GX 1985

OLLSCOIL NA hÉIREANN, GAILLIMH

NATIONAL UNIVERSITY OF IRELAND, GALWAY

SEMESTER I EXAMINATIONS, 2003/2004

THIRD ARTS, ENGINEERING AND SCIENCE EXAMINATION

Module Code:	**MA337**
Module:	**STATISTICS**
External Examiner	Dr. T. C. Bailey
Internal Examiner	Prof. J. P. Hinde

Instructions:

Duration: Two Hours.

Answer any *Three* questions.

All questions, but not necessarily parts therein, carry equal marks.

Relevant tables and formulæ are supplied.

Requirements:

Graph Paper

Question 1 is on the next page

1. (a) The data below are gross incomes (€1000s) of a random sample of 11 income tax returns.

9.7	93.1	33.0	21.2	81.4	51.1
43.5	10.6	12.8	7.8	18.1	

i) Construct a letter value display for these data and use the mid-summaries to assess whether the data are symmetric or not. What can you say about the shape of the distribution?

If the data were approximately symmetric, explain how you would use the letter values to assess approximate normality.

ii) Produce a normal scores plot for these data by plotting the sorted data against the normal scores $(\Phi^{-1}(i/(n+1)),\ i = 1, \ldots, n$, where Φ is the cumulative distribution function of a standard normal distribution, $N(0, 1))$. Does the normal distribution provide a reasonable model here?

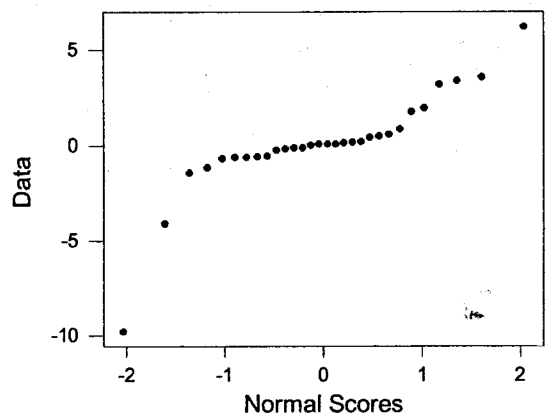
What sort of transformation would you suggest to improve the symmetry, and thus normality, of these data?

(b) The following page has normal scores plots for four different sets of data. For each plot, comment on whether the data appear normal, and, if not, describe how the data departs from normality. Where the data appear to be normal use the plot to obtain a rough estimate of the mean and standard deviation.

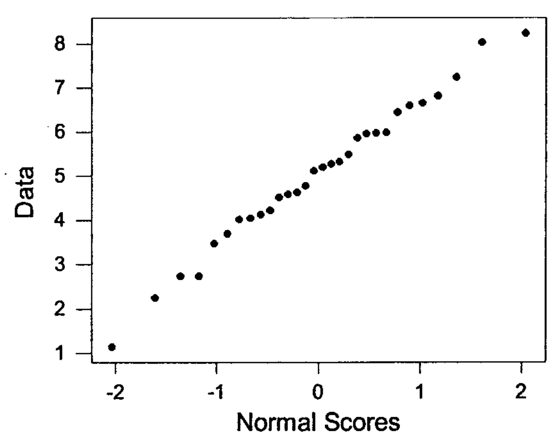
Why are we interested in checking whether data are normal? Are there any circumstances where normality of the original data is not so important?

Plots are on the next page

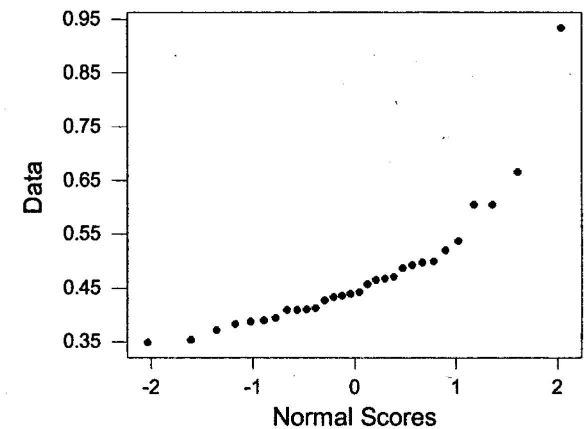
Plot 1



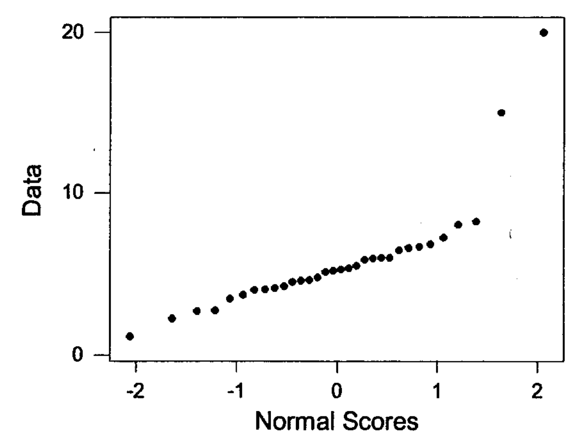
Plot 2



Plot 3



Plot 4



2. (a) The survival time of heart transplant patients can be described by an exponential distribution with mean 10 years, i.e. probability density function

$$f(x) = 0.1e^{-0.1x} \qquad x \geq 0.$$

What is the probability that a particular transplant patient survives for more than

i) 10 years;

ii) 20 years?

Explain the relationship between your answers.

(b) Write brief notes on the *Central Limit Theorem*, including:

- a clear statement of the result.
- why, in certain circumstances, it allows the normal distribution to be used as an approximation to the Poisson and the binomial distributions?
- why it is useful in statistics for inferences on population means based on a sample of size n?

For a random sample of size n from an exponential distribution with mean μ, use the Central Limit Theorem to show that $\overline{X}$, the sample mean, has an approximate $N\left(\mu, \frac{\mu^2}{n}\right)$ distribution.

Use this approximation to show that

$$\left(\frac{\overline{x}}{1 + \frac{z_{\alpha/2}}{\sqrt{n}}}, \frac{\overline{x}}{1 - \frac{z_{\alpha/2}}{\sqrt{n}}}\right)$$

is an approximate $100(1-\alpha)\%$ confidence interval for μ, where $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of the standard normal distribution.

(c) Data from a random sample of 100 heart transplant patients give a mean survival time of 12 years. Assuming that the survival times are exponentially distributed, calculate an approximate 90% confidence interval for the true mean survival time.

Are the data consistent with a claimed mean survival time of 15 years?

3. (a) State the relationships between the Normal, χ^2, t, and F distributions.
Hypothesis tests and confidence intervals on population means and variances commonly involve the Normal, χ^2, t, and F distributions. Identify which distribution is used in each of the following situations:

i) constructing a confidence interval for the variance of a normal population;
ii) comparing the means of two populations each with known variance;
iii) constructing a confidence interval for the difference between population means using paired sample data;
iv) testing the means of two populations with equal but unknown variances;
v) testing the equality of variances for two normal populations with unknown means.

The χ^2, t, and F distributions all depend on *degrees of freedom*; how are these degrees of freedom calculated in each of the above situations? State these in terms of the sample sizes, briefly explaining your answers.

(b) A coffee machine is supposed to dispense 6 fluid ounces (fl oz) of coffee into a paper cup. In reality, the amounts dispensed vary from cup to cup. However, if the machine is working properly, most of the cups will contain within 10% of the advertised 6 fl oz. So approximately the standard deviation of the amounts dispensed should be less than 0.2 fl oz. A random sample of 15 cups provided the following data, in fluid ounces:

5.90	5.82	6.20	6.09	5.93
6.18	5.99	5.79	6.28	6.16
6.00	5.85	6.13	6.09	6.18

Note: The summary statistics from these data are $\bar{x} = 6.039$ and $s = 0.154$.
Calculate a 95% confidence interval for the population mean amount of coffee dispensed, stating any assumptions that you make.
Find a 95% confidence interval for the variance of the population.
At the 5% significance level, do the data provide evidence to conclude that the standard deviation of the amounts being dispensed is less than 0.2 fl oz? Use an appropriate one-sided test on the variance, stating clearly null and alternative hypotheses, and your conclusions.

Note: The table at the end of the paper only gives upper tail percentage points of the χ^2 distribution. You may find some of the following lower tail percentage points also useful here:

$$\chi^2_{14;\,0.975} = 5.629 \qquad \chi^2_{14;\,0.95} = 6.571$$
$$\chi^2_{15;\,0.975} = 6.262 \qquad \chi^2_{15;\,0.95} = 7.261$$

4. (a) Explain briefly the following terms used in hypothesis testing:
   - null hypothesis;
   - test statistic;
   - significance level;
   - power;
   - p-value.

(b) Let X be a normal random variable with variance $\sigma^2 = 16$. Using a sample of size $n = 25$ and a significance level $\alpha = 0.05$, and writing the sample mean as $\overline{X}$, determine the rejection region for testing the hypothesis $H_0 : \mu = 10$ versus $H_1 : \mu > 10$.

Show that the power function for this test is

$$1 - \Phi\left(\frac{11.316 - \mu}{0.8}\right)$$

where $\Phi(t) = P(Z \leq t)$ for $Z \sim N(0, 1)$.
Calculate the power for $\mu = 12$.

(c) The Minitab output on the following page relates to the comparison of two different methods of measuring hormone levels in patients. Blood samples were taken from a random sample of 10 patients and then analysed by each of the methods.

Describe each of the analyses presented on the following page, giving details of the hypotheses being tested, and the conclusions.

Use the plot to help explain your answer and the differences between the results of the two test procedures. Which analysis is more appropriate in this situation?

The two-sample t-test assumes equal variances – use an appropriate hypothesis test at the 5% significance level to establish whether this assumption is reasonable.

Minitab output is on the next page