Mathematics for Data Science Tutorial 6 (week 12)

Semester 2, 2019

1. Consider the discrete random variable X described by the following table

k	1	3	4	6	8	9
$\Pr(X=k)$	0.37	0.08	q	0.06	0.09	0.21

where X can take on only those values of k shown.

(a) What is the value of q?

(b) Determine $\mathbb{E}[X]$.

(c) Determine $\mathrm{Var}(X)$.

Solution:

(a) Since the probabilities must sum to one we require $1 = 0.81 + q$ and thus $q = 0.19$.

(b) We have

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=-\infty}^{\infty} k \Pr(X=k) \\ &= 1 \times \Pr(X=1) + 3 \times \Pr(X=3) + 4 \times \Pr(X=4) \\ &\quad + 6 \times \Pr(X=6) + 8 \times \Pr(X=8) + 9 \times \Pr(X=9) \\ &= 0.37 + 0.24 + 0.76 + 0.36 + 0.72 + 1.89 = 4.34\,.\end{aligned}$$

(c) We have

$$\begin{aligned}\mathrm{Var}(X) &= \sum_{k=-\infty}^{\infty} (k - \mathbb{E}[X])^2 \Pr(X=k) \\ &= 3.34^2 \times \Pr(X=1) + 1.34^2 \times \Pr(X=3) + 0.34^2 \times \Pr(X=4) \\ &\quad + 1.66^2 \times \Pr(X=6) + 3.66^2 \times \Pr(X=8) + 4.66^2 \times \Pr(X=9) \\ &= 4.127572 + 0.143648 + 0.021964 + 0.165336 + 1.205604 + 4.560276 \\ &= 10.2244\,.\end{aligned}$$

Alternatively

$$\begin{aligned}\mathbb{E}[X^2] &= \sum_{k=-\infty}^{\infty} k^2 \Pr(X=k) \\ &= 1 \times \Pr(X=1) + 9 \times \Pr(X=3) + 16 \times \Pr(X=4) \\ &\quad + 36 \times \Pr(X=6) + 64 \times \Pr(X=8) + 81 \times \Pr(X=9) \\ &= 0.37 + 0.72 + 3.04 + 2.16 + 5.76 + 17.01 = 29.06\,,\end{aligned}$$

and then $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 29.06 - 4.34^2 = 10.2244$.

2. Consider a complex simulation which is performed n times, each instance operating independently and providing a 'positive' result with probability p (and a 'negative' result otherwise). Suppose a consensus is determined by taking the result which is the majority of the simulations.

   (a) Consider performing the simulation 3 times, what is the probability of a 'positive' consensus?

   (b) Consider performing the simulation 5 times, what is the probability of a 'positive' consensus?

   (c) For which values of p is performing the simulation 5 times more likely to produce a 'positive' consensus then performing the simulation 3 times?

   (d) Suppose that the first three 3 simulations gave a negative consensus, but then, after an additional two simulations, a positive consensus is reached. What value of p maximises the probability of this event?

   (e) What is the expectation and variance for the number of 'positive' results in n simulations?

Solution:

(a) Let X_3 be the (discrete) random variable describing the number of 'positive' results from 3 simulations. Clearly X_3 is a binomial random variable with $n = 3$ and probability p of success. A 'positive' consensus is the event $X_3 > 1$, or $X_3 \geq 2$. We then have

$$\begin{aligned}\Pr(X_3 \geq 2) &= \Pr(X_3 = 2) + \Pr(X_3 = 3) \\ &= \binom{3}{2} p^2 (1-p)^1 + \binom{3}{3} p^3 (1-p)^0 \\ &= 3p^2(1-p) + p^3 = 3p^2 - 2p^3 \,.\end{aligned}$$

(b) The random variable X_5 for the number of 'positive' results from 5 simulations is again binomial distributed but with $n = 5$. The event of interest is $X_5 \geq 3$ for which we have We then have

$$\begin{aligned}\Pr(X_5 \geq 3) &= \Pr(X_5 = 3) + \Pr(X_5 = 4) + \Pr(X_5 = 5) \\ &= \binom{5}{3} p^3 (1-p)^2 + \binom{5}{4} p^4 (1-p)^1 + \binom{5}{5} p^5 (1-p)^0 \\ &= 10p^3(1-p)^2 + 5p^4(1-p) + p^5 = 10p^3 - 15p^4 + 6p^5 \,.\end{aligned}$$

(c) This question is asking for which p do we have $\Pr(X_5 \geq 3) > \Pr(X_3 \geq 2)$, that is

$$10p^3 - 15p^4 + 6p^5 > 3p^2 - 2p^3$$

or equivalently

$$-3p^2 + 12p^3 - 15p^4 + 6p^5 = 3p^2(1-p)^2(2p-1) > 0\,.$$

From this we see that $\Pr(X_5 \geq 3) > \Pr(X_3 \geq 2)$ only when $p > 1/2$ (noting we must have $p \in [0,1]$).
(Note: this seems obvious in hindsight if you consider that more simulations are going to be more likely to give a 'positive' consensus only if 'positive' results are more likely to occur than 'negative' ones).

(d) For this event to occur is is necessary that exactly one of the first three simulations are positive, then both of the two additional simulations are positive. The probability of this event is

$$\Pr(X_3 = 1)p^2 = 3p^4(1-p)\,.$$

This has a turning point where

$$0 = \frac{d}{dp}3p^4(1-p) = 12p^3 - 15p^4 = 3p^3(4-5p)\,,$$

and since we can rule out $p = 0$ it must be $p = 4/5 = 0.8$ which maximises the probability of this event (at which $\Pr(X_3 = 1)p^2 = 0.24576$).

(e) Let X_n be the random variable for the number of positive outcomes from n simulations, then since X_n is binomial distributed we have immediately that $\mathbb{E}[X_n] = np$ and $\mathrm{Var}(X) = np(1-p)$.

---

3. Consider a random variable X with probability density function

$$f(x) = \begin{cases} \dfrac{x}{32} & \text{for } 2 < x < 6 \\ \dfrac{1}{8} & \text{for } 6 < x < 8 \\ a(100 - x^2) & \text{for } 8 < x < 10 \\ 0 & \text{elsewhere} \end{cases}$$

(a) What is the value of a?

(b) Determine $\mathbb{E}[X]$.

(c) Determine $\mathrm{Var}[X]$.

Solution:

(a) We will use the fact that the total probability must be 1, that is

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty} f(x)\,dx \\
&= \int_2^6 f(x)\,dx + \int_6^8 f(x)\,dx + \int_8^{10} f(x)\,dx \\
&= \int_2^6 \frac{x}{32}\,dx + \int_6^8 \frac{1}{8}\,dx + \int_8^{10} a(100-x^2)\,dx \\
&= \left[\frac{x^2}{64}\right]_{x=2}^{6} + \left[\frac{x}{8}\right]_{x=6}^{8} + a\left[100x - \frac{x^3}{3}\right]_{x=8}^{10} \\
&= \frac{36-4}{64} + \frac{8-6}{8} + a\frac{3000-1000-2400+512}{3} \\
&= \frac{3}{4} + a\frac{112}{3}\,.
\end{aligned}$$

Re-arranging gives $a = 3/448$.

(b) The expectation is given by

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f(x)\,dx \\
&= \int_2^6 \frac{x^2}{32}\,dx + \int_6^8 \frac{x}{8}\,dx + \int_8^{10} \frac{3x}{448}(100-x^2)\,dx \\
&= \left[\frac{x^3}{96}\right]_{x=2}^{6} + \left[\frac{x^2}{16}\right]_{x=6}^{8} + \frac{3}{448}\left[\frac{200x^2-x^4}{4}\right]_{x=8}^{10} \\
&= \frac{216-8}{96} + \frac{64-36}{16} + \frac{3(20000-10000-12800+4096)}{4\times 448} \\
&= \frac{208}{96} + \frac{28}{16} + \frac{3\times 1296}{4\times 448} \\
&= \frac{47}{12} + \frac{243}{112} = \frac{2045}{336} \approx 6.0863\,.
\end{aligned}$$

(c) This is most easily done via the formula $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$. We have

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_{-\infty}^{\infty} x^2 f(x)\,dx \\
&= \int_2^6 \frac{x^3}{32}\,dx + \int_6^8 \frac{x^2}{8}\,dx + \int_8^{10} \frac{3x^2}{448}(100-x^2)\,dx \\
&= \left[\frac{x^4}{128}\right]_{x=2}^{6} + \left[\frac{x^3}{24}\right]_{x=6}^{8} + \frac{3}{448}\left[\frac{500x^3-3x^5}{15}\right]_{x=8}^{10} \\
&= \frac{1296-16}{128} + \frac{512-216}{24} + \frac{500000-300000-256000+98304}{2240} \\
&= \frac{1280}{128} + \frac{296}{24} + \frac{42304}{2240} \\
&= 10 + \frac{37}{3} + \frac{661}{35} = \frac{4328}{105} \approx 41.219\,.
\end{aligned}$$

It follows that

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{4328}{105} - \left(\frac{2045}{336}\right)^2 \approx 4.1759\,.$$

---

4. Suppose an analysis of social media posts on a particular website reveals that the length of time between any two posts can be modelled with a probability distribution function of the form

$$f(t) = c2^{-t}\,, \qquad t \geq 0\,.$$

(Implicitly take $f(t) = 0$ for $t < 0$ in this context, i.e. there cannot be negative time in between posts.)

   (a) What must c be for this to be a valid probability distribution?

   (b) What is the mean and variance in the time between posts?

Solution:

   (a) Observe we can write $f(t) = ce^{-\log(2)t}$ which makes it clear this must be an exponential distribution with parameter $\lambda = \log(2)$, and therefore $c = \log(2)$.
   Alternatively, we know the total probability must be one, and thus

$$1 = \int_0^\infty c2^{-t}\,dt = \left[-\frac{c}{\log(2)}2^{-t}\right]_{t=0}^\infty = \frac{c}{\log(2)}\,,$$

   and thus $c = \log(2)$.

   (b) If you have recognised that this is an exponential distribution with parameter $\lambda = \log(2)$ it is immediately clear that the mean is $\lambda^{-1} = 1/\log(2)$ and the variance is $\lambda^{-2} = 1/\log(2)^2$.
   Alternatively we would need to use integration by parts to evaluate the expectation, i.e.

$$\begin{aligned}\int_0^\infty t\log(2)2^{-t}\,dt &= \left[-t2^{-t}\right]_{t=0}^\infty - \int_0^\infty -2^{-t}\,dt \\ &= 0 - \left[\frac{1}{\log(2)}2^{-t}\right]_{t=0}^\infty \\ &= \frac{1}{\log(2)}\,,\end{aligned}$$

   and similarly for the variance

$$\begin{aligned}\int_0^\infty t^2\log(2)2^{-t}\,dt &= \left[-t^2 2^{-t}\right]_{t=0}^\infty - \int_0^\infty -t2^{-t}\,dt \\ &= 0 + \frac{1}{\log(2)}\int_0^\infty t\log(2)2^{-t}\,dt \\ &= \frac{1}{\log(2)}\frac{1}{\log(2)} = \frac{1}{\log(2)^2}\,.\end{aligned}$$

5. Suppose X, Y are random variables and that $Y = -1 + 4X - X^2$.

   (a) If $\mathbb{E}[X] = 2$ and $\mathbb{E}[Y] = 1$, what is $\mathrm{Var}(X)$?

   (b) Why is it not possible that $\mathbb{E}[X] \leq 0$ if $\mathbb{E}[Y] = 1$ (assuming X, Y are real valued)?

Solution:

(a) Using the linear property of expectation we note that

$$\begin{aligned} 1 &= \mathbb{E}[Y] \\ &= \mathbb{E}[-1 + 4X - X^2] \\ &= -1 + 4\mathbb{E}[X] - \mathbb{E}[X^2]\,. \end{aligned}$$

Re-arranging we then obtain

$$\mathbb{E}[X^2] = -1 + 2\mathbb{E}[X] - 1 = -1 + 8 - 1 = 6\,.$$

Therefore we have

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 6 - 2^2 = 2\,.$$

(b) Observe that $\mathbb{E}[X] \leq 0$ implies

$$1 = \mathbb{E}[Y] = -1 + 4\mathbb{E}[X] - \mathbb{E}[X^2] \leq -1 - \mathbb{E}[X^2]\,,$$

which implies $\mathbb{E}[X^2] \leq -2$ which is not possible (since one must always have $X^2 \geq 0$ and thus its expectation must also be non-negative).
Alternatively, we can repeat the first part with a general $\mathbb{E}[X] = x$ to obtain

$$1 = \mathbb{E}[Y] = -1 + 4\mathbb{E}[X] - \mathbb{E}[X^2] = -1 + 4x - \mathbb{E}[X^2]\,,$$

and therefore $\mathbb{E}[X^2] = 4x - 2$. This then implies that $x \geq 1/2$ (since $X^2 \geq 0$). However, it then follows that

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 4x - 2 - x^2\,.$$

Since it is necessary that $\mathrm{Var}(X) \geq 0$ then we must have $2 - \sqrt{2} \leq x \leq 2 + \sqrt{2}$ (obtain the roots via the quadratic formula). Clearly $x \leq 0$ is outside of this interval.