2B Expected value

Class problems

2B1 (Expectation of a power) Random variable X has uniform distribution over closed unit interval $[0, 1]$:

$$f(x) = \begin{cases} 1, & 0 \le x \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

Let n be a positive integer.

(a) Determine the cumulative distribution function, and then the density function, of the random variable $Y = X^n$. Hint: What values can X^n take? The CDF $F_Y(t)$ measures the probability of an event. What event? When does that event occur?

(b) Calculate $\mathrm{E}(X^n)$ from the density calculated in (a).

(c) Calculate $\mathrm{E}(X^n)$ using the transformation formula $\mathrm{E}(g(X)) = \int g(x)f(x)dx$ (see e.g. Ross's section 4.5).

(d) Using the formula you now have, calculate $\mathrm{E}(X^n)$ for $n = 1, 2, 3, 4$.

Solution.

(a) $Y = X^n$ can take any values in the interval $[0, 1]$; indeed, we have $Y = t$ if and only if $X = t^{1/n}$. For each $0 \le t \le 1$ we have

$$F_Y(t) = \mathrm{P}(X^n \le t) = \mathrm{P}(X \le t^{1/n}) = \int_0^{t^{1/n}} 1\, ds = t^{1/n}.$$

So the cumulative distribution function of $Y = X^n$ is

$$F_Y(t) = \begin{cases} 0, & t < 0, \\ t^{1/n}, & 0 \le t \le 1, \\ 1, & t > 1. \end{cases}$$

Taking its derivative, we obtain the density function

$$f_Y(t) = \begin{cases} \frac{1}{n}\, t^{(1/n)-1}, & 0 \le t \le 1, \\ 0, & \text{otherwise.} \end{cases}$$

(b) Using the density function f_Y, we have

$$\mathrm{E}(Y) = \int_{-\infty}^{\infty} t f_Y(t)\, dt = (1/n) \int_0^1 t^{1/n}\, dt = (1/n) \times \left[\frac{t^{(1/n)+1}}{(1/n)+1} \right]_{t=0}^{1} = \frac{1}{1+n}.$$

Here the notation $\left[F(t)\right]_{t=a}^{b}$ means $F(b) - F(a)$. Finns may be more used to the big slash notation $\Big/_{a}^{b} F(t)$. If you do not remember how to take derivatives and integrals, and simplify fractions, now is a good time to refresh.

(c) We are using the transformation function $g(x) = x^n$, so that $Y = g(X)$. Then we have

$$\mathrm{E}(X^n) \;=\; \mathrm{E}(g(X)) \;=\; \int_{-\infty}^{\infty} g(t)f(t)\,dt \;=\; \int_{-\infty}^{\infty} t^n f(t)\,dt \;=\; \int_0^1 t^n\,dt \;=\; \frac{1}{1+n}.$$

(d) $\mathrm{E}(X^1) = 1/2$ (as it should, because $X^1 = X$)
$\mathrm{E}(X^2) = 1/3$
$\mathrm{E}(X^3) = 1/4$
$\mathrm{E}(X^4) = 1/5$.

For extra exercise, you could try generating 1000 random numbers in $[0, 1]$, and computing the average of their nth powers. Does it agree with the calculations?

2B2 (Waiting time paradox) Buses arrive to your local bus stop at fixed times, three buses per hour. You arrive to the bus stop at X minutes after 9 o'clock, where X has continuous uniform distribution over the open interval $]0, 60[$ (which does not include the endpoints).

(a) If the buses arrive regularly at 20-minute intervals, what is your expected waiting time for the next bus?

(b) Now from the timetable you learn that the buses arrive at 9:00, 9:10, 9:30, 10:00, ... and so on. Represent your waiting time W for the next bus as a function $W = g(X)$ of your arrival time, and draw a graph of the function. (Hint: define the function by breaking into cases.)

(c) Calculate $\mathrm{E}(W)$. Hint: Transformation formula.

(d) Compare results from (a) and (c), and explain with common sense.

Solution.

(a) 10 minutes.

(b) Break the interval $]0, 60[$ into three cases (subintervals). In each case, if the next bus arrives b minutes past nine, your waiting time is $b - x$, which (of course) depends on the value of x.

$$g(x) = \begin{cases} 10 - x & \text{if } 0 < x \le 10 \\ 30 - x & \text{if } 10 < x \le 30 \\ 60 - x & \text{if } 30 < x < 60. \end{cases}$$

(Graph not shown here but it is a zigzag.)

(c) Just calculate the integral in three parts. The density of X is constant $f(x) = \frac{1}{60}$ all over $]0, 60[$.

$$\begin{aligned}
\mathrm{E}(W) &= \int_0^{60} g(x)f(x)\,dx = \frac{1}{60}\int_0^{60} g(x)\,dx \\
&= \frac{1}{60}\int_0^{10}(10-x)\,dx \;+\; \frac{1}{60}\int_{10}^{30}(30-x)\,dx \;+\; \frac{1}{60}\int_{30}^{60}(60-x)\,dx \\
&\approx 0.8333 + 3.3333 + 7.5000 = \mathbf{11.6666}.
\end{aligned}$$

Instead of doing the integrals, you *could* reason as follows: Your probabilities for arriving in the three intervals are 1/6, 2/6 and 3/6. *If* you arrive in the first interval, then your expected waiting time is 5 minutes; similarly 10 and 15 minutes for the 2nd and 3rd interval. Then calculate weighted average $\frac{1}{6} \cdot 5 + \frac{2}{6} \cdot 10 + \frac{3}{6} \cdot 15 \approx 11.667$ minutes. However, this requires the notion of "conditional expected value" which we haven't learned on the course. Taking the weighted averages is then "law of total expectation", similar to the law of total probability which we had on Lecture 1A.

(d) In (c) your expected waiting time is longer. This is because you will more probably arrive during a long interval than during a short interval between buses.

This *waiting time paradox* is an example of a more general phenomenon called *size bias*, where you take a "random" sample by some process whose sampling probabilities are affected by some quantity, which may be in fact the same quantity that you are trying to estimate (or a closely related one).

In our problem, perhaps you keep a log of your waiting times every day. From observing that your average waiting time is $11\frac{2}{3}$ minutes, you might believe that buses are running only once in $23\frac{1}{3}$ minutes. In reality they run once in 20 minutes on average. Your estimate is *up-biased* because you are sampling the longer intervals at higher rate.

Real-world examples of size bias occur in many fields of science. For example, in textile industry you might want to estimate the mean length of fibers. You use a mechanical pincer that takes a sample of fibers, and you measure their lengths. But shorter fibers are more likely to be missed by your pincer, so the fibers you collect are biased towards the longer ones. For this and other examples see e.g. Arratia, Goldstein & Kochman: "Size bias for one and all", `https://arxiv.org/abs/1308.2729`.

Home problems

2B3 (Repairing the printer) Repairing a jammed printer takes a random time X (in hours) that has density function

$$f(x) = \begin{cases} 1 - x/6, & 2 < x < 4, \\ 0, & \text{otherwise.} \end{cases}$$

(Always between 2 and 4 hours; more probably near the bottom end.) Cost of repairs (in euros) is $g(x) = 100 - 40x + 10x^2$ if repair time is x.

(a) Calculate the expected repair time $\mathrm{E}(X)$.

(b) Calculate the expected repair cost $\mathrm{E}(g(X))$.

(c) Calculate the repair cost *in the case that* the repair time happens to hit its expected value, that is, calculate $g(x)$ if $x = \mathrm{E}(X)$. Is it numerically the same as in (b)? Try to explain with common sense why / why not.

Note. g is not a linear transformation so you cannot use the "shift and scale" formulas. Hint: Lecture 2A; Ross §4.5.

Grading.

(a) 0.5 points.

(b) 1 p, with 0.5 from formulating the correct integral, and 0.5 from solving it. Zero points for working like the transformation was linear. If integral limits are not 2 and 4, max 0.5 points.

(c) 0.5 points for correct calculation and observing that they are different. No fancy explanations required.

Rounded up to integer.

Solution.

(a) Directly from definition of (continuous) expectation,

$$\mathrm{E}(X) = \int_2^4 x \cdot (1 - x/6) dx = \int_2^4 (x - x^2/6) dx = \Big[x^2/2 - x^3/18\Big]_2^4 = \mathbf{26/9} \approx \mathbf{2.889}.$$

(b) Using the (continuous) transformation formula

$$\begin{aligned} \mathrm{E}(g(X)) &= \int_2^4 g(x) f(x)\, dx = \int_2^4 (100 - 40x + 10x^2) \cdot (1 - x/6)\, dx \\ &= \int_2^4 \big((-5/3)x^3 + (50/3)x^2 - (170/3)x + 100\big)\, dx \\ &= \Big[(-5/12)x^4 + (50/9)x^3 - (170/6)x^2 + 100x\Big]_2^4 \approx \mathbf{71.111}. \end{aligned}$$

(c) If $x = 26/9$, then $g(x) = g(26/9) \approx \mathbf{67.901}$. This is less than the result in (b). The reason is the nonlinearity of the g function (plot it!). The quadratic term causes higher costs *if* the repairs happen to be long. This is not seen at all if you just calculate the costs at the *expected* repair time 2.889 hours.

2B4 (Peer grading) 30 students arrive at an exercise session, and each gives their answer sheet to the assistant. The assistant shuffles the sheets thoroughly, and deals them back to the students for grading, one sheet to each student. Determine the expected value of the number of students that receive their own paper.

Hint. Define indicator variables

$$X_i = \begin{cases} 1, & \text{if the } i\text{th student receives his/her own paper,} \\ 0, & \text{otherwise.} \end{cases}$$

Recall expectation of a sum of random variables.

Grading. Points are earned by the following things (0.5 points each, total rounded up to an integer):

- Representing the number as a sum of indicator variables.
- Using linearity to find the expectation of the sum.
- Finding the expectation of one indicator variable $E(X_i) = P(X_i = 1)$
- Finding the probability of the event "ith student receives his/her own paper"

Solution. The number of students receiving their own papers is a random variable that can be expressed as a sum

$$X = X_1 + \cdots + X_{30},$$

where X_i is as hinted. By "linearity of expectation" (taking the expectation of a sum)

$$E(X) = E(X_1) + \cdots + E(X_{20}).$$

Because the indicator X_i takes only two values (0 and 1), its expected value is

$$\begin{aligned} E(X_i) &= 0 \times P(X_i = 0) + 1 \times P(X_i = 1) \\ &= P(X_i = 1). \end{aligned}$$

We still need to find $P(X_i = 1)$, probability of ith student receiving own paper. Because the papers were shuffled, the student receives each paper with equal probability; in particular, his/her own paper with probability 1/30. Putting all things together, we have

$$E(X) = \sum_{i=1}^{30} P(X_i = 1) = 30 \times \frac{1}{30} = 1.$$

It should be easy to see that we have the same expected value regardless of the number of students. It may be a good idea to check this in some simple extreme cases. What happens if there is only one student in class? Two students? — Note that finding the exact *distribution* of X would be more complicated. Here we only calculated the expected value, that is, just one aspect of the full distribution.