Probability and Statistics

## 2015 Examination - Solutions

1. (a) $\Pr(A \cup B) = \Pr(A) + \Pr(B) = 0.2 + 0.4 = 0.6.$ $= 3/5$

   (b) $\Pr(A \cap C) = \Pr(A)\Pr(C) = (0.2)(0.5) = 0.1.$ $= 1/10$

   (c) $\Pr(A \cup C) = \Pr(A) + \Pr(C) - \Pr(A \cap C) = 0.2 + 0.5 - 0.1 = 0.6.$ $= 3/5$

   (d)

$$\begin{aligned}
\Pr(A \cup B|C) &= \frac{\Pr((A \cup B) \cap C)}{\Pr(C)} = \frac{\Pr((A \cap C) \cup (B \cap C))}{\Pr(C)} \\
&= \frac{\Pr(A \cap C) + \Pr(B \cap C)}{\Pr(C)} = \frac{0.1 + 0.3}{0.5} = 0.8. \quad = 4/5
\end{aligned}$$

[Alternatively $= \Pr(A|C) + \Pr(B|C) - \Pr(A \cap B|C) = \Pr(A) + \Pr(B \cap C)/\Pr(C)$]

2. (a) The binomial $B(n, p)$ distribution, with $n = 20$, $p = 0.34$.

   (b)

$$E(X) = np = 20 \times 0.34 = 6.8.$$

   (c)

$$\text{var}(X) = npq = 20 \times 0.34 \times 0.66 = 4.488$$

   Hence the standard deviation is $\sqrt{4.488} = 2.118$.

   (d) Using Table 1 of L&S, with $n = 20$ and $p = 0.34$,

$$\Pr(X \leq 5) = F_5 = 0.2758 = 0.276 \text{ to 3 d.p.}$$

   (e)

$$\Pr(X \geq 10) = 1 - F_9 = 1 - 0.8968 = 0.103 \text{ to 3 d.p.}$$

3. (a) A Poisson distribution with parameter/mean $\mu = 15/6 = 2.5$.

   (b) Using Table 2 of L&S with $\mu = 2.5$,

$$\Pr(X = 0) = F_0 = 0.0821 = 0.082 \text{ to 3 d.p.}$$

   (c)

$$\Pr(X = 3) = F_3 - F_2 = 0.7576 - 0.5438 = 0.214 \text{ to 3 d.p.}$$

   (d)

$$\Pr(X \geq 4) = 1 - F_3 = 1 - 0.7576 = 0.242 \text{ to 3 d.p.}$$

4. (a) Let $X$ denote the weight of a randomly chosen bag.

$$Z = \frac{X - 1001}{2} \sim N(0,1)$$

Hence

$$\begin{aligned} \Pr(X > 1000) &= \Pr\left(\frac{X-1001}{2} > \frac{1000-1001}{2}\right) \\ &= 1 - \Phi(-0.5) = \Phi(0.5) = 0.6915, \end{aligned}$$

using Table 4 of L&S.

(b) Let $X_i$ denote the weights of the bags. $\sum_{i=1}^{5} X_i$ is normally distributed with mean $1001 \times 5 = 5005$ and standard deviation $\sqrt{5 \times 2^2} = 2\sqrt{5}$.

$$\begin{aligned} \Pr(\sum_{i=1}^{5} X_i > 5000) &= \Pr\left(\frac{\sum_{i=1}^{5} X_i - 5005}{2\sqrt{5}} > \frac{5000-5005}{2\sqrt{5}}\right) \\ &= 1 - \Phi(-\sqrt{5}/2) = \Phi(\sqrt{5}/2) = \Phi(1.118) = 0.8682, \end{aligned}$$

using interpolation in Table 4.

5. (a) A 95% confidence interval for $\mu$ is given by

$$\bar{x} \pm t_{23}(2.5)s/\sqrt{24},$$

i.e.,

$$193.54 \pm 2.069\sqrt{2620.17}/\sqrt{24},$$

i.e., $193.54 \pm 21.62$, i.e., $(171.9, 215.2)$

(b) A 95% confidence interval for $\sigma^2$ is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{23}(2.5)}, \frac{(n-1)s^2}{\chi^2_{23}(97.5)}\right) = \left(\frac{23 \times 2620.17}{38.08}, \frac{23 \times 2620.17}{11.69}\right) = (1582.6, 5155.2)$$

6. (a)

$$\hat{p} = \frac{34+36}{40+60} = 0.7.$$

(b) $\hat{p}_A = 34/40 = 0.85$. $\hat{p}_B = 36/60 = 0.6$. The test statistics is

$$\frac{\hat{p}_A - \hat{p}_B}{\sqrt{\hat{p}(1-\hat{p})(1/n_A + 1/n_B)}} = \frac{0.85 - 0.6}{\sqrt{(0.7)(0.3)(1/40 + 1/60)}} = 2.6726.$$

This is a z-statistic, with a standard normal distribution under the null hypothesis. For a two-tail test, from Table 4, using interpolation, the corresponding p-value is $2(1 - 0.99624) = 0.0075$. This is significant evidence, significant at the 1% level, that there is a difference between System A and System B in their ability to detect the packages. System A appears to perform better.

7. (a) The expected frequencies $E_{ij} = R_i C_j / n$ under the null hypothesis are tabulated below with the row and column totals, $R_i$ and $C_j$.

| | Material A | Material B | Material C | total |
|---|---|---|---|---|
| Number crumbled | 37.6 | 25.1 | 31.3 | 94 |
| Number not crumbled | 82.4 | 54.9 | 68.7 | 206 |
| total | 120 | 80 | 100 | 300 |

(b) Under the null hypothesis the test statistic has the chi-square distribution with (3 - 1)(2 - 1) = 2 degrees of freedom.

From Table 7, the c.d.f. F(4.86) = 0.91 to 2 d.p, so that p = 0.09.

There is no strong evidence to reject the null hypothesis that the chances of crumbling are the same for all three materials.

8. We may carry out the sign test.

Under the null hypothesis that individuals are equally likely to prefer the old and new recipes, ignoring the 2 individuals who did not have a preference, the number $S$ of individuals who prefer the old recipe to the new would have the binomial $B(18, 1/2)$ distribution. Assuming the viewpoint of the research team, we may use a one-sided alternative that overall the individuals in the population have a lower preference for the old recipe. Using Table 1 with $n = 18, p = 0.5$, the p-value of the observed value $S = 5$ is given by

$$p = F_5 = 0.0481.$$

This is significant at the 5% level, so there is (strong) evidence that individuals in the population tend to have a preference for the new recipe.

9. (a)

$$\Pr(H_i|E) = \frac{\Pr(H_i)\Pr(E|H_i)}{\sum_{j=0}^{k}\Pr(H_j)\Pr(E|H_j)} \qquad (0 \le i \le k).$$

(b) Let $H_0$ denote the hypothesis that the tested batch is satisfactory and $H_1$ the hypothesis that the impurity level is too high. Let $E$ denote the event that a high impurity reading is obtained from the instrument. From the information provided, we may take it that the prior probabilities are given by

$$\Pr(H_1) = \frac{1}{100}, \qquad \Pr(H_0) = 1 - \frac{1}{100} = \frac{99}{100}.$$

i)

$$\Pr(E|H_0) = \frac{3}{100}, \qquad \Pr(E|H_1) = 1 - \frac{2}{100} = \frac{98}{100}.$$

Substituting into Bayes' Theorem,

$$\begin{aligned}
\Pr(H_1|E) &= \frac{\Pr(H_1)\Pr(E|H_1)}{\Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1)} \\
&= \frac{\frac{1}{100}\frac{98}{100}}{\frac{99}{100}\frac{3}{100} + \frac{1}{100}\frac{98}{100}} \\
&= \frac{98}{(99 \times 3) + 98} = 0.2481.
\end{aligned}$$

ii)

$$\Pr(\bar{E}|H_0) = 1 - \frac{3}{100} = \frac{97}{100}, \qquad \Pr(\bar{E}|H_1) = \frac{2}{100}.$$

Substituting into Bayes' Theorem,

$$\begin{aligned}
\Pr(H_0|\bar{E}) &= \frac{\Pr(H_0)\Pr(\bar{E}|H_0)}{\Pr(H_0)\Pr(\bar{E}|H_0) + \Pr(H_1)\Pr(\bar{E}|H_1)} \\
&= \frac{\frac{99}{100}\frac{97}{100}}{\frac{99}{100}\frac{97}{100} + \frac{1}{100}\frac{2}{100}} \\
&= \frac{99 \times 97}{(99 \times 97) + 2} = 0.9998.
\end{aligned}$$

iii) It is true that the probability $\Pr(E|H_1) = 0.98$ of a high impurity reading given a high level of impurity is high, but that is not the same as the probability $\Pr(H_1|E) = 0.2481$ of a high level of impurity given a high impurity reading. (To confuse the two is "the prosecutor's fallacy".) It is the low prior probability $\Pr(H_1) = 1/100$ for a high level of impurity that makes the value of the posterior probability $\Pr(H_1|E)$ relatively small.

10. (a)

$$\begin{aligned}\Pr(1 \le X \le 2) &= F(2) - F(1) \\ &= \frac{6}{12} - \frac{2}{12} = \frac{1}{3}.\end{aligned}$$

(b) The median is the value $m$ such that $F(m) = 1/2$.
But from the calculations for Part (a) we see that $F(2) = 1/2$. Hence $m = 2$.

(c)

$$\begin{aligned}f(x) &= \frac{dF}{dx} \\ &= \frac{1+2x}{12} \qquad (0 \le x \le 3).\end{aligned}$$

(d)

$$\begin{aligned}E(X) &= \int_0^3 x f(x) dx \\ &= \int_0^3 \frac{x(1+2x)}{12} dx \\ &= \frac{1}{12}\left[\frac{1}{2}x^2 + \frac{2}{3}x^3\right]_0^3 \\ &= \frac{1}{12}\left(\frac{9}{2} + 18\right) = \frac{15}{8}.\end{aligned}$$

2

(e)

$$\begin{aligned}E(X^2) &= \int_0^3 x^2 f(x) dx \\ &= \int_0^3 \frac{x^2(1+2x)}{12} dx \\ &= \frac{1}{12}\left[\frac{1}{3}x^3 + \frac{2}{4}x^4\right]_0^3 \\ &= \frac{1}{12}\left(9 + \frac{81}{2}\right) = \frac{33}{8}.\end{aligned}$$

2

$$\begin{aligned}\mathrm{var}(X) &= E(X^2) - [E(X)]^2 \\ &= \frac{33}{8} - \left(\frac{15}{8}\right)^2 \\ &= \frac{264 - 225}{8^2} = \frac{39}{64}.\end{aligned}$$

11. (a) The first procedure is for a "paired comparisons" or "matched pairs" design.

i) Let $d_i = x_i - y_i$ $(1 \le i \le 12)$. We assume that $d_1, d_2, \ldots, d_{12}$ is a random sample from a $N(\mu_D, \sigma_D^2)$ distribution, where $\mu_D$ and $\sigma_D^2$ are unknown. We test the null hypothesis $H_0 : \mu_D = 0$ against the two-sided alternative $H_1 : \mu_D \neq 0$.

ii) If $\bar{d}$ is the sample mean and $s_D$ the sample standard deviation of the $d_i$ then the test statistic is $t = \sqrt{12}\bar{d}/s_D$, which under $H_0$ has the t-distribution with 11 degrees of freedom.

The second procedure is a two-sample t-test.

i) We assume that $x_1, x_2, \ldots, x_{12}$ is a random sample from a $N(\mu_1, \sigma^2)$ distribution and, independently of the first sample, $y_1, y_2, \ldots, y_{12}$ is a random sample comes from a $N(\mu_2, \sigma^2)$ distribution, where $\mu_1$, $\mu_2$ and $\sigma^2$ are unknown.
We test the null hypothesis $H_0 : \mu_1 = \mu_2$ against the two-sided alternative $H_1 : \mu_1 \neq \mu_2$ .

ii) The *pooled estimate* of the variance $\sigma^2$ is given by

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{i=1}^{m}(y_i - \bar{y})^2}{22} .$$

The test statistic is

$$\frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{12} + \frac{1}{12}\right)}} ,$$

which under $H_0$ has the t-distribution with 22 degrees of freedom.

(b) We have a "paired comparison" design. A pair of measurements $(x_i, y_i)$ has been taken from each of 12 subjects. We do not have two independent samples $x_1, x_2, \ldots, x_{12}$ and $y_1, y_2, \ldots, y_{12}$. Hence the appropriate procedure is the first one, and not the second.

(c) The p-value of 0.475 is certainly not significant at the 5% significance level, or even at the 40% level. There is no significant evidence for any effect of aspirin on clotting time.

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

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

where $k$ is the number of cells in the table, in this case 7 after amalgamation, 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 $d$ is the number of fitted parameters, 5 degrees of freedom in the present case.

ii) The p-value is 0.858, which is not significant. There is no evidence to reject the hypothesis that the numbers of bacterial cells per square follow a Poisson distribution.

(b) i) 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 frequencies somewhat less than 5 may be allowed. In this case the warning comment is that there is just one expected frequency less than 5 after the amalgamation of cells, but it is not too much less than 5.

ii)

$$\begin{aligned} O_{\{\geq 5\}} &= 5 + 3 = 8, \\ p_{\{\geq 5\}} &= 0.067637 + 0.042861 = 0.110498, \\ E_{\{\geq 5\}} &= 5.4110 + 3.4289 = 8.8399. \end{aligned}$$

The corresponding contribution to the chi-square statistic is

$$\frac{(8 - 8.8399)^2}{8.8399} = 0.079801.$$

This replaces the contributions 0.031212 for the category "5" and 0.053649 for the category "$\geq 6$" . Hence the chi-square test-statistic is now

$$1.93753 + 0.079801 - 0.031212 - 0.053649 = 1.93247,$$

with 4 degrees of freedom. Using interpolation from Table 7,

$$F(1.932) \approx 0.1734 + (4.38/5)(0.2642 - 0.1734) = 0.2529.$$

Hence $p = 1 - 0.2529 = 0.747$ to 3 decimal places.

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

ii) $I = \sum_{i=1}^{n}(x_i - \bar{x})^2/\bar{x}$, where $n = 80$ in the present case.

iii) In the present case, $I = 70.8408 \sim \chi^2_{79}$, with p $= 0.732$, which is not significant. There is no evidence to reject the hypothesis that the numbers of yeast cells per square follow a Poisson distribution. 32