HONG KONG STATISTICAL SOCIETY

2016 EXAMINATIONS − SOLUTIONS

HIGHER CERTIFICATE – MODULE 6

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.

1. (i) Full randomisation would mean randomly allocating subjects to each package/office combination, but in this case the office they work at is fixed so randomisation means randomly allocating the three subjects in each office to the three packages. **(1)**
This increases the chances of a fair comparison between the packages by reducing the chance of any systematic allocation method biasing the results due to unforeseen connections between it and any (possibly unmeasured, underlying) variable that might affect learning quality. **(1)**
If this is not done then the results could easily be misleading if the non-random method of allocation of subjects to package has led to people who might learn more or less being unevenly distributed. **(1)**
For example if the oldest subject is allocated to A and the youngest to C then this confounds the effect of package with that of age and/or experience. **(1)**

(ii) The experimental units (subjects) are divided into blocks where there is reason to believe that observations from within the same block will be more similar than those from different blocks. **(1)**
In this case blocking occurs automatically as people will be tested in their own office, and randomisation to package is performed separately for each block. **(1)**
Different offices might well differ systematically from each other, and blocks can be included as an extra term in the analysis, thus reducing the residual variance and hence increasing the power of the tests to distinguish between packages. **(1)**
In addition it is highly likely that past experience or job type will also affect how well people learn with each package, so that if all of the subjects who do one job end up allocated to package A and all those who do another job are allocated to package B then the effect of package could be confounded with that of job. **(1)**
Hence job type could be treated as another type of blocking to ensure fair allocation of different jobs to different packages, although this would require more observations or a latin square-type design. **(1)**
NOTE: Any sensible explanation, using jobs or experience or similar, is fine here.

(iii) An appropriate model is

$$Y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where Y_{ij} is the score for the individual from the j-th office using the i-th package, μ is the overall mean, τ_i the effect of the i-th package, β_j the effect of the j-th office **(1)**
and $\epsilon_{ij} \sim N(0, \sigma^2)$ is the random variation term, i.e. random variation is normally distributed with constant variance. **(1)**
The assumption of normality and unconstrained parameters may be dubious for a score out of 100 because there are fixed upper and lower limits, but if the score is the sum over many questions and the results are not too close to 0 or 100 then the CLT should make normality reasonable and the lack of constraints on the parameter estimates will be irrelevant. **(1)**

(iv)

$$TSS = 50357 - cfm = 50357 - \frac{837^2}{15} = 50357 - 46704.60 = 3652.40 \quad \textbf{(1)}$$

$$TrSS = \frac{340^2 + 233^2 + 264^2}{5} - cfm = 47917.00 - 46704.60 = 1212.40 \quad (\mathbf{1})$$

$$BlSS = \frac{156^2 + 183^2 + 165^2 + 181^2 + 152^2}{3} - cfm = 46971.67 - 46704.60 = 267.07 \quad (\mathbf{1})$$

$$RSS = TSS - TrSS - BlSS = 3652.40 - 1212.40 - 267.07 = 2172.93 \quad (\mathbf{1})$$

Hence

Source	Sum of squares	d.f.	Mean Square	F ratio
Treatments	1212.40	2	606.20	2.23
Blocks	267.07	4	66.77	0.25
Residual	2172.93	8	271.62	
Total	3652.40	14		

(d.f.) **(1)**

To test $H_0 : \tau_1 = \tau_2 = \tau_3$ versus $H_1 :$ not H_0 **(1)**
at the 5% level we compare $F_{obs} = 2.23$ to the upper 5% point $F_{crit} = F_{2,8} = 4.46$ **(1)**.
Here $F_{obs} < F_{crit}$ so we do not reject H_0 at the 5% level and conclude that there is no evidence of any differences in mean test score between the packages. **(1)**

2. (a) (i) The obvious design is a 4 by 4 factorial, 4 diets by 4 intervals. Each combination will be allocated to one or more individuals, making it a (full) factorial design. **(1)**

Each observation helps to compare both diets and intervals, whereas one-factor-at-a-time experimention would only compare diets or intervals but not both. **(1)**

A factorial structure therefore allows several factors to be investigated at once using the same resources, so compared to one-factor-at-a-time comparisons, needs fewer observations to make the same comparisons at the same level of precision. **(1)**

In addition, if there are interactions between the factors then they can be detected. **(1)**

Hence in this case a factorial design will give more powerful tests to distinguish between different diets and intervals, as well as allowing discovery of whether some intervals between consultations are more or less helpful with different diets. **(1)**

(ii) An interaction between a diet and an interval would be an effect (positive or negative) of a diet/interval combination over and above the main effects of that diet and that interval. **(1)**

These can only be estimated if there are two or more observations for each combination. **(1)**

This is because the within-cell variation can be used to discriminate between interactions and high residual variance, which cannot be done with a single observation per cell. **(1)**

NOTE: Overlap of answers between (i) and (ii) is hard to avoid, so marks will be awarded wherever the comments are made.

(iii) Blinding is when the trial subject does not know which trial group they are in, such as the new treatment or a standard treatment. Hence the knowledge of this will not affect the results, helping to ensure a fair comparison. **(1)**

However, in this case the trial subject cannot be blind to consultation interval, and it would be very difficult to blind them to diet unless the difference between the diets is tiny. Hence blinding is unlikely to be possible in this case. **(1)**

(b) The (unstandardised) residual for the i-th point is defined as

$$e_i = y_i - \hat{y}_i$$

where $\hat{y}_i$ is the fitted value from the model. **(1)**

These have zero mean, but the variance of e_i is not constant. Hence in practice we usually use standardised residuals r_i, derived by dividing e_i by its estimated standard deviation. Hence if the model is a good fit then approximately $r_i \sim N(0, 1)$ independently. **(1)**

The following marks are still available if the candidate did not mention standardisation.

Common diagnostic plots are:

Residuals versus fitted values. This is usually the most important. There should be a random scatter of residuals about zero throughout the range of fitted values.

(1)

If, for example, the residuals are mostly negative at first, then mostly positive, then mostly negative, then this suggests that a linear fit in at least one variable should in fact be a quadratic fit. **(1)**

Another common problem is that the residuals have a 'trumpet' shape, with the variance increasing for larger fitted values. This shows that the assumption of common variance is incorrect. **(1)**

This can often be solved by a data transformation, commonly using $\log Y$ or $\sqrt{Y}$ as the response variable. In other cases taking a log transform of all response and predictor variables will convert an additive to a multiplicative relationship.

Residuals versus predictors. Similar reasoning to the above, but just looking at one predictor variable. **(1)**

Residuals versus observation number. This only makes sense if the observation number represents something, commonly the order in which observations were taken. **(1)**

We hope for a random scatter of residuals about zero throughout the range of observation numbers. If, for example, the residuals are mostly negative at first and positive at the end then this suggests that the model is overestimating at first and underestimating at the end. This may be due to the effect of an unobserved variable which changes over the period. **(1)**

Normal probability plot of residuals. The ordered residuals are plotted against the cumulative distribution function of the standard normal distribution, so that a straight line of $y = x$ supports the assumption of normality. **(1)**

Non-normality often manifests as points being below the line at one end and above at the other, usually indicating that the distribution of the residuals is more skewed than expected under normality. **(1)**

A histogram of the residuals performs a similar job.

(Marks also for equivalent examples/illustrations)

3. (i) If the system is in control then the mass of a component is described by the random variable X where $X \sim N(132, 4^2)$. Hence the random variable $\bar{X}$ describing the mean length of a sample of size 5 has variance $4^2/5 = 3.2 = 1.789^2$, so that $\bar{X} \sim N(132, 1.789^2)$. **(1)**

Warning limits are defined such that, if the system is in control, a fraction 0.05 of observations will be outside these limits. Therefore if an observation from an in-control system is selected at random then the probability of both it and the next observation being outside the warning limits is $0.05^2 = 0.0025$ (assuming independence). Hence two successive observations outside the warning limits suggest a possible problem. **(1)**

Action limits are defined such that, if the system is in control, a fraction 0.001 of observations will be outside these limits. Hence any observation outside the action limits suggests a possible problem. **(1)**

Therefore the 95% warning limits are

$$132 \pm Z_{0.975} \times 1.789 = 132 \pm 1.96 \times 1.789 = 132 \pm 3.506 \quad (\mathbf{1})$$

while the 99.9% action limits are

$$132 \pm Z_{0.9995} \times 1.789 = 132 \pm 3.29 \times 1.789 = 132 \pm 5.885 \quad (\mathbf{1})$$

The system is assumed to be in control otherwise, so that we try to avoid false positives for out of control, and hence avoid trying to correct for problems that are not in fact there. **(1)**

(ii) The 3nd batch mean (136.29) is the first to be outside the warning limits, but the 4th (135.41) is just inside them so no conclusions are reached. **(1)**
Similarly the 5th (136.44) is just outside and the 6th (135.17) is just inside, so again no conclusions are reached despite having two outside the warning limits, because they were not consecutive. **(1)**
However, batch 7 (138.01) is outside the action limits so we would deem the system out of control from batch 7 onwards. **(1)**

(iii) A cusum chart plots the cumulative sum of the difference between actual and target values of some quantity against time or observation number. **(1)**

Hence for a target value of k for some quantity x, it plots $\sum_{i=1}^{t}(x_i - k)$ or $\sum_{i=1}^{t}(\bar{x}_i - k)$ versus $t = 1 \ldots$. **(1)**

If the mean of the system really is k then the chart simply shows random variation about k, so that $\sum_{i=1}^{t}(\bar{x}_i - k)$ randomly varies about zero, **(1)**

but if it moves away from k then this should show up fairly quickly because every value collected contributes to the cumulative sum, driving the running mean away from k and hence the value of $\sum_{i=1}^{t}(\bar{x}_i - k)$ away from zero. **(1)**

The steepness of the gradient of the graph indicates how far the mean is away from k while increasing steepness shows that the mean is moving further away from k. However, early values well away from k will stay in the calculation even

after the system is brought under control, so the calculation may need to be reset sometimes. **(1)**

If the system is producing items which are on average very slightly longer than k, but within acceptable tolerances, then the cusum will still show $\sum_{i=1}^{t}(\bar{x}_i - k)$ steadily increasing against t (or similarly for the mean slightly below k). **(1)**

In comparison to non-cumulative plots of means, such as Shewhart charts, a cusum chart will tend to indicate a small change in mean more quickly, but may be slightly slower to pick up a sudden large change. **(1)**

NOTE: Explanation can be in terms of x or $\bar{x}$.

(iv) The cumulative scores are

$$0.87, 2.39, 6.68, 10.09, 14.53, 17.70, 23.71, 28.26, 34.12 \quad (\mathbf{1})$$

which are plotted against 1-9 respectively. **(1)**

These are strictly increasing, showing that every mean is above the target $k = 132$, so that the system is clearly not randomly varying about $k = 132$, and the graph of cusum against time shows a slightly increasing slope, suggesting that the problem may be getting worse. **(1)**

Diagram. **(1)**

NOTE: There is inevitable overlap between the answers so the candidate should receive the marks wherever the appropriate comment appears.

4. (i) The random variable ϵ represents the level of random variation around the presumed linear relationship, **(1)**
and we assume $\epsilon \sim N(0, \sigma^2)$. To be more precise, let ϵ_i be the random variable which represents random variation around the i-th point, then we assume that $\epsilon_i \sim N(0, \sigma^2)$ independently. **(1)**

(ii) In ordinary least squares estimation (approximation) we estimate α, β, γ by $\hat{\alpha}, \hat{\beta}, \hat{\gamma}$, the values which minimise the residual sum of squares

$$\begin{aligned} R &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \alpha - \beta w_i - \gamma x_i)^2 \quad \textbf{(1)} \end{aligned}$$

Hence differentiate with respect to each parameter and set to zero, thus forming the three normal equations to be solved.

$$\frac{\partial R}{\partial \alpha} = -2 \sum_{i=1}^{n} (y_i - \alpha - \beta w_i - \gamma x_i) \quad \textbf{(1)}$$

$$\frac{\partial R}{\partial \beta} = -2 \sum_{i=1}^{n} w_i (y_i - \alpha - \beta w_i - \gamma x_i) \quad \textbf{(1)}$$

$$\frac{\partial R}{\partial \gamma} = -2 \sum_{i=1}^{n} x_i (y_i - \alpha - \beta w_i - \gamma x_i) \quad \textbf{(1)}$$

Simplifying and equating to zero, we have the equations

$$\begin{aligned} \sum y_i - n\hat{\alpha} - \hat{\beta} \sum w_i - \hat{\gamma} \sum x_i &= 0 \quad \textbf{(1)} \\ \sum w_i y_i - \hat{\alpha} \sum w_i - \hat{\beta} \sum w_i^2 - \hat{\gamma} \sum w_i x_i &= 0 \quad \textbf{(1)} \\ \sum x_i y_i - \hat{\alpha} \sum x_i - \hat{\beta} \sum w_i x_i - \hat{\gamma} \sum x_i^2 &= 0 \quad \textbf{(1)} \end{aligned}$$

(iii) Indicator or dummy variables only take the values zero or one and are used to represent categorical variables. It takes the value one when an observation belongs to the appropriate category (e.g. group) or 0 otherwise. **(1)**
A categorical variable with k categories can hence be modelled using $k-1$ dummy variables, where usually one category (commonly first or last) is treated as the reference category and membership of it is indicated by all the indicator variables being zero. **(1)**
Hence the one-way ANOVA model can be written in terms of indicator variables, with group 3 as the reference category, by defining w_i to be 1 when the i-th observation belongs to group 1, zero otherwise, **(1)**
and x_i to be 1 when the i-th observation belongs to group 2, zero otherwise. **(1)**

(iv) Hence, if n_j is the number of observations in group j, so that $\sum w_i = n_1$ and $\sum x_i = n_2$, and $\sum y_{ij}$ is the sum of the observations in group j, then

$$\begin{aligned} \sum y_i - n\hat{\alpha} - \hat{\beta} n_1 - \hat{\gamma} n_2 &= 0 \quad \textbf{(1)} \\ \sum y_{i1} - \hat{\alpha} n_1 - \hat{\beta} n_1 &= 0 \quad \textbf{(1)} \\ \sum y_{i2} - \hat{\alpha} n_2 - \hat{\gamma} n_2 &= 0 \quad \textbf{(1)} \end{aligned}$$

so that from the second and third equations

$$\begin{aligned} \bar{y}_1 &= \hat{\alpha} + \hat{\beta} \quad (\mathbf{1}) \\ \bar{y}_2 &= \hat{\alpha} + \hat{\gamma} \quad (\mathbf{1}) \end{aligned}$$

so that plugging back into the first equation and using $\sum n_j = n$

$$\begin{aligned} \sum y_i &= n\hat{\alpha} + (\bar{y}_1 - \hat{\alpha})n_1 + (\bar{y}_2 - \hat{\alpha})n_2 \\ &= n_3\hat{\alpha} + \sum y_{i1} + \sum y_{i2} \quad (\mathbf{1}) \\ \Rightarrow \sum y_{i3} &= n_3\hat{\alpha} \\ \Rightarrow \hat{\alpha} &= \bar{y}_3 \\ \Rightarrow \hat{\beta} &= \bar{y}_1 - \bar{y}_3 \\ \hat{\gamma} &= \bar{y}_2 - \bar{y}_3 \quad (\mathbf{1}) \end{aligned}$$

as expected.