EXAMINERS' REPORT

April 2010 Examinations

Subject CT3 — Probability and Mathematical Statistics Core Technical

Introduction

The attached subject report has been written by the Principal Examiner with the aim of helping candidates. The questions and comments are based around Core Reading as the interpretation of the syllabus to which the examiners are working. They have however given credit for any alternative approach or interpretation which they consider to be reasonable.

R D Muckart
Chairman of the Board of Examiners

July 2010

1 Let p be the proportion of women.

Then, using a weighted average, $1.671p + 1.758(1-p) = 1.712$
$\Rightarrow 0.087p = 0.046 \Rightarrow p = 0.529$ so percentage is 52.9%

2 $P(\text{all 3 on male lives}) = \frac{7}{10}\times\frac{6}{9}\times\frac{5}{8} = \frac{7}{24} = 0.292$

[OR $\binom{7}{3}/\binom{10}{3} = 35/120 = 7/24$]

3
$$\begin{aligned} E\left[S^2\right] &= \frac{1}{n-1}\{\Sigma E(X_i^2) - nE(\bar{X}^2)\} \\ &= \frac{1}{n-1}\{\Sigma(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\} \\ &= \frac{1}{n-1}\{n(\sigma^2+\mu^2) - \sigma^2 - n\mu^2\} \\ &= \frac{1}{n-1}\{(n-1)\sigma^2\} = \sigma^2 \end{aligned}$$

4 Approximate 95% CI for $\lambda_m - \lambda_f$ is $(\bar{x}_m - \bar{x}_f) \pm 1.96\sqrt{\frac{\bar{x}_m}{120}+\frac{\bar{x}_f}{80}}$

$\Rightarrow \quad (0.24-0.15) \pm 1.96\sqrt{\frac{0.24}{120}+\frac{0.15}{80}}$

$\Rightarrow \quad 0.09 \pm 1.96(0.062) \quad \Rightarrow \quad 0.09 \pm 0.122 \quad \text{or} \quad \Rightarrow \quad (-0.032, 0.212)$

5 $S = \Sigma X_i$ where X_i has a uniform distribution on 1, 2, 3, 4, 5, with mean 3 and variance $(25-1)/12 = 2$ (result known, or calculated via $E[X^2] = 11$, or from book of formulae, p10, with $a = 1$, $b = 5$, $h = 1$).

So $S \sim N(300, 200)$ approximately

$$P(280 \le S \le 320) = P\left(\frac{279.5-300}{\sqrt{200}} < Z < \frac{320.5-300}{\sqrt{200}}\right)$$

$$= P\left(-1.450 < Z < 1.450\right) = 0.853$$

6 (i) (a) $\Sigma X_i \sim$ gamma($4n$, λ)

(b) If $Y \sim$ gamma(α, λ) and 2α is an integer, then $2\lambda Y \sim \chi^2_{2\alpha}$ (from book of formulae, p12)

So $2\lambda n\bar{X} \sim \chi^2$ with df $8n$.

(ii) $P(\chi^2_{40}(97.5) < 10\lambda\bar{X} < \chi^2_{40}(2.5)) = 0.95$

giving the 95% CI as $\left(\frac{\chi^2_{40}(97.5)}{10\bar{X}}, \frac{\chi^2_{40}(2.5)}{10\bar{X}}\right)$

Data $\Rightarrow$ $\left(\frac{24.43}{10(17.5)}, \frac{59.34}{10(17.5)}\right) = (0.140, 0.339)$

7 The 95% CI for the population percentage p is $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

giving $|p - \hat{p}| \leq 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

For the margin of error to be less than 0.5% we need to solve

$$0.005 = 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \Rightarrow n = \frac{1.96^2\,\hat{p}(1-\hat{p})}{0.005^2}\ .$$

Using the percentage from the previous study as the value for $\hat{p}$, i.e. $\hat{p} = 0.06$, we obtain n = 8,666.6.

So we need a sample of (at least) 8667 people.

(OR, solution can be based on $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ and $P(-0.005 < \hat{p} - p < 0.005) > 0.95$ without referring to the CI.)

8 (i) $\Sigma f = 100, \quad \Sigma fx = 27, \quad \Sigma fx^2 = 35$

$$\bar{x} = \frac{27}{100} = 0.27$$

$$s^2 = \frac{1}{99}\{35 - \frac{27^2}{100}\} = 0.2799 \quad \therefore s = 0.529$$

Third moment about mean is

$$m_3 = \frac{1}{100}\{76(0-0.27)^3 + 22(1-0.27)^3 + (2-0.27)^3 + (3-0.27)^3\} = 0.3259$$

[OR: using $\Sigma fx^3 = 57, \quad m_3 = \frac{1}{100}\{57 - 3(0.27)(35) + 2(100)(0.27)^3\}$]

So coefficient of skewness is $\dfrac{0.3259}{(0.2799)^{3/2}} = 2.20$

[OR: can use $m_2 = 0.2771$ in denominator to give 2.23]

(ii) (a) $\hat{\mu} = \bar{x} = 0.27$

(b) Coefficient of skewness is $\dfrac{1}{\sqrt{0.27}} = 1.92$ (from book of formulae, p7)

so, the data distribution is slightly more positively skewed than the fitted Poisson.

9 (i) $E[N] = \dfrac{k(1-p)}{p} = \dfrac{2(0.2)}{0.8} = 0.5$ and $V[N] = \dfrac{k(1-p)}{p^2} = \dfrac{2(0.2)}{0.8^2} = 0.625$

$$E[X] = \frac{1}{\lambda} = \frac{1}{2} = 0.5 \quad \text{and} \quad V[X] = \frac{1}{\lambda^2} = \frac{1}{2^2} = 0.25$$

$$E[S] = E[N]E[X] = 0.5 \times 0.5 = 0.25, \text{ i.e. } £250$$

$$V[S] = E[N]V[X] + V[N]\{E[X]\}^2 = 0.5 \times 0.25 + 0.625 \times 0.5^2 = 0.28125$$

$$\therefore \quad SD[S] = 0.530, \text{ i.e. } £530$$

(ii) $E[N] = V[N] = \mu = 0.5$

$E[X] = \frac{\alpha}{\lambda} = \frac{2}{4} = 0.5$ and $V[X] = \frac{\alpha}{\lambda^2} = \frac{2}{4^2} = 0.125$

$E[S] = E[N]E[X] = 0.5 \times 0.5 = 0.25$, i.e. £250

$V[S] = E[N]V[X] + V[N]\{E[X]\}^2 = 0.5 \times 0.125 + 0.5 \times 0.5^2 = 0.1875$

$\therefore \quad SD[S] = 0.433$, i.e. £433

(iii) As expected the means are the same,
but the standard deviation in (i) is larger than that in (ii) due to the fact that both *N* and *X* have larger variances.

10 (i) We have:

$$E[X] = \int_c^\infty x f_X(x)dx = \int_c^\infty x \frac{ac^a}{x^{a+1}}dx = ac^a \int_c^\infty x^{-a}dx = -\frac{ac^a}{a-1}\left[x^{-(a-1)}\right]_c^\infty$$

and for $a > 1$

$$E[X] = -\frac{ac^a}{a-1}(0 - c^{-a+1}) = \frac{ac}{a-1}.$$

(ii) $F_X(x) = \int_c^x f_X(t)dt = \int_c^x \frac{ac^a}{t^{a+1}}dt$

which gives

$$F_X(x) = -c^a\left[t^{-a}\right]_c^x = -c^a(x^{-a} - c^{-a}) = 1 - \left(\frac{c}{x}\right)^a, \quad x \geq c$$

[*OR* differentiate $F_X(x)$ to obtain $f_X(x)$]

(iii) The likelihood function is given by:

$$L(a) = \prod_{i=1}^{n} f_X(x_i) = \prod_{i=1}^{n} \frac{ac^a}{x_i^{a+1}} = a^n c^{na} \prod_{i=1}^{n} x_i^{-(a+1)}$$

and

$$l(a) = n\log(a) + na\log(c) - (a+1)\sum_{i=1}^{n} \log(x_i)$$

For the MLE:

$$l'(a) = 0 \Rightarrow \frac{n}{a} + n\log(c) - \sum_{i=1}^{n} \log(x_i) = 0$$

$$\Rightarrow \hat{a} = \frac{n}{\sum_{i=1}^{n} \log(x_i) - n\log(c)} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{c}\right)},$$

and for $c = 2.5$, $\hat{a} = \dfrac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{2.5}\right)}$

(iv) For the asymptotic variance we use the Cramer-Rao lower bound:

$$l''(a) = -\frac{n}{a^2}, \text{ and } E\left[l''(a)\right] = -\frac{n}{a^2}$$

giving

$$V\left[\hat{a}\right] = -\left\{E\left[l''(a)\right]\right\}^{-1} = \frac{a^2}{n}.$$

Hence, asymptotically, $\hat{a} \sim N(a, a^2/n)$.

(v) MLE is

$$\hat{a} = \frac{n}{\sum_{i=1}^{n} \log\left(\frac{x_i}{c}\right)} = \frac{n}{\sum_{i=1}^{n} \log(x_i) - n\log(c)} = \frac{30}{32.9 - 30 \times \log(2.5)} = 5.544.$$

Using the asymptotic normal distribution given above, an approximate 95% CI is given by

$$\hat{a} \pm 1.96\sqrt{\frac{a^2}{n}} = \hat{a} \pm 1.96\frac{\hat{a}}{\sqrt{n}}$$

i.e. $5.544 \pm 1.96\frac{5.544}{\sqrt{30}}$, giving (3.560, 7.528).

(vi) Size of claim in the following year will be given by $1.05X$

So we want $P(1.05X > 4) = P\left(X > \frac{4}{1.05}\right) = 1 - F_X\left(\frac{4}{1.05}\right)$

and using F_X given in the question

$$P(1.05X > 4) = \left(\frac{1.05 \times 2.5}{4}\right)^6 = 0.0799.$$

11 (i) (a) $\bar{x} = 19.513, s^2 = \frac{1}{14}\left(5778.69 - \frac{292.7^2}{15}\right) = 4.7955$

(b) Test statistic is $\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$

Here $t = (19.513 - 18)/(4.7955/15)^{1/2} = 2.68$

P-value = $P(t_{14} > 2.68)$, which is just less than 0.01 (1%)

We reject H_0 and accept "$\mu > 18$" at the 1% level of testing.

(ii) (a) Here $t = (19.867 - 18)/(19.432/15)^{1/2} = 1.64$

P-value = $P(t_{14} > 1.64)$, which is between 0.05 and 0.1.

P-value exceeds 5% and so we cannot reject H_0, so "$\mu = 18$" can stand.

(b) Sample 2 does not provide enough evidence to justify rejecting H_0, despite having the same size and a similar mean to Sample 1.

The reason for the loss of significance is the much greater variation in the data in Sample 2 – the variance is four times bigger than in Sample 1 (19.432 v 4.7955)

– this greatly increases the standard error of estimation and reduces the value of the *t*-statistic (1.64 v 2.68).

(iii) (a) Here $t = (19.644 - 18)/(5.275/25)^{1/2} = 3.58$

P-value = $P(t_{24} > 3.58)$, which is less than 0.001 (0.1%)

We reject H_0 and accept "$\mu > 18$" at a level lower than 0.1%.

(b) Sample 3 provides even stronger evidence against H_0, despite having a similar mean and variance to Sample 1.

The main reason for the much greater level of significance is the increased sample size (25 v 15)
– this decreases the standard error of estimation and increases the value of the *t*-statistic considerably (3.58 v 2.68).

12 (i)

- the three sets of points are positioned at different levels (the means are shown), so there is a prima facie case for suggesting that the underlying means are different (i.e. there are country effects)
- the means are in the order England (highest), Scotland, Wales (lowest)
- the variation in the data for Scotland is perhaps lower than that for England, but with only 5 observations for each country, we cannot be sure that there is a real underlying difference in variance

(ii) (a) $SS_T = 1316.63 - 137.1^2/15 = 63.536$, $SS_B = (55.6^2 + 36.8^2 + 44.7^2)/5 - 137.1^2/15 = 35.644$

$\therefore SS_R = 63.536 - 35.644 = 27.892$

Source of variation	*Df*	*SS*	*MSS*
Between countries	2	35.644	17.82
Residual	12	27.892	2.324
Total	14		

Under H_0: no country effects $F = 17.82/2.324 = 7.67$ on (2,12) df

P-value of $F = 7.67$ is less than 0.01, so we reject H_0 and conclude that there are differences among the population means of the average sum insured

(b) We have strong evidence that country effects exist – the means appear to be in the order England (highest), Scotland, Wales (lowest).

(iii) (a) $S_{xx} = 7543 - 329^2/15 = 326.9333$, $S_{yy} = 63.536$ (from (i)(b) above)

$S_{xy} = 3091.7 - 329 \times 137.1/15 = 84.64$

$\hat{\beta} = 84.64 / 326.9333 = 0.25889$, $\hat{\alpha} = 137.1/15 - \hat{\beta} \times (329/15) = 3.4617$

So fitted line is $y = 3.462 + 0.2589x$

(b) $R^2 = S_{xy}^2/(S_{xx}S_{yy}) = 84.64^2/(326.9333 \times 63.536) = 0.34488$ so 34.5%

(c) $\text{SSRES} = S_{yy} - S_{xy}^2/S_{xx} = 63.536 - 84.64^2/326.9333 = 41.62349$

$\Rightarrow \hat{\sigma}^2 = 41.62349/13 = 3.201807$

$\Rightarrow s.e.(\hat{\beta}) = (3.201807/326.9333)^{1/2} = 0.09896$

(iv) From the plot we see that the relationship between "index" and "average sum insured" is weak, positive (and possibly linear) – the percentage of the variation in "average sum insured" explained by the relationship with "index" is only 34.5%.

So "index" is of some, but limited, use as a predictor of "average sum insured".

(v) We should try a "multiple regression" model which includes "country" and "index" in the model.

[*Note:* although not explicitly in the syllabus, a comment to the effect that "Country" should be included as a qualitative variable (a "factor") e.g. by using a text vector (with entries "*E*", "*W*", "*S*" say) or a pair of (Bernoulli) dummy variables, may attract a bonus for a borderline candidate.]

END OF EXAMINERS' REPORT