Probability and Statistics 2014 (Answers)

Question 1

(a) If A and B are mutually exclusive then $\Pr(A \cap B) = 0$. Here $\Pr(A \cap B) \neq 0$, and so A and B are not mutually exclusive.

(b) In general, $\Pr(A) + \Pr(B) = \Pr(A \cup B) + \Pr(A \cap B)$. Thus

$$\Pr(B) = \Pr(A \cup B) + \Pr(A \cap B) - \Pr(A) = 0.8 + 0.3 - 0.6 = 0.5$$

(c) A and B are statistically independent if $\Pr(A) \times \Pr(B) = \Pr(A \cap B)$ or if $Pr(A) = Pr(A|B)$ or $Pr(B) = Pr(B|A)$.

Since $\Pr(A) \times \Pr(B) = 0.3 = \Pr(A \cap B)$, it follows that A and B are statistically independent.

Question 2

(a) The binomial $B(n, p)$ distribution, with $n = 8$, $p = 0.4$. (May be expressed as $X \sim B(8, 0.4)$.)

(b) $\Pr(X = 3) = F(3) - F(2) = 0.5941 - 0.3154 = 0.279$ to 3 decimal places.

Alternatively: $\Pr(X = 3) = \binom{8}{3} \times 0.4^3 \times (1 - 0.5)^5 = 0.279$ to 3 decimal places.

Question 3

(a) A Poisson distribution with parameter/mean $\mu = 15 \div \frac{72}{60} = 12.5$
$X \sim Poisson(12.5)$

(b) Using Table 2 of L&S interpolating between $\mu = 12.4$ and $\mu = 12.6$,

$$\Pr(X \leq 18) = F_{18}$$

which is between 0.9513 and 0.9448, giving $= 0.948$ to 3 d.p..

Question 4

(a) Let X denote the length of a randomly chosen squirrel's tail in cm. Then $X \sim N(21, 1.2^2)$.

$$\begin{aligned} Z &= \frac{X - 21}{1.2} \sim N(0, 1). \\ Pr(Z > 20) &= 1 - \phi\left(\frac{20 - 21}{1.2}\right), \\ &= \phi\left(\frac{5}{6}\right) \\ &= 0.7976, \end{aligned}$$

by extrapolation between values for 0.83 and 0.84 in Table 4 of L&S.

(b) $\bar{X} \sim N(21, \frac{1.2^2}{25})$, $\bar{Z} = \frac{\bar{X}-21}{\frac{1.2}{5}}$.

$$\begin{aligned}
\Pr(20.5 < \overline{Z} < 21.5) &= \Pr\left(-\frac{0.5}{0.24} < \overline{Z} < \frac{0.5}{0.24}\right), \\
&= \Pr(-2.08 < \overline{Z} < 2.08), \\
&= 2\Phi(2.08) - 1, \\
&= 2 \times 0.98124 - 1, \\
&= 0.962.
\end{aligned}$$

Question 5

(a) For a random variable with the continuous distribution, the median η is the value of x for which $Pr(X \leq x) = \frac{1}{2}$. Now

$$\begin{aligned}
\int_{-\infty}^{\eta} f(x)\,\mathrm{d}x &= \int_{-\infty}^{\eta} \frac{6}{x^2}\,\mathrm{d}x, \\
&= \left[-\frac{6}{x}\right]_{-\infty}^{\eta}, \\
&= \left[-\frac{6}{x}\right]_{6}^{\eta} \quad \text{(Since } x \geq 6\text{)}, \\
&= -\frac{6}{\eta} + 1.
\end{aligned}$$

Thus $-\frac{6}{\eta} + 1 = \frac{1}{2}$, and so $\eta = 12$.

(b) Noting that the support for $f(x)$ is $x \geq 6$,

$$\begin{aligned}
\Pr(3 < X < 9) &= \int_{6}^{9} \frac{6}{x^2}\mathrm{d}x, \\
&= \left[-\frac{6}{x}\right]_{6}^{9}, \\
&= -\frac{6}{9} + 1, \\
&= \frac{1}{3}.
\end{aligned}$$

integral $\int \frac{6}{x^2} dx$ (M1)

Question 6

(a) The expected frequencies under the null hypothesis are 10 for each minute. The chi-square test statistic is

$$(11-10)^2/10 + (26-10)^2/10 + (13-10)^2/10 + (7-10)^2/10 + (9-10)^2/10 +$$

$$(7-10)^2/10 + (6-10)^2/10 + (8-10)^2/10 + (1-10)^2/10 + (12-10)^2/10.$$

$$= \frac{1 + 256 + 9 + 9 + 1 + 9 + 16 + 4 + 81 + 4}{10} = 39.$$

(b) Under the null hypothesis it has the chi-square distribution with 9 degrees of freedom.

From Table 8 of L&S, the critical point for 1% when $\nu = 9$ is 21.67. Therefore the probability of observing this test statistic is less than 1%.

Question 7

(a) 78 families contain older boys and 82 contain older girls. The ratio of boys to girls overall may be assumed to be 1:1 or, better, taken from the sample to be 159:161. The information that there are more older boys than older girls is not relevant, since the question under consideration is whether there is association between the genders of the older and younger.

Assuming no association between between the genders of the older and younger children, the expected frequencies are given in the following table.

	older boy	older girl	total
younger boy	39.49	41.51	81
younger girl	38.51	40.49	79
Total	78	82	160

B1

(b) The chi-square test statistic is given by

$$\begin{aligned} X^2 &= \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}, \\ &= \frac{(45-39.49)^2}{39.49}+\frac{(36-41.51)^2}{41.51}+\frac{(33-38.51)^2}{38.51}+\frac{(46-40.49)^2}{40.49}, \\ &= 0.7688+.7314+0.7884+0.7498, \\ &= 3.038. \end{aligned}$$

M1

A1

There is 1 degree of freedom, since the totals with older boys and older girls are constrained separately. [7]

From Table 8, $\chi^2_{1,5\%} = 3.841$. There is no strong evidence of an association between the genders of older and younger children in two-child families at the 5% level of significance. A1

Question 8

(a) Binomial(10,0.5)

(b) To give evidence at the 10% level, $F_r \geq 0.9$
$r = 7$, from Table 1, for $n = 10$, L&S.

(c) (i) For a Signed Rank Test, the distribution under the null hypothesis must be symmetrical.

(ii) For a one sample t-test, the distribution of the differences between scores in the first turn and the second are approximately normal under the null hypothesis.

Question 9

(a) (i) $\Pr(A) = \sum_{i=1}^{k} \Pr(B_i)\Pr(A|B_i)$. B2

(ii) For $1 \le j \le k$, $\Pr(B_j|A) = \dfrac{\Pr(B_j)\Pr(A|B_j)}{\sum_{i=1}^{k}\Pr(B_i)\Pr(A|B_i)}$. B1 B1

(b) Tabulate the expected numbers of faults by car and type by multiplying the numbers of cars by the rate of faults.

Make	electrical	mechanical	total
Amethyst	40	10	50
Baxter	16	48	64
Chianti	6	36	42
Total	62	94	156

Sight of 156 B1 somewhere

M1 → (i) $\dfrac{36}{156} = \dfrac{3}{13}$. A1 (36 B1)

M1 → (ii) $\dfrac{16}{64} = \dfrac{1}{4}$. A1 (16 B1)

M1 → (iii) $\dfrac{42}{156} = \dfrac{7}{26}$. A1 A1 (42 B1)

M1 → (iv) $\dfrac{62}{156} = \dfrac{31}{78}$. A1 A1 (62 B1)

(v) Let B_i denote the event that the faulty car is of make i, where ($i = Am, Ba, Ch$). Let A be the event that the fault is mechanical. We have the prior probabilities $\Pr(B_{Am}) = 50/156, \Pr(B_{Ba}) = 64/156, \Pr(B_{Ch}) = 42/156$. We have the conditional probabilities [illegible]

M1

$$\begin{aligned}\Pr(A|B_{Am}) &= 10/50 = 1/5 && \text{A1}\\ \Pr(A|B_{Ba}) &= 48/64 = 3/4 && \text{A1}\\ \Pr(A|B_{Ch}) &= 36/42 = 6/7 && \text{A1}\end{aligned}$$

Using Bayes' Theorem,

$$\Pr(B_{Am}|A) = \frac{(50/156)(1/5)}{(50/156)(1/5) + (64/156)(3/4) + (42/156)(6/7)} = \quad \text{M1}$$

$$\frac{10/156}{94/156} = 10/94 = 5/47 \quad \text{A1}$$

$$\Pr(B_{Ba}|A) = \frac{(64/156)(3/4)}{94/156} = \frac{48}{94} = 24/47 \quad \text{A1}$$

$$\Pr(B_{Ch}|A) = \frac{(42/156)(6/7)}{94/156} = \frac{36}{94} = 18/47 \quad \text{A1}$$

Question 10

(a) We require that $\int_0^4 c(4-x)^2\,dx = 1$. Thus M1

$$-\frac{c}{3}\left[(4-x)^3\right]_0^4 = 1. \quad \text{M1}$$

Hence $\frac{64c}{3} = 1$, and so $c = \frac{3}{64}$.

(b)

$$\begin{aligned} E(X) &= \int_0^4 x\frac{3}{64}(4-x)^2\,dx, && \text{M1}\\ &= \frac{3}{64}\int_0^4 \left(16x - 8x^2 + x^3\right)\,dx, \\ &= \frac{3}{64}\left[8x^2 - \frac{8}{3}x^3 + \frac{1}{4}x^4\right]_0^4 = 1. && \text{M1 A1} \end{aligned}$$

(c)

$$\begin{aligned} E(X^2) &= \int_0^4 x^2\frac{3}{64}(4-x)^2\,dx. && \text{M1}\\ &= \frac{3}{64}\int_0^4 \left(16x^2 - 8x^3 + x^4\right)\,dx, \\ &= \frac{3}{64}\left[\frac{16}{3}x^3 - 2x^4 + \frac{1}{5}x^5\right]_0^4 = \frac{8}{5}. && \text{A1} \end{aligned}$$

Hence $var(X) = E(X^2) - (E(X))^2 = \frac{8}{5} - 1 = \frac{3}{5}$. M1 A1✓

(d) For $0 \le x \le 4$,

$$\begin{aligned} F(x) &= \int_0^x \frac{3}{64}(4-u)^2\,du, && \text{M1}\\ &= \left[\frac{3}{64}\frac{(4-u)^3}{-3}\right]_0^x, \\ &= 1 - \frac{1}{64}(4-x)^3. && \text{A1} \end{aligned}$$

Thus

$$F(x) = \begin{cases} 0, & x < 0, \quad \text{∴ B1} \\ 1 - \frac{1}{64}(4-x)^3, & 0 \le x \le 4, \quad \text{B1} \\ 1, & x > 4. \end{cases}$$

$\frac{x^3}{64} - \frac{3}{16}x^2 + \frac{3x}{4}$

(e) The median η satisfies $\int_0^\eta f(x)\,dx = \frac{1}{2}$. That is, $F(\eta) = \frac{1}{2}$. Thus M1

$$\begin{aligned} 1 - \frac{(4-\eta)^3}{64} &= \frac{1}{2}. && \text{M1}\\ \text{Hence } \eta &= 4 - \sqrt[3]{32}, \\ &\approx 0.8252. && \text{A1} \end{aligned}$$

B1

(f) Since the median is less than the mean the probability density function is positively skewed. B1

(g) M1

$$\begin{aligned}\Pr(1 < X < 2) &= F(2) - F(1),\\ &= \left(1 - \frac{2^3}{64}\right) - \left(1 - \frac{3^3}{64}\right),\\ &= \frac{19}{64}.\end{aligned}$$

A1

Question 11

(a) (i) We are using a small-sample comparison of means or 'two-sample t-test.' B1 We have a random sample $x_{1i}, i = 1, 2, \ldots, 16$ of size 16 of smokers and, B1 independently of the first sample, a random sample $x_{2j}, j = 1, 2, \ldots, 20$ of size 20 of non-smokers. We assume that the first sample comes from a B1 $N(\mu_1, \sigma^2)$ distribution and that the second sample comes from a $N(\mu_2, \sigma^2)$ B1 distribution, where μ_1, μ_2 and σ^2 are unknown.

We test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the one-sided alternative $H_1 : \mu_1 < \mu_2$. B1

(ii) In general, if n_1 and n_2 are the two sample sizes, respectively, the *pooled estimate* of the variance σ^2 is given by

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

B2

$$s^2 = \frac{\sum_{i=1}^{n_1}(x_{1i} - \bar{x}_1)^2 + \sum_{j=1}^{n_2}(x_{2j} - \bar{x}_2)^2}{n_1 + n_2 - 2}.$$

3.

B1

(iii) The p-value of 0.003 is strongly significant, beyond the 0.5% significance level. There is strong significant evidence that haemoglobin levels among smokers are lower than among non-smokers. B1

(b) (i) Two measurements exist for each experimental subject. These measure- B1 ments are not independent, which is a condition of the two-sample test. There may also be wide variation between experimental subjects but only small variation between pairs of before and after measurements. The test in part (a) would therefore have greater standard errors and less power to detect a significant difference between means.

(ii) It is assumed that the paired observations are randomly selected from normally distributed populations. B1 We test the null hypothesis $H_0 : \mu_{x_1 - x_2} = 0$ B1 against the two-sided alternative $H_1 : \mu_{x_1 - x_2} \neq 0$. B1

(iii) A two-sided alternative hypothesis does not allow for the prior knowledge B1 that an iron supplement has potential to increase haemoglobin but is B1 unlikely to reduce it. A one-sided alternative is justified. The p value compares the test statistic $t \overset{B1}{=} -1.67$ against the t-distribution with 13 B1 degrees of freedom. The p-value for a one-sided hypothesis is $1 - Pr(X < -t)$, which, for a symmetrical distribution, is half the value given in the two-tailed Minitab output.

B1 0.118

B1 0.059

Question 12

B1

(a) (i) The null hypothesis is that the data are a random sample from a Poisson distribution. The test statistic is

$$X^2 = \sum_{r=1}^{k} \frac{(O_r - E_r)^2}{E_r}\,, \quad B1$$

where k is the number of cells, in this case 4, the O_r are the observed frequencies, and the E_r are the expected frequencies under the null hypothesis. Its distribution under the null hypothesis is (approximately) the chi-square distribution with $k - 1 - d$ degrees of freedom, where d B1 is the number of fitted parameters, 2 degrees of freedom in the present case.

(ii) It is commonly asserted that for the chi-square approximation to be valid, the expected frequencies should be greater than 5, although 1 or 2 expected B1 frequencies somewhat less than 5 may be allowed. In this case B1 the amalgamation of the frequencies has ensured that all of the expected frequencies are greater than 5. Without amalgamation, 6 of the expected B1 frequencies would be less than 5, some of them much less than 5.

(iii) The p-value is too large to provide significant evidence of a lack of fit. B1.

(b) (i) Binomial distribution with parameters (n, p) where $n = 8$, the number B1 of wild cards each year, and $p = 0.16875$, the probability of a single win, given 54 wins out of a possible 320. B1

(ii)

$$Pr(R = 0) = \binom{8}{0} 0.16875^0 (1 - 0.16875)^8 = 0.228 \quad \cancel{A}\ M1$$

$$E_0 = Pr(R = 0) \times 40 = 9.12 \quad A1$$

The contribution to the test statistic is

$$\frac{(O_0 - E_0)^2}{E_0}$$

$$= (15 - 9.12)^2/9.12 = 3.79 \quad B1$$

Already, the test statistic has exceeded that under the Poisson distribution. The test will have no more degrees of freedom than under the B1 Poisson distribution ($Pr(R \leq 3) = 0.95$, so after $r = 3$ all further cells will need to be amalgamated) so the analysis will not choose to pursue the binomial distribution.

(c) (i) For a Poisson distribution, the variance is equal to the mean, $\sigma^2 = \mu$. B1

(ii) $I = \sum_{i=1}^{n} (x_i - \bar{x})^2/\bar{x}$, where $n = 40$ in the present case.

B1

(iii) In the present case, $I = 74.89 \sim \chi^2_{39}$, with $p = 0.00048$, significant B1
even at the 0.1% level. There is significant evidence to reject the hy-
pothesis that the annual number wild-card entrants winning first-round B1
matches follows a Poisson distribution. This is much stronger evidence
than was provided by the first goodness of fit test. Without the need to
amalgamate, the index of dispersion test makes more powerful use of the B1
available information.