ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

**THE ROYAL STATISTICAL SOCIETY**

**2015 EXAMINATIONS − SOLUTIONS**

**GRADUATE DIPLOMA – MODULE 5**

The Society is providing these solutions to assist candidates preparing for the examinations in 2017.

The solutions are intended as learning aids and should not be seen as "model answers".

Users of the solutions should always be aware that in many cases there are valid alternative methods. Also, in the many cases where discussion is called for, there may be other valid points that could be made.

While every care has been taken with the preparation of these solutions, the Society will not be responsible for any errors or omissions.

The Society will not enter into any correspondence in respect of these solutions.

## Question 1

| | |
|---|---|
| **(i)** | The plots suggest that subjects with higher Sodium and Potassium levels and lower Carbon Dioxide levels are more likely to be Type B, but there is not much difference in Age between the groups [1 mark]. Some plots show good separation between the groups, notably Potassium vs Age and Potassium vs Carbon Dioxide [1 mark]. It appears that (some of) these variables will be useful for classification [1 mark]. |
| **(ii)** | With leaving-one-out cross-validation, each subject in turn is omitted from the dataset, and the discriminant function estimated from the remaining subjects' data is used to predict the class of the omitted one [1 mark]. The misclassification rates are estimated by comparing the true class of each subject with the predicted class when it is left out [1 mark]. This procedure reduces the downward bias in misclassification results that is found when the discriminant is estimated and used on the same dataset [1 mark]. In this example, linear discrimination seems likely to be very useful since the misclassification rates obtained, even with cross-validation, are low [1 mark]. Type B individuals are particularly well classified [1 mark]. |
| **(iii)** | LD-A = 2254.04 [1 mark].<br>LD-B = 2254.93 [1 mark].<br>Since LD-A < LD-B, classify the new individual as Type B [1 mark]. |
| **(iv)** | Linear discriminant analysis (LDA) assumes that the classification variables jointly have a multivariate normal distribution [1 mark], separately within the populations of Type A and Type B individuals [1 mark], though the procedure is robust against minor departures from this assumption [1 mark]. The population covariance matrices are assumed to be equal [1 mark]. Quadratic discriminant analysis (QDA) does not require the population covariance matrices to be equal [1 mark]. |
| **(v)** | Without cross-validation, the results with QDA seem better than with LDA [1 mark]. With cross-validation, however, the results with QDA are worse than with LDA especially for Type B individuals [1 mark]. Recommend using LDA [1 mark], since the cross-validated misclassification rates give a better guide to how the discriminants will perform with new patients [1 mark]. |

## Question 2

| | |
|---|---|
| **(i)** | The purpose of cluster analysis is to determine whether a set of objects may be meaningfully divided into different classes [1 mark] and, if so, to determine those classes [1 mark] using similarities and dissimilarities between the objects [1 mark]. |
| **(ii)** | Euclidean distance: $\left(\sum_{i=1}^{p}(x_i - y_i)^2\right)^{1/2}$ [1 mark].<br>*Candidates should define and discuss two of the following, for 2 marks each.*<br>Squared Euclidean distance $=\sum_{i=1}^{p}(x_i - y_i)^2$ [1 mark]. This will have the effect of increasing the distances between pairs of observations, making them less likely to cluster together and hence increase the number of clusters [1 mark].<br>Manhattan distance $=\sum_{i=1}^{p}\lvert x_i - y_i \rvert$ [1 mark]. This will have the effect of decreasing the distances between pairs of observations, making them more likely to cluster together and hence decrease the number of clusters [1 mark].<br>Chebyshev distance $=\max_i \lvert x_i - y_i \rvert$ [1 mark]. This will have the effect of increasing the distances between pairs of observations, making them less likely to cluster together and hence increase the number of clusters [1 mark].<br>Mahalanobis distance $=\left([\underline{\boldsymbol{x}} - \underline{\boldsymbol{y}}]^{\mathrm{T}}\boldsymbol{S}^{-1}[\underline{\boldsymbol{x}} - \underline{\boldsymbol{y}}]\right)^{1/2}$ [1 mark]. It is not clear in general what effect this will have on the number of clusters as that will depend on the pattern of correlations between variables [1 mark]. |
| **(iii)** | Both methods of clustering begin with all objects in their own cluster [1 mark]. At the first step, pairs of objects whose distance is less than a certain threshold value are then combined into a new cluster [1 mark] and this process continues until no distances between clusters are below the threshold [1 mark]. In single-linkage, at later steps the distance between two clusters is determined by the distance of the two closest objects (nearest neighbours) in the different clusters [1 mark]. In complete-linkage, the distance between two clusters is determined by the greatest distance between any two objects in the different clusters [1 mark]. As a result, single-linkage will usually give more clusters than complete-linkage [1 mark]. |
| **(iv)** | *k* means clustering starts by randomly generating *k* clusters [1 mark], determining the location of the cluster centre then assigning each point to the cluster whose centre is closest to it [1 mark], iteratively re-computing the new cluster centres until convergence occurs [1 mark], which is identified by point-cluster assignments no longer changing [1 mark]. The idea is to move objects between clusters with the goal of minimising variability within clusters and maximising variability between clusters. A major disadvantage of this method is that *k* must be set before the cluster analysis is started [1 mark], but it is often computationally faster that hierarchical methods [1 mark]. |

## Question 3

**(i)**

$$S(t) = \int_t^{\infty} \lambda\gamma t^{\gamma-1} e^{-\lambda t^{\gamma}} dt \qquad [1\text{ mark}]$$

$$= \int_{\lambda t^{\gamma}}^{\infty} e^{-u} du \qquad \text{setting } u = \lambda t^{\gamma} \qquad [1\text{ mark}]$$

$$= \left[-e^{-u}\right]_{\lambda t^{\gamma}}^{\infty} = e^{-\lambda t^{\gamma}} \qquad [1\text{ mark}]$$

**(ii)**

$$\log S(t) = \log\left(e^{-\lambda t^{\gamma}}\right) = -\lambda t^{\gamma} \qquad [1\text{ mark}]$$

$$\log\left[-\log S(t)\right] = \log\left[\lambda t^{\gamma}\right] \qquad [1\text{ mark}]$$

$$= \log(\lambda) + \gamma \log(t) \qquad [1\text{ mark}]$$

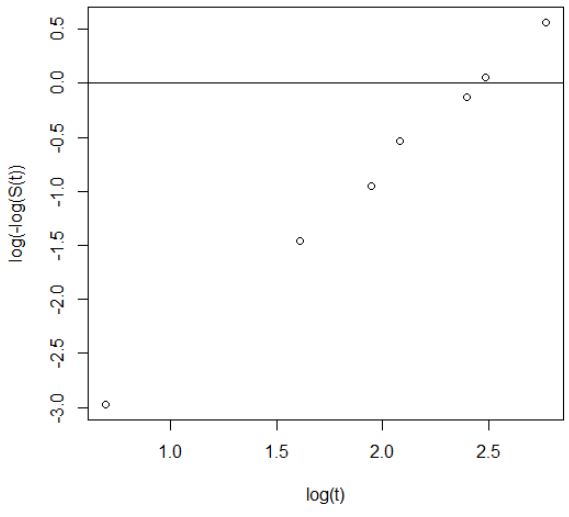
So, the intercept would be log(λ) [1 mark] and the slope γ [1 mark].

**(iii)**

| $t_i$ | $n_i$ | $d_i$ | $\frac{n_i - d_i}{n_i}$ | $\hat{S}(t_i)$ | $\log\left\{-\log\left[\hat{S}(t_i)\right]\right\}$ |
|---|---|---|---|---|---|
| 2 | 20 | 1 | 19/20 | 0.950 | -2.97 |
| 5 | 18 | 3 | 15/18 | 0.792 | -1.46 |
| 7 | 14 | 2 | 12/14 | 0.679 | -0.949 |
| 8 | 11 | 2 | 9/11 | 0.555 | -0.530 |
| 11 | 8 | 2 | 6/ 8 | 0.416 | -0.131 |
| 12 | 6 | 1 | 5/ 6 | 0.347 | 0.057 |
| 16 | 2 | 1 | 1/ 2 | 0.173 | 0.562 |

[The final column of the table above is not required until part (iv). ½ mark for method + ½ mark for each correct row leading to a correct value of $S(t)$.]

**(iv)**



## Question 3

| | |
|---|---|
| **(iv)** | **… continued**<br><br>[**1 mark** for correct method of calculating log(-log($S(t)$)) values, **1 mark** for correct values, **1 mark** for appropriate graph, **1 mark** for correct points on graph]<br><br>The points lie approximately on a straight line, so a Weibull model seems appropriate [1 mark]. |
| **(v)** | An estimate of $\gamma$, the slope, would be about 3.5/2.5 = 1.4 [1 mark]. An estimate of the intercept, log($\lambda$), would be about -3.0 – 0.7 = -3.7 [1 mark], giving an estimate of $\lambda$ of about 0.025 [1 mark]. |

## Question 4

**(i)**

The model is:

$$h(t;\underline{z}) = h_0(t).exp\{\underline{\beta}^T \underline{z}\} = h_0(t).exp\{\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4\} \quad \text{[1 mark]}$$

Here, $h_0(t)$ is the baseline hazard function for subjects who are given standard care, are male and not white, and have 0% burns [1 mark]. $z_1$ is an indicator variable for Treatment, so $\beta_1$ summarises the average effect of being treated with additional care rather than standard care. $z_2$ is an indicator variable for Gender, so $\beta_2$ summarises the average effect of being Female rather than Male. $z_3$ is an indicator variable for Race, so $\beta_3$ summarises the effect of being White rather than Non White. *Idea of an indicator variable, 1 mark; idea of an average difference in effect, 1 mark.* $z_4$ is the percentage burns so $\beta_4$ summarises the average effect of each additional % of body surface area affected [1 mark]. *Alternative, equivalent parameterisations should also be given full marks.*

**(ii)**

$b_1/\text{se}(b_1) = -0.606/0.296 = -2.05 < -1.96$, so Treatment is a useful explanatory variable [1 mark]. Since the sign of $b_1$ is negative, additional care will reduce the hazard of infection (all other factors being equal) [1 mark].

$b_2/\text{se}(b_2) = -0.631/0.390 = -1.62 > -1.96$, so Gender is not a useful explanatory variable [1 mark]. There is no significant difference on average between the hazard for Male and Female subjects (all other factors being equal) [1 mark].

$b_3/\text{se}(b_3) = 2.12/1.01 = 2.10 > 1.96$, so Race is a useful explanatory variable [1 mark]. Since the sign of $b_3$ is positive, White subjects have a higher hazard of infection (all other factors being equal) [1 mark].

$b_4/\text{se}(b_4) = 0.00404/0.00703 = 0.57 < 1.96$, so Severity is not a useful explanatory variable [1 mark]. There is no significant difference on average between the hazard for subjects with different initial severity levels (all other factors being equal) [1 mark].

**(iii)**

The hazard ratio is $\exp\{\beta_1\}$ [1 mark]. A 95% CI for $\beta_1$ is:

$b_1 \pm 1.96\ \text{se}(b_1)$, i.e. $-0.606 \pm (1.96 \times 0.296)$, i.e. $(-1.186, -0.026)$ [1 mark]

Therefore, a 95% CI for the hazard ratio is (0.306, 0.974) [1 mark].

**(iv)**

Add 3 new indicator variables to the model, corresponding to any 3 of the causes [1 mark]. For example, one suitable variable would be coded 1 if the cause was chemical and 0 otherwise [1 mark].

**(v)**

This might have been done because there were so few instances of each of the other causes, or because there were clinical reasons for thinking that the effects of the other kinds of burn would be similar [1 mark]. The effect of the new variable is not significant: $-0.481/0.330 = -1.46 > -1.96$ [1 mark].

## Question 5

**(i)**

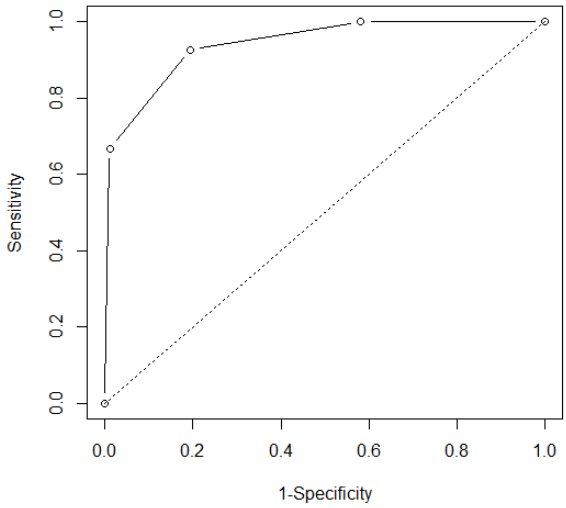
| | True + | True - |
|---|---|---|
| **Test +** | 18 | 1 |
| **Test -** | 9 | 92 |

Sensitivity = $^{18}/_{27}$ = 0.667 [1 mark]
Specificity = $^{92}/_{93}$ = 0.989 [1 mark]
PPV = $^{18}/_{19}$ = 0.947 [1 mark]
NPV = $^{92}/_{101}$ = 0.911 [1 mark]

PPV and NPV are both high, meaning that patients classified as positive or negative for hypothyroidism are very likely to be positive or negative respectively [1 mark], which is important for patients [1 mark]. Specificity is high, meaning that almost all patients who are truly negative are classified as negative by this test [1 mark]. Sensitivity is only moderate so almost one-third of truly positive patients will be classified as negative and this makes the test less useful than it might be [1 mark].

**(ii)**

*c* = 7: sensitivity = 25/27 = 0.926, specificity = 75/93 = 0.806 [2 marks]
*c* = 9: sensitivity = 27/27 = 1 , specificity = 39/93 = 0.419 [2 marks]




[1 mark for correct method, 1 mark for correct axis labels, 1 mark for correct points, 1 mark for $y=x$ line]

Recommend c ≈ 7 [1 mark] as that achieves the best balance between sensitivity and specificity [1 mark] in the absence of information about costs of misdiagnosis and prevalence.

**(iii)**

The "ideal" ROC curve has AUC = 1 [1 mark], so a value of 0.86 suggests that this procedure is potentially useful for diagnosis [1 mark].

## Question 6

**(i)** In a case-control study, cases (observed to have an outcome of interest) are identified, along with appropriate controls (observed to be free of the outcome) [1 mark]. Enquiries are them made to find out whether or not each subject in each group had previously been exposed to a certain risk factor, and the frequency of exposure is compared between the groups [1 mark]. One advantage of case-control studies relative to prospective studies (such as cohort studies) is that results can be obtained more quickly, without having to wait for some subjects to develop the outcome [1 mark]. Case-control studies are particularly beneficial for studying rare outcomes, since other types of studies would need a large cohort to be enrolled in order to generate sufficient outcomes for analysis [1 mark].

**(ii)** The main advantage of a matched case-control study is an improvement in the precision for estimating effect sizes (or an increase in power for statistical tests) [1 mark]. This arises because the subjects in a matched pair are similar in respect of key variables and the matching reduces residual variance [1 mark]. A disadvantage is that it can be difficult to find controls who exactly match each case [1 mark] and this can increase the cost or extend the time required to carry out the study [1 mark].

**(iii)** For the unmatched analysis, require the following table.

| | **Smoker** | **Non-Smoker** | |
|---|---|---|---|
| **Control** | 37 | 130 | |
| **Case** | 55 | 112 | [1 mark for table, possibly implied] |

Unmatched OR=$\frac{55/112}{37/130}$=1.73 [1 mark]

Matched OR=$\frac{40}{22}$=1.82 [1 mark for method, 1 mark for correct value]

**(iv)** $H_0$: there is no association between mother's smoking behaviour and low birth weight [1 mark]

For the unmatched analysis, expected frequencies are: 46, 121, 46, 121 [1 mark].

So: $\chi^2=9^2\left(\frac{2}{46}+\frac{2}{121}\right)$=4.86 [1 mark for formula, 1 mark for correct final value]

Since $\chi^2 > 3.841$ reject $H_0$ at 5% level [1 mark].

For McNemar's Test, $\chi^2=\left(\frac{|40-22|-1}{40+22}\right)^2$ =4.66 [1 mark for formula, 1 mark for correct final value]

Since $\chi^2 > 3.841$ reject $H_0$ at 5% level [1 mark].

## Question 7

**(i)** Consider any element, $x'$, in the population. Number of different possible samples that include $x'$ = $\binom{N-1}{n-1}$ [1 mark]. Each of these samples has probability $\binom{N}{n}$ of being chosen. So, probability that $x'$ is chosen = $\frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$ [1 mark].

**(ii)**

$$E(\bar{x}) = E\left\{\frac{1}{n}\sum_{i=1}^{n} x_i\right\} = \sum_{j=1}^{N} \frac{x_j}{n} P(x_j \text{ in sample}) \qquad [1\text{ mark}]$$

$$= \sum_{j=1}^{N} \frac{x_j}{n}\frac{n}{N} = \frac{1}{N}\sum_{j=1}^{N} x_j = \mu \qquad [1\text{ mark}]$$

**(iii)**

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\mathrm{Var}\left\{\sum_{i=1}^{n} x_i\right\} \qquad [1\text{ mark}]$$

$$= \frac{1}{n^2}\left\{n\sigma^2 + n(n-1)\tau\right\}, \text{where } \tau = \mathrm{Cov}(x', x) \qquad [1\text{ mark}]$$

Now $\tau = E(x'x) - \mu^2$ [1 mark]

and $E\left\{(x')^2\right\} = \sigma^2 + \mu^2$ [1 mark]

But $E\left\{(x')^2\right\} = E\left\{x'\left[N\mu - \sum_{x_j \neq x'} x_j\right]\right\}$

$= N\mu^2 - (N-1)E(x'x)$ [1 mark]

So $\sigma^2 + \mu^2 = N\mu^2 - (N-1)E(x'x)$

i.e. $E(x'x) = \mu^2 - \sigma^2/(N-1)$ [1 mark]

i.e. $\tau = -\sigma^2/(N-1)$ [1 mark]

$$\mathrm{Var}(\bar{x}) = \frac{1}{n^2}\left\{n\sigma^2 - \frac{n(n-1)\sigma^2}{N-1}\right\} = \frac{\sigma^2}{n}\frac{N-n}{N-1} \qquad [1\text{ mark}]$$

## Question 7

**(iv)**

$$2\sqrt{\text{Var}(\bar{x})} = B$$

i.e. $$2\sqrt{\frac{\sigma^2}{n}}\sqrt{\frac{N-n}{N-1}} = B$$ [1 mark]

i.e. $$\frac{\sigma^2}{n}\frac{N-n}{N-1} = \frac{B^2}{4}$$ [1 mark]

i.e. $$\sigma^2 N - \sigma^2 n = \frac{B^2}{4}(N-1)n$$

i.e. $$\sigma^2 N = \left\{\frac{B^2}{4}(N-1) + \sigma^2\right\}n$$ [1 mark]

i.e. $$n = \frac{4\sigma^2 N}{B^2(N-1) + 4\sigma^2}$$ [1 mark]

**(v)**

This is an example of part (iv) with $N = 1000$, $\sigma^2 = 100$, $B = 2$ [1 mark].

So: $$n = \frac{4 \text{ x } 100 \text{ x } 1000}{(4 \text{ x } 999) + (4 \text{ x } 100)} = 90.99$$ [1 mark substitution, 1 mark value]

Need to sample at least 91 chickens. [1 mark]

## Question 8

**(i)**

| $i$ | $N_i$ | $\sigma_i$ | $c_i$ | $N_i\sigma_i/\sqrt{c_i}$ |
|---|---|---|---|---|
| 1 | 1000 | 1550 | 1.5 | 1265570 |
| 2 | 1100 | 2150 | 1.5 | 1931014 |
| 3 | 1600 | 2950 | 1.5 | 3853864 |
| 4 | 600 | 1100 | 1 | 660000 |
| Total | 4300 | | | 7710448 |

For proportional allocation, $n_i \approx \frac{N_i}{N} n$ [1 mark], giving 116, 128, 186, 70 [1 mark].

For optimal allocation, $n_i \approx \frac{N_i \sigma_i / \sqrt{c_i}}{\sum N_i \sigma_i / \sqrt{c_i}} n$ [1 mark], giving 82, 125, 250, 43

[1 mark for $N_i\sigma_i/\sqrt{c_i}$ values, 1 mark for $n_i$ values].

**(ii)**

$Var(\bar{x}_{prop}) = \sum \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i} = 10297.86$ [1 mark formula, 1 mark value]

$Var(\bar{x}_{opt}) = \sum \left(\frac{N_i}{N}\right)^2 \frac{\sigma_i^2}{n_i} = 9372.01$ [1 mark value]

$Var(\bar{x}_{srs}) = \frac{\sigma^2}{n} = \frac{3750^2}{500} = 28125$ [1 mark formula, 1 mark value]

$eff(\bar{x}_{prop}) = \frac{28125}{10297.86} = 2.73$ [1 mark formula, 1 mark value]

$eff(\bar{x}_{opt}) = \frac{28125}{9372.01} = 3.00$ [1 mark value]

Prefer optimal allocation which has the highest relative efficiency [1 mark].

**(iii)**

Without knowing the $N$ and $N_i$'s of interest in advance, the correct sample sizes cannot be calculated for either stratified sampling method [1 mark]. Using the wrong weights would introduce bias into the estimation of the overall mean income [1 mark]. It might be better to use simple random sampling which is unbiased [1 mark].

**(iv)**

It seems likely that graduate level jobs will have systematically higher salaries, so a further improvement in relative efficiency could be made by taking this categorisation into account [1 mark]. If the proportion of graduate-level jobs in each stratum were known in advance, then post-hoc stratification could be applied to the data collected [1 mark]. However, it seems that this information is not available and it would be dangerous to assume that the sample proportions would be good estimates of the population proportions, due to potential self-selection bias, so this further stratification cannot sensibly be carried out [1 mark].