MATH184001

This question paper consists
of 9 printed pages, each
of which is identified by the
reference **MATH184001**.

Required statistical
tables are included.
Only approved basic
calculators may be used.

Examination for the Module MATH1840
(May/June 2005)

INTRODUCTION TO STATISTICAL MODELLING

Time allowed: **2 hours**

Attempt ALL questions in Section A and THREE questions from Section B.

Section A is worth 40% of the examinations marks.

The marks for each question in Section A are indicated in square brackets in the margin.
All questions in Section B carry equal marks.

Section A

A1. Below are several items pertaining to individuals: [2]

a)	height	e)	number of siblings
b)	weight	f)	religion
c)	age last birthday	g)	place of birth
d)	sex	h)	high school class rank

Which involves continuous variables?

A: f) and g)   B: a) and c)   C: h)   D: a) and b)   E: c) and e)

A2. The sample mean of the following data: [3]

79   516   24   265   41   15   411

is

A: 7   B: 15   C: 79   D: 71.6   E: 193

A3. Why is the sample standard deviation preferable to the sample range as a measure of dispersion? [2]

A4. For the following dataset: [6]

78	104	84	70	96
73	87	85	76	94

Find the **median** and the (approximate) **quartiles**.

A5. According to a census, the age distribution of employees is as shown in the following table. [8]

Age (years)	People (thousands)
16–19	6,549
20–24	13,690
25–34	28,149
35–44	20,879
45–54	15,923
55–64	11,414
65 & over	2,923

An employed person is selected at random. Let

A = event the person is under 20
B = event the person is between 20 and 54, inclusive
C = event the person is under 45
D = event the person is at least 55

QUESTION A5 CONTINUED...

a) Determine the number of persons who comprise the event "A or D"

b) List all the pairs of events which are mutually exclusive

c) What is the probability that the randomly selected person is at least 55?

A6. Suppose that 60% of adults over 30 in a certain community are graduates. Furthermore, suppose that 80% of the graduates over 30 have incomes over $£15,000$. What percentage of adults over 30 in this community are graduates and have incomes over $£15,000$? [3]

A: 48 B: 20 C: 40 D: 12 E: 32

A7. Suppose that the probability of being born on Christmas day is $1/365$. If we have a group of 100 people (whose birthdays are independent) then what is the probability that no one in the group has a Christmas birthday: [4]

A: 0.209 B: 0.969 C: 0.791 D: 0.031 E: 0.760

A8. Briefly explain what is meant by the term *inference* and give an example. [5]

A9. State the Central Limit Theorem, and explain why it is important. [7]

Section B

B1. Suppose two fair dice (say red and green in colour) are rolled. Let R denote the score on the red die and G denote the score on the green die. Let $T = R + G$ denote the total score, and let $L = \max(R, G)$ denote the maximum score.

(a) Consider the events: [8]

A:	R is even	C:	T is 10
B:	G is odd	D:	T is even

i) Compute $\mathsf{P}(A), \mathsf{P}(B), \mathsf{P}(C), \mathsf{P}(D)$

ii) Compute $\mathsf{P}(D \mid A)$

iii) Are the events A and D independent? Why?

(b) Consider now the random variables T and L. [12]

i) What is the mean of T?

ii) What is the sample space of L?

iii) By considering the possible combinations of R and G, or otherwise, obtain the probability distribution of L.

iv) Show that the mean of L is $4\frac{17}{36}$

B2. Data were collected on the magnitude of all earthquakes occurring which registered at least 4.0 on the Richter scale. There were 123 such events since 1st January, 2000.

(a) Considering the graphical displays shown below [10]

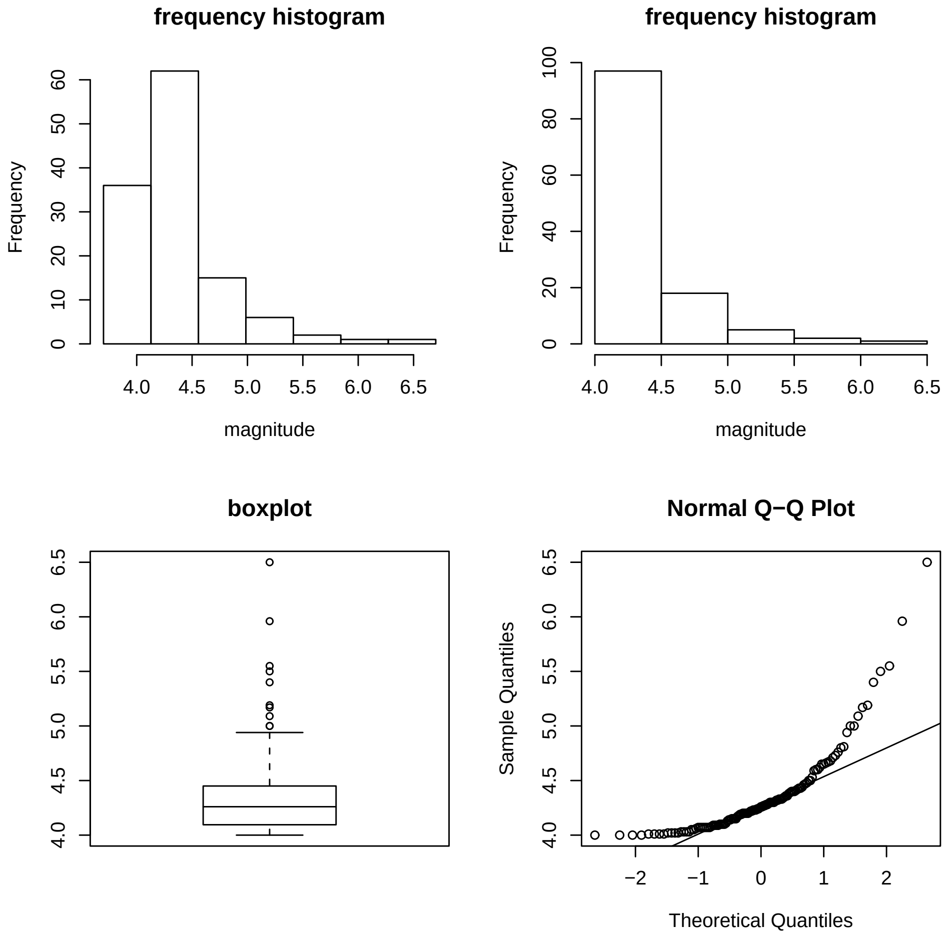
i) explain which histogram you prefer, and why. Describe the distribution.

ii) explain how the boxplot is constructed (if drawing "by hand") and the information this boxplot conveys.

iii) What does the "Normal Q-Q Plot" show?

(b) Under what circumstances might a transformation of the data be useful in data analysis? Illustrate your answer by an example. [6]

(c) What transformation could usefully be used for the earthquake data? [4]



B3. The following data represent porosity measurements (%) on ten samples of Tensleep Sandstone, Pennsylvanian, from the Bighorn Basin, Wyoming.

$$13, 17, 15, 23, 27, 29, 18, 27, 20, 24$$

Calculate a 95% confidence interval for the mean. How do you interpret this interval? [8]

State what assumptions are being made about the data, and what results are being used. Describe how you would check your assumptions. [8]

Without doing any further calculations, do the data support the hypothesis that the mean porosity is 24.0% (which was claimed in a previous study). [4]

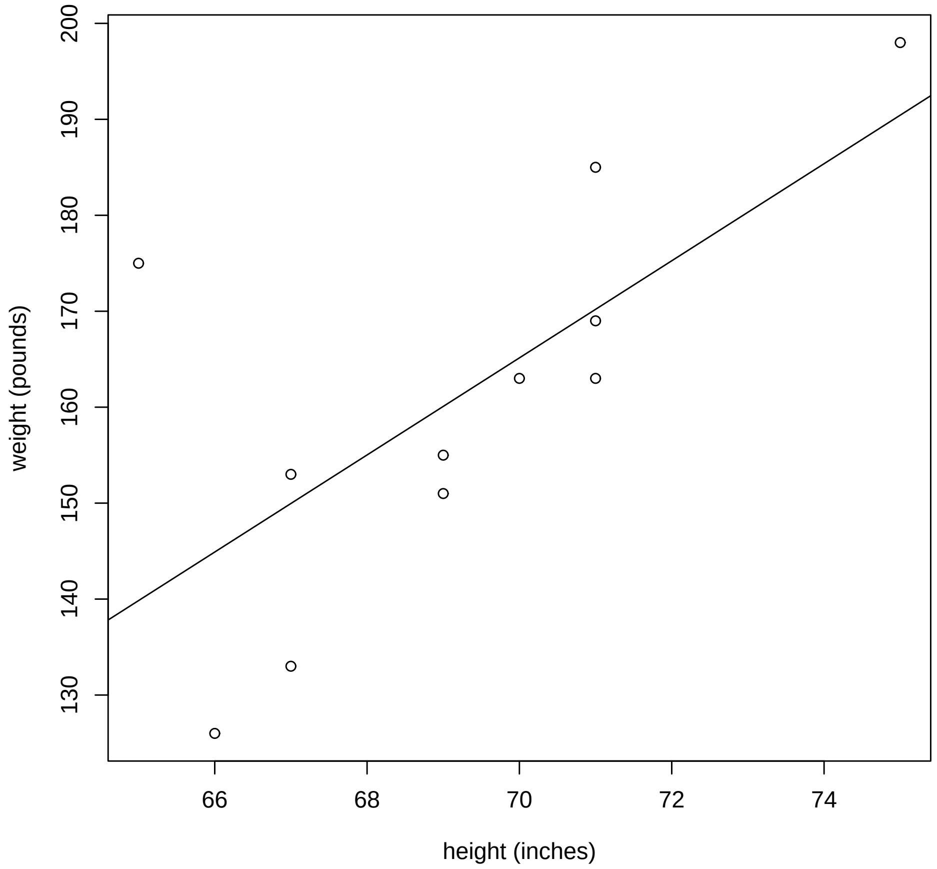
B4. The height (x) and weight (y) of 11 males was measured.

(a) If x is measured in inches, and y is measured in pounds, then what are the units of the correlation between x and y? What does the correlation between x and y measure? In what range does the correlation lie? [4]

(b) The data, and some analysis in R, are shown below. (Some of the output has been deliberately removed.) [16]

```
> x = c(65, 67, 71, 71, 66, 75, 67, 70, 71, 69, 69)
> y = c(175, 133, 185, 163, 126, 198, 153, 163, 169, 151, 155)
> cor(x, y)
[1] 0.6873831
> var(x, y)
[1] 41.3
> plot(x, y, xlab = "height (inches)", ylab = "weight (pounds)")
> lm1=lm(y ˜ x)
> abline(lm1)
> summary(lm1)

Call:
lm(formula = y ˜ x)

Residuals:
    Min      1Q  Median      3Q     Max
-18.903  -8.139  -2.139   5.302  35.156

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -188.992    123.364  -1.532   0.1599
x              5.059      1.782   2.839   0.0194 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.1 on 9 degrees of freedom
Multiple R-Squared: 0.4725,Adjusted R-squared: 0.4139
F-statistic: 8.061 on 1 and 9 DF,  p-value: 0.01943
```

QUESTION B4 CONTINUED...



Using the above output and plot to help, what is

i) the correlation between x and y

ii) the equation of the fitted regression line. How do you interpret the intercept and the slope?

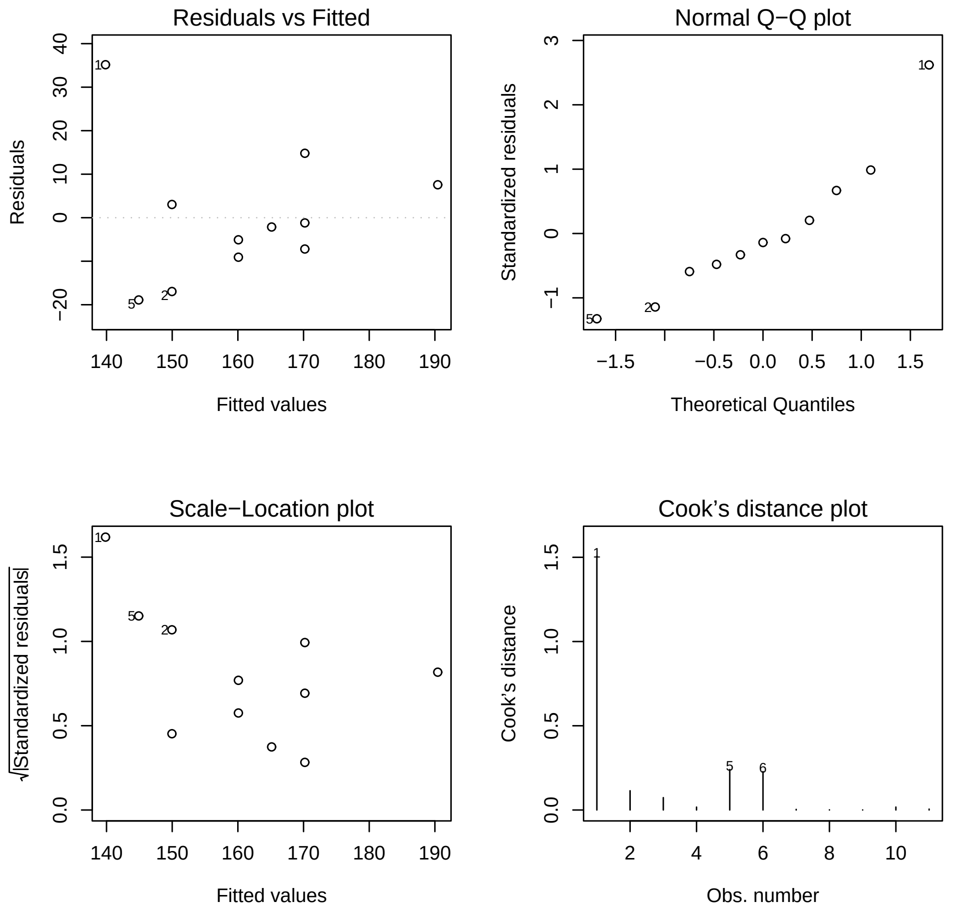
iii) the 95% confidence interval of the slope

What would you predict the weight for someone of height 73 inches?

What do you conclude from the plot overleaf which was obtained from the command

```
> plot(lm1)
```

and what further analysis would you suggest on these data?



Residuals vs Fitted

Normal Q−Q plot

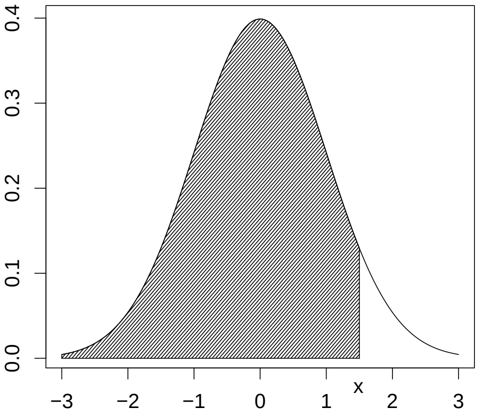
Scale−Location plot

Cook’s distance plot

Normal Distribution Function Tables

The first table gives

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{1}{2}t^2} dt$$



and this corresponds to the shaded area in the figure to the right. $\Phi(x)$ is the probability that a random variable, normally distributed with zero mean amd unit variance, will be less than or equal to x. When $x < 0$ use $\Phi(x) = 1 - \Phi(-x)$, as the normal density with mean zero is symmetric about zero. To interpolate, use the formula

$$\Phi(x) \approx \Phi(x_1) + \frac{x - x_1}{x_2 - x_1} \big(\Phi(x_2) - \Phi(x_1)\big)$$

Table 1

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0.00	0.5000	**0.50**	0.6915	**1.00**	0.8413	**1.50**	0.9332	**2.00**	0.9772	**2.50**	0.9938
0.05	0.5199	**0.55**	0.7088	**1.05**	0.8531	**1.55**	0.9394	**2.05**	0.9798	**2.55**	0.9946
0.10	0.5398	**0.60**	0.7257	**1.10**	0.8643	**1.60**	0.9452	**2.10**	0.9821	**2.60**	0.9953
0.15	0.5596	**0.65**	0.7422	**1.15**	0.8749	**1.65**	0.9505	**2.15**	0.9842	**2.65**	0.9960
0.20	0.5793	**0.70**	0.7580	**1.20**	0.8849	**1.70**	0.9554	**2.20**	0.9861	**2.70**	0.9965
0.25	0.5987	**0.75**	0.7734	**1.25**	0.8944	**1.75**	0.9599	**2.25**	0.9878	**2.75**	0.9970
0.30	0.6179	**0.80**	0.7881	**1.30**	0.9032	**1.80**	0.9641	**2.30**	0.9893	**2.80**	0.9974
0.35	0.6368	**0.85**	0.8023	**1.35**	0.9115	**1.85**	0.9678	**2.35**	0.9906	**2.85**	0.9978
0.40	0.6554	**0.90**	0.8159	**1.40**	0.9192	**1.90**	0.9713	**2.40**	0.9918	**2.90**	0.9981
0.45	0.6736	**0.95**	0.8289	**1.45**	0.9265	**1.95**	0.9744	**2.45**	0.9929	**2.95**	0.9984
0.50	0.6915	**1.00**	0.8413	**1.50**	0.9332	**2.00**	0.9772	**2.50**	0.9938	**3.00**	0.9987

The inverse function $\Phi^{-1}(p)$ is tabulated below for various values of p.

Table 2

p	**0.900**	**0.950**	**0.975**	**0.990**	**0.995**	**0.999**	**0.9995**
$\Phi^{-1}(p)$	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905

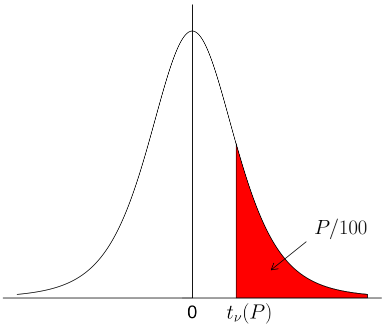
CONTINUED...

Percentage Points of the t-Distribution

This table gives the percentage points $t_\nu(P)$ for various values of P and degrees of freedom ν, as indicated by the figure to the right.

The lower percentage points are given by symmetry as $-t_\nu(P)$, and the probability that $|t| \geq t_\nu(P)$ is $2P/100$.

The limiting distribution of t as $\nu \to \infty$ is the normal distribution with zero mean and unit variance.



	Percentage points P						
ν	**10**	**5**	**2.5**	**1**	**0.5**	**0.1**	**0.05**
1	3.078	6.314	12.706	31.821	63.657	318.309	636.619
2	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.120	2.583	2.921	3.686	4.015
18	1.330	1.734	2.101	2.552	2.878	3.610	3.922
21	1.323	1.721	2.080	2.518	2.831	3.527	3.819
25	1.316	1.708	2.060	2.485	2.787	3.450	3.725
30	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	1.299	1.676	2.009	2.403	2.678	3.261	3.496
70	1.294	1.667	1.994	2.381	2.648	3.211	3.435
100	1.290	1.660	1.984	2.364	2.626	3.174	3.390
∞	1.282	1.645	1.960	2.326	2.576	3.090	3.291