# Solutions sheet 0

Remembering IA Maths for NST
Foundations of Data Science—DJW—2019/2020

**Question 1.** 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.

> *Copied from Maths for NST A page 185. Writing*
>
> $$\begin{aligned} &H \equiv \textit{hearts}, \quad D \equiv \textit{Diamonds}, \quad C \equiv \textit{Clubs}, \quad S \equiv \textit{Spaces}, \\ &a \equiv \textit{ace}, \quad k \equiv \textit{king}, \quad q \equiv \textit{queen}, j \equiv \textit{jack}, \quad 10 \equiv \textit{10}, \ldots \end{aligned}$$
>
> *we can write the events as*
>
> $$\begin{aligned} A &= \{H_a, D_a, C_a, S_a\} & \mathbb{P}(A) &= \frac{4}{52} = \frac{1}{13} \\ B &= \{S_a, S_k, S_q, S_j, S_{10}, \ldots, S_2\} & \mathbb{P}(B) &= \frac{13}{52} = \frac{1}{4} \\ C &= \{H_a, H_k, H_q, H_j, H_{10}, D_a, D_k, \ldots\} & \mathbb{P}(C) &= \frac{4 \times 5}{52} = \frac{5}{13}. \end{aligned}$$
>
> *(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.** 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?

> *Copied from Maths for NST A page 187. Write $\Omega$ 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.** Consider drawing 2 balls out of a bag of 5 balls: 1 red, 2 green, 2 blue. What is the probability of the second ball drawn from the bag being blue given that the first ball was blue if (i) the first ball is replaced, (ii) the first ball is not replaced?

> *Adapted from Maths for NST A page 188.*
>
> *In scenario (i), after the first ball has been replaced then the bag still has 1 red, 2 green, 2 blue, so the probability that the second ball is blue is* $2/5$.
>
> *In scenario (ii), after the first ball has been removed, then the bag now has 1 red, 2 green, 1 blue, so the probability that the second ball is blue is* $1/4$.
>
> *We could answer the question more formally, by setting up a sample space containing all possible outcomes for the pair (1st draw, 2nd draw),*
> $$\Omega = \big\{(r,r),(r,g),(r,b),(g,r),(g,g),(g,b),(b,r),(b,g),(b,b)\big\},$$
> *then figuring out the probability of each of these outcomes, then using conditional probability to find* $\mathbb{P}(B_2 \mid B_1)$ *where*
> $$\begin{aligned} B_1 &= \{\textit{blue on 1st draw}\} = \{(b,r),(b,g),(b,b)\} \\ B_2 &= \{\textit{blue on 2nd draw}\} = \{(r,b),(g,b),(b,b)\}. \end{aligned}$$
>
> *Alternatively we could imagine that the balls have identities as well as colours, for example* $g_1 \equiv$ *green ball 1,* $g_2 \equiv$ *green ball 2, and consider the sample space*
> $$\Omega' = \big\{(r,r),(r,g_1),(r,g_2),\dots\big\}.$$
> *This second sample space makes it easier to reason about the probability of each outcome.*
>
> *In the language of conditional probability and independence, it turns out the events* $B_1$ *and* $B_2$ *are independent in scenario (i), but they are not independent in scenario (ii).*

**Question 4.** Two cards are drawn from a deck of cards. What is the probability of drawing two queens, given that the first card is not replaced?

> *Adapted from IA Maths for NST, page 189.*
>
> *Let's answer this question formally, by setting up a full sample space of all possible outcomes, i.e. let* $\Omega$ *consist of all pairs (1st card, 2nd card). Every outcome is equally likely, except for outcome where the 1st and 2nd cards are identical; these have probability 0, since the first card is not replaced. Then*
> $$\{\textit{draw two queens}\} = \big\{(H_q, D_q),(H_q,S_q),(D_q,H_q),\dots\big\}$$
> *so*
> $$\mathbb{P}(\textit{draw two queens}) = \frac{\textit{num. possible outcomes with two queens}}{\textit{num. possible outcomes}} = \frac{4\times 3}{52\times 52 - 52} = \frac{1}{221}.$$

**Question 5.** A screening test is 99% effective in detecting a certain disease when a person has the disease. The test yields a 'false positive' for 0.5% of healthy persons tested. Suppose 0.2% of the population has the disease. (i) What is the probability that a person whose test is positive has the disease? (ii) What is the probability that a person whose test is negative actually has the disease after all?

> *Taken from IA Maths page 192. Define the events*
> $$A = \{\textit{person has disease}\}, \qquad B = \{\textit{test came back positive}\}.$$
> *The question tells us several conditional probabilities:*
> $$\begin{aligned} &\mathbb{P}(B \mid A) = 0.99, && \mathbb{P}(B \mid \neg A) = 0.005, \\ &\mathbb{P}(\neg B \mid A) = 0.01, && \mathbb{P}(\neg B \mid \neg A) = 0.995. \end{aligned}$$
> *It also tells us about the probability of* $A$ *absent any diagnostic information:*
> $$\begin{aligned} &\mathbb{P}(A) = 0.002, \\ &\mathbb{P}(\neg A) = 0.998. \end{aligned}$$
> *For (i), Applying Bayes's rule,*
> $$\begin{aligned} \mathbb{P}(A \mid B) = \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(B)} &= \frac{\mathbb{P}(A)\,\mathbb{P}(B|A)}{\mathbb{P}(A)\,\mathbb{P}(B|A) + \mathbb{P}(\neg A)\,\mathbb{P}(B|\neg A)} \\ &= \frac{0.002\times 0.99}{0.002\times 0.99 + 0.998\times 0.005} \\ &\approx 0.284. \end{aligned}$$

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

$$\begin{aligned}\mathbb{P}(A \mid \neg B) = \frac{\mathbb{P}(A)\,\mathbb{P}(\neg B \mid A)}{\mathbb{P}(\neg B)} &= \frac{\mathbb{P}(A)\,\mathbb{P}(\neg B \mid A)}{\mathbb{P}(A)\,\mathbb{P}(\neg B \mid A) + \mathbb{P}(\neg A)\,\mathbb{P}(\neg B \mid \neg A)} \\ &= \frac{0.002 \times 0.01}{0.002 \times 0.01 + 0.998 \times 0.995} \\ &\approx 2.01 \times 10^{-5}.\end{aligned}$$

**Question 6.** What is the probability that in a room of $r$ people at least two have the same birthday?

*Taken from IA Maths page 194.*

*Assume 365 days in a year. Let the event $A$ be "two or more people have birthdays on the same day", so $\neg A$ is "all people have different birthdays". The number of arrangements of birthdays in which all birthdays are distinct is*

$$N_{\neg A} = 365 \times 364 \times \cdots \times (365 - r + 1) = \frac{365!}{(365-r)!}.$$

*The total number of arrangements allowing repeats is*

$$N = 365 \times 365 \times \cdots \times 365 = 365^r.$$

*So*

$$\mathbb{P}(\neg A) = \frac{N_{\neg A}}{N} = \frac{365!}{(365-r)!365^r},$$

*and*

$$\mathbb{P}(A) = 1 - \mathbb{P}(\neg A) = 1 - \frac{365!}{(365-r)!365^r}.$$

**Question 7.** Out of 10 physics professors and 12 chemistry professors, a committee of 5 people must be chosen in which each subject has at least 2 representatives. In how many ways can this be done?

*There are 10 physics professors (P) and 12 chemistry professors (C), and the committee needs 5 people, so it can be either $2P + 3C$ or $3P + 2C$. Writing $N_{2P}$ for the number of ways to pick 2 physics professors, etc.,*

$$N_{2P} = \binom{10}{2} = 45, \quad N_{3C} = \binom{12}{3} = 220, \quad N_{2P+3C} = N_{2P}N_{3C} = 45 \times 220 = 9900$$

*and*

$$N_{3P} = \binom{10}{3} = 120, \quad N_{2C} = \binom{12}{2} = 66, \quad N_{3P+2C} = N_{3P}N_{2C} = 120 \times 66 = 7920.$$

*The total number of possible committees is*

$$N_{2p+3C} + N_{3P+2C} = 9900 + 7920 = 17820.$$

**Question 8.** What is the probability of throwing exactly 3 heads out of 6 tosses of a fair coin? How about at least one head?

*Taken from IA Maths page 200. Let $p$ be the probability of heads, $1-p$ the probability of tails. The coin is fair so $p = 1/2$. The question asks for*

$$\mathbb{P}(\textit{3 heads out of 6 tosses}) = \binom{6}{3} p^3(1-p)^3 = \frac{6!}{3!3!}({}^1\!/_2)^3({}^1\!/_2)^3 = \frac{20}{64} = \frac{5}{16}.$$

> *It also asks for*
>
> $$\mathbb{P}(\textit{at least one head}) = 1 - \mathbb{P}(\textit{zero heads}) = 1 - \binom{6}{0} p^0 (1-p)^6 = 1 - \frac{1}{64} = \frac{63}{64}.$$

**Question 9.** A bag contains 6 blue balls and 4 red balls. Three balls are drawn from the bag without replacement. Let $X$ be the number of these three balls that are red. Find the density function $f(x) = \mathbb{P}(X = x)$.

> *Taken from IA Maths page 204. This question uses the term "density function", when "probability mass function" is technically better. But given that so many machine learning formulae apply equually to probability mass functions (in the discrete case) and density functions (in the continuous case), it's nice to have a single common wording for both cases.*
>
> $$\begin{aligned} \mathbb{P}(X=0) &= \mathbb{P}(bbb) = \frac{6}{10} \times \frac{5}{9} \times \frac{4}{8} = \frac{120}{720} = \frac{1}{6}. \\ \mathbb{P}(X=1) &= \mathbb{P}(rbb) + \mathbb{P}(brb) + \mathbb{P}(bbr) \\ &= \frac{4}{10} \times \frac{6}{9} \times \frac{5}{8} + \frac{6}{10} \times \frac{4}{9} \times \frac{5}{8} + \frac{6}{10} \times \frac{5}{9} \times \frac{4}{8} = \frac{120+120+120}{720} = \frac{1}{2}. \\ \mathbb{P}(X=2) &= \mathbb{P}(rrb) + \mathbb{P}(rbr) + \mathbb{P}(bbr) \\ &= \frac{4}{10} \times \frac{3}{9} \times \frac{6}{8} + \frac{4}{10} \times \frac{6}{9} \times \frac{3}{8} + \frac{6}{10} \times \frac{4}{9} \times \frac{3}{8} = \frac{72+72+72}{720} = \frac{3}{10} \\ \mathbb{P}(X=3) &= \mathbb{P}(rrr) = \frac{4}{10} \times \frac{3}{9} \times \frac{3}{8} = \frac{24}{720} = \frac{1}{30}. \end{aligned}$$
>
> *As a sanity check, let's verify that probabilities sum to one:*
>
> $$\mathbb{P}(X=0) + \mathbb{P}(X=1) + \mathbb{P}(X=2) + \mathbb{P}(X=3) = \frac{120}{720} + \frac{360}{720} + \frac{216}{720} + \frac{24}{720} = 1.$$

**Question 10.** 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 11.** Find 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 sum to one?*
>
> $$\int_{-\infty}^{\infty} f(x)\, dx = \int_0^{\infty} \lambda e^{-\lambda x}\, dx = \big[-e^{-\lambda x}\big]_0^{\infty} = -0 - (-1) = 1.$$

> *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*
>
> $$\begin{aligned} \mathbb{E}\,X^2 &= \int_0^{\infty} x^2\,\lambda e^{-\lambda x}\,dx \\ &= \left[x^2(-e^{-\lambda x})\right]_0^{\infty} + \int_0^{\infty} 2x\,e^{-\lambda x}\,dx \quad \textit{using integration by parts} \\ &= 0 + \frac{2}{\lambda}\int_0^{\infty} x\,\lambda e^{-\lambda x}\,dx \quad \textit{inserting } \lambda/\lambda \textit{ so we can use the } \mathbb{E}\,X \textit{ formula} \\ &= \frac{2}{\lambda} \times \frac{1}{\lambda} = \frac{2}{\lambda^2}, \end{aligned}$$
>
> *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 12.** 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?

> *From IA Maths page 234. For part (i),*
>
> $$\mathbb{P}(\textit{B gets a first throw then wins}) = \mathbb{P}\big(\textit{A gets \{4,5,6\} then B gets \{1,2\}}\big) = \frac{3}{6} \times \frac{2}{6} = \frac{1}{6}.$$

*For part (ii),*

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

*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".*

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

*Now,*

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

*Putting this all together,*

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

**Question 13.** 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 page 118. 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 we learnt a test using the Hessian to see if the stationary point is indeed a minimum:*

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

*so $A$ has a local minimum at this point.*

**Question 14.** *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 $\log\bigl(p^a(1-p)^b\bigr)$, where $a$ and $b$ are both positive.

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

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

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

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

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

*[Note. On the printed handout, the formulae for (c) and (d) were not what I intended. I've updated them here.]*

*(a) The derivative is*

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

*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*

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

*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):*

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

*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 $\sigma$ involves $\mu$; but since it's straightforward to evaluate $\mu$ we'll let this be.*

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

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

*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 15.** Bayes's rule says that, for any events $A$ and $B$ with $\mathbb{P}(B) > 0$,

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

There are four core definitions and laws in probability theory. Derive Bayes's rule from them, and explain carefully which of the four you are using at each step.

(a) $\mathbb{P}(\Omega) = 1$ where $\Omega$ is the entire sample space

(b) Conditional probability: $\mathbb{P}(A \mid B) = \mathbb{P}(A \cap B)/\,\mathbb{P}(B)$, when $\mathbb{P}(B) > 0$

(c) Sum rule: If $\{B_1, B_2, \dots\}$ partition $\Omega$ then $\mathbb{P}(A) = \sum_i \mathbb{P}(A \cap B_i)$
Law of total probability: $\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)$

(d) $A$ and $B$ are said to be *independent* if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$

*Note: "$\{B_1, B_2, \dots\}$ partition $\Omega$" means the $B_i$ are mutually exclusive and $\bigcup_i B_i = \Omega$.*

> *By the definition of conditional probability, definition (b), 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, law (c): since $A$ and $\neg A$ partition $\Omega$,*
>
> $$\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) + \mathbb{P}(B \mid \neg A)\,\mathbb{P}(\neg A).$$
>
> *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.)*