)$ is generated. Let A be its area.

1. Find the mean and variance of A, without first finding the CDF or PDF of A.
2. Find the CDF and PDF of A.

Hint: For part 2. look up Exercise S5E2.

Solution 5. 1. For calculating both the expectation and the variance, we can use that A is just a function of R, as $A = \pi R^2$. Then by the linearity of expectation and by the law of the unconscious statistician

$$E(A) = E(\pi R^2) = \pi\int_0^1 r^2\,\mathrm{d}r = \pi\left[\frac{r^3}{3}\right]_{r=0}^{1} = \frac{\pi}{3}.$$

To calculate the variance $Var(A) = E(A^2) - (E(A))^2$, we also need $E(A^2)$, that is by similar steps

$$E(A^2) = E(\pi^2 R^4) = \pi^2 E(R^4) = \int_0^1 r^4\,\mathrm{d}r = \pi^2\left[\frac{r^5}{5}\right]_{r=0}^{1} = \frac{\pi^2}{5},$$

therefore

$$Var(A) = E(A^2) - (E(A))^2 = \pi^2\frac{9-5}{45} = \pi^2\frac{4}{45}.$$

2. Similarly to exercise S5E2, the CDF of A can be found via transforming the CDF of A, to an expression corresponding to the CDF of R:

$$F_A(a) = P(A \leq a) = P(\pi R^2 \leq a) = P\left(R^2 \leq \frac{a}{\pi}\right) = P\left(R \leq \sqrt{\frac{a}{\pi}}\right) = F_R\left(\sqrt{\frac{a}{\pi}}\right),$$

where for the second to last equality we also used that $P(R < 0) = 0$. Since R is uniformly distributed on $(0, 1)$,

$$F_R(r) = \begin{cases} 0 & \text{if } r \leq 0, \\ r & \text{if } 0 < r < 1, \\ 1 & \text{if } r \geq 1. \end{cases}$$

Therefore

$$F_A(a) = \begin{cases} 0 & \text{if } a \leq 0, \\ \sqrt{\frac{a}{\pi}} & \text{if } 0 < a < \pi, \\ 1 & \text{if } a \geq \pi. \end{cases}$$

Then using the definition of the PDF, i.e. $f(x) = F'(x)$, we have

$$f_A(a) = \begin{cases} \frac{1}{2\sqrt{\pi a}} & \text{if } 0 < a < \pi, \\ 0 & \text{otherwise.} \end{cases}$$

Exercise 6. Let $U \sim Unif(0, 1)$. As a function of U, create an r.v. X with CDF $F(x) = 1 - e^{-x^3}$ for $x > 0$.

Hint: See Example 5.3.4 in the book for some motivation.

Solution 6. By the universality of the uniform (Thm 5.3.1 in the book), we have $F_X^{-1}(U) \sim X$, so let us find the inverse of the CDF.

$$\begin{aligned} u = F_X(x) &= 1 - e^{-x^3} \\ &\iff e^{-x^3} = 1 - u \\ &\iff x^3 = log(1 - u) \\ &\iff x^3 = -log(1 - u) \\ &\iff x = (-log(1 - u))^{1/3}, \end{aligned}$$

therefore the inverse of the CDF of X is $F_X^{-1}(x) = (-log(1 - x))^{1/3}$. Then by the universality of the uniform, if we plug in U it follows that $F_X^{-1}(U) = (-log(1 - U))^{1/3}$ has the desired distribution.

Remark: It is not a bad practice exercise to show that if $U \sim Unif(0, 1)$ then $1 - U \sim Unif(0, 1)$ as well. Using this, the expression can be simplified to $(-log(U))^{1/3}$ and it still has the same target distribution.

Exercise 7. A woman is pregnant, with a due date of January 10, 2020. Of course, the actual date on which she will give birth is not necessarily the due date. On a timeline, define time 0 to be the instant when January 10, 2020 begins. Suppose that the time T when the woman gives birth has a Normal distribution, centered at 0 and with standard deviation 8 days. What is the probability that she gives birth on her due date? (Your answer should be in terms of Φ, and simplified.)

Solution 7. On the timeline described in the exercise, the time of birth T is distributed as $\mathcal{N}(0, 8^2)$. The probability of interest is, the probability of the event, that the woman gave birth between the start of January 10, 2020, i.e. time 0, and the start of January 11, 2020, i.e. time 1, so using notation

$$P(0 \leq T \leq 1).$$

Using the fundamental theorem of calculus, and the definition of the CDF (see Proposition 5.1.3 and Remark 5.1.4 in the book if in doubt)

$$P(0 \leq T \leq 1) = F_T(1) - F_T(0).$$

Unfortunately, this expression uses the CDF of T and not the CDF of the standard normal distribution (Φ), so we are not done yet. We have to standardize T, which we can achieve either by remembering Exercise S5E2, by Thm 5.4.4 from the book, or by hand. For the latter, let $X \sim \mathcal{N}(0, 1)$. Note that by the definition 5.4.3 in the book $T = 0 + 8 \cdot X$, so

$$F_T(t) = P(T \leq t) = P\left(\frac{T-0}{8} \leq \frac{t-0}{8}\right) = P\left(X \leq \frac{t-0}{8}\right) = \Phi\left(\frac{t-0}{8}\right).$$

Plugging it back to the probability

$$P(0 \leq T \leq 1) = \Phi\left(\frac{1}{8}\right) - \Phi(0) \approx 0.05.$$

Exercise 8. Let $Z \sim N(0, 1)$. A measuring device is used to observe Z, but the device can only handle positive values, and gives a reading of 0 if $Z \leq 0$; this is an example of censored data. So assume that $X = Z \cdot I_{Z>0}$ is observed rather than Z, where $I_{Z>0}$ is the indicator of $Z > 0$. Find $E(X)$ and $Var(X)$

Hint: Example 5.4.7 can be a good starting point.

Solution 8. By the law of the unconscious statistician,

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} I_{z>0} z e^{-z^2/2} dz = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} z e^{-z^2/2} dz,$$

as on the interval $(-\infty, 0)$ the integrand is 0 due to the indicator function. We can use the change of variable formula with letting $u = \frac{z^2}{2}$ and $\mathrm{d}u = z\mathrm{d}z$, so we have

$$E(X) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-u} du = \frac{1}{\sqrt{2\pi}}.$$

To obtain the variance, note that by the law of the unconscious statistician, and from the fact that $z^2 e^{-z^2/2}$ is symmetric around 0,

$$E(X^2) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-z^2/2} dz = \frac{1}{2},$$

since a $\mathcal{N}(0, 1)$ random variable has variance 1. Thus,

$$Var(X) = E(X^2) - (E(X))^2 = \frac{1}{2} - \frac{1}{2\pi}.$$

Note that X is neither purely discrete nor purely continuous, since $X = 0$ with probability $1/2$ and $P(X = x) = 0$ for $x \neq 0$. So X has neither a PDF nor a PMF; but LOTUS still works, allowing us to work with the PDF of Z to study expected values of functions of Z. *Sanity check*: The variance is positive, as it should be. It also makes sense that the variance is substantially less than 1 (which is the variance of Z), since we are reducing variability by making the r.v. 0 half the time, and making it nonnegative rather than roaming over the entire real line.