Solutions sheet 0

Prerequisites

IB Data Science—DJW—2021/2022

Question 1 (Elementary probability). A card is drawn at random from a pack. Event A is 'the card is an ace', event B is 'the card is a spade', event C is 'the card is either an ace, or a king, or a queen, or a jack, or a 10'. Compute the probability that the card has (i) one of these properties, (ii) all of these properties.

(Taken from Maths for NST A). Writing

$$H \equiv \textit{hearts}, \quad D \equiv \textit{Diamonds}, \quad C \equiv \textit{Clubs}, \quad S \equiv \textit{Spaces},$$
$$a \equiv \textit{ace}, \quad k \equiv \textit{king}, \quad q \equiv \textit{queen}, j \equiv \textit{jack}, \quad 10 \equiv 10, \ldots$$

we can write the events as

$$A = \{H_a, D_a, C_a, S_a\} \qquad \mathbb{P}(A) = \frac{4}{52} = \frac{1}{13}$$
$$B = \{S_a, S_k, S_q, S_j, S_{10}, \ldots, S_2\} \qquad \mathbb{P}(B) = \frac{13}{52} = \frac{1}{4}$$
$$C = \{H_a, H_k, H_q, H_j, H_{10}, D_a, D_k, \ldots\} \qquad \mathbb{P}(C) = \frac{4 \times 5}{52} = \frac{5}{13}.$$

(i). At least one of these properties?

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) &= 1 - \mathbb{P}(\overline{A \cup B \cup C}) \\ &= 1 - \mathbb{P}\big\{H_2, \ldots, H_9, D_2, \ldots, D_9, C_2, \ldots, C_9\}\big) \\ &= 1 - \frac{3 \times 8}{52} = \frac{7}{13}.\end{aligned}$$

(ii). All of these properties?

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(S_a) = \frac{1}{52}.$$

Question 2 (Elementary probability). A biased die has probabilities p, $2p$, $3p$, $4p$, $5p$, $6p$ of throwing 1, 2, 3, 4, 5, 6 respectively. Find p. What is the probability of throwing an even number?

(Taken from Maths for NST A.). Write Ω for the entire sample space, $\Omega = \{1, 2, 3, 4, 5, 6\}$. It's always the case that $\mathbb{P}(\Omega) = 1$, i.e. it's certain that the outcome will be something *from the sample space. So*

$$\begin{aligned}1 = \mathbb{P}(\Omega) &= \mathbb{P}(1) + \mathbb{P}(2) + \mathbb{P}(3) + \mathbb{P}(4) + \mathbb{P}(5) + \mathbb{P}(6) \\ &= p + 2p + 3p + 4p + 5p + 6p \\ &= 21p\end{aligned}$$

hence $p = 1/21$. Also,

$$\mathbb{P}(\textit{even}) = \mathbb{P}\big(\{2, 4, 6\}\big) = 2p + 4p + 6p = \frac{12}{21} = \frac{4}{7}.$$

Question 3 (Independence). We roll a die twice, and get the answers X and Y. Assume the two rolls are independent. Consider the events $E = \{X = 1\}$, $F = \{Y = 6\}$, and $G = \{X + Y = 7\}$. (i) Are E and F independent? (ii) Are E and G independent? (iii) Are E, F, and G independent?

(Taken from IA Probability lecture 2.) First, note that the sample space consists of all 36 possible pairs of values for (X, Y), each outcome equally likely, and the events are

$$E = \{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)\}$$
$$F = \{(1,6),(2,6),(3,6),(4,6),(5,6),(6,6)\}$$
$$G = \{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}.$$

(i) Yes, because $\mathbb{P}(E) = 1/6$, $\mathbb{P}(F) = 1/6$, $\mathbb{P}(E \cap F) = 1/36$. (ii) Yes, because $\mathbb{P}(E) = 1/6$, $\mathbb{P}(G) = 1/6$, $\mathbb{P}(E \cap G) = 1/36$. (iii) No, because $\mathbb{P}(E) = 1/6$, $\mathbb{P}(F) = 1/6$, $\mathbb{P}(E \cap F \cap G) = 1/36 \neq 1/216$.

Question 4 (Law of total probability). There are three boxes each containing a different number of lightbulbs. The first box has 10 bulbs of which 4 are dead, the second has 6 bulbs of which 1 is dead, and the third has 8 bulbs of which 3 are dead. What is the probability of a dead bulb being selected when a bulb is chosen at random from one of the three boxes?

(Taken from IA Probability *lecture 2.)*
Let event E be 'a dead bulb is picked', and let F_i be 'box i is chosen'. The question tells us

$$\mathbb{P}(E \mid F_1) = 4/10, \qquad \mathbb{P}(E \mid F_2) = 1/6, \qquad \mathbb{P}(E \mid F_3) = 3/8$$

and

$$\mathbb{P}(F_i) = 1/3, \qquad i \in \{1,2,3\}.$$

It asks us to compute $\mathbb{P}(E)$. By the law of total probability,

$$\mathbb{P}(E) = \sum_{i=1}^{3} \mathbb{P}(E \mid F_i)\,\mathbb{P}(F_i) = 4/10 \times 1/3 + 1/6 \times 1/3 + 3/8 \times 1/3 \approx 0.31.$$

Question 5 (Conditional probability). A deck of 52 playing cards is randomly divided into 4 piles of 13 cards each. What is the probability that each pile has exactly one ace?

(Taken from IA Probability lecture 2. This type of calculation is sometimes called the 'chain rule' of probability.)
We can split the overall event 'each pile has one ace' into a series of tests that build upon each other in sequence: is ♢ in a different pile to ♡; then in addition is ♠ in yet another pile to these two; then in addition is ♣ also in a different pile? Let's define the events

$$E_2 = \text{ace}\diamondsuit \text{ is in different pile to ace}\heartsuit$$
$$E_3 = \text{ace}\spadesuit, \text{ ace}\diamondsuit, \text{ and ace}\heartsuit \text{ in different piles}$$
$$E_4 = \text{all aces in different piles}$$

Then, by counting the 'slots' in which each ace might be,

$$\mathbb{P}(E_2) = 1 - \frac{12}{51}$$
$$\mathbb{P}(E_3 \mid E_2) = 1 - \frac{24}{50}$$
$$\mathbb{P}(E_4 \mid E_3 \text{ and } E_2) = 1 - \frac{36}{49}.$$

With these definitions,

$$\begin{aligned}
\mathbb{P}(E_4) &= \mathbb{P}(E_4 \text{ and } E_3 \text{ and } E_2) && \text{since } E_4 \text{ contains the other events} \\
&= \mathbb{P}(E_4 \mid E_3 \text{ and } E_2)\,\mathbb{P}(E_3 \text{ and } E_2) && \text{by defn of cond.prob} \\
&= \mathbb{P}(E_4 \mid E_3 \text{ and } E_2)\,\mathbb{P}(E_3 \mid E_2)\,\mathbb{P}(E_2) && \text{by defn of cond.prob} \\
&= \frac{39}{51} \times \frac{26}{50} \times \frac{13}{49} \approx 0.105.
\end{aligned}$$

Question 6 (Advanced elementary probability). Players A and B roll a six-sided die in turn. If a player rolls 1 or 2 that player wins and the game ends; if a player rolls 3 the other player wins and the game ends; otherwise the

turn passes to the other player. A has the first roll. What is the probability (i) that B gets a first throw and wins on it? (ii) that A wins before A's second throw? (iii) that A wins, if the game is played until there is a winner?

(Taken From IA Maths for NST.) For part (i),

$$\mathbb{P}(B \text{ gets a first throw then wins}) = \mathbb{P}\big(A \text{ gets } \{4,5,6\} \text{ then } B \text{ gets } \{1,2\}\big) = \frac{3}{6} \times \frac{2}{6} = \frac{1}{6}.$$

For part (ii),

$$\mathbb{P}\big(A \text{ wins before } A\text{'s second throw}\big) = \mathbb{P}\big(A \text{ gets } \{1,2\}\big) + \mathbb{P}\big(A \text{ gets } \{4,5,6\} \text{ then } B \text{ gets } 3\big) = \frac{2}{6} + \frac{3}{6} \times \frac{1}{6} = \frac{5}{12}.$$

For part (iii), we'll break it down by the number of rounds played, where a single round is "A throws once and then maybe B throws once".

$$\begin{aligned}\mathbb{P}(A \text{ wins game}) = {} & \mathbb{P}(A \text{ wins first round}) \\ & + \mathbb{P}(\text{play passes } ABA \text{ then } A \text{ wins round}) \\ & + \mathbb{P}(\text{play passes } ABABA \text{ then } A \text{ wins round}) \\ & + \cdots\end{aligned}$$

Now,

$$\mathbb{P}\big(\text{play passes } (AB)^n A\big) = \Big[\mathbb{P}\big(A \text{ gets } \{4,5,6\} \text{ then } B \text{ gets } \{4,5,6\}\big)\Big]^n = \Big(\frac{3}{6} \times \frac{3}{6}\Big)^n = \frac{1}{4^n}.$$

Putting this all together,

$$\begin{aligned}\mathbb{P}\big(A \text{ wins game}\big) &= \sum_{n=0}^{\infty} \mathbb{P}\big(\text{play passes } (AB)^n A\big)\, \mathbb{P}(A \text{ wins this round}) \\ &= \sum_{n=0}^{\infty} \frac{1}{4^n} \times \frac{5}{12} \quad \text{using answer to (ii)} \\ &= \frac{5}{12} \times \frac{1}{1-1/4} \quad \text{since } 1 + r + r^2 + \cdots = 1/(1-r) \text{ for } |r| < 1 \\ &= \frac{5}{12} \times \frac{4}{3} = \frac{5}{9}.\end{aligned}$$

Question 7 (Bayes's rule). A screening test is 94% effective in detecting COVID, when the person has the disease. The test yields a 'false positive' for 1% of healthy persons tested. Suppose 0.4% of the population has the disease. (i) What is the probability you actually have COVID, if you test positive? (ii) What is the probability you actually have COVID, if you test negative?

(Taken from IA Probability lecture 2.) Define the events

$$A = \{\text{person has disease}\}, \qquad B = \{\text{test came back positive}\}.$$

The question tells us several conditional probabilities:

$$\begin{aligned}&\mathbb{P}(B \mid A) = 0.94, && \mathbb{P}(B \mid A^{\complement}) = 0.01, \\ &\mathbb{P}(B^{\complement} \mid A) = 0.06, && \mathbb{P}(B^{\complement} \mid A^{\complement}) = 0.99.\end{aligned}$$

It also tells us about the probability of A absent any diagnostic information:

$$\begin{aligned}&\mathbb{P}(A) = 0.004, \\ &\mathbb{P}(A^{\complement}) = 0.996.\end{aligned}$$

For (i), Applying Bayes's rule,

$$\begin{aligned}\mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A)\ \mathbb{P}(B|A)}{\mathbb{P}(B)} = \frac{\mathbb{P}(A)\ \mathbb{P}(B|A)}{\mathbb{P}(A)\,\mathbb{P}(B|A) + \mathbb{P}(A^{\complement})\,\mathbb{P}(B|A^{\complement})} \\ &= \frac{0.004 \times 0.94}{0.004 \times 0.94 + 0.996 \times 0.01} \\ &\approx 0.27.\end{aligned}$$

For (ii), applying Bayes's rule again,

$$\begin{aligned}\mathbb{P}(A \mid B^{\complement}) &= \frac{\mathbb{P}(A)\,\mathbb{P}(B^{\complement} \mid A)}{\mathbb{P}(B^{\complement})} = \frac{\mathbb{P}(A)\,\mathbb{P}(B^{\complement} \mid A)}{\mathbb{P}(A)\,\mathbb{P}(B^{\complement} \mid A) + \mathbb{P}(A^{\complement})\,\mathbb{P}(B^{\complement} \mid A^{\complement})} \\ &= \frac{0.004 \times 0.06}{0.004 \times 0.06 + 0.996 \times 0.99} \\ &\approx 2.4 \times 10^{-4}.\end{aligned}$$

Question 8 (Elementary probability). Derive Bayes's rule from the other three core definitions and laws of probability. Explain carefully which of the three you are using at each step.

By the definition of conditional probability, applied to $(A \mid B)$,

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}\,A \cap B}{\mathbb{P}(B)} \quad \textit{when } \mathbb{P}(B) > 0.$$

Applying it again to $(B \mid A)$, and rearranging,

$$\mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\,\mathbb{P}(A).$$

(This applies when $\mathbb{P}(A) > 0$ by the definition of conditional probability, and it applies when $\mathbb{P}(A) = 0$ since then both sides are equal to zero.) Putting these two together, we get the first version of Bayes's rule:

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\,\mathbb{P}(B \mid A)}{\mathbb{P}(B)}.$$

Next, by the law of total probability: since A and $A^{\complement}$ partition Ω,

$$\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) + \mathbb{P}(B \mid A^{\complement})\,\mathbb{P}(A^{\complement}).$$

Substituting this into the first version of Bayes's rule gives the second version of Bayes's rule. (Note that the law of total probability is just an application of the sum rule combined with the definition of conditional probability, so we could have derived it that way instead.)

Question 9 (Expectation). Let X be a random variable. Show that $\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$.

Note: by convention, $\mathbb{E}$ is taken to have lower precedence than multiplication and power, and higher precedence than addition and subtraction. So the expression of interest is $(\mathbb{E}(X^2)) - (\mathbb{E}(X))^2$.

The variance is defined to be

$$\operatorname{Var} X = \mathbb{E}\big[(X - \mu)^2\big] \quad \textit{where} \quad \mu = \mathbb{E}\,X.$$

Expanding the square,

$$\begin{aligned}\operatorname{Var} X &= \mathbb{E}\big[(X^2 - 2X\mu + \mu^2)\big] \\ &= \mathbb{E}(X^2) - \mathbb{E}(2X\mu) + \mathbb{E}(\mu^2) \quad \textit{since } \mathbb{E}(A + B) = \mathbb{E}\,A + \mathbb{E}\,B \\ &= \mathbb{E}\,X^2 - 2\mu\,\mathbb{E}\,X + \mu^2 \quad \textit{since } \mu \textit{ is a constant} \\ &= \mathbb{E}\,X^2 - 2\mu\mu + \mu^2 \\ &= \mathbb{E}\,X^2 - \mu^2\end{aligned}$$

which is the expression required in the question.

Question 10 (cdf and pdf). Derive the mean and variance of the Exponential distribution, which has density

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that X takes a value in excess of two standard deviations from the mean?

First, a sanity check. Does this density make sense, i.e. does it integrate to one?

$$\int_{-\infty}^{\infty} f(x)\,dx = \int_0^{\infty} \lambda e^{-\lambda x}\,dx = \left[-e^{-\lambda x}\right]_0^{\infty} = -0 - (-1) = 1.$$

Now for the mean:

$$\begin{aligned}
\mathbb{E}\,X &= \int_{-\infty}^{\infty} x f(x)\,dx \quad \textit{by definition of mean} \\
&= \int_0^{\infty} x\,\lambda e^{-\lambda x}\,dx \\
&= \left[x(-e^{-\lambda x})\right]_0^{\infty} - \int_0^{\infty} 1 \times (-e^{-\lambda x})\,dx \quad \textit{using integration by parts} \\
&= 0 + \int_0^{\infty} e^{-\lambda x}\,dx \\
&= \left[-\frac{1}{\lambda} e^{-\lambda x}\right]_0^{\infty} \\
&= \frac{1}{\lambda}.
\end{aligned}$$

For the variance, let's use the formula from the last question, $\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2$. *The first term is*

$$\begin{aligned}
\mathbb{E}\,X^2 &= \int_0^{\infty} x^2\,\lambda e^{-\lambda x}\,dx \quad \textit{by the Law of the Unconscious Statistician} \\
&= \left[x^2(-e^{-\lambda x})\right]_0^{\infty} + \int_0^{\infty} 2x\,e^{-\lambda x}\,dx \quad \textit{using integration by parts} \\
&= 0 + \frac{2}{\lambda}\int_0^{\infty} x\,\lambda e^{-\lambda x}\,dx \quad \textit{inserting } \lambda/\lambda \textit{ so we can use the } \mathbb{E}\,X \textit{ formula} \\
&= \frac{2}{\lambda} \times \frac{1}{\lambda} = \frac{2}{\lambda^2},
\end{aligned}$$

giving

$$\operatorname{Var} X = \mathbb{E}\,X^2 - (\mathbb{E}\,X)^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

The question also asks for the probability of a value in excess of two standard deviations from the mean. The wording is ambiguous, but let's interpret it as

$$\begin{aligned}
&\mathbb{P}\big(|X - \mu| > 2\sigma\big) \quad \textit{where } \mu = \mathbb{E}\,X = \frac{1}{\lambda},\ \sigma = \sqrt{\operatorname{Var} X} = \frac{1}{\lambda} \\
&= \mathbb{P}(X < \mu - 2\sigma) + \mathbb{P}(X > \mu + 2\sigma) \\
&= \mathbb{P}\Big(X < \frac{1}{\lambda} - \frac{2}{\lambda}\Big) + \mathbb{P}\Big(X > \frac{1}{\lambda} + \frac{2}{\lambda}\Big) \\
&= \mathbb{P}\Big(X < -\frac{1}{\lambda}\Big) + \mathbb{P}\Big(X > \frac{3}{\lambda}\Big) \\
&= \mathbb{P}\Big(X > \frac{3}{\lambda}\Big) \quad \textit{since } \mathbb{P}(X < 0) = 0 \\
&= \int_{3/\lambda}^{\infty} \lambda e^{-\lambda x}\,dx \\
&= \left[-e^{-\lambda x}\right]_{3/\lambda}^{\infty} \\
&= e^{-\lambda 3/\lambda} = e^{-3} \approx 5\%.
\end{aligned}$$

Question 11 (Law of the Unconscious Statistician). The University bus arrives at the Computer Lab bus stop at 7am, 7:15am, and so on at 15 minute intervals. You arrive at the bus stop at a time uniformly distributed in the interval [1pm, 1:20pm]. Let W be the length of time you wait. Find the expected value of W.

Let T be the time we arrive, measured in minues after 1pm, $T \sim U[0, 20]$. The waiting time is a function of T,

$$W = h(T) = \begin{cases} 15 - T & \text{if } T \in (0, 15) \\ 15 - (T - 15) & \text{if } T \in (15, 20). \end{cases}$$

(It doesn't matter how we define $h(T)$ at $T = 0$ or $T = 15$ or $T = 20$. Since T is a continuous random variable, the value of $h(T)$ for those T values does not affect the integral below.) So

$$\begin{aligned} \mathbb{E}\, W &= \int_{t=0}^{20} h(t)\, \mathrm{pdf}(t)\, dt \quad \text{by the Law of the Unconscious Statistician} \\ &= \int_{t=0}^{20} h(t) \frac{1}{20}\, dt \quad \text{since } T \sim U[0, 20] \\ &= \int_{t=0}^{15} 15 - t\, dt + \int_{t=15}^{20} 30 - t\, dt \\ &= \frac{1}{20}\Big\{15 \times 15 - [t^2/2]_0^{15} + 30 \times 5 - [t^2/2]_{15}^{20}\Big\} = 8.75. \end{aligned}$$

Question 12. A coal bunker is to be constructed on the side of a house. Assuming that it is a cuboid of given volume V, find the shape that minimizes the external surface area.

(From IA Maths.) Let the three sides be x, y, and z, where x is measured horizontally going perpendicular to the house side, y is measured horizontally and parallel to the house side, and z is measured vertically. Then the volume V and the external surface area A are

$$V = xyz, \quad A = xy + yz + 2xz.$$

We can eliminate z by writing $z = V/xy$, and considering

$$A(x, y) = xy + \frac{V}{x} + \frac{2V}{y}.$$

To find the stationary points of A, set the partial derivatives to zero:

$$\frac{\partial A}{\partial x} = y - \frac{V}{x^2} = 0, \qquad \frac{\partial A}{\partial y} = x - \frac{2V}{y^2} = 0.$$

From the first equation $y = V/x^2$; substituting into the second $x = 2x^4/V$. Since $x \neq 0$, we obtain $V = 2x^3$ and so

$$x = \Big(\frac{V}{2}\Big)^{1/3}, \quad y = \frac{V}{x^2} = 2x, \quad z = \frac{V}{xy} = x.$$

The optimal shape is therefore 1:2:1.

To check that A is minimized by this solution, we should check the Hessian, i.e. the matrix of second derivatives. (For IB Data Science, this amount of calculation is overkill, and we'll prefer numerical optimization. For Part II Machine Learning and Bayesian Inference, this Hessian calculation is something you should be comfortable with.) The Hessian consists of

$$\frac{\partial^2 A}{\partial x^2} = \frac{2V}{x^3} = 4, \quad \frac{\partial^2 A}{\partial y^2} = \frac{4V}{y^3} = 1, \quad \frac{\partial^2 A}{\partial x \partial y} = 1.$$

In IA Maths for NST we learnt a test using the Hessian to see if the stationary point is indeed a minimum:

$$\frac{\partial^2 A}{\partial x^2}\frac{\partial^2 A}{\partial y^2} > \Big(\frac{\partial^2 A}{\partial x \partial y}\Big)^2 \quad \text{with} \quad \frac{\partial^2 A}{\partial x^2} > 0 \quad \text{and} \quad \frac{\partial^2 A}{\partial y^2} > 0$$

so A has a local minimum at this point.

Question 13. *More examples of finding stationary points. These two specific cases crop up again and again in data science. You should be able to solve them blindfolded.*

(a) Find the value of $p \in [0, 1]$ that maximizes $p^a(1 - p)^b$, where a and b are both positive.

(b) Find the value of $p \in [0,1]$ that maximizes $\log\big(p^a(1-p)^b\big)$, where a and b are both positive.

(c) Find the values of $\mu \in \mathbb{R}$ and $\sigma > 0$ that jointly maximize

$$\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\Big)$$

(d) Find the values of $\mu \in \mathbb{R}$ and $\rho > 0$ that jointly maximize

$$\Big(\frac{1}{\sqrt{2\pi\rho}}\Big)^n \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\rho}\Big)$$

What do you notice about the solutions to (a) *versus* (b), and about the solutions to (c) *versus* (d)?

(a) The derivative is

$$\frac{d}{dp}p^a(1-p)^b = ap^{a-1}(1-p)^b - bp^a(1-p)^{b-1} = p^{a-1}(1-p)^{b-1}\big(a(1-p)-bp\big).$$

The maximum is at

$$\frac{d}{dp}=0 \quad\Longrightarrow\quad a(1-p)-bp=0 \quad\Longrightarrow\quad p(a+b)=a \quad\Longrightarrow\quad p=\frac{a}{a+b}.$$

(b) The derivative is

$$\frac{d}{dp}\log\big(p^a(1-p)^b\big) = \frac{d}{dp}\big(a\log p + b\log(1-p)\big) = \frac{a}{p} - \frac{b}{1-p}.$$

The maximum is at

$$\frac{d}{dp}=0 \quad\Longrightarrow\quad \frac{a}{p}=\frac{b}{1-p} \quad\Longrightarrow\quad a(1-p)=bp \quad\Longrightarrow\quad p=\frac{a}{a+b}.$$

These two must *give the same answer. The only difference is a stretch of the y-axis, so of course the maximum is at the same p value. (This works because $x \mapsto \log x$ is a strictly increasing function.)*

(c) We might as well take logs to make the maximization simpler, as in part (b):

$$\log\Big\{\Big(\frac{1}{\sqrt{2\pi\sigma^2}}\Big)^n \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\sigma^2}\Big)\Big\} = -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)^2.$$

The maximum is when $\partial/\partial\mu = 0$ and $\partial/\partial\sigma = 0$, i.e. when

$$\frac{\partial}{\partial\mu}: \quad -\frac{2}{2\sigma^2}\sum_{i=1}^n (x_i-\mu)=0 \quad\Rightarrow\quad \sum(x_i-\mu)=0 \quad\Rightarrow\quad \sum x_i = n\mu \quad\Rightarrow\quad \mu=\frac{\sum x_i}{n}$$

$$\frac{\partial}{\partial\sigma}: \quad -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^n (x_i-\mu)^2=0 \quad\Rightarrow\quad n\sigma^2=\sum(x_i-\mu)^2 \quad\Rightarrow\quad \sigma=\sqrt{\frac{1}{n}\sum(x_i-\mu)^2}.$$

Normally it's considered bad form to leave of the variables on the right hand side, in the way that the formula for σ involves μ; but since it's straightforward to evaluate μ we'll let this be.

(d) As for part (c) we'll take logs first:

$$\log\Big\{\Big(\frac{1}{\sqrt{2\pi\rho}}\Big)^n \exp\Big(-\sum_{i=1}^n \frac{(x_i-\mu)^2}{2\rho}\Big)\Big\} = -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\rho - \frac{1}{2\rho}\sum_{i=1}^n (x_i-\mu)^2.$$

The maximum is when $\partial/\partial\mu = 0$ and $\partial/\partial\sigma = 0$, i.e. when

$$\frac{\partial}{\partial\mu}=0 \quad\Rightarrow\quad \mu=\frac{\sum x_i}{n} \quad \textit{exactly as before}$$

$$\frac{\partial}{\partial\rho}: \quad -\frac{n}{2\rho}-\frac{1}{2\rho^2}\sum(x_i-\mu)^2=0 \quad\Rightarrow\quad \rho=\frac{1}{n}\sum(x_i-\mu)^2.$$

This is the same answer as (c), with the substitution $\rho = \sigma^2$. There is a one-to-one mapping between $\sigma > 0$ and $\rho > 0$, so it doesn't matter which form of the optimization we do, we'll find the same maximum either way.

Question 14. Replace the `for` loop with vectorized operations. Here `size` is an integer.

```
def rgalaxies(size, p=[0.28, 0.54, 0.18], μ=[9740, 21300, 15000], σ=[340, 1700, 10600]):
    res = []
    for _ in range(size):
        cluster = np.random.choice([0,1,2], p=p)
        μi,σi = μ[cluster], σ[cluster]
        x = np.random.normal(loc=μi, scale=σi)
        res.append(x)
    return res
```

```
def rgalaxies(size, p, μ, σ):
    cluster = np.random.choice([0,1,2], p=p, size=size)
    return np.random.normal(loc=np.array(μ)[cluster], scale=np.array(σ)[cluster])
```

Question 15. Given a vector $[x_1, \ldots, x_n]$, and parameters p, μ, and σ as in question 14, write numpy vectorized code to compute

$$\sum_{i=1}^{n} \log\Big[\sum_{k=1}^{3} p_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{-(x_i-\mu_k)^2/2\sigma_k^2}\Big]$$

Your code should be vectorized over $i = 1, \ldots, n$, but it need not be vectorized over $k = 1, 2, 3$. (Vectorizing over both i and k needs numpy's 'broadcast semantics', which are frequently needed in deep learning, but are overkill for IB Data Science.)

```
ϕ = scipy.stats.norm.pdf
res = np.zeros(len(x))
for pk,μk,σk in zip(p,μ,σ):
    res += pk * ϕ(x, loc=μk, scale=σk)
np.sum(np.log(res))
```