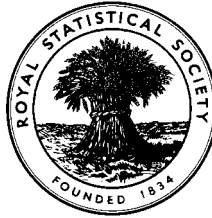


EXAMINATIONS OF THE ROYAL STATISTICAL SOCIETY
(formerly the Examinations of the Institute of Statisticians)



HIGHER CERTIFICATE IN STATISTICS, 1996

Paper III: Statistical Applications and Practice

Time Allowed: Three Hours

*Candidates should answer **FIVE** questions.*

All questions carry equal marks.

Graph paper and Official tables are provided.

Candidates may use silent, cordless, non-programmable electronic calculators.

*Where a calculator is used the **method** of calculation should be stated in full.*

Note that $\binom{n}{r}$ is the same as ${}^n C_r$, and that \ln stands for \log_e .

1. An experiment to investigate the factors that influence the strength of concrete recorded the number of days the concrete was allowed to set before being tested and the amount of a binder added to the concrete mix. Two batches of concrete were made, one containing the equivalent of 400 kg/m^3 of the binder and the other containing the equivalent of 550 kg/m^3 of binder. Each batch was then cast into 12 cubes and allowed to set under identical controlled conditions. After 7 days 4 cubes were selected at random from each batch and their strengths were determined using a standard industrial testing machine. After 14 days another 4 cubes were selected randomly from each batch and tested. The remaining 4 cubes in each batch were tested after 28 days. The strength test results are shown in the table.

Strength test results (N/mm^2)

| Binder content | Setting time (days) | | |
|----------------------|---------------------|------------|------------|
| | 7 | 14 | 28 |
| 400 kg/m^3 | 26.2, 24.5 | 32.4, 33.1 | 42.0, 42.5 |
| | 26.1, 25.4 | 31.8, 30.7 | 42.1, 40.6 |
| 550 kg/m^3 | 31.0, 33.2 | 51.0, 55.1 | 64.5, 61.5 |
| | 29.4, 30.6 | 52.0, 47.0 | 60.4, 58.8 |

An analysis of variance table for these results is reproduced, in part, below.

Part of the analysis of variance table

| Source of variation | Sum of squares |
|--|----------------|
| setting time | 2178.0 |
| binder content | |
| setting time \times binder content interaction | |
| Residual | 65.4 |
| Total | 3807.35 |

The sum of strength test results for all 24 cubes is 971.9 and the corresponding sum of the squares is 43165.25.

Produce a complete analysis of variance table. Carry out the statistical tests which you think are appropriate, explaining what hypotheses are being tested.

Produce a diagram of means and their standard errors which makes clear the nature of any interaction that there may be between the two factors. Write a short report (4 or 5 informative sentences) which explains the findings of your analysis in non-technical language for the investigator who carried out the experiments.

2. The failure time, T , of a particular type of device is believed to follow a distribution which has the probability density function

$$f(t) = \lambda^2 t e^{-\lambda t}, \quad t \geq 0.$$

and cumulative distribution function

$$P(T \leq t) = 1 - e^{-\lambda t} (1 + \lambda t), \quad t \geq 0.$$

Show that

$$E(T) = \frac{2}{\lambda}$$

The failure times of 215 of these devices were recorded and the following descriptive statistics were obtained.

| N | Mean | Median | StDev | SEMean | Min | Max |
|-----|--------|--------|--------|--------|------|--------|
| 215 | 198.14 | 164.00 | 146.37 | 9.98 | 5.00 | 721.00 |

Under the assumption that the failure times do follow the distribution with the pdf given above, expected frequencies were calculated. The complete set of observed frequencies and *some* of the expected frequencies for the 215 failure times are given in the grouped frequency distribution below.

| Failure time in days (t) | Observed frequency | Expected frequency |
|------------------------------|--------------------|--------------------|
| $t \leq 50$ | 27 | 19.70 |
| $50 < t \leq 100$ | 43 | |
| $100 < t \leq 150$ | 31 | 38.52 |
| $150 < t \leq 200$ | 30 | |
| $200 < t \leq 250$ | 17 | |
| $250 < t \leq 300$ | 20 | 18.82 |
| $300 < t \leq 350$ | 12 | |
| $350 < t \leq 400$ | 13 | 9.37 |
| $400 < t \leq 500$ | 11 | 10.75 |
| $t > 500$ | 11 | 8.36 |

Estimate λ and complete the table of expected frequencies. Test whether the data are consistent with the assumption that these failure times follow the assumed distribution.

3. Hodgkin's disease is a form of cancer of the lymph nodes and it is classified into four histological types depending on the appearance under the microscope of tissue sections from lymph node biopsies. The types are Lymphocyte predominant (LP), Nodular sclerosing (NS), Mixed cellularity (MC) and Lymphocyte depleted (LD). In a study of how histological type affects response to treatment, 538 patients with Hodgkin's disease were treated with chemotherapy. Response was recorded as either Positive, Partial or None according to well established criteria.

The table below gives the histological type and the response of each patient 3 months after the start of treatment.

| | | Histological type | | | | Total |
|-----------|----------|-------------------|----|-----|----|-------|
| | | LP | NS | MC | LD | |
| Response: | Positive | 74 | 68 | 154 | 18 | 314 |
| | Partial | 18 | 16 | 54 | 10 | 98 |
| | None | 12 | 12 | 58 | 44 | 126 |
| Total | | 104 | 96 | 266 | 72 | 538 |

Investigate whether response to treatment is related to histological type and, if so, how it is related. You should calculate the proportions of patients of each histological type who show a positive response, and the standard errors of these proportions, as part of your explanation of how response and histological type are related.

4. A paediatrician is investigating a complaint among new-born infants known as severe respiratory distress syndrome, (SRDS). This complaint is often fatal and its cause is unclear. She has collected data on birthweights of infants who developed SRDS and has classified it according to whether the infant died or survived.

Complete data on birthweights (kg) of infants with SRDS

| Infants who died | | | | Infants who survived | | | |
|------------------|-------|-------|-------|----------------------|-------|-------|-------|
| 1.030 | 1.262 | 1.720 | 2.270 | 1.130 | 1.930 | 2.570 | 3.400 |
| 1.050 | 1.295 | 1.750 | 2.275 | 1.410 | 2.015 | 2.600 | 3.640 |
| 1.100 | 1.300 | 1.770 | 2.440 | 1.575 | 2.040 | 2.700 | |
| 1.175 | 1.310 | 1.820 | 2.500 | 1.680 | 2.090 | 2.830 | |
| 1.185 | 1.500 | 1.890 | 2.560 | 1.715 | 2.200 | 2.950 | |
| 1.225 | 1.550 | 1.940 | 2.730 | 1.720 | 2.400 | 3.005 | |
| 1.230 | 1.600 | 2.200 | | 1.760 | 2.550 | 3.160 | |

Descriptive statistics for the two set of birthweights are given below, Q1 and Q3 denote the lower and upper quartiles, respectively.

Descriptive statistics for birth weights of infants who died.

| N | Mean | Median | StDev | SEMean | Min | Max | Q1 | Q3 |
|----|--------|--------|--------|--------|-------|-------|-------|-------|
| 27 | 1.6917 | 1.600 | 0.5176 | 0.0996 | 1.030 | 2.730 | 1.230 | 2.200 |

Descriptive statistics for birth weights of infants who survived.

| N | Mean | Median | StDev | SEMean | Min | Max | Q1 | Q3 |
|----|-------|--------|-------|--------|-------|-------|-------|-------|
| 23 | 2.307 | 2.200 | 0.665 | 0.139 | 1.130 | 3.640 | 1.720 | 2.830 |

The paediatrician is proposing to perform a two-sample *t*-test on the data. What question would this answer? Draw box and whisker plots of the data and explain how they demonstrate that the data conform reasonably well to the assumptions necessary for valid application of the *t*-test.

Carry out the *t*-test and state your conclusions in language that the pediatrician will understand.

The paediatrician asks you whether the result of the *t*-test indicates that birthweight can be used as a predictor of infants who are likely to die if they develop SRDS. How would you advise her?

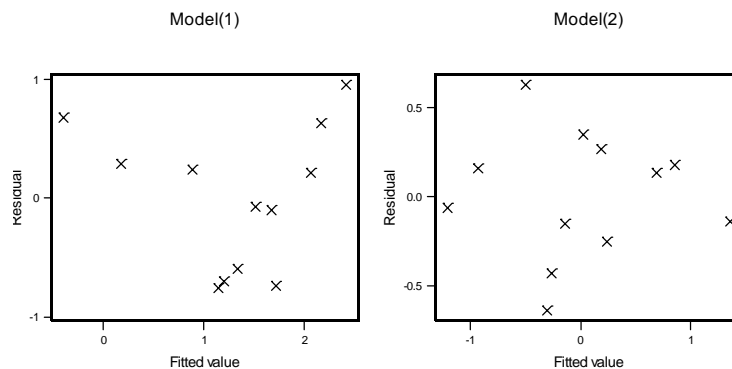
5. An environmental scientist investigated the relationship between the oxygenation rate for a stream and the depth of the stream. The mean oxygenation rate (y) and the depth of the stream (x) for 12 different streams are given in the table below together with the natural logarithms of these quantities.

| y (ppm per day) | x (metres) | $\log(y)$ | $\log(x)$ |
|-------------------|--------------|-----------|-----------|
| 1.44 | 1.54 | 0.3646 | 0.4317 |
| 2.27 | 0.99 | 0.8197 | -0.0100 |
| 0.98 | 1.34 | -0.0202 | 0.2926 |
| 0.50 | 1.86 | -0.6931 | 0.6205 |
| 0.74 | 1.72 | -0.3011 | 0.5423 |
| 1.13 | 2.17 | 0.1222 | 0.7747 |
| 0.28 | 3.46 | -1.2729 | 1.2412 |
| 3.36 | 0.64 | 1.2119 | -0.4462 |
| 2.79 | 0.89 | 1.0260 | -0.1165 |
| 1.57 | 1.38 | 0.4510 | 0.3220 |
| 0.46 | 2.89 | -0.7765 | 1.0612 |
| 0.39 | 1.91 | -0.9416 | 0.6471 |

| Intermediate summary statistics for y and x | Intermediate summary statistics for $w = \log(y)$ and $v = \log(x)$ |
|--|---|
| $\Sigma y = 15.910$, $\Sigma x = 20.790$, $\Sigma y^2 = 32.2421$, $\Sigma x^2 = 43.3521$, $\Sigma xy = 20.276$ | $\Sigma w = -0.00986$, $\Sigma v = 5.3608$, $\Sigma w^2 = 7.2266$, $\Sigma v^2 = 4.9534$, $\Sigma vw = -3.8841$. |

A statistical package was used to fit (1) a linear equation relating y to x , (2) a linear equation relating $\log(y)$ to $\log(x)$. Plots of the residuals against the fitted values are shown in the figure.

Plots of residuals against fitted values



(Question continued on next page)

Explain why these plots indicate that a power law equation

$$y = A x^b$$

provides a better representation of the data than the linear equation

$$y = a + b x.$$

Use the appropriate intermediate summary statistics to determine the coefficients in the power law equation.

Plot mean oxygenation rate against depth and draw in the power law curve on the same graph.

6. (a) Describe the circumstances in which simple exponential smoothing may be used to provide forecasts of future values of a time series X_t .
- (b) If $\hat{x}(t, 1)$ denotes the one step ahead forecast at time t where

$$\hat{x}(t, 1) = \alpha x_t + \alpha(1 - \alpha)x_{t-1} + \alpha(1 - \alpha)^2 x_{t-2} + \alpha(1 - \alpha)^3 x_{t-3} + \dots$$

show that

$$\hat{x}(t, 1) = \alpha(x_t - \hat{x}(t-1, 1)) + \hat{x}(t-1, 1).$$

- (c) The data below give the weekly sales of a commodity over a 13 week period where no special sales promotions or higher than usual levels of advertising took place. Values are to be read along the rows.

| | | | | | | |
|------|------|------|------|------|------|------|
| 5105 | 5618 | 5514 | 5423 | 6044 | 5790 | 5437 |
| 5848 | 6253 | 5835 | 5740 | 6063 | 5340 | |

Plot the data as a time series.

Use simple exponential smoothing with $\alpha = 0.25$ and the first week's sales as the forecast for the second week's sales to forecast one week ahead at each time point.

Plot the set of forecasts on the same graph as the original series and comment on their adequacy.

Suppose a similar set of forecasts had been calculated for some other value of α . How would you compare the adequacy of the two sets of forecasts?

7. When selecting a sample from a finite population three methods which could be used are
- (a) systematic sampling,
 - (b) quota sampling,
 - (c) simple random sampling.

Describe the essential features of each method and the information that the survey practitioner must have in order to select a sample using each method.

Which of the three methods provides the greatest statistical validity for the survey results and why is this so? Why are the other methods used?

What factors should be taken into consideration when deciding on sample size?

8. A psychologist has conducted an investigation of the critical flicker frequency (cff) and its possible relationship to eye colour. The critical flicker frequency is the highest frequency, in cycles per second, at which the flicker in a light source can be detected by an observer. The investigation produced the following data.

| Eye colour | critical flicker frequency |
|------------|--|
| Blue | 25.7, 27.2, 29.9, 28.5, 29.4, 28.3 |
| Brown | 26.8, 27.9, 23.7, 25.0, 26.3, 24.8, 25.7, 24.5 |
| Green | 26.4, 24.2, 28.0, 26.9, 29.1 |

The sum of squares of all values is 13659.67.

Display the data in an appropriate way.

Carry out an analysis of variance. State clearly the assumptions made and the conclusions. If you find a “significant effect” you should investigate whether mean cff differs between all three eye colours or whether one eye colour differs from the other two.

How would you expect the individuals acting as experimental subjects to have been chosen to ensure statistical validity of the results of the analysis?