**Probability and Statistics**

## 2013 Examination - Solutions

1. (a) If $A$ and $B$ are mutually exclusive then $\Pr(A \cup B) = \Pr(A) + \Pr(B)$. Hence $p = \Pr(B) = \Pr(A \cup B) - \Pr(A) = 0.9 - 0.5 = 0.4$.

   (b) Generally

$$\begin{aligned}\Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\ &= \Pr(A) + \Pr(B) - \Pr(A)\Pr(B)\end{aligned}$$

   if $A$ and $B$ are independent. Hence $0.9 = 0.5 + p - 0.5p$, from which it follows that $p = 0.8$.

2. (a) The binomial $B(n, p)$ distribution, with $n = 18$, $p = 0.4$.

   (b)

$$E(X) = np = 18 \times 0.4 = 7.2.$$

   (c)

$$\mathrm{var}(X) = npq = 18 \times 0.4 \times 0.6 = 4.32$$

   Hence the standard deviation is $\sqrt{4.32} = 2.078$.

   (d) Using Table 1 of L&S, with $n = 18$ and $p = 0.4$,

$$\Pr(X \geq 10) = 1 - F_9 = 1 - 0.8653 = 0.135 \text{ to 3 d.p.}$$

3. (a) A Poisson distribution with parameter/mean $\mu = 0.0013 \times 8000 = 10.4$.

   (b) Using Table 2 of L&S with $\mu = 10.4$,

$$\Pr(X \leq 5) = F_5 = 0.0534 = 0.053 \text{ to 3 d.p.}$$

   (c)

$$\Pr(X > 10) = 1 - F_{10} = 1 - 0.5331 = 0.467 \text{ to 3 d.p.}$$

   (d)

$$\Pr(6 \leq X \leq 10) = F_{10} - F_5 = 0.5331 - 0.0534 = 0.480 \text{ to 3 d.p.}$$

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

$$Z = \frac{X - 12}{0.2} \sim N(0,1)$$

Hence

$$\begin{aligned} \Pr(X > 12.1) &= \Pr\left(\frac{X-12}{0.2} > \frac{12.1 - 12}{0.2}\right) \\ &= 1 - \Phi(0.5) = 1 - 0.6915 = 0.3085, \end{aligned}$$

using Table 4 of L&S.

(b) Let $X_i$ denote the weights of the blocks. $\sum_{i=1}^{10} X_i$ is normally distributed with mean $10 \times 12 = 120$ and standard deviation $\sqrt{10 \times 0.2^2} = \sqrt{0.4}$.

$$\begin{aligned} \Pr(\sum_{i=1}^{10} X_i > 121) &= \Pr\left(\frac{\sum_{i=1}^{10} X_i - 120}{\sqrt{0.4}} > \frac{121 - 120}{\sqrt{0.4}}\right) \\ &= 1 - \Phi(1.5811) = 1 - 0.9430 = 0.0570. \end{aligned}$$

5. (a)

$$s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2} = \frac{(9 \times 0.019^2) + (12 \times 0.016^2)}{21} = 0.000301.$$

(b)

$$\frac{\bar{x} - \bar{y}}{\sqrt{s^2\left(\frac{1}{n} + \frac{1}{m}\right)}} = \frac{0.041 - 0.026}{\sqrt{0.000301\left(\frac{1}{10} + \frac{1}{13}\right)}} = 2.055.$$

This is a t-statistic with 21 degrees of freedom. $t_{21}(5) = 1.721$. So, for a one-tail test, the t-statistic is significant at the 5% significance level. There is strong evidence to reject the null hypothesis of equal population means. There is strong evidence that the mean level of DDT is greater for juveniles than for nestlings.

6. (a) The sample proportion is $87/200 = 0.435$. A 95% confidence interval is given by

$$0.435 \pm 1.96\sqrt{\frac{(0.435)(0.565)}{200}}$$

i.e.

$$0.435 \pm 0.069 \quad \text{i.e.} \quad (0.366, 0.504).$$

(b) The longest interval occurs when the sample proportion is 0.5. We require

$$1.96\sqrt{\frac{(0.5)(0.5)}{n}} = 0.01$$

i.e.

$$n = 196^2/4 = 98^2 = 9604.$$

To the nearest hundred, we require n = 9600.

7. (a) The expected frequencies under the null hypothesis are 50 for each quarter.
   (b) The chi-square test statistic is

$$(62-50)^2/50+(48-50)^2/50+(44-50)^2/50+(46-50)^2/50=4.00.$$

   Under the null hypothesis it has the chi-square distribution with 3 degrees of freedom.

   (c) Using Table 7 of Lindley and Scott, $p = 1 - F(4) = 1 - 0.7385 = 0.2615$. There is no strong evidence to reject the null hypothesis that birthdays are uniformly distributed throughout the year.

8. Under the null hypothesis of no association between between income category and attitude towards the legislation, the expected frequencies are given in the following table.

| | low income | high income | total |
|---|---|---|---|
| for the legislation | 102 | 51 | 153 |
| against the legislation | 98 | 49 | 147 |
| Total | 200 | 100 | 300 |

The chi-square test statistic is given by

$$X^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}} = (14)^2\left(\frac{1}{102}+\frac{1}{51}+\frac{1}{98}+\frac{1}{49}\right) = 11.76\ .$$

From Table 8, $\chi_1^2(0.1) = 10.83$. There is very significant evidence, at the 0.1% level, that there is an association between income category and attitude towards the legislation.

Low income voters tend to be more for the legislation and high income voters against the legislation.

9. (a) i)

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \qquad [\Pr(B) > 0].$$

Similarly,

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \qquad [\Pr(A) > 0].$$

Hence

$$\Pr(B)\Pr(A|B) = \Pr(A \cap B) = \Pr(A)\Pr(B|A).$$

Rearranging, we obtain the required result.

ii)

$$\Pr(B_j|A) = \frac{\Pr(B_j)\Pr(A|B_j)}{\sum_{i=1}^{k}\Pr(B_i)\Pr(A|B_i)} \qquad (j = 1, 2, \ldots, k).$$

(b) Let $B_i$ denote the event that Drug $i$ is the injected one ($i = 1, 2, 3$). Let $A$ be the event that no antitoxin forms. We have the prior probabilities $\Pr(B_1) = 1/3, \Pr(B_2) = 1/2, \Pr(B_3) = 1/6$. We have the conditional probabilities

$$\begin{aligned} \Pr(A|B_1) &= 1 - 1/4 = 3/4 \\ \Pr(A|B_2) &= 1 - 1/8 = 7/8 \\ \Pr(A|B_3) &= 1 - 1/3 = 2/3 \end{aligned}$$

Using Bayes' Theorem,

$$\Pr(B_2|A) = \frac{(1/2)(7/8)}{(1/3)(3/4) + (1/2)(7/8) + (1/6)(2/3)} = 0.548.$$

(c) Let $H_0$ denote the hypothesis that the arrested man is innocent and $H_1$ the hypothesis that he is guilty. Let $E$ denote the evidence of the DNA match. From the information provided, we may take it that the prior probabilities are given by

$$\Pr(H_1) = \frac{1}{200,000}, \qquad \Pr(H_0) = 1 - \frac{1}{200,000}.$$

Furthermore, we may take it that

$$\Pr(E|H_0) = \frac{1}{20,000,000} \qquad \Pr(E|H_1) = 1.$$

Substituting into Bayes' Theorem,

$$\begin{aligned} \Pr(H_1|E) &= \frac{\Pr(H_1)\Pr(E|H_1)}{\Pr(H_0)\Pr(E|H_0) + \Pr(H_1)\Pr(E|H_1)} \\ &= \frac{\frac{1}{200,000}}{(1 - \frac{1}{200,000})\frac{1}{20,000,000} + \frac{1}{200,000}} = 0.990. \end{aligned}$$

10. (a) Using the standard normal density function,

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 1.$$

Hence

$$\int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{2},$$

and the result follows.

(b)

$$\begin{aligned} E(X) &= \int_0^\infty x^2 \exp\left(-\frac{x^2}{2}\right) dx \\ &= \left[-x \exp\left(-\frac{x^2}{2}\right)\right]_0^\infty + \int_0^\infty \exp\left(-\frac{x^2}{2}\right) dx = \sqrt{\frac{\pi}{2}}. \end{aligned}$$

(c)

$$\begin{aligned} \mu_k &= \int_0^\infty x^{k+1} \exp\left(-\frac{x^2}{2}\right) dx \\ &= \left[-x^k \exp\left(-\frac{x^2}{2}\right)\right]_0^\infty + k \int_0^\infty x^{k-1} \exp\left(-\frac{x^2}{2}\right) dx = k\mu_{k-2}. \end{aligned}$$

(d) From (c), $\mu_2 = 2\mu_0 = 2$. Hence, using (b),

$$\operatorname{var}(X) = \mu_2 - \mu_1^2 = 2 - \pi/2 = (4 - \pi)/2.$$

(e) For $x \geq 0$,

$$F(x) = \int_0^x u \exp\left(-\frac{u^2}{2}\right) du = \left[-\exp\left(-\frac{u^2}{2}\right)\right]_0^x = 1 - \exp\left(-\frac{x^2}{2}\right).$$

(f) The median is the number $m$ such that $F(m) = 1/2$. Hence $\exp(-m^2/2) = 1/2$, from which it follows that $\exp(m^2/2) = 2$ and so $m = \sqrt{2 \ln 2}$.

Calculating to three decimal places, $\mu = 1.253$ and $m = 1.177$. Because $f$ is positively skewed, we expect the mean to be greater than the median.

11. (a) i) (We have a "matched pairs" or "paired comparison" design.) Let $x_i$ and $y_i$ denote the measured diameter as given by Caliper 1 and Caliper 2, respectively, for Inspector $i$. Let $d_i = x_i - y_i$ ($1 \leq i \leq 12$). 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 $n$ is the number of pairs and $\bar{d}$ is the sample mean and $s_D$ the sample standard deviation of the $d_i$ then the test statistic is $t = \sqrt{n}\bar{d}/s_D$, which under $H_0$ has the t-distribution with $n-1$ degrees of freedom, where $n = 12$ in the present case.

iii) The p-value of 0.674 is certainly not significant at the 5% significance level, nor even remotely significant at any level that we would normally consider. There is no significant evidence that there is any difference in the mean measurements given by the two types of caliper.

(b) In the case of the nonparametric Wilcoxon signed-rank test:

i) It is assumed that the differences $d_i$ are symmetrically distributed about their median value $\eta$. We test the null hypothesis $H_0 : \eta = 0$ against the two-sided alternative $H_1 : \eta \neq 0$.

ii) The calculations of the signed ranks are shown in the table below.

| Inspector | Caliper 1 $x_i$ | Caliper 2 $y_i$ | difference $d_i$ | sign | rank of $\lvert d_i \rvert$ | signed rank |
|---|---|---|---|---|---|---|
| 1 | 0.265 | 0.264 | 0.001 | $+$ | 2 | $+2$ |
| 2 | 0.265 | 0.265 | 0 | na | | |
| 3 | 0.266 | 0.264 | 0.002 | $+$ | 4.5 | $+4.5$ |
| 4 | 0.267 | 0.266 | 0.001 | $+$ | 2 | $+2$ |
| 5 | 0.267 | 0.267 | 0 | na | | |
| 6 | 0.265 | 0.268 | -0.003 | $-$ | 6.5 | $-6.5$ |
| 7 | 0.267 | 0.264 | 0.003 | $+$ | 6.5 | $+6.5$ |
| 8 | 0.267 | 0.265 | 0.002 | $+$ | 4.5 | $+4.5$ |
| 9 | 0.265 | 0.265 | 0 | na | | |
| 10 | 0.268 | 0.267 | 0.001 | $+$ | 2 | $+2$ |
| 11 | 0.268 | 0.268 | 0 | na | | |
| 12 | 0.265 | 0.269 | -0.004 | $-$ | 8 | $-8$ |

The Wilcoxon statistic is given by the sum of the positive signed ranks, $2 + 4.5 + 2 + 6.5 + 4.5 + 2 = 21.5$.

(c) With the parametric t-test it is assumed that the differences are normally distributed, but with the non-parametric Wilcoxon test no normality assumption is made.

Apparently by coincidence, the p-values for the two procedures are exactly the same to 3 decimal places, and so the conclusions to be drawn are the same in both cases. This is a little surprising. We would expect the non-parametric test to be less powerful than the parametric test and so give a larger p-value.

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 0.187. There is no significant evidence, not even at the 10% level, to reject the hypothesis that the numbers of yeast 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 amalgamation of the frequencies has ensured that all but one of the expected frequencies are greater than 5. The one that is less than 5 is not much less. Without amalgamation, 4 of the expected frequencies would be less than 5, some of them much less than 5.

ii)

$$\begin{aligned} O_{\{\geq 5\}} &= 13 + 4 = 17, \\ p_{\{\geq 5\}} &= 0.026029 + 0.010378 = 0.036407, \\ E_{\{\geq 5\}} &= 10.411 + 4.151 = 14.562. \end{aligned}$$

The corresponding contribution to the chi-square statistic is

$$\frac{(17 - 14.562)^2}{14.562} = 0.40817.$$

This replaces the contributions 0.64358 for the category "5" and 0.00551 for the category "$\geq 6$" . Hence the chi-square test-statistic is now

$$7.47796 + 0.40817 - 0.64358 - 0.00551 = 7.23704,$$

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

$$F(7.237) \approx 0.8641 + (2.37/5)(0.8883 - 0.8641) = 0.8756.$$

Hence $p = 1 - 0.8756 = 0.124$ 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 $n = 400$ in the present case.

iii) In the present case, $I = 435.556 \sim \chi^2_{399}$, with p = 0.1003, not significant even at the 10% level. There is no significant evidence to reject the hypothesis that the numbers of yeast cells per square follow a Poisson distribution.