MATHEMATICAL TRIPOS: PART IB (1995)

Statistics – Example Sheet 3 **AY**

[*Third sheet of 3*]

In questions 1–5 below, you may safely assume that the underlying populations are normal.

1. A sample of nine pieces of steel wire of type A yields the following breaking strengths: 11.99, 12.02, 12.03, 12.09, 12.14, 12.16, 12.16, 12.16, 12.23. Use a t-test to test for any difference in mean breaking strength between wire A and a standard type whose mean breaking strength is known to be 12.10 units.

2. The following figures give the % extension under a given load for two random samples of lengths of yarn, the first sample being taken before washing and the second after six washings.

Before washing:	12.3	13.7	10.4	11.4	14.9	12.6	
After 6 washings:	15.7	10.3	12.6	14.5	12.6	13.8	11.9

Do they provide any justification for concluding that extensibility is affected by washing? (You may assume that the standard deviation is unaltered by the washings.)

3. In another experiment with the same type of yarn, six lengths of yarn were selected at random and each length cut into two halves. One of the halves was tested for extension without washing, and the other after six washings, giving the following % extensions:

Length:	1	2	3	4	5	6
Before washing:	13.9	12.5	11.0	11.8	10.8	14.6
After 6 washings:	14.7	12.1	13.2	13.6	11.5	15.4

What evidence do these provide as to the effect of washing on extensibility? Why should this experiment be analysed differently from the previous one?

4. A butter-packing machine in a dairy packs butter in 250g packages though the actual weights packed vary. In a trial run, the following deviations from the nominal 250g were recorded:

$$5, 8, 0, 3, -1, 1, 6, 5, 8, 4, 9, 0, 4.$$

It is proposed to replace the machine with a new model which, it is claimed, has smaller variability in package weights. A trial run on the new machine produced the following deviations from the nominal 250g:

$$3, -2, 1, 4, 0, 2, 1, 5, 3, 2, 1, 4.$$

Does the new machine show a significant improvement over the old?

5. Consider dataset 10 from the specimen datasets. Is there any evidence to suggest that temperature affects density?

6. Carry out a linear regression analysis on the following data:

x	-3	-3	-2	-1	0	1	1	2	2	3
y	114	110	111	107	107	108	104	105	103	101

Find a 95% confidence interval for β, the true slope of the regression line, and for the mean value of Y when $x = 1$.

7. The range (in m) of a howitzer with muzzle velocity V m/s fired at angle of elevation α is $g^{-1}V^2 \sin 2\alpha$, where $g = 9.81$. Suppose that 9 shells are fired at different angles of elevation, with the following results:

α	5	10	15	20	25	30	35	40	45
range	4860	9580	14080	18100	21550	24350	26400	27700	28300

(Ranges are given in m, and angles in degrees.) Estimate V from these data.

8. A beaker of liquid is heated, and then removed from the source of heat and left to cool. The temperature of the liquid is measured at the moment it is removed from the heat, and at intervals of 1 minute thereafter. The following results are observed:

time	0	1	2	3	4	5	6
temperature	100	82	60	50	40	32	28

Set up, and study, an appropriate regression model.

9.* Define the *multivariate normal* distribution as follows. If $\mathbf{Y}$ is a p-vector with p-dimensional pdf:

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{|2\pi\Sigma|^{\frac{1}{2}}} \exp -\tfrac{1}{2}\{(\mathbf{y}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\}, \mathbf{y} \in \mathbb{R}^p ,$$

where $\boldsymbol{\mu}$ is a p-vector and Σ a symmetric, positive definite $p \times p$ matrix, then $\mathbf{Y}$ is said to have a multivariate normal distribution, and we write

$$\mathbf{Y} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma) .$$

Prove the following:

(i) $Y_1, \ldots, Y_p$ are independent if and only if Σ is diagonal. Also show that when $\Sigma = kI$, $\boldsymbol{\mu} = \mu\mathbf{1}$ then $Y_1, \ldots, Y_p$ are IID.

(ii) If $\mathbf{Y} \sim \mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$, then $\mathbf{Z} = A\mathbf{Y} \sim \mathrm{N}_p(A\boldsymbol{\mu}, A\Sigma A^T)$. (Note the special case: Lemma 3.4.1 from notes.)

(iii) Show that $\Sigma^{-\frac{1}{2}}(\mathbf{Y}-\boldsymbol{\mu}) \sim \mathrm{N}_p(0, I)$.

(iv) From (iii), deduce that $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu}$, and that the ij^{th} entry of Σ, σ_{ij} is $\mathrm{cov}\,(Y_i, Y_j)$.

10. Suppose (X_i, Y_i) $1 \le i \le n$ are IID from a bivariate normal population with mean $\boldsymbol{\mu} = (\mu_X, \mu_Y)^T$, and $\Sigma = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$.

(a) Calculate the conditional distribution of Y_i given X_i.

(b) Suppose we attempt to fit the linear model

$$Y_i = \alpha + \beta(x_i - \overline{x}) + \varepsilon_i\,.$$

From (a), argue that the hypotheses $H_0 : \beta = 0$ and $H_0 : \rho = 0$ are equivalent.

(c) Letting $S^2 = (n-2)^{-1}\sum_{i=1}^n (Y_i - \hat{\alpha} - \hat{\beta}(x_i - \overline{x}))^2$, and $S_{XX} = \sum_{i=1}^n (x_i - \overline{x})^2$, show that

$$\frac{\sqrt{S_{XX}}\hat{\beta}}{S} = \sqrt{n-2}\,\frac{r}{\sqrt{1-r^2}}\,.$$

where r is the sample covariance, $\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x}) / (\Sigma(y_i - \overline{y})^2 \Sigma(y_i - \overline{x})^2)^{\frac{1}{2}}$. Therefore construct a test of $H_0 : \beta = 0$ depending only on r^2 and n.

11.* A random sample of 59 people from the planet Krypton yielded the following results:

		Eye-colour	
		1 (Blue)	2 (Brown)
Sex	1 (Male)	19	10
	2 (Female)	9	21

Professor A had believed that sex and eye-colour are independent factors on Krypton. After doing a χ^2 test of his hypothesis against the alternative

$$H_1 : \sum_i \sum_j p_{ij} = 1,$$

he finds that he has to reject it at the 5% level and even at the 1% level.

Professor B has always believed the much stronger hypothesis that $p_{ij} = \frac{1}{4}$ for all i and j. After doing a χ^2 test of his hypothesis against H_1, he finds that he does not need to reject it at the 5% level.

You are asked for expert statistical advice: firstly, to check the calculations, and secondly, to comment on whether or not Statistics is an absurd subject. Submit your report to your supervisor.

12.* (Strong Likelihood Principle) On the Planet Altair A, all nuclear power stations were installed exactly 20 years ago by ANFL. It is known that cataclysmic accidents occur at a Poisson process of constant (but unknown) rate λ per year. The first cataclysmic accident has just happened. ANFL argue that it is not fair to apply statistics just after an accident has occurred. However GreenAlt, the local conservation movement, argues as follows:

1°. The likelihood of getting precisely 1 accident (some time) in a 20-year period is

$$(20\lambda)e^{-20\lambda}$$

by the formula for the Poisson distribution.

2°. The likelihood that the first accident occurs at time 20 is

$$\lambda e^{-20\lambda}$$

by the formula for the exponential distribution.

3°. Since these likelihoods are proportional and since the constant of proportionality, 20, will drop out of every statistical consideration, any inferences about λ made on the basis that the first accident has just occurred must be the same as those based on the fact that precisely 1 accident has occurred in all in 20 years.

Most of the people-in-the-street on Altair A believe that there is substance in ANFL's argument and that GreenAlt are trying to overturn common sense with silly mathematics. What do you think? The Problem is not unknown on Planet Earth; and GreenAlt's argument, if valid, remains valid if 1 accident is replaced by 3 or 4.