Mathematics 206 Probability & Statistics II Solutions

Professor Peter M. Higgins

November 1, 2018

Solutions and Comments for the Problems

Problem Set 1

1(a) For 3 people, the chances of 3 different birthdays is $1 \times \frac{6}{7} \times \frac{5}{7} = \frac{30}{49} > \frac{1}{2}$ but for 4 persons we get $\frac{30}{49} \times \frac{4}{7} = \frac{120}{343} < \frac{1}{2}$. Hence we require 4 people to make it more likely than not that we will have one or more coincidences of birthdays on the same day of the week.

(b) In general, if we have n people, the probability p_n that they all have *different* birthdays is

$$p_n = 1 \times \frac{364}{365} \times \frac{363}{365} \times \cdots \times \frac{(365 - n + 1)}{365}.$$

We require the least value of n for which $p_n < \frac{1}{2}$. By successive calculation we find that $n = 23$.

Comment Since $4 > \frac{7}{2}$ we might guess that the answer to (b) is 183, that being the least integer that exceeds $\frac{365}{2}$, yet the correct answer is only 23. It follows that for a class of 30 children it is quite likely that at least 2 of them share the same birthday.

2(a) If A leads all the way, (which is possible as $p > q$) then the path of the count will always lie above the x-axis, apart from the initial point $(0, 0)$, ending at the point $(p + q, p - q)$ and vice-versa.

(b) The paths of counts where A never falls behind their opponent are exactly those that never lie below the x-axis, although the path may touch that axis (corresponding to points in the count where the vote is tied).

(c) The total number of counts is $\binom{p+q}{p} = \binom{p+q}{q}$.

3. The path of the reverse count can be seen by making $(p+q, p-q)$ the new origin and inverting the path, so that the positive x-direction now goes from right to left. A count in which A leads throughout reverses to a count where A's winning margin (which is $p-q$) is never attained until the final vote is counted. For counts that never cross the x-axis, the reverse counts are exactly those for which the lead of A never exceeds his eventual winning margin.

4. Consider a path from $(0, a)$ to (b, c) that first meets the x-axis at $(d, 0)$. If we reflect that initial portion of the path in the x-axis the result is a path from $(0, -a)$ to (b, c). Conversely, if we reflect the initial segment of such a path in the x-axis the result is a path from $(0, a)$ to (b, c) that first meets the x-axis at $(d, 0)$. Since these two operations are then inverses of one another, it follows that these operations are bijections between the underlying sets of paths. In particular the number of paths in each collection is the same.

5. There are b edges in the path with the number u of up-edges exceeding the number d of down-edges by $a + c$, so that $u = d + a + c$ and

$$u + d = b \Rightarrow d + a + c + d = b$$

$$d = \frac{b-a-c}{2},$$

$$u = \frac{b-a-c}{2} + a + c = \frac{a+b+c}{2}.$$

Hence the number of paths between $(0,-a)$ and (b,c) is

$$\binom{b}{\frac{a+b+c}{2}}.$$

Comment: note that a, b, c cannot be arbitrary positive integers but must satisfy the constraint that $a + c \leq b$ so that $\frac{a+b+c}{2} \leq \frac{2b}{2} = b$. Moreover, reducing the equation $a + c + 2d = b$ modulo 2 gives that $b - (a + c)$ is even and therefore so is $b - (a + c) + 2(a + c) = a + b + c$.

6. The number of counts in which B never leads is equal to the number of counts from $(0,0)$ to $(p+q, p-q)$ that do not touch or cross the line $y = -1$. By the Reflection principle, the number of such paths which begin at the origin yet violate this condition is the number of paths from $(0,-2)$ to $(p+q, p-q)$. Here we have $a = 2$, $b = p + q$ and $c = p - q$ so applying the result of Question 5 gives

$$\frac{a+b+c}{2} = \frac{2+p+q+p-q}{2} = p+1$$

and so the number of counts in which B does lead at some point is:

$$\binom{p+q}{p+1}.$$

Therefore the probability that B never leads is:

$$1 - \frac{\binom{p+q}{p+1}}{\binom{p+q}{p}} = 1 - \frac{(p+q)!}{(q-1)!(p+1)!} \cdot \frac{p!q!}{(p+q)!} = 1 - \frac{q}{p+1}$$

$$= \frac{p+1-q}{p+1}.$$

7. For A to lead all through the night, the first counted vote must be for A. The total number of paths from $(1,1)$ to $(p-q, p+q)$ is:

$$\binom{p+q-1}{q}.$$

The numerator of the probability that A leads throughout, given that he takes the first vote, is the number of paths from $(1,1)$ to $(p+q, p-q)$ that do not meet the x-axis. The complementary set of paths are those from $(1,1)$ that do meet the x-axis which, by the Reflection principle, equals the number of paths from $(1,-1)$ to $(p+q, p-q)$. Such a path has $p - q + 1$ more up-edges u than down edges d and $u + d = p + q - 1$. Hence

$$p - q + 1 + 2d = p + q - 1 \Rightarrow d = q - 1,\ u = (q-1) + (p - q + 1) = p.$$

Hence, given that the first vote is for A, the probability that A leads throughout is:

$$\left(1 - \frac{\binom{p+q-1}{p}}{\binom{p+q-1}{q}} = 1 - \frac{(p+q-1)!}{p!(q-1)!} \cdot \frac{q!(p-1)!}{(p+q-1)!} = 1 - \frac{q}{p} = \frac{p-q}{p}.\right.$$

Therefore the probability that A leads thoughout is this ratio multiplied by $\frac{p}{p+q}$, the probability that the first vote counted is for A, which then gives

$$\frac{p}{p+q} \cdot \frac{p-q}{p} = \frac{p-q}{p+q}.$$

8. There are three disjoint possibilities for the count. Either A leads all the way, or A never falls behind but there is a tie during the count, or B leads at some point. In the latter case however there must a tie after this point as A eventually leads (and wins). Hence our required probability P satisfies:

$$P = 1 - \mathrm{P}(\text{vote is tied at some point in the count}).$$

The set of counts that feature at least one tie comprises two disjoint subsets: those in which the first vote is for A, and those for which the first vote is for B. However, by the Reflection principle, both these sets have the same size for there is a bijection between the first set and the second defined by reflecting the first part of the count up to the point of the first tie. On the other hand, the second set is just the set of counts in which the first vote is for B, as all those counts lead to a tie later on. Hence the probability of a count from each of these sets is simply $\frac{q}{p+q}$, the probability that the first vote is for B. Therefore we obtain the result:

$$P = 1 - 2\frac{q}{p+q} = \frac{p+q-2q}{p+q} = \frac{p-q}{p+q}.$$

9(a). We prove this by an easy induction: initially $n = y = 0$ so the claim holds. Suppose that after n steps ($n \geq 0$) it is the case that $n \equiv y \pmod 2$. After $n+1$ steps the path has co-ordinates of either $(n+1, y-1)$ or $n+1, y+1)$. In either event the parity of n and y remain equal, and so the induction continues.

Comment In particular, the walk may only return to the origin after an even number of steps.

(b) The walk will return to the origin if and only if there have been an equal number, n in this case, of up-steps and down-steps. Since there are 2^{2n} walks of length $2n$, the required probability u_{2n} is given by:

$$u_{2n} = \frac{\binom{2n}{n}}{2^{2n}} = \frac{(2n)!}{(2^n n!)^2} = \frac{(2n)(2n-1)(2n-2)(2n-3)\cdots 1}{((2n)(2n-2)(2n-4)\cdots 2)^2}$$

$$= \frac{(2n-1)(2n-3)(2n-5)\cdots 1}{(2n)(2n-2)(2n-4)\cdots 2}.$$

Comment It may be shown using *Stirling's formula for factorials* that $u_{2n} \sim \frac{1}{\sqrt{\pi n}}$. In particular it follows that the complementary probability of at least one

return to the origin approaches 1 and $n \to \infty$, which is to say that ultimately the random walk must return to it origin. However, further analysis shows that returns to the origin become increasing rare as the walk progressees and are governed by the *arcsine distribution*.

10. Let W be a walk that returns to the origin (perhaps not for the first time) after $2n$ steps. Let the leftmost minimum point of W be $M = (k, m)$. Reflect the section from the origin to M along the vertical line $y = k$ and slide the reflected portion to the point $(2n, 0)$ of W. Taking M as the origin of a new coordinate system, the new path W' leads from the origin M to the point $(2n, 2m)$ and has all vertices on or above the x-axis. The mapping $W \mapsto W'$ is then a bijection from the set S_{2n} of all walks that return to the origin after $2n$ steps, to T_{2n}, the set of all walks of $2n$ steps that do not pass below the x-axis: the reason for this is that if $W_1, W_2 \in S_{2n}$ with $W_1 \neq W_2$ then $W_1' \neq W_2'$ so the mapping defined by $W \mapsto W'$ is one-to-one and since W can be retrieved unambigously from W', it follows that the mapping may be reversed, and so $u_{2n} = |S_{2n}| = |T_{2n}|$.

Now let U_t denote the set of walks of length t that have no returns to the origin. Clearly U_t is the disjoint union of U_t^+ and U_t^-, which are the collections of all members of U_t in which the first step is positive or negative respectively. A walk $W \in U_{2n}^+$ consists of an initial positive step followed by a walk $X \in T_{2n-1}$ and conversely any walk that is the concatentation of an initial positive step and a member of T_{2n-1} is a member of U_{2n}^+. Hence $|U_{2n}^+| = |T_{2n-1}|$.

Since no member of T_{2n-1} ends with a y-coordinate of 0 (i.e. does not return to the origin because $2n-1$ is odd), it follows that any member of T_{2n-1} may be extended in 2 ways, either by a positive or negative step, and the result is a member of T_{2n}. It follows that $|T_{2n}| = 2|T_{2n-1}|$. We conclude that

$$|U_{2n}| = 2|U_{2n}^+| = 2|T_{2n-1}| = |T_{2n}| = |S_{2n}|;$$

which is to say that the number of walks of length $2n$ that never return to the origin equals the number of walks that do return after $2n$ steps. The probability of each is, by Question 9(b), equal to $u_{2n} = \frac{\binom{2n}{n}}{2^{2n}}$.

Problem Set 2

1.

$$L(\theta; x_1, \cdots, x_n) = \big(\frac{1}{\sqrt{2\pi}}\big)^n \exp\big[-\sum_{i=1}^{n}(x_i - \theta)^2/2\big].$$

To find the maximum we work with $\log L$ instead, which has the same maximum, and put the first derivative equal to zero:

$$\frac{d\ln(\theta : x_1, \cdots, x_n)}{d\theta} = \sum_{i=1}^{n}(x_i - \theta) = 0.$$

Hence we set $u(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i$, the sample mean is the maximum likelihood estimate of the the actual mean and $\hat{\theta} = \bar{X}$.

2. In this case

$$L(\theta; x_1, \cdots, x_n) = \frac{\theta^{\sum_1^n x_i} e^{-n\theta}}{\Pi x_i!} \Rightarrow \ln L = (\ln \theta)\sum_{i=1}^{n} x_i - n\theta - \sum_{i=1}^{n}\sum_{j=1}^{x_i} j$$

$$\Rightarrow \frac{d(\ln L)}{d\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} - n$$

$$\Rightarrow \frac{n\hat{\theta}}{\sum_{i=1}^{n} x_i} = 1 \Rightarrow u(x_1, \cdots, x_n) = \frac{\sum_{i=1}^{n} x_i}{n};$$

hence $\hat{\theta} = \bar{X}$.

3.

$$\ln L = \ln(\theta^n (\Pi_{i=1}^{n} x_i)^{\theta - 1} = n \ln \theta + (\theta - 1)\sum_{i=1}^{n} \ln x_i$$

$$\Rightarrow \frac{d(\ln L)}{d\theta} = \frac{n}{\theta} + \sum_{i=1}^{n} \ln x_i, \text{ putting this equal to 0 gives:}$$

$$\hat{\theta} = -\frac{n}{\sum_{i=1}^{n} \ln X_i} = -\frac{n}{\ln(X_1 X_2 \cdots X_n)}.$$

4.

$$\ln L = \ln\big(\frac{e^{-\frac{1}{\theta}\sum_{i=1^{\cdot}}^{n} x_i}}{\theta^n}\big) = -\frac{1}{\theta}\sum_{i=1}^{n} x_i - n \ln \theta$$

$$\Rightarrow \frac{d \ln L}{d\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta^2} - \frac{n}{\theta}, \text{ putting this equals to 0 gives:}$$

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i}{n}$$

so that $\hat{\theta} = \bar{X}$.

5.

$$\ln L = \ln\big(e^{-\sum_{i=1}^{n} x_i + n\theta}\big) = n\theta - \sum_{i=1}^{n} x_i$$

$$\Rightarrow \frac{d \ln L}{d\theta} = n > 0;$$

Since this derivative is strictly increasing, we maximize the likelihood by making θ as large as possible. Since we have $\theta < x$, it follows that $\hat{\theta}$ is the *first order statistic* $\min(X_i)$.

6.

$$\ln L = \ln \frac{1}{2^n}\big(e^{-\sum_{i=1}^{n} |x_i - \theta|}\big) = -n \ln 2 - \sum_{i=1}^{n} |x_i - \theta|;$$

this quantity is maximized by minimizing $S(\theta) = \sum_{i=1}^{n} |x_i - \theta|$. This is achieved by taking the *median*. To see this, by re-labelling the data we may assume that

$x_1 \leq x_2 \leq \cdots \leq x_n$. Since $S(x_1) < S(\theta)$ for all $\theta < x_1$ and similarly $S(x_n) < S(\theta)$ for all $\theta > x_n$, an optimal value of θ lies in the interval $[x_1, x_n]$. Suppose that $x_i \leq \theta \leq x_{i+1}$ say. By putting $\theta = x_i$ we change S by $(n-i)(\theta - x_i) - i(\theta - x_i) = (n-2i)(\theta - x_i) \geq 0$ for θ optimal, whence $n - 2i \geq 0 \Leftrightarrow i \leq \frac{n}{2}$. By putting $\theta = x_{i+1}$ we change S by $i(x_{i+1} - \theta) - (n-i)(x_{i+1} - \theta) = (2i - n)(x_{i+1} - \theta) \geq 0$ for optimal θ so we also require also that $n \leq 2i \Leftrightarrow i \geq \frac{n}{2}$. If n is even, this gives $i = \frac{n}{2}$ and $x_{\frac{n}{2}} \leq \theta \leq x_{\frac{n+2}{2}}$; any value of θ in this interval (including the median) yields an optimal $\hat{\theta}$. If n is odd, the previous argument shows that moving θ to one of x_i or x_{i+1} will decrease $S(\theta)$; clearly the optimal value for θ in these circumstances is unique and is the median $x_{\frac{n+1}{2}}$.

7.

$$L(\theta; x_1, \cdots, x_n) = \frac{1}{\theta^n},\ 0 < x_i \leq \theta,$$

which is decreasing in θ. We therefore maximize the likelihood by taking θ as small as possible. Now $\theta \geq x_i$ so that L can be made no larger than

$$\frac{1}{[\max(x_i)]^n}$$

so that the unique maximum likelihood statistic for θ is the nth *order statistic* $\hat{\theta} = \max(X_i)$.

8. Let $X = \max(X_i)(0 \leq i \leq n)$ and let $f(x)$ denote the pdf of X. Then

$$\Pr(X \leq x) = \left(\tfrac{x}{\theta}\right)^n,\ (0 \leq x \leq \theta)$$

$$\Rightarrow f(x) = \frac{n}{\theta^n} x^{n-1}\ (0 \leq x \leq \theta)$$

$$E(X) = \int_0^\theta x f(x)\, dx = \frac{n}{\theta^n} \int_0^\theta x^n\, dx = \frac{n}{\theta^n} \left[\frac{x^{n+1}}{n+1}\right]_0^\theta$$

$$= \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1}\theta \neq \theta,$$

so that $\hat{\theta} = \frac{n\theta}{n+1}$ has mean less than the parameter θ that it is estimating.

9. In this case

$$L(\theta; x_1, x_2, \cdots, x_n) = 1,\ \theta - \frac{1}{2} \leq x_i \leq \theta + \frac{1}{2},$$

and 0 elsewhere. Thus L attains its maximum provided only that θ-$\frac{1}{2} \leq \min(x_i)$ and $\max(x_i) \leq \theta + \frac{1}{2}$ and so every statistic $u(X_1, \cdots, X_n)$ such that

$$\max(X_i) - \tfrac{1}{2} \leq u(X_1, \cdots, X_n) \leq \min(X_i) + \tfrac{1}{2}$$

is a maximum likelihood statistic. One such statistic is $[\min(X_i) + \max(X_i)]/2$.

10.

$$\ln(L(\theta; x_1, x_2, \cdots, x_n)) = \ln(\theta^{\sum_{i=1}^n x_i} (1-\theta)^{n - \sum_{i=1}^n x_i})$$

$$= (\ln\theta)(\sum_{i=1}^{n} x_i) + \ln((1-\theta))(n - \sum_{i=1}^{n} x_i)$$

$$\Rightarrow \frac{d\ln L}{d\theta} = \frac{\sum_{i=1}^{n} x_i}{\theta} + \frac{\sum_{i=1}^{n} x_i - n}{1-\theta}, \text{ putting this equal to 0 gives:}$$

$$\frac{(1-\theta)\sum_{i=1}^{n} x_i + (\sum_{i=1}^{n} x_i - n)\theta}{\theta(1-\theta)} = 0$$

$$\Rightarrow \sum_{i=1}^{n} x_i - n\theta = 0,$$

so that once again we find $\hat{\theta} = \bar{X}$.

Problem Set 3

1. If H_0 is true then $X \sim N(150, 100)$. Our *test statistic* is

$$z = \frac{x-\mu}{\sigma} = \frac{172-150}{10} = 2.2.$$

Since $|z| = 2 \cdot 2 > 1 \cdot 96$ we reject H_0 (as $2 \cdot 5\%$ of the normal distribution lies to the right of $x = 1 \cdot 96$ and the distribution is symmetric about the origin).

2. We have $X \sim \text{Bin}(n, p)$ with $n = 100$ and under $H_0 : p = 0 \cdot 9$. We use a one-tailed test as most approrpriate. We have $\mu = np = 100 \times 0 \cdot 9 = 90$ and $\sigma^2 = npq = 100(0 \cdot 9)(0 \cdot 1) = 9$. Our approximating distribution under H_0 is then $X \sim N(90, 9)$. We will reject H_0 if $z < -1 \cdot 645$ (as this represents the lower 5% of the standard normal distribution). Now with the usual continuity correction we compute

$$z = \frac{x - np}{\sqrt{npq}} = \frac{83 \cdot 5 - 90}{3} = -2 \cdot 17 < -1 \cdot 645,$$

and so we reject H_0. At the 5% significance level the claim of the distributor is rejected.

3. Under H_0 we have $\overline{X} \sim N(420, \frac{12^2}{100})$. Our test statistic is $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ so we calculate

$$z = \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} = \frac{423 - 420}{1 \cdot 2} = 2 \cdot 5.$$

Since $|z| > 1 \cdot 96$ we reject H_0 and conclude that the mean length of pipes produced has changed.

4. Here we have $\overline{X} \sim N(6, \frac{0 \cdot 8^2}{50})$ so that

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{\overline{X} - 6}{0 \cdot 113}.$$

Now $H_0 : \mu = 6$ as opposed to $H_1 : \mu \neq 6$ so at the 5% level we require that

$$-1 \cdot 96 \leq \frac{\overline{x} - 6}{0 \cdot 113} \leq 1 \cdot 96$$

$$\Leftrightarrow 5 \cdot 78 \leq \overline{x} \leq 6 \cdot 22.$$

Therefore, to accept H_0 the mean mass of the 50 components needs to lie in the interval ($5 \cdot 78$cm, $6 \cdot 22$cm).

5. (i) We have $X \sim \text{Bin}(n, p)$ with $n = 120$. Under H_0 we have $np = (120)(\frac{1}{6}) = 20$ and $npq = (120)(\frac{1}{2})(\frac{1}{2}) = 30$ so that $X \sim N(60, 30)$ is the approximating normal distribution. Allowing for the continuity correction we have

$$P(50 \leq X \leq 70) \to P(49 \cdot 5 \leq X \leq 70 \cdot 5) = P\big(\frac{49 \cdot 5 - 60}{\sqrt{30}} \leq X \leq \frac{70 \cdot 5 - 60}{\sqrt{30}}\big)$$

$$= P(-1 \cdot 917 \leq Z \leq 1 \cdot 917) = 0 \cdot 9446 \text{ (from standard normal table).}$$

Hence the probability of a Type I error is $1 - 0 \cdot 9446 = 0 \cdot 0554$.

Comment It is worth knowing that when subtracting a decimal from 1, the sum of the corresponding digits adds to 9 (here $9 + 0 = 9, 4 + 5 = 9$, and so on) except the final pair add to 10 (here $6 + 4 = 10$).

(ii) Here we have H_0 : coin is fair ($p = 0 \cdot 5$) while H_1 : coin is biased ($p = 0 \cdot 6$).

$P(\text{Type II error}) = P(\text{accepting } H_0 | H_1 \text{ is true}) = P(49 \cdot 5 \leq X \leq 70 \cdot 5 | p = 0 \cdot 6)$.

For $p = 0 \cdot 6$ we have $np = 72$ and $npq = 28 \cdot 8$ and so $X \sim N(72, 28 \cdot 8)$.

$$P(49 \cdot 5 \leq X \leq 70 \cdot 5) = P\big(\frac{49 \cdot 5 - 72}{\sqrt{28 \cdot 8}} \leq \frac{X - 72}{\sqrt{28 \cdot 8}} \leq \frac{70 \cdot 5 - 72}{\sqrt{28 \cdot 8}}\big)$$

$$= P(-4 \cdot 193 \leq Z \leq -0 \cdot 2795) = 0 \cdot 390 \text{ (3 sf).}$$

Therefore, with this decision rule there is a fairly high probability (39%) that the coin will be accepted as fair although in fact $p = 0 \cdot 6$.

6 (i) We are performng a *two-tailed test*, so we reject H_0 at the 5% level if $P(X \leq 2) \leq 0 \cdot 025$ (which is the appropriate direction as $2 < 6 \cdot 5$). From tables or direct calculation ($e^{-6 \cdot 5}(1 + 6 \cdot 5 + \frac{6 \cdot 5}{2})$) we see that under H_0, $P(X \leq 2) = 0 \cdot 043 > 0 \cdot 025$. Therefore we do not reject H_0 and accept that $\lambda = 6 \cdot 5$.

(ii) We are performing a *one-tailed test* and reject H_0 only if $P(X \leq 2) \leq 0 \cdot 05$, which it is (from part (i)) and so we reject H_0 in favour of H_1: $\lambda < 6 \cdot 5$.

7. Over a 12 month period the expected number of breakdowns in $12 \times 3 = 36$ so we are dealing with a $X \sim \text{Po}(36)$ distribution. We need to find the least n such that $P(X > n) \leq 0 \cdot 02$. We approximate X by the corresponding normal distribution $Y \sim N(36, 36)$ and with the continuity correction we solve

$$P(Y > n) \leq 0 \cdot 02$$

$$\Leftrightarrow P(\frac{Y-36}{6} \geq \frac{n+0\cdot 5-36}{6}) = P(Z \geq \frac{n-35\cdot 5}{6}) \leq 0\cdot 02$$
$$\Rightarrow \frac{n-35\cdot 5}{6} \geq 2\cdot 055 \Rightarrow n \geq 47\cdot 83,$$

and since n is integral, the least value of n that satisfies our requirements is $n = 48$. In conclusion, if the supplier guarantees a refund if 48 or more breakdowns occur in a given year, then the probability of a refund is less than $0\cdot 02$.

8. Under H_0 she is dealing with an r.v. $X \sim \text{Bin}(13, 0\cdot 5)$ and $P(X = x) = \binom{13}{x}(0\cdot 5)^x$. The observed value of x is $x = 10$. She performs a one-tailed test and is justified in rejecting H_0 at the 5% level if $P(X \geq 10) \leq 0\cdot 05$. Now

$$P(X \geq 10) = (0\cdot 5)^{13}\Big(\binom{13}{10} + \binom{13}{11} + \binom{13}{12} + \binom{13}{13}\Big) \approx 0\cdot 046\cdots .$$

Since $P(X \geq 10) \leq 0\cdot 05$ she rejects H_0 and concludes that there is evidence that the students performed better on the final examination than the mock examination.

9. Under the null hypothesis we have $\mu_1 - \mu_2 = 0$ so that under H_0 we have:

$$\overline{X}_1 - \overline{X}_2 \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) = N(0, \frac{40}{100} + \frac{30}{80}) = N(0, 0\cdot 775).$$

Hence $Z = \frac{\overline{X}_1 - \overline{X}_2 - 0}{\sqrt{0\cdot 775}} = \frac{\overline{X}_1 - \overline{X}_2}{0\cdot 880\cdots}$. We use a two-tailed test ($H_1 : \mu_1 \neq \mu_2$) and so at the 5% level we reject H_0 in favour of H_1 if $|z| \geq 1\cdot 96$, where

$$z = \frac{38\cdot 3 - 40\cdot 1}{0\cdot 880} = -2\cdot 04 < -1\cdot 96.$$

Therefore we conclude that there is evidence of a difference between the population means at the 5% significance level.

10. We find that $\hat{\sigma}^2 = \frac{100}{99}(14\cdot 5)^2 = 212\cdot 3...$ so that $\hat{\sigma} = 14\cdot 57...$

Comment For large values of n the fraction $\frac{n}{n-1}$ makes little difference. There are theoretical justifications for using this multiplier however, including the fact that the expectation of $\frac{nS^2}{n-1}$ is indeed σ^2.

Now $\overline{X} \sim N(\mu, \frac{\hat{\sigma}^2}{n})$, with $\hat{\sigma} = 14\cdot 57$, $n = 100$. Under $H_0 : \mu = 50$ we get $\overline{X} \sim N(50, \frac{14\cdot 57^2}{100})$. With a one-tailed test ($H_1 : \mu > 50$) at the 1% level we reject H_0 if $z \geq 2\cdot 326$ where

$$z = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{52\cdot 6 - 50}{14\cdot 57/\sqrt{100}} = 1\cdot 784;$$

since this does not exceed the critical value of $z = 2\cdot 326$ we accept the null hypothesis that $\mu = 50$.

Problem Set 4

1. Clearly $f(x,y) \geq 2-1-1=0$. Moreover

$$\int_0^1 \int_0^1 (2-x-y)\,dxdy = \int_0^1 [2x - \frac{x^2}{2} - xy]_0^1\,dy = \int_0^1 (2-\frac{1}{2}-y)\,dy$$

$$= \frac{3}{2} - [\frac{y^2}{2}]_0^1 = \frac{3}{2} - \frac{1}{2} = 1;$$

and so $f(x,y)$ is a pdf. Next

$$\int_0^1 (2-x-y)\,dy = [2y - xy - \frac{y^2}{2}]_0^1 = [(2-x-\frac{1}{2}) - 0] = \frac{3}{2} - x;$$

$$\Rightarrow f_1(x) = \frac{3}{2} - x,\ 0 \leq x \leq 1.$$

By symmetry we also get $f_1(y) = \frac{3}{2} - y$, $0 \leq y \leq 1$.

2. We have $0 < f(x,y)$ and

$$\int_0^2 \int_0^x \frac{xy}{2}\,dydx = \int_0^2 [\frac{xy^2}{4}]_0^{y=x}\,dx = \int_0^2 \frac{x^3}{4}\,dx$$

$$= [\frac{x^4}{16}]_0^2 = \frac{16}{16} - 0 = 1;$$

so that $f(x,y)$ is a pdf. Next we also have

$$\int_0^{y=x} \frac{xy}{2}\,dy = \frac{x^3}{4}$$

$$\Rightarrow f_1(x) = \frac{x^3}{4},\ 0 < x < 2.$$

To obtain $g(y)$ we change the order of integration of $f(x,y)$:

$$f_1(y) = \int_y^2 \frac{xy}{2}\,dx = [\frac{x^2 y}{4}]_{x=y}^{x=2} = y - \frac{y^3}{4}$$

$$\Rightarrow f_1(y) = \frac{y}{4}(4-y^2)\ 0 < y < 2.$$

Comment Note that $g(y) \geq 0$ and that $\int_0^2 g(y)\,dy = [\frac{y^2}{2} - \frac{y^4}{16}]_0^2 = (2-1) - (0-0) = 1$, as required for a pdf.

3. Since $f_1(x)$ and $f(x,y)$ are non-negative (and we are assuming $f_1(x) \neq 0$) we have that $\frac{f(x,y)}{f_1(x)} \geq 0$. Moreover

$$\int_{-\infty}^{\infty} \frac{f(x,y)}{f_1(x)}\,dy = \frac{1}{f_1(x)} \int_{-\infty}^{\infty} f(x,y)\,dy = \frac{f_1(x)}{f_1(x)} = 1,$$

as required to show that $f(y|x)$ is a pdf.

4.

$$f_1(x) = \int_x^1 2\,dy = 2(1-x),\ 0 < x < 1;$$

$$f_1(y) = \int_0^y 2\,dx = 2y,\ 0 < y < 1;$$

$$\Rightarrow f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{2}{2(1-x)} = \frac{1}{1-x},\ 0 < x < y < 1;$$

$$f(x|y) = \frac{f(x,y)}{f_1(y)} = \frac{2}{2y} = \frac{1}{y},\ 0 < x < y,\ 0 < y < 1.$$

5. We require

$$f(y|x) = \frac{f(x,y)}{f_1(x)} = \frac{xy}{2} \cdot \frac{4}{x^3} = \frac{2y}{x^2},\ 0 < x < 2,\ 0 < y < x.$$

$$E(Y|X) = \int_0^x y f(y|x)\,dy = \int_0^x \frac{2y^2}{x^2}\,dy = [\frac{2y^3}{3x^2}]_0^x = \frac{2x^3}{3x^2} = \frac{2x}{3},\ 0 < x < 2.$$

6.

$$P(a \le X \le b, c \le Y \le d) = \int_c^d \int_a^b f(x,y)\,dxdy = \int_c^d \int_a^b g(x)h(y)\,dxdy$$

$$= \int_c^d h(y)(\int_a^b g(x)\,dx)\,dy = P(a \le X \le b) \int_c^d h(y)\,dy$$

$$= P(a \le X \le b)P(c \le Y \le d).$$

7.

$$E(X+Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x+y)f(x,y)\,dxdy$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xf(x,y)\,dxdy + \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} yf(x,y)\,dxdy$$

$$= \int_{-\infty}^{\infty} x(\int_{-\infty}^{\infty} f(x,y)\,dy)\,dx + \int_{-\infty}^{\infty} y(\int_{-\infty}^{\infty} f(x,y)\,dx)\,dy$$

$$= \int_{-\infty}^{\infty} xf_1(x)\,dx + \int_{-\infty}^{\infty} yf_1(y)\,dy$$

$$= E(X) + E(Y).$$

Comment It is important that this result holds whether or not the random variables are independent. It is a consequence of linearity of the integral operator, and so extends to linear combinations of the random variables, meaning that $E(aX + bY) = aE(X) + bE(Y)$.

8. Writing μ_X and μ_Y for $E(X)$ and $E(Y)$ respectively we have through using the result of the previous question that

$$E((X-\mu_X)(Y-\mu_Y)) = E(XY - \mu_X Y - \mu_Y X + \mu_X\mu_Y)$$

$$E(XY) - \mu_X E(Y) - \mu_Y E(X) + \mu_X\mu_Y = E(XY) - 2\mu_X\mu_Y + \mu_X\mu_Y$$
$$= E(XY) - E(X)E(Y) = \mathrm{Cov}(X,Y).$$

9. Using the result of the previous question we obtain

$$\mathrm{Var}(X+Y) = E(X+Y)^2 - (E(X+Y))^2$$

$$= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x+y)^2 f(x,y)\,dxdy - (\mu_X+\mu_Y)^2$$

$$= (\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}x^2 f(x,y)\,dxdy - \mu_X^2) + (\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}y^2 f(x,y)\,dxdy - \mu_Y^2) + 2(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xyf(x,y)\,dxdy - \mu_X\mu_Y)$$

$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).$$

Replacing Y by $-Y$ in the previous formula; note that

$$\mathrm{Var}(-Y) = E((-Y)^2) - (E(-Y))^2 = E(Y^2) - (-1)^2(E(Y))^2 = E(Y^2) - (E(Y))^2 = \mathrm{Var}(Y);$$

similarly

$$\mathrm{Cov}(X-Y) = E(X(-Y)) - E(X)(E(-Y) = -E(XY) + E(X)E(Y) = -\mathrm{Cov}(X,Y)$$

$$\therefore \mathrm{Var}(X-Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X,Y).$$

10. Let $Z = X + tY$ $(t \in \mathbb{R})$. Then, since $\mathrm{Var}(tY) = t^2\mathrm{Var}Y$ and $\mathrm{Cov}(X,tY) = t\mathrm{Cov}(X,Y)$ we have by Question 9:

$$\mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(tY) + 2\mathrm{Cov}(X,tY) = t^2\mathrm{Var}(Y) + 2t\mathrm{Cov}(X,Y) + \mathrm{Var}(X) \geq 0$$

which we shall write as $at^2 + bt + c \geq 0$. This quadratic must then have no root or exactly one real root, so that its discriminant must be non-positive, which is to say

$$b^2 - 4ac = 4(\mathrm{Cov}(X,Y))^2 - 4\mathrm{Var}(X)\mathrm{Var}(Y) \leq 0$$
$$\Leftrightarrow \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\mathrm{Var}(Y)} \leq 1$$
$$\Leftrightarrow -1 \leq \rho(X,Y) \leq 1.$$

Problem Set 5

1. We have $H_0 : p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$ for a *multinomial distribution* involving $n = 120 + 48 + 36 + 13 = 217$ cells. Under H_0 the expected frequencies, correct to the nearest integer are given by:

n_i	120	48	36	13
e_i	122	41	41	13

For example, $e_1 = \frac{9}{16} \times 217 = 122 \cdot 06$. Hence

$$\chi^2 = \frac{(120-122)^2}{122} + \frac{(48-41)^2}{41} + \frac{(36-41)^2}{41} + \frac{(13-14)^2}{14} = 1 \cdot 9.$$

Now the 5% critical value of χ^2 on $k-1 = 4-1 = 3$ degrees of freedom is from tables $\chi^2_3 = 7 \cdot 8$; consequently the result is not significant and so H_0 is accepted, which is to say that the outcome is consistent with the stated theory.

2. Here we have $H_0 : p_1 = \cdots = p_5 = \frac{1}{5}$. Our χ^2 statistic will be on $5-1 = 4$ degrees of freedom and its value is given by:

$$\chi^2 = \frac{1}{40}\big((54-40)^2 + (44-40)^2 + (40-40)^2 + (35-40)^2 + (27-40)^2\big) = 10 \cdot 1.$$

This compares to the 5% critical level of $\chi^2_4 = 9 \cdot 5$ and so this is a statistically significant result. We can conclude at the 5% level that this cohort is less fit than expected.

3. The probability of hitting interval A_1 is $p_1 = \int_0^{\frac{1}{4}} 2x\,dx = \frac{1}{16}$. Similarly we find that $p_2 = \frac{3}{16}$, $p_3 = \frac{5}{16}$ and so, by subtraction we get $p_4 = \frac{7}{16}$. Given that $n = 80$ the expected values of our observations in each interval are

$$e_1 = \frac{80}{16} = 5,\ e_2 = 3 \times 5 = 15,\ e_3 = 5 \times 5 = 25,\ e_4 = 7 \times 5 = 35.$$

Our test statistic is χ^2 on $4-1 = 3$ degrees of freedom and takes the value:

$$\chi^2 = \frac{(6-5)^2}{5} + \frac{(18-15)^2}{15} + \frac{(20-25)^2}{25} + \frac{(36-35)^2}{35} = \frac{64}{35} = 1 \cdot 83.$$

However the critical value for χ^2_3 as the $0 \cdot 025$ level in $9 \cdot 35$ and so these observations are consistent with the model.

4. Here we have $n = 70 + 36 + 38 = 144$. Hence we have $k - 1 = 2$. The expected probablitites for each colour are $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{4}{16} = \frac{1}{4}$. The expected values for each colour in this experiment are then $e_1 = \frac{9 \times 144}{16} = 81$, $e_2 = \frac{3 \times 166}{16} = 27$ and $e_3 = \frac{144}{4} = 36$. Hence

$$\chi^2 = \frac{(81-70)^2}{81} + \frac{(27-36)^2}{27} + \frac{(36-38)^2}{36} = \frac{121}{81} + \frac{81}{27} + \frac{4}{36} = 4 \cdot 60.$$

This compares with the 5% critical value of χ^2_2 which is $5 \cdot 99$ and so the results are consistent with the Mendelian genetic theory.

5. First we calculate

$$\hat{q} = \frac{40 + \frac{50}{2}}{40 + 50 + 20} = \frac{65}{110} = \frac{13}{22} = 0 \cdot 5909, \, \hat{p} = \frac{9}{22} = 0 \cdot 4091.$$

The expected proportions are then

$$e_1 = \hat{q}^2 n = 38 \cdot 41, \; e_2 = 2\hat{p}\hat{q}n = 53 \cdot 18, \, e_3 = n\hat{p}^2 = 18 \cdot 41.$$

$$\chi^2 = \frac{(38 \cdot 41 - 40)^2}{38 \cdot 41} + \frac{(53 \cdot 18 - 50)^2}{53 \cdot 18} + \frac{(18 \cdot 41 - 20)^2}{18 \cdot 41} = 0{\cdot}0066 + 0{\cdot}1902 + 0{\cdot}1373 = 0{\cdot}3341.$$

Now χ^2 has $3 - 2 = 1$ degree of freedom. The critical value at $0 \cdot 25\%$ of χ^2_1 is $5 \cdot 412$ so the data is certainly consistent with the Hardy-Weinberg formula.

6(a) Since each variance is based on a sample size of $n = 5$, each contributes a χ^2_4 statistic and the sum of 5 such variables is a χ^2_{20} statistic.

(b) We find $\sum n_i s_i^2 = 9235$ and from the χ^2_{20} table we find that $\chi^2_1 = 9 \cdot 237$ and $\chi^2_2 = 35 \cdot 02$. Hence we obtain for a 96% confidence interval that:

$$\frac{9235}{35 \cdot 01} < \sigma^2 < \frac{9235}{9 \cdot 237} \Leftrightarrow 264 < \sigma^2 < 1000.$$

7.

$$t = \frac{\overline{x}\sqrt{n-1}}{s} = \frac{1 \cdot 24\sqrt{9}}{1 \cdot 45} = 2 \cdot 57, \; \nu = 9;$$

the probability that $T > 2 \cdot 57$ is approximately $0 \cdot 017$ and so the outcome is significant at the 5% level.

8.

$$\overline{X} \pm t_{0{\cdot}05} | \frac{S}{\sqrt{n-1}} = 1 \cdot 24 \pm \frac{1.45}{\sqrt{10}} = 1 \cdot 24 \pm 0 \cdot 46 = (0 \cdot 78, 1 \cdot 69).$$

9. Direct calculation gives $\overline{x} = 6 \cdot 0$ and $\overline{y} = 5 \cdot 7$.

$$n_X s_X^2 = \frac{10}{10-1} \sum_{i=1}^{10} (x_i - \overline{x})^2 = 0 \cdot 64, \; n_Y s_Y^2 = \frac{10}{10-1} \sum_{i=1}^{10} (y_i - \overline{y})^2 = 0 \cdot 24.$$

10. Hence

$$t = \frac{0 \cdot 3}{\sqrt{0 \cdot 64 + 0 \cdot 24}} \sqrt{\frac{100(18)}{20}} = 3 \cdot 03, \; \nu = 10 + 10 - 2 = 18.$$

The critical ·005 value of T is $t = 2 \cdot 878$ using only the right tail because of H_1. Hence the result is significant and the hypothesis of no increase in yield is rejected.

Problem Set 6

1. Writing the line as $x = cy + d$ we interchanging x and y in these so called *normal equations* to obtain $\sum y = nc + d\sum y$ and $\sum xy = c\sum y + d\sum y^2$. Solving the first equation for y on x gives

$$\frac{\sum y}{n} = a + b\frac{\sum x}{n} \Leftrightarrow \overline{y} = a + b\overline{x};$$

by symmetry in the variables we obtain $\overline{x} = c + d\overline{y}$ and so the point $(\overline{x}, \overline{y})$ satisfies both equations and so represents the intersection of the regression lines.

2. We have $n = 7$. Substituting the values in the table the normal equations become:

$$89 = 7a + 38b$$

$$495 = 38a + 270b$$

$$\Rightarrow 3382 = 266a + 1444b$$

$$3465 = 266a + 1890b$$

$$\Rightarrow 83 = 446b \Rightarrow b = \frac{83}{446} = 0 \cdot 18609...$$

Substituting for b then gives

$$a = \frac{89 - 38b}{7} = 11 \cdot 704...$$

Taking a and b to three significant figures gives:

$$y = 11 \cdot 7 + 0 \cdot 186x.$$

3. The normal equations now take on the form:

$$38 = 7c + 89d$$

$$495 = 89c + 1147d$$

$$\Rightarrow 3382 = 623c + 7921d$$

$$3465 = 623c + 8029d$$

$$\Rightarrow d = \frac{83}{108} = 0 \cdot 7685...$$

Substituting for d then gives

$$c = \frac{38 - 89d}{7} = -4 \cdot 342...$$

Taking c and d to three significant figures gives:

$$x = -4 \cdot 34 + 0 \cdot 769y.$$

4. We have $y = a + 0 \cdot 6x$. We also know that $(\overline{x}, \overline{y}) = (10, 4)$ lies on the line so that

$$4 = a + (0 \cdot 6)(10) \Rightarrow a = 4 - 6 = -2.$$

Substituting $x = 12$ then gives

$$y = -2 + 0 \cdot 6(12) = 5 \cdot 2.$$

5.

$$s_{xy} = \frac{1}{n}\sum(x - \overline{x})(y - \overline{y}) = \frac{\sum xy}{n} - \overline{x}\frac{\sum y}{n} - \overline{y}\frac{\sum x}{n} + \frac{n\overline{xy}}{n}$$

$$= \frac{\sum xy}{n} - \overline{xy} - \overline{yx} + \overline{xy} = \frac{\sum xy}{n} - \overline{xy}.$$

6. We eliminate a from the normal equations:

$$\sum x \sum y = na \sum x + b(\sum x)^2$$

$$n \sum xy = na \sum x + nb \sum x^2$$

$$\Rightarrow b(n \sum x^2 - (\sum x)^2) = n \sum xy - \sum x \sum y$$

$$\therefore b = \frac{\frac{\sum xy}{n} - \frac{\sum x}{n} \cdot \frac{\sum y}{n}}{\frac{\sum x^2}{n} - \left(\frac{\sum x}{n}\right)^2} = \frac{s_{xy}}{s_x^2}.$$

Hence the equation of the regression line is $y = a + \frac{s_{xy}}{s_x^2}x$. Since $(\overline{x}, \overline{y})$ is a known point on the line, its equation has the form $\frac{y - \overline{y}}{x - \overline{x}} = b$;

$$\therefore y - \overline{y} = \frac{s_{xy}}{s_x^2}(x - \overline{x}).$$

7. We calculate:

$$\overline{x} = \frac{\sum x}{n} = \frac{509}{12} = 42 \cdot 417; \overline{y} = \frac{\sum y}{n} = \frac{733}{12} = 61 \cdot 083...$$

$$s_{xy} = \frac{\sum xy}{n} - \overline{xy} = \frac{37249}{12} - \left(\frac{509}{12}\right)\left(\frac{733}{12}\right) = 513 \cdot 13...$$

$$s_x^2 = \frac{\sum x^2}{n} - \overline{x}^2 = \frac{27963}{12} - \left(\frac{509}{12}\right)^2 = 531 \cdot 07...$$

$$b = \frac{s_{xy}}{s_x^2} = 0 \cdot 966...$$

$$\Rightarrow y - 61 \cdot 083 = 0 \cdot 966(x - 42 \cdot 417)$$

$$\therefore y = 20 \cdot 1 + 0 \cdot 966x.$$

8(a) Putting $x = 82$ in our equation returns

$$y = 20 \cdot 1 + (0 \cdot 966)(82) = 99 \cdot 3;$$

so, rounding down, 99 customers are expected.

(b) Putting $y = 120$ into our equation returns the value

$$x = \frac{120 - 20 \cdot 1}{0 \cdot 966} = 103 \cdot 4;$$

rounding down we obtain that we need103 advanced ticket sales before the model predicts a full house.

9 & 10.

$$Q(a,b) = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - a - b(x_i - \overline{x}))^2$$

$$\Rightarrow \frac{\partial Q}{\partial a} = -2\sum (y_i - a - bx_i + b\overline{x}) = 0$$

$$\Rightarrow na = n\overline{y} - bn\overline{x} + bn\overline{x}$$

$$\therefore a = \overline{y}.$$

Substituting in Q gives:

$$Q(b) = \sum (y_i - \overline{y} - b(x_i - \overline{x}))^2$$

$$\Rightarrow \frac{dQ}{db} = -2\sum (y_i - \overline{y} - b(x_i - \overline{x}))(x_i - \overline{x}) = 0$$

$$\Rightarrow \sum (x_i - \overline{x})(y_i - \overline{y}) = b\sum (x_i - \overline{x})^2$$

$$\therefore b = \frac{s_{xy}}{s_x^2}.$$

Problem Set 7

1.

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2}\sqrt{\sum (y_i - \overline{y})^2}} = \frac{\sum a_i b_i}{\sqrt{\sum a_i^2}\sqrt{\sum b_i^2}}.$$

2. We have

$$\sum (z^2 a_i^2 + 2z a_i b_i + b_i^2) = (\sum a_i^2)z^2 + (2\sum a_i b_i)z + \sum b_i^2 > 0.$$

It follows that this quadratic in z has a negative determinant:

$$4(\sum a_i b_i)^2 - 4(\sum a_i^2)(\sum b_i^2) < 0$$

$$\Rightarrow (\sum a_i b_i)^2 < (\sum a_i^2)(\sum b_i^2) \Rightarrow 0 < \frac{(\sum a_i b_i)^2}{(\sum a_i^2)(\sum b_i^2)} < 1$$

and taking square roots gives:

$$\therefore |r| < 1.$$

3. On the other hand if $\sum(za_i + b_i)^2 = 0$ then $za_i + b_i = 0$ for all i and in particular there is a unique value of z satisfying the quadratic, the discriminant of which must be 0. Hence we infer that $r^2 = 1$ so that $r = \pm 1$.

4. We find $\sum x = 528$, $\sum y = 666$, $\sum x^2 = 43,464$ $\sum y^2 = 46,820$, $\sum xy = 38,640$. With $n = 10$ we find $\overline{x} = 52 \cdot 8$, $\overline{y} = 66 \cdot 6$,

$$s_{xy} = \frac{\sum xy}{n} - \overline{xy} = \frac{38,640}{10} - (52 \cdot 8)(66 \cdot 6) = 347 \cdot 52$$

$$s_x^2 = \frac{\sum x^2}{n} - \overline{x}^2 = \frac{34,464}{10} - (52{\cdot}8)^2 = 658{\cdot}56,\ s_y^2 = \frac{\sum y^2}{n} - \overline{y}^2 = \frac{46,820}{10} - 66{\cdot}6^2 = 246{\cdot}44;$$

$$r = \frac{s_{xy}}{s_x s_y} = \frac{(347 \cdot 52)}{\sqrt{658 \cdot 56}\sqrt{246 \cdot 44}} = 0 \cdot 8626...$$

Therefore $r = 0{\cdot}86$ (2 d.p.) indicating a very strong positive correlation between the two marks.

5. To calculate r we also need $s_y^2 = 5980 \cdot 91$ so that $s_y = 77 \cdot 336$. Also $s_x = \sqrt{531 \cdot 07} = 23 \cdot 045$. Hence

$$r = \frac{s_{xy}}{s_x s_y} = \frac{513 \cdot 13}{(23 \cdot 045)(77 \cdot 336)} = 0 \cdot 2879.$$

6. (a) We have $b = \frac{s_{xy}}{s_x^2}$, $d = \frac{s_{xy}}{s_y^2}$ so that

$$bd = \frac{s_{xy}^2}{s_x^2 s_y^2} = r^2.$$

(b) Since s_x^2 and s_y^2 are both positive, it follows that b and d have the same sign as s_{xy}.

7. From Question 2 Set 6 we have $y = 11 \cdot 7 + 0 \cdot 186x$ $x = -4 \cdot 34 + 0 \cdot 769y$. Now since both b and d are positive, then $r \geq 0$. Hence

$$r = +\sqrt{(0 \cdot 186)(0 \cdot 769)} = 0 \cdot 38 \text{ (2 d.p)}$$

indicating weak positive correlation between x and y.

8. If $r = 0$ then $s_{xy} = 0$ whence we also have $b = d = 0$ (and conversely). Hence the regression lines have the respective forms $y = a$ and $x = c$, which are constants and the corresponding lines, being parallel to the x- and y-axes respectively, are orthogonal.

9.

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (0 \cdot 38)\sqrt{\frac{12-2}{1-(0 \cdot 38)^2}} = 1 \cdot 4045.$$

This compares with the 95% two-tailed t-statistic on $\nu = 10$ degrees of freedom, which is (from tables) $2 \cdot 228 > 1 \cdot 4045$. We conclude that the value of the regression statistic is too low to reject the null hypothesis of no correlation at the 95% confidence level.

10. From the same table value we obtain a 95% confidence interval, to 2 d.p, for ρ of

$$r \pm 2 \cdot 228 = (0 \cdot 38 - 2 \cdot 228, 0 \cdot 38 + 2 \cdot 228) = (-1 \cdot 85, 2 \cdot 66).$$

We note that the confidence interval straddles 0, which is consistent with Question 9, which did not reject the hypothesis of zero correlation between advanced sales and door sales.

Comments Since $1- \leq r \leq 1$ we may write uniquely as $r = \cos\phi$ $(0 \leq \phi \leq \pi)$, which is a function of the angle between the regression lines of y on x and x on y. More details can be found by searching on the notion of *cosine similarity*.

Data sets may of course have an underlying relationship that is not linear. Power laws is a type that often arises: $y = ax^b$ $(a > 0)$. By taking logs this equation may be recast as $\log y = \log a + b \log x$ and so such a power law is equivalent to a linear relationship between the logs of the variables x and y. That relationship may then be tested and measured using the techniques above.

Problem Set 8

1. The ith row of P is $\sum_{j=1}^{n} p_{ij} = 1$ as the sum represent the probability that the process will be in one of the allowable states E_j at the next stage. Therefore the $\{p_{ij}\}_{1 \leq j \leq n}$ is a probability distribution so that all p_{ij} and non-negative and sum to 1 over j.

2. For $n = 0$ the claim is that $x^{(0)} = x^{(0)}P^0$, which is trivially true. Suppose that for some $n \geq 0$ we have $x^{(n)} = x^{(0)}P^n$. We then have

$$x^{(n+1)} = x^{(n)}P = (x^{(0)}P^n)P = x^{(0)}(P^nP) = x^{(0)}P^{n+1}$$

and so the induction continues (because of associativity of matrix multiplication).

3. By continuity x is fixed point of the transformation defined by right multiplication by P, which is to say that $xP = P$ so that x is an eigenvector of P with eigenvalue 1.

4. (a) With the order 1 = dry, 2 = wet we have as transition matrix:

$$P = \begin{bmatrix} 0 \cdot 8 & 0 \cdot 2 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix}.$$

(b) Since $n = 4$ we calculate:

$$P^2 = \begin{bmatrix} 0 \cdot 8 & 0 \cdot 2 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix} \begin{bmatrix} 0 \cdot 8 & 0 \cdot 2 \\ 0 \cdot 4 & 0 \cdot 6 \end{bmatrix} = \begin{bmatrix} 0 \cdot 72 & 0 \cdot 28 \\ 0 \cdot 56 & 0 \cdot 44 \end{bmatrix}$$

$$P^4 = (P^2)^2 = \begin{bmatrix} 0\cdot 72 & 0\cdot 28 \\ 0\cdot 56 & 0\cdot 44 \end{bmatrix}\begin{bmatrix} 0\cdot 72 & 0\cdot 28 \\ 0\cdot 56 & 0\cdot 44 \end{bmatrix} = \begin{bmatrix} 0\cdot 6752 & 0\cdot 3248 \\ 0\cdot 6496 & 0\cdot 3504 \end{bmatrix}.$$

Comment: Note that P^n is also a transition matrix so, once again, the row probabilities sum to 1.

We are given that Wednesday is dry so that $x^{(0)} = (1\ 0)$ and so we want $(1\ 0)P^4 = 0\cdot 6752$.

(c) Since Wednesday is a wet day we now have $x^{(0)} = (0\ 1)$ and so our answer is $(0\ 1)P^4 = 0.6496$.

(d) We solve $x(P-I) = 0$ so that, writing $x = [a\ 1-a]$ we have

$$\begin{bmatrix} a & 1-a \end{bmatrix}\begin{bmatrix} -0\cdot 2 & 0\cdot 2 \\ 0\cdot 4 & -0\cdot 6 \end{bmatrix} = [0\,0]$$

$$\Rightarrow -0\cdot 2a + 0\cdot 4 - 0\cdot 4a = 0$$

$$\Rightarrow a = \frac{4}{6} = \frac{2}{3},$$

so, in the long term, the proportion of dry days is 2 out of 3.

5(a) If a customer is in state Pow, then their state probabilites are $(p_{pp}, p_{pz}) = (0\cdot 75, 0\cdot 25)$; similarly the state probabilities for a Zap customer is $(p_{zp}, p_{zz}) = (0\cdot 1, \cdot 0\cdot 9)$. Hence our transition matrix P is given by:

$$P = \begin{bmatrix} 0\cdot 75 & 0\cdot 25 \\ 0\cdot 10 & 0\cdot 90 \end{bmatrix} \Rightarrow P - I = \begin{bmatrix} -0\cdot 25 & 0\cdot 25 \\ 0\cdot 10 & -0\cdot 10 \end{bmatrix}.$$

Hence the steady state vector $x = [a\,1-a]$, known as the *stationary distribution*, where a is the long term probability of a customer is with Pow customer satisfies:

$$-0\cdot 25a + 0\cdot 10 - 0\cdot 10a = 0 \Rightarrow a = \frac{0\cdot 10}{0\cdot 35} = \frac{2}{7}.$$

Hence the steady state proportion of Zap customers is $1 - \frac{2}{7} = \frac{5}{7}$.

(b) Again we search for a normalised eigenvector with eigenvalue 1:

$$P - I = \begin{bmatrix} -\alpha & \alpha \\ \beta & -\beta \end{bmatrix};$$

hence for the stationary distribution $[a, 1-a]$ we have

$$-a\alpha + \beta(1-a) = 0$$

$$\Rightarrow -a(\alpha+\beta) = -\beta$$

$$\Rightarrow a = \frac{\beta}{\alpha+\beta}.$$

$$1 - a = 1 - \frac{\beta}{\alpha+\beta} = \frac{\alpha}{\alpha+\beta}.$$

Therefore the stationary distribution vector is $(\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$.

6. We again find the probability eigenvector with eigenvalue 1:

$$P - I = \begin{bmatrix} -\frac{2}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{2} & -\frac{3}{4} & \frac{1}{4} \\ \frac{1}{5} & \frac{3}{5} & -\frac{4}{5} \end{bmatrix}.$$

Writing $x = (a\, b\, c)$ we have $a + b + c = 1$. Using row operations on $P - I$ we have:

$$\begin{bmatrix} -2 & 1 & 1 \\ 2 & -3 & 1 \\ 1 & 3 & -4 \end{bmatrix} \to \begin{bmatrix} -2 & 1 & 1 \\ 0 & -2 & 2 \\ 0 & \frac{7}{2} & -\frac{7}{2} \end{bmatrix} \to \begin{bmatrix} -2 & 1 & 1 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix} \to \begin{bmatrix} -2 & 0 & 2 \\ 0 & 1 & -1 \\ 0 & 1 & -1 \end{bmatrix}$$

$$\to \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix};$$

which gives $a = b = c$ so the steady state vector is $x = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$.

7. (a) We may pass from E_i to E_j in the first transition, with probability p_{ij} or we may first pass to state E_k with $k \neq j$, with probability p_{ik}, from which point the mean number of steps to E_j is m_{kj}. This gives the equation:

$$m_{ij} = p_{ij} + \sum_{k=1}^{n} p_{ik}(1 + m_{kj}) = \sum_{k=1}^{n} p_{ik} + \sum_{k \neq j} p_{ik} m_{kj}$$

$$= 1 + \sum_{k \neq j} p_{ik} m_{kj}.$$

(b) If we take $n = 2$ the previous equation becomes for $i \neq j$:

$$m_{ij} = 1 + p_{ii} m_{ij} \Rightarrow m_{ij}(1 - p_{ii}) = 1$$

$$\therefore m_{ij} = \frac{1}{1 - p_{ii}}.$$

(c) In Question 5(b) we have $p_{11} = 1 - \alpha, \pi_1 = \frac{\beta}{\alpha+\beta}$, $p_{22} = 1 - \beta$, and $p_{22} = \frac{\alpha}{\alpha+\beta}$ and so by part (b):

$$m_{12} = \frac{1}{1 - (1 - \alpha)} = \frac{1}{\alpha}; \text{ similarly } m_{21} = \frac{1}{\beta}.$$

Hence we obtain

$$M = \begin{bmatrix} 1 + \frac{\alpha}{\beta} & \frac{1}{\alpha} \\ \frac{1}{\beta} & 1 + \frac{\beta}{\alpha} \end{bmatrix}.$$

(d) For the Utility companies problem we have $\alpha = 0 \cdot 25$, $\beta = 0 \cdot 1$, $\pi_1 = \frac{2}{7}$ and $\pi_2 = \frac{5}{7}$ so we obtain

$$M = \begin{bmatrix} 3 \cdot 5 & 4 \\ 10 & 1 \cdot 4 \end{bmatrix}.$$

This means, for instance, Zap customers will switch to Pow, on average, once every ten years.

8. Since $f_{jj}^{(0)} = p_{jj}^{(0)} = 1$, the base $n = 0$ case holds. Now suppose that $n \geq 1$. Let $q_{j,n,m}$ denote the probability of a return to E_j in n steps having first returned to E_j in m steps ($1 \leq m \leq n-1$). We see that

$$q_{j,n,k} = f_{jj}^{(m)} p_{jj}^{(n-m)}.$$

Now the event of return to E_j in n steps is a disjoint union of the events of the type return to E_j in n steps having first returned to E_j in m steps ($1 \leq m \leq n$), which is to say

$$p_{jj}^{(n)} = \sum_{m=1}^{n} f_{jj}^{(m)} p_{jj}^{(n-m)}$$

from which (1) now follows (as $p_{jj}^{(0)} = 1$).

9.

$$P^2 = \begin{bmatrix} 0 \cdot 72 & 0 \cdot 28 \\ 0 \cdot 56 & 0 \cdot 44 \end{bmatrix};$$

hence $p_{11}^{(2)} = 0 \cdot 72$, $p_{22}^{(2)} = 0 \cdot 44$, $f_{11}^{(1)} = 0 \cdot 8 = p_{11}^{(1)}$, $f_{22}^{(1)} = 0 \cdot 6 = p_{22}^{(1)}$. Hence

$$f_{11}^{(2)} = p_{11}^{(2)} - f_{11}^{(1)} p_{11}^{(1)} = 0 \cdot 72 - (0 \cdot 8)(0 \cdot 8) = 0 \cdot 08;$$

$$f_{22}^{(2)} = p_{22}^{(2)} - f_{22}^{(1)} p_{22}^{(1)} = 0 \cdot 44 - (0 \cdot 6)(0 \cdot 6) = 0 \cdot 08.$$

10. Consider the Markov chain M with transition matrix:

$$P = \begin{bmatrix} 0 \cdot 5 & 0 \cdot 5 & 0 \\ 0 & 0 \cdot 3 & 0 \cdot 7 \\ 0 & 0 \cdot 6 & 0 \cdot 4 \end{bmatrix}$$

(a) Since $p_{21} = p_{31} = 0$, the state E_1 will be inacessible once it has been left. It follows that $f_{11}^{(1)} = p_{11}^{(1)} = 0 \cdot 5$ but that $f_{11}^{(n)} = 0$ for all $n \geq 2$. Hence $F_{11} = f_{11}^{(1)} = \frac{1}{2}$.

Comment We call a state such as E_1 a *transient state* because $F_{11} < 1$, meaning that, once left, the process is not certain to ever return to that state.

(b) If the process begins at E_1, the probability that we are in in E_1 after n steps is $\frac{1}{2^n} \to 0$ as $n \to \infty$. Moreover, if we begin at E_2 or E_3 the probability of ever entering E_1 is 0. It follows that $\pi_1 = 0$ as the process must eventually leave E_1, never to return. Hence we can apply the result of Question 5(b) to conclude that $(\pi_2, \pi_3) = (\frac{\beta}{\alpha+\beta}, \frac{\alpha}{\alpha+\beta})$, where, in this case, $\alpha = 0 \cdot 7$ and $\beta = 0 \cdot 6$.

$$\therefore \pi = (0, \frac{0 \cdot 6}{0 \cdot 7 + 0 \cdot 6}, \frac{0 \cdot 7}{0 \cdot 6 + 0 \cdot 6}) = (0, \frac{6}{13}, \frac{7}{13}).$$

Problem Set 9

1.

$$\sum_{n=0}^{\infty} |p_n z^n| = \sum_{n=0}^{\infty} p_n |z|^n \leq \sum_{n=0}^{\infty} p_n = 1$$

and so $P(z)$ certainly coverges if $|z| \leq 1$.

2.

$$P(0) = p_0 + \sum_{n=1}^{\infty} (p_n) 0^n = p_0; \; P(1) = \sum_{n=0}^{\infty} p_n 1^n = \sum_{n=0}^{\infty} p_n = 1.$$

3.

$$P^{(k)}(z) = \sum_{n=0}^{\infty} n(n-1)\cdots(n-k+1) p_n z^{n-k} = \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1) p_n z^{n-k}$$

$$= \sum_{n=0}^{\infty} (n+k)(n+k-1)\cdots(n+1) p_{n+k} z^n$$

$$\Rightarrow P^{(k)}(0) = k! p_k \Rightarrow p_k = \frac{P^{(k)}(0)}{k!}.$$

4.

$$P'(z) = \sum_{n=0}^{\infty} n p_n z^{n-1} \Rightarrow z P'(z) = \sum_{n=0}^{\infty} n p_n z^n = \mathbb{E}(n)$$

$$\Rightarrow \mathbb{E}(n) = P'(1).$$

$$P^{(2)}(z) = \sum_{n=0}^{\infty} n(n-1) p_n z^{n-2} \Rightarrow z^2 P^{(2)}(z) = \sum_{n=0}^{\infty} n(n-1) p_n z^n$$

$$\Rightarrow \mathbb{E}(n(n-1)) = P^{(2)}(1).$$

$$\mathbb{V}ar(n) = \mathbb{E}(n(n-1)) - \mathbb{E}^2(n) + \mathbb{E}(n)$$

$$= P^{(2)}(1) + P'(1) - (P'(1))^2.$$

5(a)

$$Z(p_{n-1}) = \sum_{n=0}^{\infty} p_{n-1} z^n = z \sum_{n=1}^{\infty} p_{n-1} z^{n-1} = z \sum_{n=0}^{\infty} p_n z^n$$

$$\Rightarrow Z(p_{n-1}) = z P(z).$$

(b)

$$Z(p_{n+1}) = \sum_{n=0}^{\infty} p_{n+1} z^n = \sum_{n=1}^{\infty} p_n z^{n-1}$$

$$\Rightarrow zZ(p_{n+1}) = \sum_{n=1}^{\infty} p_n z^n = (Z(p_n) - p_0)$$

$$\Rightarrow Z(p_{n+1}) = \frac{1}{z}(Z(p_n) - p_0).$$

(c)

$$Z(ap_n + bq_n) = \sum_{n=0}^{\infty}(ap_n + bq_n)z^n = a\sum_{n=0}^{\infty} p_n z^n + b\sum_{n=0}^{\infty} q_n z^n$$

$$= aZ(p_n) + bP(q_n).$$

6.

$$Z(y_n) = \sum_{n=0}^{\infty}(\sum_{k=0}^{n} q_k p_{n-k} z^n) = (\sum_{k=0}^{\infty} q_k z^k)(\sum_{n=k}^{\infty} p_{n-k} z^{n-k})$$

$$= (\sum_{k=0}^{\infty} q_k z^k)(\sum_{n=0}^{\infty} p_n z^n) = Z(q_n)Z(p_n).$$

7(a)

$$Z(p_n) = \sum_{n=0}^{\infty} e^{-\lambda}\frac{\lambda^n}{n!} z^n = e^{-\lambda}\sum_{n=0}^{\infty}\frac{(\lambda z)^n}{n!} = e^{-\lambda}e^{\lambda z}$$

$$\Rightarrow Z(p_n) = e^{\lambda(z-1)}.$$

(b) We have $Z'(p_n) = \lambda e^{\lambda(z-1)}$. By Question 4 we obtain:

$$\mathbb{E}(X) = P'(1) = \lambda e^{\lambda(1-1)} = \lambda e^0 = \lambda.$$

$P^{(2)}(z) = \lambda^2 e^{\lambda(z-1)}$. Again by Question 4:

$$\mathbb{V}ar(X) = P^{(2)}(1) + P'(1) - (P'(1))^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

8. Put $y_n = q_n * p_n$ and so by Questions 6 and 7 we obtain:

$$Z(y_n) = Z(p_n)Z(q_n) = e^{\lambda_1(z-1)}e^{\lambda_2(z-1)} = e^{-(\lambda_1+\lambda_2)(z-1)};$$

which we recognize as the z-transform of the Poisson random variable $X = Po(\lambda_1 + \lambda_2)$.

9. With $q = 1 - p$ we have:

$$P(z) = \sum_{k=0}^{n}\binom{n}{k}p^k q^{n-k} z^k = \sum_{k=0}^{n}\binom{n}{k}(pz)^k q^{n-k}$$

$$= (q + pz)^n.$$

Hence we obtain $P'(z) = np(q + pz)^{n-1}$ so that

$$\mathbb{E}(X) = P'(1) = np(q + p)^{n-1} = np.$$

Also $P^{(2)}(z) = n(n-1)p^2(q+pz)^{n-1}$ so that

$$\mathbb{V}ar(X) = P^{(2)}(1) + P'(1) - (P'(1))^2 = n(n-1)p^2(q+p)^{n-1} + np - (np)^2$$

$$= n(n-1)p^2 + np - n^2p^2 = np - np^2 = np(1-p) = npq.$$

10. We have, since $x_0 = 0$,

$$Z(x_n) = \sum_{n=0}^{\infty} x_n z^n = x_1 z + \sum_{n=2}^{\infty} (ax_{n-1} + by_n) z^n$$

$$= y_1 z + az \sum_{n=1}^{\infty} x_n z^n + b \sum_{n=2}^{\infty} y_n z^n$$

$$= y_1 z + azZ(x_n) + b(Z(y_n) - y_1 z)$$

$$\Rightarrow Z(x_n)(1-az) = bZ(y_n) + (1-b)y_1 z$$

$$\Rightarrow Z(x_n) = \frac{b}{1-az} \cdot Z(y_n) + (1-b) \cdot \frac{y_1 z}{1-az}$$

Now

$$\frac{b}{1-az} = b \sum_{n=0}^{\infty} a^n z^n = bZ(a^n);$$

$$\frac{(1-b)y_1 z}{1-az} = (1-b)y_1 \sum_{n=0}^{\infty} a^n z^{n+1} = (1-b)y_1 (\sum_{n=1}^{\infty} a^{n-1} z^n)$$

$$= (1-b)y_1 (Z(a^{n-1}) + a^{-1}) = \frac{1-b}{a} y_1 + (1-b)y_1 Z(a^{n-1}).$$

Hence

$$Z(x_n) = Z(ba^n) * Z(y_n) + (1-b)y_1 Z(a^{n-1})$$

$$\therefore x_n = b(y_1 a^{n-1} + y_2 a^{n-2} + \cdots + y_n) + (1-b)y_1 a^{n-1}.$$

Comment We have introduced the z-transform as a similar creature to the Generating function of Set 1. The use of complex variable z has not been exploited here but that feature makes the transform useful in other fields, particularly signal theory. For example $p_n = \frac{1}{2\pi n} \oint \frac{P(z)}{z^{n+1}}\, dz$ for any closed contour containing the origin but none of the poles of $P(z)$. Methods such as partial fractions and the use of the inverse transform can be used to find distributions.

Problem Set 10

1.

$$(1+x)^{n_1+n_2} = \sum_{k=0}^{n_1+n_2} \binom{n_1+n_2}{k} x^k = (1+x)^{n_1}(1+x)^{n_2}$$

$$= (\sum_{k=0}^{n_1} \binom{n_1}{k} x^k)(\sum_{l=0}^{n_2} \binom{n_2}{l} x^l).$$

Now the coefficient of x^m from the first summation is $\binom{n_1+n_2}{m}$, while expanding the product of the latter two sums gives that same coefficient in the form:

$$\sum_{k+l=m} \binom{n_1}{k}\binom{n_2}{l} = \sum_{k=0}^{m} \binom{n_1}{k}\binom{n_2}{m-k}$$

$$\therefore \sum_{k=0}^{m} \binom{n_1}{k}\binom{n_2}{m-k} = \binom{n_1+n_2}{m}.$$

Comment The identity is named after Alexandre-Théophile Vandermonde (1772), although it was already known in 1303 by the Chinese mathematician Zhu Shijie (Chu Shi-Chieh).

2(a)

$$P(Y=m) = \sum_{k=0}^{m} \binom{n_1}{k} p^k q^{n_1-k} \cdot \binom{n_2}{m-k} p^{m-k} q^{n_2-m+k}$$

$$= \sum_{k=0}^{m} \binom{n_1}{k}\binom{n_2}{m-k} p^m q^{n_1+n_2-m} = p^m q^{n_1+n_2-m} \sum_{k=0}^{m} \binom{n_1}{k}\binom{n_2}{m-k} = \binom{n_1+n_2}{m} p^m q^{n_1+n_2-m}$$

with the last equality coming from Vandermonde's identity. Therefore $Y \sim B(n_1+n_2, p)$.

(b) Let $Y = (X_1 + \cdots + X_{m-1}) + X_m$. By induction we have the random variable X in brackets had distribution $B(n_1 + \cdots + n_{m-1}, p)$. Then $Y = X + X_m \sim B(n_1 + \cdots + n_m, p)$ by the $m = 2$ case of part (a).

3(a)

$$P(Y=n) = \sum_{k=0}^{n} P(X_1 = k) P(X_2 = n-k)$$

$$= \sum_{k=0}^{n} \frac{e^{-\lambda_1}\lambda_1^k}{k!} \cdot \frac{e^{-\lambda_2}\lambda_2^{n-k}}{(n-k)!} = e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{\lambda_1^k \lambda_2^{n-k}}{k!(n-k)!}$$

$$= \frac{e^{-(\lambda_1+\lambda_2)}}{n!} \sum_{k=0}^{n} \binom{n}{k} \lambda_1^k \lambda_2^{n-k} = \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1+\lambda_2)^n}{n!}.$$

$$\therefore Y \sim \text{Po}(\lambda_1 + \lambda_2).$$

(b) Again by induction this reduces to the $m = 2$ case above so $X_1 + \cdots + X_m \sim \text{Po}(\lambda_1 + \cdots + \lambda_m)$.

4. Since $X_1, X_2 \geq 1$, the convolution sum runs between $k = 1$ and $k = n-1$ as the contributions from $k = 0$ and $k = n$ are then both zero. Hence with $q = 1 - p$ being the failure probability we obtain:

$$P(Y = n) = \sum_{k=0}^{n} P(X_1 = k)P(X_2 = n - k) = \sum_{k=1}^{n-1} pq^{k-1} \cdot pq^{n-k-1}$$

$$= \sum_{k=1}^{n-1} p^2 q^{n-2} = p^2 q^{n-2} \sum_{k=1}^{n-1} 1 = (n-1)p^2(1-p)^{n-2},\ n = 2, 3, \cdots .$$

5. This time the sum takes the form:

$$P(Y = n) = \sum_{k=1}^{n-1} p_1 q_1^{k-1} p_2 q_2^{n-k-1} = p_1 p_2 q_1^{-1} q_2^{n-1} \sum_{k=1}^{n-1} \frac{q_1^k}{q_2^k}$$

$$= \frac{p_1 p_2 q_2^{n-1}}{q_1} \Big(\frac{1 - (\frac{q_1}{q_2})^n}{1 - \frac{q_1}{q_2}} - 1 \Big) = \frac{p_1 p_2}{q_1 q_2} \Big(\frac{q_2^n - q_1^n}{q_2 - q_1} \cdot q_2 - q_2^n \Big)$$

$q_2 - q_1 = (1 - p_2) - (1 - p_1) = p_1 - p_2$. Hence

$$\frac{q_2^{n+1} - q_1^n q_2 - q_2^{n+1} + q_1 q_2^n}{q_2 - q_1} = \frac{q_1 q_2 (q_2^{n-1} - q_1^{n-1})}{p_1 - p_2}$$

$$\therefore P(Y = n) = \frac{p_1 p_2}{p_1 - p_2} \cdot \Big((1 - p_2)^{n-1} - (1 - p_1)^{n-1} \Big),\ n = 2, 3, \cdots .$$

Comment Note that if $p_1 = p_2$ then the formula used for the sum of the finite geometric series no longer applies, the sum just becoming $\sum_{k=1}^{n-1} 1 = n - 1$, as in part (a).

6. In the definition of $f(x) * g(x)$ we substitute $u = x - t$ so that $t = x - u$ and $dt = -du$. Also $t \to \pm\infty$ as $u \to \mp\infty$. We then have

$$f(x) * g(x) = \int_{-\infty}^{\infty} f(t)g(x - t)\, dt = -\int_{\infty}^{-\infty} f(x - u)g(u)\, du$$

$$= \int_{-\infty}^{\infty} g(u)f(x - u)\, du = g(x) * f(x).$$

7. For a given value of x, when is the integrand of $f(x) * g(x)$ non-zero? We require $0 \leq t \leq 1$ and $0 \leq x - t \leq 1$ simultaneously. Now

$$0 \leq x - t \leq 1 \Leftrightarrow -1 \leq t - x \leq 0 \Leftrightarrow x - 1 \leq t \leq x.$$

Then $x - 1 \leq t \leq 1$ implies $x \leq 2$, while $0 \leq t \leq x$, so that $0 \leq x \leq 2$. Now $x - 1 \leq t$ is true for all $0 \leq t \leq 1$ if and only if $x - 1 \leq 0 \Leftrightarrow x \leq 1$. On the other hand $t \leq x$ is true for all $0 \leq t \leq 1$ if and only if $1 \leq x$. Hence we examine the two case $x \leq 1$ and $1 \leq x$ separately.

Suppose that $x \leq 1$. Then for $0 \leq t \leq 1$, $g(x-t) = 1$ if and only if $t \leq x$ so that in this case:

$$f(x) * g(x) = \int_0^x dt = x.$$

On the other hand, if $1 \leq x$ then for $0 \leq t \leq 1$, $g(x-t) = 1$ if and only if $x - 1 \leq t$ and we obtain:

$$f(x) * g(x) = \int_{x-1}^1 dt = 1 - (x-1) = 2 - x.$$

Therefore $f_Z(x) = x$ if $0 \leq x \leq 1$ and $f_Z(x) = 2 - x$ if $1 \leq x \leq 2$ (the graph of which is a right-angled triangle).

8. In this case we have both terms in the product in the integrand are non-zero if and only if $t \geq 0$ and $x - t \geq 0$, which is to say $t \leq x$. Hence we require $0 \leq t \leq x$, so that $x \geq 0$ and we obtain:

$$f(x) * g(x) = \int_0^x \lambda e^{-\lambda t} \lambda e^{-\lambda(x-t)}\, dt = \lambda^2 e^{-\lambda x} \int_o^x dt = \lambda^2 x e^{-\lambda x},\ x \geq 0.$$

9. Here $f(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ $-\infty < x < \infty$. Hence the convolution is:

$$f(x) * g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} e^{-\frac{(x-t)^2}{2}}\, dt$$

Now

$$t^2 + (x-t)^2 = 2t^2 - 2tx + x^2 = 2(t^2 - tx + \frac{x^2}{2}) = 2(t - \frac{x}{2})^2 + \frac{x^2}{4}$$

$$\Rightarrow f(x) * g(x) = \frac{e^{-\frac{x^2}{4}}}{2\pi} \int_{-\infty}^{\infty} e^{-(t-\frac{x}{2})^2}\, dt$$

For our integral, put $u = \sqrt{2}(t - \frac{x}{2})$ so that $(t - \frac{x}{2})^2 = \frac{u^2}{2}$ and $dt = \frac{du}{\sqrt{2}}$, so our integral becomes:

$$\frac{e^{-\frac{x^2}{4}}}{2\sqrt{2}\pi} \int_{-\infty}^{\infty} e^{-\frac{u^2}{2}}\, du = \frac{e^{-\frac{x^2}{4}}}{2\sqrt{2}\pi} \cdot \sqrt{2\pi} = \frac{e^{-\frac{x^2}{4}}}{\sqrt{4\pi}}.$$

Therefore $Z = X + Y \sim N(0, 2)$ as $2\sigma_Z^2 = 4$ so $\sigma_Z^2 = 2$.

10(a) The same considerations apply as in Question 7 as each density is only non-zero in the unit interval. Suppose that $x \leq 1$. Then for $0 \leq t \leq 1$, $g(x-t) = 2 - 2(x-t) = 2(1 + t - x)$ if and only if $t \leq x$ so that in this case:

$$f_Z(x) = f(x) * g(x) = 4 \int_0^x t(1 + t - x)\, dt = 4 \int_0^x t(1-x) + t^2)\, dt$$

$$= 4[\frac{t^2}{2}(1-x) + \frac{t^3}{3}]_0^{t=x} = 4(\frac{x^2}{2}(1-x) + \frac{x^3}{3}) = 2x^2(1 - x + \frac{2x}{3}) = 2x^2(1 - \frac{x}{3})\ (0 \leq x \leq 1).$$

By symmetry, the pdf $f_Z(x)$ of Z is symmetric in the line $x = 1$, which is to say that

$$f_Z(1+x) = f_Z(1-x) \ (0 \le x \le 1)$$

so that for $1 \le x \le 2$ we have

$$f_Z(x) = f_Z(1+(x-1)) = f_Z(1-(x-1)) = f_Z(2-x)$$

$$= 2(2-x)^2(1-\frac{2-x}{3}) = 2(2-x)^2(\frac{1+x}{3}) \ (1 \le x \le 2).$$

(b) Clearly $h(x) \ge 0$ for all $0 \le x \le 2$. We verify that $\int_0^1 f_Z(x)\,dx = \frac{1}{2}$. Then by symmetry the same is true of $\int_1^2 f_Z(x)\,dx$ so that $\int_{\mathbb{R}} f(x) * g(x)\,dx = 1$, as required.

$$\int_0^1 2x^2(1-\frac{x}{3})\,dx = [\frac{2x^3}{3} - \frac{2x^4}{12}]_0^1 = 2(\frac{1}{3} - \frac{1}{12}) = 2(\frac{3}{12}) = \frac{1}{2}.$$