ROYAL STATISTICAL SOCIETY

DATA | EVIDENCE | DECISIONS

THE ROYAL STATISTICAL SOCIETY

2016 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.

MODULE 5: Topics in Applied Statistics

1. Consider a study of the association between a binary outcome (disease versus no disease) and a binary risk factor (exposure to a risk factor versus no exposure). Let D and $\overline{D}$ denote the presence and absence of disease and E and $\overline{E}$ denote the presence and absence of exposure.

   (i) Define the following concepts using conditional probabilities.

   (a) Relative risk for disease.

   (2)

$$\text{relative risk for disease} = \frac{\Pr(D \mid E)}{\Pr(D \mid \overline{E})}$$

Marks are indicated in square brackets, e.g., **[1]** denotes one mark.
[1] for correct numerator of right-hand side.
[1] for correct denominator of right-hand side.

   (b) Odds ratio for a case-control study.

   (2)

$$\frac{\text{odds}(E \mid D)}{\text{odds}(E \mid \overline{D})} = \frac{\Pr(E \mid D)/\Pr(\overline{E} \mid D)}{\Pr(E \mid \overline{D})/\Pr(\overline{E} \mid \overline{D})}$$

[1] for correct numerator of right-hand side.
[1] for correct denominator of right-hand side.
lose one mark if only left-hand side is stated.

   (c) Odds ratio for a cohort study.

   (2)

$$\frac{\text{odds}(D \mid E)}{\text{odds}(D \mid \overline{E})} = \frac{\Pr(D \mid E)/\Pr(\overline{D} \mid E)}{\Pr(D \mid \overline{E})/\Pr(\overline{D} \mid \overline{E})}$$

[1] for correct numerator of right-hand side.
[1] for correct denominator of right-hand side
lose one mark if only left-hand side is stated.

   (ii) Prove that the odds ratio for a case-control study equals the odds ratio for a cohort study. Explain why this result is important in medical studies of the association between a dichotomous outcome and a dichotomous risk factor.

   (8)

The odds ratio for a case-control study is:

$$\frac{\text{odds}(E \mid D)}{\text{odds}(E \mid \overline{D})} = \frac{\Pr(E \mid D)/\Pr(\overline{E} \mid D)}{\Pr(E \mid \overline{D})/\Pr(\overline{E} \mid \overline{D})} = \frac{\Pr(E,D)/\Pr(\overline{E},D)}{\Pr(E,\overline{D})/\Pr(\overline{E},\overline{D})} = \frac{\Pr(E,D)\Pr(\overline{E},\overline{D})}{\Pr(E,\overline{D})\Pr(\overline{E},D)}$$

which is the cross-product ratio of the cell probabilities of the two-by-two table.

[1] for correct numerator of right-hand side.

[1] for correct denominator of right-hand side.

The odds ratio for a cohort study is:

$$\frac{\text{odds}(D \mid E)}{\text{odds}(D \mid \overline{E})} = \frac{\Pr(D \mid E)/\Pr(\overline{D} \mid E)}{\Pr(D \mid \overline{E})/\Pr(\overline{D} \mid \overline{E})} = \frac{\Pr(D,E)/\Pr(\overline{D},E)}{\Pr(D,\overline{E})/\Pr(\overline{D},\overline{E})} = \frac{\Pr(D,E)\Pr(\bar{E},\overline{D})}{\Pr(D,\overline{E})\Pr(\overline{D},E)}$$

which is the cross-product ratio of the cell probabilities of the two-by-two table.

[1] for correct numerator of right-hand side.

[1] for correct denominator of right-hand side.

Hence, the result is obtained as these two odds ratios are equal.

Explain why this result is important in medical studies of the association between a dichotomous outcome and a dichotomous risk factor.

- Result shows that one is estimating the same measure of association of disease and exposure in the these two types of studies. **[1]**
- Result shows that it is not necessary to distinguish which of the two variables is considered to be the outcome and which is considered to be the explanatory (predictor) variable. **[1]**
- Result permits the study of disorders with long time lag between exposure and disease **[1]**, and rare diseases **[1]**.

(iii) Consider cross-sectional, case-control and cohort studies. For which of these studies can the relative risk be estimated directly? For which of these studies can the odds ratio be estimated directly?

(5)

Relative risk - estimated directly only in cross-sectional **[1]** and cohort studies **[1]**.
Odds ratio - estimated directly in cross-sectional **[1]**, case-control **[1]** and cohort studies **[1]**.

(iv) When can the odds ratio be used as an approximation to the relative risk?

(1)

- For rare diseases, the odds ratio can be interpreted as an approximation to the relative risk **[1]**.

2. (i) Define the standardised mortality ratio (SMR). When comparing mortality in two populations, give one advantage and one disadvantage of using SMRs versus using crude death rates or age-specific rates. Give three examples of how the SMR is used in practice.

(7)

$$\text{SMR} = \frac{\text{observed (actual) number of deaths in study population}}{\text{expected number of deaths in study population}}$$

Usually this is multiplied by 100.

[1] for correct definition of SMR as ratio of observed to expected deaths.
[1] for correct definition of how the expected number is calculated.

The expected number of deaths in the study population (population of interest) is calculated as $\sum r_i^S n_i$, where n_i is the number of persons in category i of the study population and r_i^S is the corresponding category-specific event rate in a standard population S. Categories are usually defined by age, gender, race, ethnicity, etc.

Advantages: (a) the SMR is probably a more appropriate comparator than comparing crude death rates as crude rates may differ markedly when the two populations have different age structures, (b) as a summary index of mortality, it is more easily compared across populations, e.g. geographic areas, than comparing many age-specific rates across populations. Disadvantages: (a) the magnitude depends on the standard population chosen, (b) it is a fictional rate, i.e. it does not apply to a real population, (c) it may disguise different patterns in specific age groups.

[1] for giving one of its advantages.
[1] for giving one of its disadvantages.

N.B. other advantages and disadvantages not listed above are allowed to gain marks.

Give three examples of how it is used in practice. Classical examples include comparing rates in geographic areas, different occupations and different cohorts.

[3] one mark for each example.

(ii) Describe indirect standardisation, stating explicitly how to calculate the indirectly standardised mortality rate. State any assumptions you make when applying this technique to compare populations.

(6)

Indirect standardisation compares the rates of the index (study) and standard (reference) populations by applying the stratum-specific rates in the standard/reference population to the index (study) population **[1]** .

Indirect standardisation involves using the index population's crude rate r^{index} and the standard population's stratum-specific rates **[1]**. The calculation involves two components: 1) calculation of the standard mortality ratio (SMR) and 2) multiplication of the SMR and the standard population's crude rate to get the indirectly standardised rate for the index population.

[1] for use of SMR as multiplication factor.
[1] for correct equation of how the expected number is calculated.
[1] for use of standard population's crude rate.

Assumption: SMRs from different index populations are not strictly comparable because they are calculated using different weighting schemes that depend on the age structures of the index populations. SMRs can be compared if you make the assumption that the ratio of rates between the index and standard population is constant. In other words, assume proportionality of rates (which is like the assumption of proportional hazards), i.e.,

$$r_i^{index} = k \times r_i^S \quad \textbf{[1]}$$

(iii) When is indirect preferred to direct standardisation?

(3)

Indirect preferred:

1) if specific rates are not available **[1]**,

2) if there are small numbers in some strata leading to imprecise specific rates **[1]**,

3) if the event is rare so that some strata may contain zeroes **[1]**.

(iv) Describe the Poisson regression model for rates which underlies the technique of indirect standardisation and state clearly how to interpret this model. You may assume that the random variables of interest, e.g. numbers of deaths, have a Poisson distribution.

(4)

- (unknown) rate in the ith stratum of the index population r_i is proportional to the *known* rate in the ith stratum of the standard population, r_i^S, i.e.

$$r_i = k\, r_i^S \quad \textbf{[1]}$$

- $Y_i \sim \text{Poisson}(N_i k r_i^S)$ **[1]** for correct mean
- $N_i =$ exposure in the ith stratum of the index population **[1]**
- known r_i^S rate in ith stratum of the standard population **[1]**
- $e_i = r_i^S N_i$ are *known* expected # deaths in the ith stratum **[1]**
- $\log E(Y_i) = \log(e_i k) = \log(e_i) + \log(k) = \log(e_i) + \beta_0$ **[1]** for correct model
- $\widehat{\text{SMR}} = \exp(\widehat{\beta}_0) = \hat{k}$ **[1]**

Allow up to 4 marks for any of the above 7 points.

3. (i) Explain the main purpose of principal component analysis. Briefly discuss the decisions that need to be made when carrying out a principal component analysis.

(5)

Main purpose is dimension reduction Principal component analysis (PCA) constructs p artificial uncorrelated variables, termed principal components (PCs), from linear combinations of a p observed variables. Principal component analysis (PCA) reduces dimensionality by variable reduction **[1]** from p to q variables, where $q < p$, by using only a smaller number of PCs that account for most of the variance in the observed variables. This can make the patterns in the data more clear, i.e., this smaller number of artificial variables can help improve our understanding **[1]** of what is going on. Typically, the analyst tries to interpret the first few principal components in terms of the original variables to gain a greater understanding of the data.

Carry out the PCA using the variance-covariance or correlation matrix? [1]

What criterion to use for dimension reduction? [1]

Interpretation of retained PCs? [1]

(ii) The examination marks of 88 students in five different subject areas of mathematics have been recorded. Each examination was marked out of 100 marks. Some summary statistics for these mathematics marks are given in the table below.

Examination	Mean	Variance	Standard Deviation	Minimum	Maximum
Mechanics	38.97	305.69	17.48	0	77
Vectors	50.59	172.84	13.15	9	82
Algebra	50.60	112.89	10.62	15	80
Analysis	46.68	220.38	14.85	9	70
Statistics	42.31	297.76	17.26	9	81

For these data, discuss whether it is appropriate to carry out the principal component analysis using the variance-covariance matrix or using the correlation matrix.

(4)

- PCA is scale (unit) dependent **[1]**.
- All mathematics marks are measured on comparable scales **[1]** of 0 to 100 marks, so PCA could be justifiably carried out on the sample covariance matrix.

- However, the marks variance differs **[1]** considerably between subjects, where the subjects with the lowest mean scores having the highest variances.
- So many analysts would want to use the standardised variables for the PCA, i.e., use the correlation matrix **[1]**.

(iii) The correlation matrix for the data is given below. Briefly describe the correlation structure in the variables.

(3)

- All the correlations are positive **[1]**, which can be interpreted as good students tend to do well in all the subjects and bad students tend to do badly in all the subjects. The size of the correlations are 'moderate' **[1]** and range from a high of 0.71, between Algebra and Analysis, to a low of 0.39, between Mechanics and Statistics.
- Otherwise, there are no other obvious patterns **[1]**.

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	1.00	0.55	0.55	0.41	0.39
Vectors	0.55	1.00	0.61	0.49	0.44
Algebra	0.55	0.61	1.00	0.71	0.66
Analysis	0.41	0.49	0.71	1.00	0.61
Statistics	0.39	0.44	0.66	0.61	1.00

(iv) An analyst has extracted the principal components and eigenvalues from the correlation matrix for the data. The coefficients for the first three principal components and the eigenvalues are given in the table below.

		Component		
	Variable	1	2	3
	Mechanics	-0.40	0.65	0.62
	Vectors	-0.43	0.44	-0.71
	Algebra	-0.50	-0.13	-0.04
	Analysis	-0.46	-0.39	-0.13
	Statistics	-0.44	-0.47	0.32
Eigenvalue		3.18	0.74	0.45

(a) Interpret the first and second principal components.

(2)

- PC1 weights each of the mathematics subjects fairly equally, so it is a measure of overall ability of the students across the five subjects **[1]**.
- PC2 is a contrast between Analysis and Statistics versus Mechanics and Vectors **[1]**.

(b) How much of the total variation in the data is explained by the first principal component? How much of the total variation in the data is explained by the first two principal components?

(2)

- Eigenvalues give variance explained by each principal component. For a correlation matrix, the eigenvalues sum to 5, the number of variables. So PC1 explains $\frac{3.18}{5} \times 100 = 63.6\%$ **[1]** of the standardised variance.
- The first two principal components explain $\frac{3.18 + .74}{5} \times 100 = 78.4\%$ **[1]** of the standardised variance.

(c) What criteria might you use to decide on the apparent dimensionality of these data? Hence comment on the apparent dimensionality of these data.

(4)

- Using the eigenvalue one criterion, these data have only one dimension **[1]** as only PC1 has an eigenvalue higher than 1.
- Using the criterion of choosing the first components that explain 70-80% of the variance, these data have two dimensions **[1]** as PC1 alone doesn't explain at least 70% of the standardised variance (only 63.6%), but the PC1 and PC2 explain greater than 70% of the standardised variance (78.4%).
- It seems reasonable to represent the data by the first two principal components as one criterion supports two principal components **[1]** and each principal component is interpretable **[1]**.

4. (i) Explain the purpose of cluster analysis and discuss briefly the decisions that need to be made when carrying out a cluster analysis.

(5)

Purpose The purpose of cluster analysis is to combine a set of objects into groups, termed clusters, such that the objects in the same cluster tend to be similar to each other in some sense and objects in different clusters tend to be dissimilar **[1]**. In other words, the resulting clusters should exhibit high internal (within-cluster) homogeneity and high external (between cluster) heterogeneity.

Decisions
- Decide to use the raw data or to transform it **[1]**? For example, decide to standardise continuous variables or to recode categorical variables?
- Decide on some measure of distance **[1]** between the different observations we are trying to cluster, e.g., squared Euclidean distance or taxicab distance.
- Decide which algorithm to use **[1]** for finding the clusters, e.g., single linkage or K-means clustering.
- Decide on the final number of clusters **[1]** using some criterion.

(ii) What is the difference between a hierarchical and a non-hierarchical method of clustering? Give an example of a non-hierarchical method.

(3)

- A hierarchical method of clustering builds a hierarchy of clusters, usually presented in a dendrogram, from which the user still needs to choose appropriate clusters **[1]**. Hierarchical clustering algorithms can be either top down (sequentially splitting clusters) or bottom up (sequentially merging clusters).
- In contrast, in a non-hierarchical method, the desired number of clusters is specified in advance **[1]**.
- Example: The most frequently used non-hierarchical method is K-mean clustering **[1]**, which assumes a fixed number of clusters, K. Others include the partitioning method.

(iii) What is a dendrogram? Why is it useful?

(2)

- A dendrogram is a tree (or branching) diagram of how and when clusters are formed in a hierarchical cluster analysis **[1]**.

- A dendrogram is useful in that distinct clusters may be recognisable by visual inspection **[1]** of the dendrogram.

(iv) The pairwise distances (dissimilarities) between five objects are as follows.

Object	1	2	3	4	5
1	0				
2	4	0			
3	6	9	0		
4	1	7	10	0	
5	6	3	5	8	0

Use single-linkage (also known as nearest neighbour) cluster analysis on the dissimilarity matrix above, and draw the associated dendrogram for your analysis.

(7)

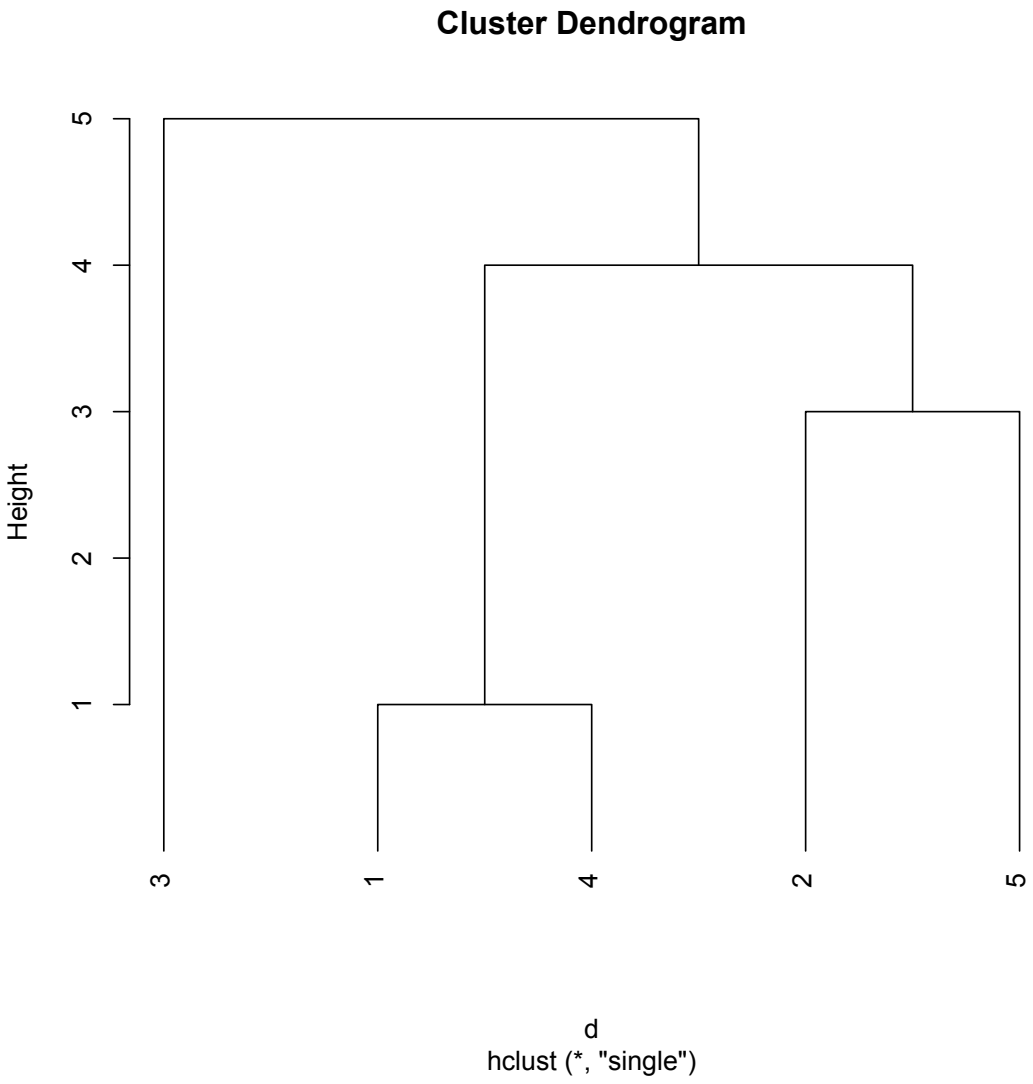
A correct dendrogram with correct distances gets a total of 7 marks (see following dendrogram drawing). One mark for each correct merger of objects into clusters, total of 3 marks. No mark for final merger. One mark for each correct distance in the dendrogram, total of 4 marks. Namely,

- merge objects 1 and 4 **[1]**, with a distance of 1 **[1]**
- merge objects 2 and 5 **[1]**, with a distance of 3 **[1]**
- merge clusters (1,4) and (2,5) **[1]**, with a distance of 4 **[1]**
- merge cluster (1,2,4,5) and object 3, with a distance of 5 **[1]**

(v) Discuss whether there is a clear cluster structure in the data.

(3)

- It looks like object 3 does not cluster with the others, i.e., is a singleton cluster **[1]**.
- Perhaps (1,4) and (2,5) should be regarded as clusters **[1]** as their nearest neighbour distance is 5, while the nearest neighbour distances between objects 1 and 4 was 1 and between objects 2 and 5 was 3.
- Overall, it is hard to be definitive **[1]**.


Cluster Dendrogram
Height
1
2
3
4
5
3
1
4
2
5
d
hclust (*, "single")

5. (i) Define the *survivor function* $S(t)$ and the *hazard function* $h(t)$ for a continuous random variable T measuring lifetime. Write down an expression for the survivor function in terms of the hazard function.

(3)

The survivor function $S(t) = \Pr(T > t)$ is the probability of surviving beyond time t.

[1] mark for correct definition of survivor function. Note that $S(t) = 1 - F(t)$, where $F(t)$ is the cumulative distribution function, is also an acceptable definition.

The hazard function $h(t) = f(t)/S(t)$, i.e., the ratio of the probability density function to the survivor function.

[1] mark for correct definition of hazard function

The survivor and hazard functions are related in three different ways:

i. $S(t) = \exp(-\int_0^t h(u)\,du)$

ii. $h(t) = -\dfrac{d}{dt}(\log S(t))$

iii. $f(t) = h(t)\,S(t)$

[1] mark for stating any of the above 3 relationships between $h(t)$ and $S(t)$

(ii) The exponential distribution has constant hazard function $h(t) = \lambda$. Write down expressions for the density of the exponential distribution and the mean of this distribution in terms of λ.

(2)

- $T \sim \text{exponential}(\lambda)$
- mean $\mu = 1/\lambda$ **[1]**
- density $f(t) = \lambda e^{-\lambda t}$ **[1]**

(iii) Explain what is meant by a *right-censored* observation. Give two different examples of ways in which a right-censored observation might arise.

(3)

An observation is right-censored if all that is known is that the observation T is above a certain value t.

[1] mark for correct definition of right-censored observation

Examples include: a) individuals alive at the end of a study, so their age at death is unknown, but greater than their age at the end of the study; b) loss-to-followup - when a subject leaves a study before experiencing the event of interest.

[2] marks for examples, 1 mark each

(iv) After a radical mastectomy for breast cancer, ten female patients were randomly assigned to one of two groups, an experimental group who received chemotherapy, and a control group who received no drugs. At the end of two years, survival times in months were recorded and are given in the table below. A right-censored observation is denoted by +, so 16+ denotes a right-censored observation at 16 months.

Experimental group	23	16+	18+	20+	24+
Control group	15	18	19	19	20

Compute the Kaplan-Meier estimate of the survivor function for each group and plot the results on one graph.

(10)

Computation: The Kaplan-Meier estimator of the survival curve $S(t)$ is computed as follows. The computation requires the ordered failure times $t_{(1)} < t_{(2)} < \cdots < t_{(r)}$ to be considered for each group separately.

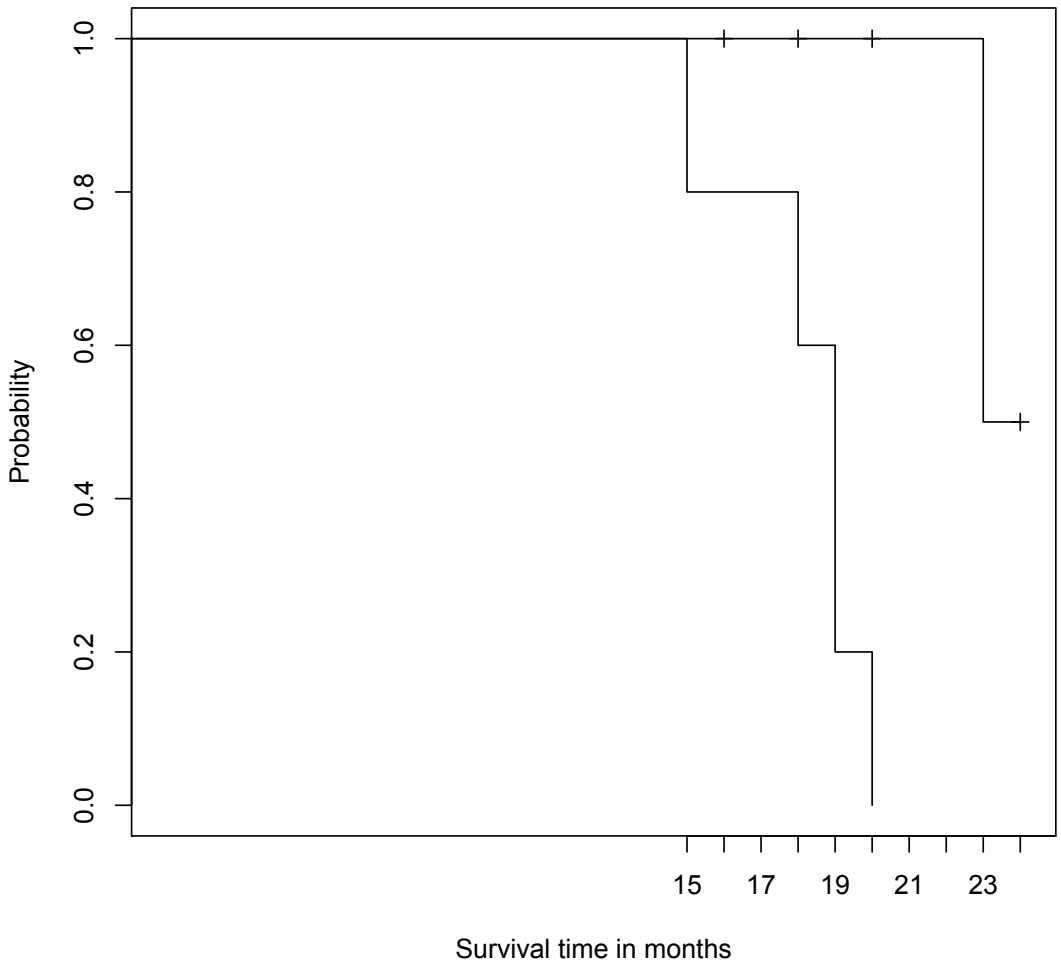

Kaplan-Meier plots for Control and Experimental groups
Probability
0.0
0.2
0.4
0.6
0.8
1.0
15
17
19
21
23
Survival time in months

Table 1: Calculation of Kaplan-Meier estimate for Control treatment

Time $t_{(j)}$	# at risk at time $t_{(j)}$ $n_{(j)}$	# of failures at time $t_{(j)}$ $d_{(j)}$	Proportion surviving $\frac{n_{(j)}-d_{(j)}}{n_{(j)}}$	Estimate $\hat{S}(t)$ at each $t_{(j)}$
15	5	1	$4/5$	0.8
18	4	1	$3/4$	$0.75 \times 0.8 = 0.6$
19	3	2	$1/3$	$0.33 \times 0.6 = 0.2$
20	1	1	0	$0.0 \times 0.2 = 0$

[5] marks for each correct proportion surviving at each failure time (four marks for the control group as there are four observed failure times, one mark for the treatment group as there is only one observed failure time)

[4] marks for the plotted survival curve for the experimental group (one mark for each correct value of the survival function at each the four observed failure times)

[1] one mark for the plotted survival curve for the control group (one mark for the correct value of the survival function at the only observed failure time)

(v) If survival times for the control group have an exponential distribution, estimate the hazard rate.

(2)

$$\text{mean } = \frac{91}{5} = 18.2 \implies \text{hazard = reciprocal of the mean} = \frac{5}{91} = 0.0549$$

Table 2: Calculation of Kaplan-Meier estimate for the experimental treatment. Rows for censored observations, marked with a plus symbol, have been included. Note that the rows for censored observations might be omitted and there is no loss of marks if these rows are omitted.

Time $t_{(j)}$	# at risk at time $t_{(j)}$ $n_{(j)}$	# of failures at time $t_{(j)}$ $d_{(j)}$	Proportion surviving $\frac{n_{(j)}-d_{(j)}}{n_{(j)}}$	Estimate at each $t_{(j)}$ $\hat{S}(t)$
16+	–	–	–	–
18+	–	–	–	–
20+	–	–	–	–
23	2	1	$1/2$	0.5
24+	–	–	–	0.5

Note: If the largest failure time has been censored, e.g., 24+ in the CMF treatment group, the survival curve is estimated up to and including the next largest value, e.g., 23 in the CMF treatment group, the value for which is then assumed to apply for all times onward.

Tick marks: Many Kaplan-Meier curves are graphed with a tick mark at each time point where there is a censored observations times. While this is good practice, no marks are lost if the tick marks are missing from the Kaplan-Meier curve for the CMF group.

Alternative solution: As there is no censoring in the control group, the Kaplan-Meier estimator is one minus the empirical distribution function. Full marks can be attained for the curve for the control group if this relationship is stated and the correct curve is graphed.

6. Consider the following experiment on visual perception using random-dot stereograms. A random-dot stereogram is composed of two rectangles placed side by side, where each rectangle appears to consist only of randomly scattered dots, without any image. When viewed with only one eye functioning, the viewer cannot see a hidden image. However, when viewed with both eyes, if a person focuses the eyes in front of or behind the pair of images, then a three-dimensional hidden image of a diamond can be seen. Sometimes the diamond image can be seen quickly, but on other occasions it can take a while before it can be perceived. Here the response variable is the time in seconds needed to perceive the diamond image.

   The experiment investigated whether giving a person prior knowledge about the shape of the image reduces the time needed to recognise it. Forty-three subjects (group NV) received just verbal information about the shape of the hidden object. Thirty-five subjects (group VV) received both verbal information and visual information, for example a drawing of the hidden object.

   (i) The response time T to perceive the diamond image has a Weibull distribution with hazard function

   $$h(t) = \lambda\,\gamma t^{\gamma-1}, \qquad t \geq 0$$

   where $\lambda > 0$ and $\gamma > 0$ are parameters to be estimated. Show that the parameters are given by the intercept and slope of the theoretical relationship between the logarithm of the cumulative hazard function plotted against the logarithm of time.

   (3)

   The cumulative hazard function for a Weibull distribution is:

   $$H(t) = \int_0^t h(u)\,du = \int_0^t \lambda\,\gamma u^{\gamma-1}\,du = \lambda\,u^{\gamma}\Big|_0^t = \lambda\,t^{\gamma}$$

   $$\log(H(t)) = \log(\lambda) + \gamma\log(t)$$

   [1] mark for correct definition of cumulative hazard function

   [1] mark for $H(t) = \lambda\,t^{\gamma}$

   [1] mark for $\log(H(t)) = \log(\lambda) + \gamma\log(t)$

   Note that as $H(t) = -\log S(t)$ this is equivalent to $\log(-\log S(t)) = \log(\lambda) + \gamma\log(t)$

   (ii) Write down the proportional hazards model for a vector of p time-constant covariates, X, assuming a Weibull baseline hazard function. Interpret each term in your model.

Write down an expression for the hazard ratio for this model.

(5)

The proportional hazards (PH) model incorporates covariates $X = (X_1, \ldots, X_p)$ in the hazard function $h(t, X)$

- hazard $h(t, X)$ modelled directly **[1]**
- p covariates X's affect hazard multiplicatively **[1]**

$$\begin{aligned} h(t,X) &= h_0(t) \times e^{X_1\beta_1 + \cdots + X_p\beta_p} \\ &= h_0(t) \times e^{X_1\beta_1} \times \cdots \times e^{X_p\beta_p} \end{aligned}$$

- $h_0(t)$ is the baseline hazard, a nonnegative function of time, but not of X's **[1]**
- The Weibull PH model can be written as:

$$h(t,X) = \lambda\, \gamma t^{\gamma-1} \times e^{X_1\beta_1 + \cdots + X_p\beta_p} \qquad \textbf{[1]}$$

- denote X's for 2 individuals as $X^* = (X_1^*, \ldots, X_p^*)$ and $X = (X_1, \ldots, X_p)$

$$\frac{h(t,X^*)}{h(t,X)} = e^{(X_1^* - X_1)\beta_1 + \cdots + (X_p^* - X_p)\beta_p}$$

baseline hazards cancel out, so hazard ratio θ is independent of time, i.e. constant

$$h(t,X^*) = \theta h(t,X) \qquad \text{where} \qquad \theta = e^{(X_1^* - X_1)\beta_1 + \cdots + (X_p^* - X_p)\beta_p} \qquad \textbf{[1]}$$

(iii) Data from the random-dot stereogram experiment were analysed using a proportional hazards model with a Weibull baseline hazard function. One explanatory variable was included in the model: GroupVV, taking the value 1 if the subject was in Group VV, or 0 if the subject was in Group NV. The following edited computer output shows the results from the fitted model.

	Estimate	Standard Error
λ	0.060	0.019
γ	1.260	0.104
GroupVV	0.552	0.233

(a) Use the fitted model to estimate the hazard ratio for a subject in the VV Group compared to a subject in the NV Group. Construct a 95% confidence interval for this hazard ratio. Which group, on average, has the shorter response times?

(3)

- The estimated hazard ratio (HR) for GroupVV = 1 versus GroupVV = 0 is

$$\widehat{HR} = e^{0.552} = 1.737 \quad \textbf{[1]}$$

- $0.552 \pm 1.96 \times 0.233$ is $(0.095, 1.009)$.
  Exponentiating these endpoints yields the 95% CI for the hazard ratio:

$$(1.100, 2.743). \quad \textbf{[1]}$$

- Response times for GroupVV are shorter on average **[1]** than the response times for GroupNV because a positive coefficient increases the hazard, therefore, decreasing response times.

(b) What is the estimated hazard function for a subject in Group NV? What is the estimated hazard function for a subject in GroupVV? How does the estimated hazard function change with time for each group?

(3)

- For Group NV: $h(t) = 0.060 \times 1.260\, t^{0.260}$ **[1]**
- For Group VV: $h(t) = (0.060 \times e^{0.552}) \times 1.260\, t^{0.260}$ **[1]**
- As the estimated Weibull parameter $\hat{\gamma} = 1.260$ is greater than one, the hazard for both groups increases with time. **[1]**

(iv) Someone suggested using an exponential distribution, instead of a Weibull distribution, to model the response times. What advice should you give regarding this idea? Justify your answer.

(3)

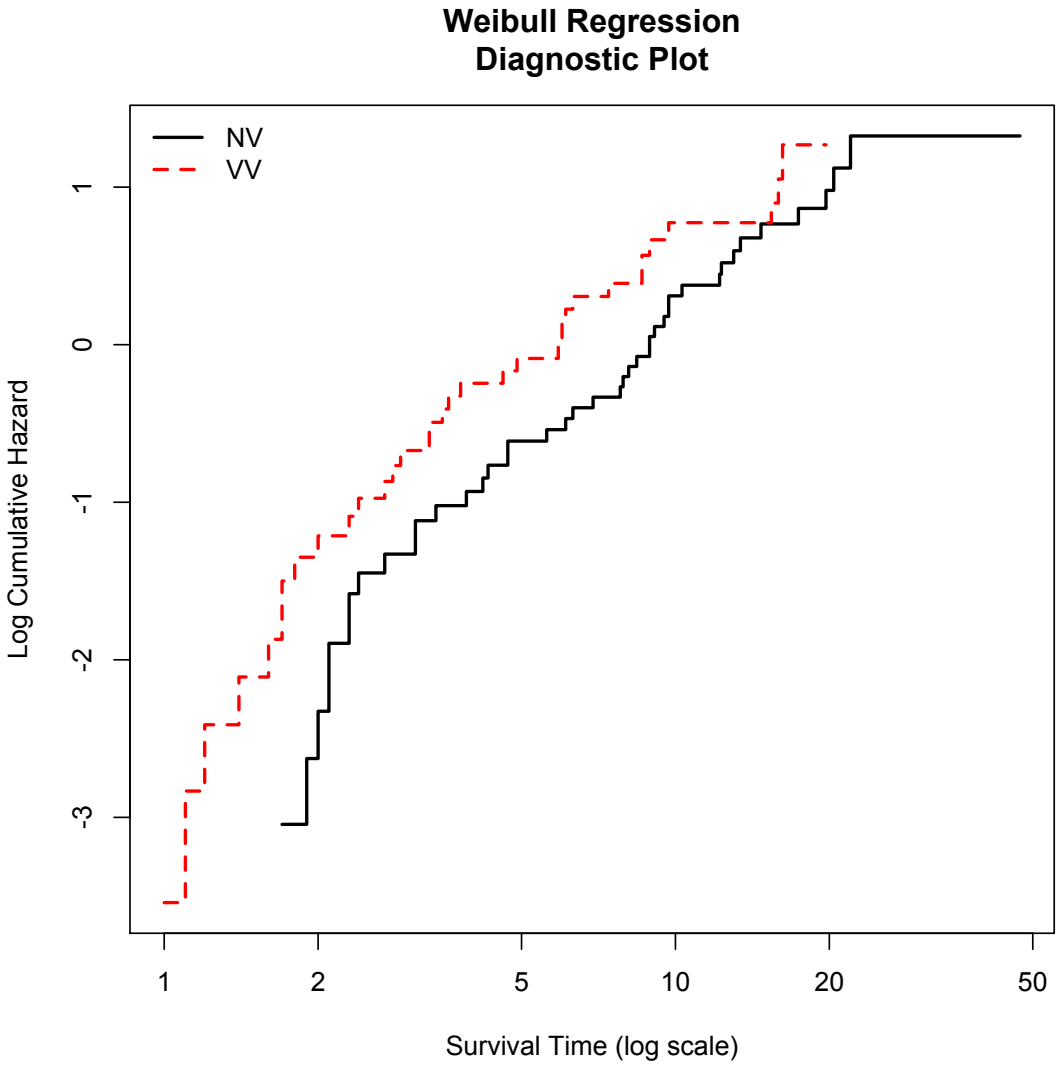
- Exponential distribution is a special case of the Weibull distribution for $\gamma = 1$. **[1]**
- The 95% CI for γ is (1.056, 1.464) **[1]** which does not include 1.
- This is a bad idea as the Weibull PH model significantly improves the fit of the model to the data compared to the exponential PH model. **[1]**

(v) The graph **on the next page** shows estimates of the log cumulative hazard function for the NV group (solid line) and the VV group (dashed line). Referring to this graph, discuss whether the Weibull proportional hazards model is appropriate for the stereogram response times. Justify your answer.

(3)

- For the two groups, the graph of log(-log(estimated Kaplan-Meier curve)) versus log of survival time should result in
  - approximately parallel curves if the hazards are proportional across the groups **[1]**
  - roughly straight lines if the survival distribution is Weibull **[1]**
  - Justification: These plots are approximately parallel and straight, so the Weibull and PH assumptions appear reasonable based on this graph. **[1]**

log(-log($\widehat{S}(t)$)) for GroupNV (solid line) and GroupVV (dashed line)



7. (i) For a simple random sample without replacement from a finite population, the sample mean is an unbiased estimator of the population mean and has variance

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\, S^2$$

Define the symbols in the above formula.

(4)

- $\bar{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$ = sample mean **[1]**
- N = finite population size **[1]**
- n = sample size **[1]**
- $S^2 = \dfrac{\sum_{i=1}^{N} (Y_i - \bar{Y})^2}{N-1}$ = finite population variance

(ii) Use the formula in part (i) to prove that

$$V(\hat{p}) = \frac{N-n}{N-1} \frac{P(1-P)}{n}$$

where $\hat{p}$ and P denote the sample and population proportion of a certain characteristic.

(3)

$$\text{Let } Y_i = \begin{cases} 1, & \text{if unit } i \text{ has the characteristic} \\ 0, & \text{otherwise} \end{cases}$$

$$S^2 = \frac{\sum_{j=1}^{N} (Y_j - \bar{Y})^2}{N-1} = \frac{\sum_{i=1}^{N} (Y_i^2 - 2Y_i\bar{Y} + \bar{Y}^2)}{N-1} = \frac{\sum_{i=1}^{N} Y_i^2 - N\bar{Y}^2}{N-1}$$

Note that $Y_i^2 = Y_i$ for $Y_i = 0, 1$, so $\sum_{i=1}^{N} Y_i^2 = \sum_{i=1}^{N} Y_i$. **[1]**
Let $A = \sum_{i=1}^{N} Y_i$ denote the total with this characteristic, so that $P = A/N = \bar{Y}$.

$$S^2 = \frac{\sum_{i=1}^{N} Y_i - N\bar{Y}^2}{N-1} = \frac{A - N(A/N)^2}{N-1} = \frac{N(A/N) - N(A/N)^2}{N-1} = \frac{NP(1-P)}{N-1} \quad \textbf{[1]}$$

$$V(\bar{y}) = (\frac{1}{n} - \frac{1}{N})\, S^2 = (\frac{1}{n} - \frac{1}{N}) \frac{NP(1-P)}{N-1} = (\frac{N-n}{Nn}) \frac{NP(1-P)}{N-1} = (\frac{N-n}{N-1}) \frac{P(1-P)}{n} \quad \textbf{[1]}$$

There are 650 Members of Parliament (MPs) in the United Kingdom (UK) House of Commons. MPs can employ staff to help run their offices. MPs may employ as many office staff as they wish, subject to a fixed budget. A 20% simple random sample was selected in order to estimate the total number of office staff currently employed by MPs. The following data on the number of staff employed were collected.

Number of employees	1	2	3
Frequency	25	80	25

(iii) Estimate the total number of office staff currently employed by MPs and use the result in part (i) to construct an approximate 95% confidence interval for this total. (5)

$$\bar{y} = \frac{25 \times 1 + 80 \times 2 + 25 \times 3}{130} = \frac{260}{130} = 2 \quad \textbf{[1]}$$

$$s^2 = \frac{(25 \times 1^2 + 80 \times 2^2 + 25 \times 3^2) - (130 \times 2^2)}{129} = \frac{570 - 520}{129} = \frac{50}{129} = 0.3875968 \quad \textbf{[1]}$$

$$\hat{T} = N\bar{y} = 650 \times 2 = 1300 \quad \textbf{[1]}$$

$$v(\hat{T}) = V(N\bar{y}) = N^2 \frac{(1-f)}{n} s^2, \text{where } f = \frac{n}{N} \text{ is the sampling fraction}$$

$$v(\hat{T}) = 650^2 \frac{(1-0.2)}{130} 0.3875968 = 2600 \times 0.3875968 = 1007.75168$$

$$\sqrt{v(\hat{T})} = \sqrt{1007.75168} = 31.74510 \quad \textbf{[1]}$$

So a 95% confidence interval is given by

$$1300 \pm 1.96 \times 31.74510$$

or

$$(1237.78, 1362.22) \quad \textbf{[1]}$$

Note that 2 is allowed to be used instead of 1.96 in the confidence interval calculation.

Shingles is a painful skin rash caused by the reactivation of the chickenpox virus in people who have previously had chickenpox. A shingles vaccine is recommended by the UK's National Health Service to people aged 70 or older.

(iv) On 30 April 2012, there were 405 members of the UK's House of Lords aged 70 or older. A researcher plans to take a simple random sample of these 405 people in order to estimate the percentage who have been vaccinated against shingles to within a margin of error of 1%. The researcher asks you to estimate the smallest achieved sample

size that would be necessary to do so with 95% confidence. Using the result in part (ii), make this estimate. What advice should you give the researcher regarding the use of the smallest sample size that you found and regarding whether their over-age-70 Lordships are a suitable source of a sample that is in some sense representative of all UK adults over 70.

(8)

- A margin of error of 1% means that the half width of the confidence interval is:
  $z_\alpha \times \sqrt{\frac{N-n}{N-1} \times \frac{P(1-P)}{n}} = 0.01$ **[1]**
- With 95% confidence $\alpha = 0.05, z_{\alpha/2} = z_{0.025} = 1.96$ **[1]**
- Use the worst case scenario of $p = 0.5$, so that $P(1-P)$ is maximised at 1/4. **[1]**
- $\frac{N-n}{N-1} \times \frac{P(1-P)}{n} = \left(\frac{0.01}{1.96}\right)^2$
- $\frac{N}{n} - 1 = \frac{N-1}{P(1-P)} \times \left(\frac{0.01}{1.96}\right)^2$
- $\frac{N}{n} = \frac{N-1}{P(1-P)} \times \left(\frac{0.01}{1.96}\right)^2 + 1$
- $n = \frac{N}{\frac{N-1}{P(1-P)} \times \left(\frac{0.01}{1.96}\right)^2 + 1}$
- $n = \frac{N}{\frac{N-1}{0.25} \times \left(\frac{0.01}{1.96}\right)^2 + 1}$
- $n = \frac{405}{\frac{404}{0.25} \times \left(\frac{0.01}{1.96}\right)^2 + 1}$
- $405/((404/0.25) * (0.01/1.96)^2 + 1) = 388.6505$
- achieved sample size should be 389 **[1]**
- as only 16 members would not be sampled, the researcher may want to consider using a census **[1]**
- Members of the House of Lords are unlikely to be representative as they are, for example, much better educated and wealthy than the general population and the percentage of female and minority members is much less than the over-age-70 UK population. Up to **[3]** marks for examples of why they are unlikely to be representative, 1 mark for each example.

8. (i) Define the term *stratified random sampling*.

(2)

The population is divided into strata **[1]**, i.e., mutually exclusive and exhaustive subgroups, before sampling. Strata are mutually exclusive, i.e., every member of the population belongs to only one stratum. Strata are exhaustive, i.e., no member of the population is excluded. Random samples are then taken separately **[1]** from each stratum.

(ii) Explain the circumstances when stratified random sampling may be expected to work well, and give an example of such a situation.

(2)

A stratified random sample can increase the precision of our estimates, i.e., reduce variance, if we stratify the population into relatively homogeneous strata so that most of the variation is between strata. **[1]** Accept any other sensible comment.

One mark for any suitable example **[1]**. Examples include:

- It is desirable to have stratum-specific estimates, e.g., for specific groups within the population.
- A stratified random sample may be easier/cheaper to administer, i.e., the cost per observation may be reduced by stratification of the population into convenient strata for sampling.

(iii) Explain in words what is meant by the expressions *stratification with proportional allocation*, *stratification with Neyman allocation* and *stratification with optimal allocation*.

In your answer, you should explain in what sense each of "Neyman allocation" and "optimal allocation" makes the most effective use of resources. Give three reasons why these allocations are only theoretical.

(11)

Proportional allocation: the sampling fraction is the same **[1]** for all strata, i.e., it assigns sample sizes to strata in proportion to the stratum population size.

Neyman allocation: this is an allocation to minimise the variance of the stratified sampling estimator of the population mean **[1]**, for a given sample size, for a fixed total cost, where costs of sampling are the same in all strata. **[1]**

Optimal allocation: this is an allocation to minimise the variance of the stratified sampling estimator of the population mean **[1]**, for a given sample size, for a fixed total cost, where costs of sampling vary from stratum to stratum. **[1]**

Note that Neyman allocation is a special case of optimal allocation when the cost of obtaining an observation from each stratum is the same.

Neyman allocation is optimal in the sense that it minimises the variance of the stratified estimator of the population mean subject to a fixed total sample size n. **[1]**

Optimal allocation is when the stratum-specific sample sizes n_h are chosen to

- minimise the variance of the stratified estimator of the population mean subject to fixed cost **[1]** or
- minimise total cost subject to fixed variance **[1]**.

These optimal allocations are only theoretical:

- as the values of S_h are typically unknown **[1]**
- direct costs can be at best estimated **[1]**
- as the solution to the optimal allocation formula typically results in non-integer solutions **[1]**

(iv) State the optimal allocation formula for the allocation of sample size in stratified random sampling. Define all the symbols used in the formula.

(5)

The optimal allocation formula, up to a constant of proportionality, is

$$n_h \propto \frac{N_h S_h / \sqrt{c_h}}{\sum_h (N_h S_h / \sqrt{c_h})} \quad \textbf{[1]}$$

where

- n_h - sample size for stratum h **[1]**
- N_h - total number of elements in stratum h **[1]**
- S_h - population standard deviation in stratum h **[1]**
- c_h - direct cost of obtaining an observation from stratum h **[1]**