EXAMINATION SOLUTIONS

STATISTICS 1

MATH 11400

(Paper Code MATH-43C)

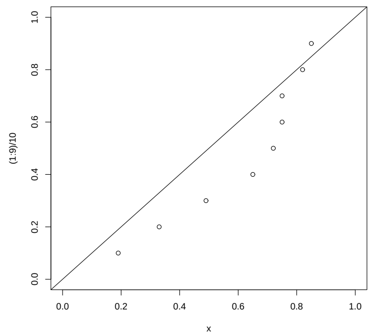
May/June 2010, 1 hours 30 minutes

A.1 (8 marks possible, standard methods, new data)

(a) (3 marks)

```
> stem(x,scale=2)
  The decimal point is 1 digit(s) to the left of the |
  1 | 9
  2 |
  3 | 3
  4 | 9
  5 |
  6 | 5
  7 | 255
  8 | 25
```

(b) (4 marks)



(c) (1 mark) The data appear to be skewed to the right compared to the uniform distribution.

A.2 (8 marks possible, seen)

(a) (3 marks) $m_1 = E(X) = \alpha/\lambda$; $m_2 = E(X^2) = \alpha(\alpha+1)/\lambda^2$.

(b) (5 marks)

$$\hat{\alpha} = \frac{m_1^2}{m_2 - m_1^2} \quad \text{and} \quad \hat{\lambda} = \frac{m_1}{m_2 - m_1^2}$$

Continued over...

A.3 (8 marks possible, first part seen)

(a) (6 marks) With $\ell(\theta)$ as the loglikeliood function, the MLE satisfies

$$\frac{\partial}{\partial\theta}\ell(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log p_X(x_i;\theta) = 0$$

i.e.

$$0 = \sum_i \left(-1 + \frac{x_i}{\theta}\right) = -n + \frac{\sum_i x_i}{\theta}$$

whose solution is evidently $\hat{\theta} = \overline{x}$. It is a *maximum* by inspection of the 2nd derivative or by noting simply that $\partial\ell/\partial\theta$ is decreasing in θ.

(b) (2 marks) By invariance, $\widehat{\exp(-\theta)} = \exp(-\hat{\theta}) = \exp(-\overline{x})$.

A.4 (8 marks possible, seen)

(a) (4 marks) $\mathrm{N}(0, 1/n)$ and χ^2_{n-1}. They may put sd not variance as second parameter of the Normal, so accept $\mathrm{N}(0, \sigma = 1/\sqrt{n})$ and penalise by 1 mark for $\mathrm{N}(0, 1/\sqrt{n})$

(b) (1 mark) Independent.

(c) (3 marks) t_{n-1}.

A.5 (8 marks possible, seen)

(a) (4 marks) Type I: rejecting the null hypothesis when in fact it is true; Type II: not rejecting the null hypothesis when in fact it is false.

(b) (4 marks) Significance level: prescribed acceptable ptobability of Type I error. Power: probability of rejecting null hypothesis under a prescribed alternative hypothesis (1− type II error probability).

Continued over...

B.1 (30 marks possible, standard methods, applied to an unseen example)

(a) (15 marks) The likelihood equation is $d\ell(\theta)/d\theta = 0$ where ℓ is the log-likelihood.

$$\ell(\theta) = n \log \theta + \log \prod_i x_i - (\theta/2) \sum_i x_i^2$$

$$0 = d\ell(\theta)/d\theta = n/\theta - (1/2) \sum_i x_i^2$$

so $\hat{\theta}_{\text{mle}} = 2n / \sum_{i=1}^n x_i^2$. That this is a maximum of the likelihood can be seen, for example, by noting that the 2nd derivative of ℓ is $-n/\theta^2$, so always negative.

(b) (6 marks) For $x > 0$, $F(x;\theta) = \int_{-\infty}^x f(u;\theta)du = 0 + \int_0^x \theta u \exp(-\theta u^2/2)du$ which after substitution $y = \theta u^2/2$ gives $\int_0^{\theta x^2/2} \exp(-y)dy = 1 - \exp(-\theta x^2/2)$ as required.
Set this to 1/2, and solve for x to get the median, which is clearly $M = \sqrt{2 \log 2/\theta}$.

(c) (3 marks) Solving $M = \sqrt{2 \log 2/\tilde{\theta}}$ clearly gives $\tilde{\theta} = 2 \log 2/M^2$.

(d) (6 marks) Both estimators seem to be centred approximately at the true value, but $\tilde{\theta}$ has greater spread and in particular a long upper tail, so is an inferior estimator compared to the MLE.

B.2 (30 marks possible, standard methods, unseen data and interpretation)

(a) (5 marks) If the two samples are paired (matched) then the paired comparison procedure is indicated; if unpaired then the two-sample test may be appropriate. Assumptions for the paired test: differences within pairs are i.i.d. Normally distributed. Assumptions for two-sample test - the two samples are independent, and each is drawn from an i.i.d. Normal population; in the standard case the variances of the two populations are equal; Welch's procedure allows an approximate test even if they are unequal.

(b) (20 marks) Pooled estimator is $S_p^2 = ((n-1)S_X^2 + (m-1)S_Y^2)/(n+m-2) = (\sum(X_i - \overline{X})^2 + \sum(Y_i - \overline{Y})^2)/(n+m-2)$. But $\sum(X_i - \overline{X})^2 = 74675 - 1219^2/20 = 376.95$ and $\sum(Y_i - \overline{Y})^2) = 80768 - 1416^2/25 = 565.76$ so $S_p^2 = (376.95 + 565.76)/43 = 21.9235$, $S_p = 4.6823$.
The t statistic is $(\overline{X} - \overline{Y})/(S_p\sqrt{1/n + 1/m}) = (60.95 - 56.64)/(4.6823 \times 0.3) = 3.0683$. The most appropriate alternative hypothesis is the two-sided one that the means are unequal, so the p-value is the probability that t on 43 degrees of freedom exceeds 3.0683 in absolute value, namely `2*pt(-3.0683,43)`= 0.0037. [*Answers around 0.004 are acceptable, as this is a reasonable limit of accuracy for visual interpolation in the supplied tables.*] So there is very strong evidence against the null hypothesis: men and women are not equally confident on average about their financial prospects.

(c) (5 marks) If data are collected from husband/wife couples, then they are obviously paired, and this should be taken care of in the analysis by using the paired comparison test; all other things being equal, this will be more sensitive since it will take account of the (very plausible) possibility that different couples will have jointly high or low confidence.

Continued over...

B.3 (30 marks possible, (a) is bookwork, (b) is standard method, unseen data)

(a) (10 marks) $\hat{\beta} \sim \mathrm{N}(\beta, \sigma^2/ss_{xx})$ and $(n-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n-2}$. These random variables are independent, so by the standard result that the ratio of a standard Normal to the square root of an independent χ^2 divided by its degrees of freedom has a t distribution with the same number of degrees of freedom, $(\hat{\beta}-\beta)/\sqrt{\hat{\sigma}^2/ss_{xx}} = (\hat{\beta}-\beta)/\sqrt{\sigma^2/ss_{xx}} \div \sqrt{\hat{\sigma}^2/\sigma^2} \sim t_{n-2}$.

(b) i. (15 marks) We have $ss_{xx} = 10686 - 294^2/10 = 2042.4$, $ss_{xy} = 388.2936 - 294 \times 13.4142/10 = -6.08388$, so $\hat{\beta} = -6.08388/2042.4 = -0.002979$. Also $\hat{\sigma}^2 = (ss_{yy} - ss^2_{xy}/ss_{xx})/(n-2)$; $ss_{yy} = 18.01371 - 13.4142^2/10 = 0.01963384$, so $\hat{\sigma}^2 = (0.01963384 - 6.08388^2/2042.4)/8 = 0.001511241/8 = 0.0001889$. The endpoints of the standard 95% confidence interval for β are $\hat{\beta} \pm t_{n-2;0.025} s_{\hat{\beta}}$ where $s_{\hat{\beta}} = \hat{\sigma}/\sqrt{ss_{xx}} = \sqrt{0.0001889/2042.4} = 0.0003041$. We have $t_{n-2;0.025} =$ `qt(0.975,8)` $=$ 2.306, so the endpoints are $-0.002979 \pm 2.306 \times 0.0003041$, or $(-0.003680, -0.002278)$.

ii. (5 marks) There is some evidence of curvature – positive residuals are associated with low and high fitted values, suggesting that the relationship between log fibrin and time is convex rather than linear.

End of solutions.