

Higher Certificate in Statistics

May 1999

SOLUTIONS

Paper I :

Statistical Theory

1 Outcome probabilities are:

(i)  $P(H/B) = 1/2$  and so  $P(T/A) = P(T/B) = 1/2$  .

Hence every entry in the body of the table is the product of marginal probabilities (by independence),  $1/2 \times 1/2 = 1/4$ .

Writing  $G_A = \text{Gain for A}$ ,  $G_B = \text{Gain for B}$  ,

$$E[G_A] = (3 \times 1/4) - (2 \times 1/4) - (2 \times 1/4) + (1 \times 1/4) = 0 = E[G_B]$$

since this is a two person game.

(ii)  $E[G_A] = (3 \times 2/9) - (2 \times 4/9) - (2 \times 1/9) + (1 \times 2/9) = -2/9$

(iii) giving  $E[G_A] = 3p_A p_B - 2p_A(1 - p_B) - 2(1 - p_A)p_B + (1 - p_A)(1 - p_B) = 1 - 3p_A - 3p_B + 8p_A p_B$

If  $p_B = 3/8$  ,  $E[G_A] = 1 - 3p_B$  from the first alternative form; in this case  $E[G_A] = 1 - 9/8 = -1/8$

Hence choose to play as B, with  $p_B = 3/8$  , which in the long run will guarantee a win of  $1/8$  (since  $E[G_B] = -E[G_A]$ )

2 (i)  $X$  is binomially distributed  $B(n, p)$  and so  $P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}$  ,  $x = 0, 1, 2 \dots n$ .

Hence

$$E[X] = \sum_{x=0}^n x p^x (1 - p)^{n-x} \frac{n!}{x!(n-x)!} = np \sum_{x=0}^n \frac{(n-1)! p^{x-1} (1-p)^{n-x}}{(x-1)!(\{n-1\} - \{x-1\})!}$$

since the term in  $E[X]$  for the value  $x = 0$  is zero. Put  $Y = (x - 1)$

Thus

$$E[X] = np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{n-y} = (p + (1-p))^{n-1} np = np$$

The expression for variance is  $V[X] = np(1-p)$ .

(ii) (a)  $Y$  is  $B(4, 1/2)$  so  $P(Y = y) = \binom{4}{y} 1/2^4$  for  $y = 0, 1, 2, 3, 4$

$$P(Y \geq 3) = P(3) + P(4) = 4 \times 1/16 + 1 \times 1/16 = 5/16$$

$$P(Y = 2) = \frac{4 \times 3 \times 1}{2 \times 1 \times 16} = 3/8.$$

(b)  $E[Y] = 2, V[Y] = 1.$

(c) Now  $Y = Z_1 + z_2 + Z_3 + Z_4$ , where each  $Z_i$  is a Bernoulli variable with mean  $p_i$  and variable  $p_i(1 - p_i)$  By independence,  $E[Y] = \sum_{i=1}^4 E[Z_i]$  and  $V[Y] = \sum_{i=1}^4 V[Z_i]$  so  $E[Y] = 3/4 + 1/3 + 2/3 + 1/4 = 2$  (the same as before since the new probabilities average to 1/2).

Also  $V[Y] = 3/4 \times 1/4 + 1/3 \times 2/3 + 2/3 \times 1/3 + 1/4 \times 3/4 = 3/16 + 2/9 + 2/9 + 3/16 = 59/72$  (less than before)

NOTE: for probabilities which average to 1/2, the first case gives the maximum sine if  $p_i \neq 1/2$  the product  $p_i(1 - p_i)$  is  $< 1/4$  for each component.

3 For boys' heights,  $n_1 = 100, X \sim N(160, 16)$ ; and for girls' heights,  $n_2 = 81, Y \sim N(150, 9)$

(i) (a)  $P(X > 156) = P(\frac{X-160}{4} > \frac{156-160}{4}) = P(Z > -1)$  where Z has the distribution  $N(0,1)$ . This is the same as  $P(Z < 1)$ , which is 0.8413.

(b)  $P(Y > 156) = P(\frac{Y-150}{3} > \frac{156-150}{3}) = P(Z > 2) = 0.0228$

(c) Probability= $P(> 156|boy)P(boy) + P(> 156|girl)P(girl) = 0.8412 \times 100/181 + 0.0228 \times 81/181$  if selection is random from the whole population. This is 0.4750.

(ii) (a) Assuming that the boy's heights are independent, the required probability is  $(0.8413)^4 = 0.5010$ .

(b) Mean height of boys  $\sim N(160, 16/4) \sim N(160, 4)$

Hence  $P(\text{mean} > 156) = P(\frac{\text{mean}-160}{2} > \frac{156-160}{2}) = P(Z > -2) = P(Z < 2)$  by symmetry=0.9772.

The assumption is likely to be reasonable except when, for example, they come from the same family.

(iii) Assuming independence again,  $X - Y \sim N(160 - 150, 16 + 9) \sim N(10, 25)$ .

$P(X - Y > 0) = P(\frac{X-Y-10}{5} > \frac{0-10}{5}) = P(Z > -2) = 0.9772$  Both X and Y are assumed chosen from the year group.

(iv) Mean height is  $\frac{n_1\bar{X}+n_2\bar{Y}}{n_1+n_2} = W$ , and  $\bar{X} \sim N(160, 16/100), \bar{Y} \sim N(150, 9/81)$ . i.e.  $\bar{X} \sim N(160, 0.16)$ ;  $\bar{Y} \sim N(150, 1/9)$  We require  $P(155.5 < W < 156.5)$  or  $P(W < 156.5) - P(W < 155.5)$

$V[W] = (\frac{n_1}{n_1+n_2})^2 V[\bar{X}] + (\frac{n_2}{n_1+n_2})^2 V[\bar{Y}] = (\frac{100}{181})^2 (0.16) + (\frac{81}{181})^2 (\frac{1}{9}) = 0.00488386 = 0.0222521 = 0.071091$  and  $E[W] = \frac{100 \times 160 + 81 \times 150}{181} = 155.52486$

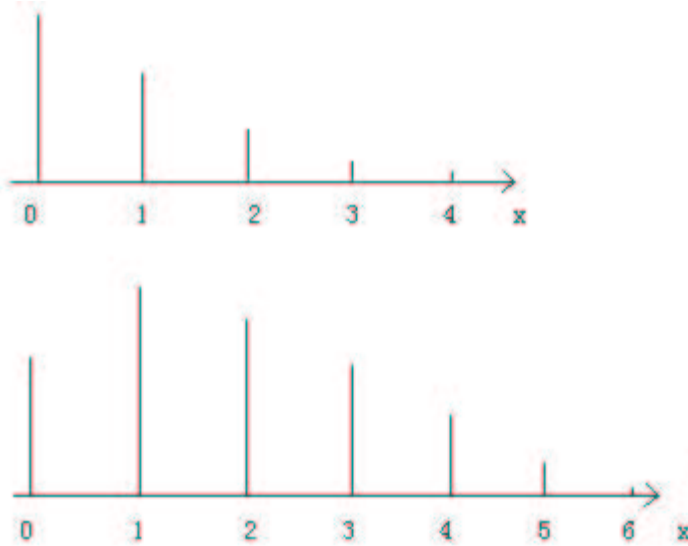
Z-value corresponding to  $W = 155.5$  is  $\frac{155.5 - 155.52486}{\sqrt{0.071091}}$  i.e.  $-\frac{0.02486}{0.26663} = -0.0932$  and for 156.5

$Z = \frac{156.5 - 155.52486}{\sqrt{0.071091}} = \frac{0.97514}{0.26663} = 3.657$

$P(Z < -0.0932) = 0.4629$  and  $P(Z < 3.657) = 0.9999$  is required probability=0.9999-0.4629=0.5370.

4 (i) For  $u = 1/2, P(0) = e_{-1/2} = 0.60653, P(1) = 0.30327, P(2) = 0.07582$  and  $P(3) = 0.01264$  with  $P(\geq 4)$  very small.

For  $u = 2, p(0) = 0.13534, p(1) = 0.27067 = p(2), p(3) = 0.18045, p(5) = 0.03609, p(6) = 0.01203$



(ii) (a) The sampling method does not collect any data for which  $y=0$ . Hence in this distribution  $\sum_{y=1}^{\infty} P(y) = 1$

But since we have a Poisson distribution originally we know that  $P(y = 0) = e^{-u}$ , and  $P(y \geq 1) = 1 - e^{-u}$ .

This the total probability for the data that have been collected is  $1 - e^{-u}$ , and each individual probability must be expressed as a proportion of this in order to make them add to 1.

Therefore  $P(Y = y) = \frac{e^{-u} u^y}{(1 - e^{-u}) y!}$  for  $y = 1, 2, 3, \dots$

$E[Y] = \sum_{y=1}^{\infty} \frac{u^y y e^{-y}}{(1 - e^{-u}) y!} = \frac{u e^{-u}}{1 - e^{-u}} \sum_{y=1}^{\infty} \frac{u^{y-1}}{(y-1)!} = \frac{u e^{-u}}{1 - e^{-u}} \sum_{s=0}^{\infty} \frac{u^s}{s!} = \frac{u}{1 - e^{-u}}$  since the  $\sum_s$  factor equals  $e^u$ .

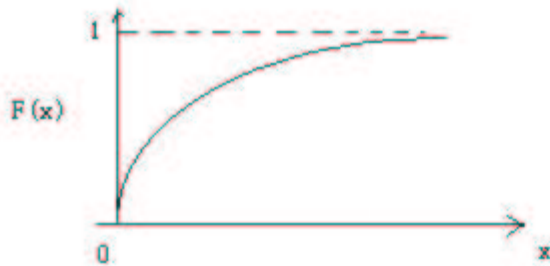
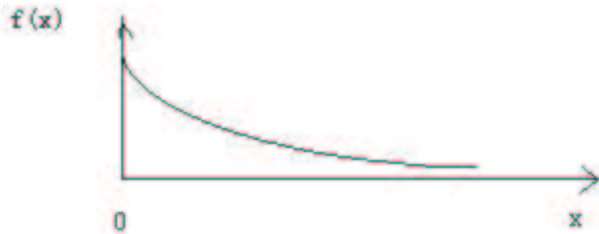
(b)  $E[Y] - P(Y = 1) = \frac{u}{1 - e^{-u}} - \frac{u e^{-u}}{1 - e^{-u}} = u$

(c) Hence the sample fraction of leaves with just one insect, subtract from the sample mean (which is an unbiased estimate of  $E[Y]$ ) gives an unbiased estimate of  $u$ .

5 (i)  $F(x) = P(X \leq x) = \int_0^x \lambda e^{-\lambda u} du = [-e^{-\lambda u}]_0^x = 1 - e^{-\lambda x} \quad (x \geq 0)$

(ii)  $L = \prod_{i=1}^n \lambda e^{-\lambda x_i} = \lambda^n \exp(-\lambda \sum_{i=1}^n x_i) = \lambda^n e^{-n\lambda \bar{x}} \quad \ln(L) = n \ln \lambda - n\lambda \bar{x}$

$\frac{d}{d\lambda}(\ln(L)) = 1/\lambda - n\bar{x} = 0$  for  $\hat{\lambda} = 1/\hat{x} \quad \frac{d^2}{d\lambda^2}(\ln(L)) = -n/\lambda^2 < 0$  for all  $\lambda$ , so there is a maximum for  $\hat{\lambda}$



(iii) Using the result just found,  $\hat{\lambda} \sim N(\lambda, \lambda^2/n) \sim N(\lambda, 1/n(\bar{x})^2)$ , and so approximately  $\frac{\hat{\lambda}-\lambda}{(\bar{x}\sqrt{n})^{-1}} \sim N(0, 1)$ , from which an approximate 95% confidence interval for  $\lambda$  is given by  $\hat{\lambda} \pm 1.96/(\bar{x}\sqrt{n})$  or  $\frac{1}{\bar{x}}(1 \pm \frac{1.96}{\sqrt{n}})$

6 (i)  $P(\text{device fails}) = P(\text{both components faulty}) = p^2$ , by independence. Hence  $P(\text{device works}) = 1 - p^2$ . Again assuming independence of performance of individual devices, R is Binomial  $(n, 1 - p^2)$

(ii) For line A,  $P(\text{device faulty}) = 1/4$ ; and for B,  $P(\text{faulty}) = 1/9$

(iii) Let event F denote failure/faulty.

$P(F|A) = 1/4$ ;  $P(F|B) = 1/9$ , from (ii)

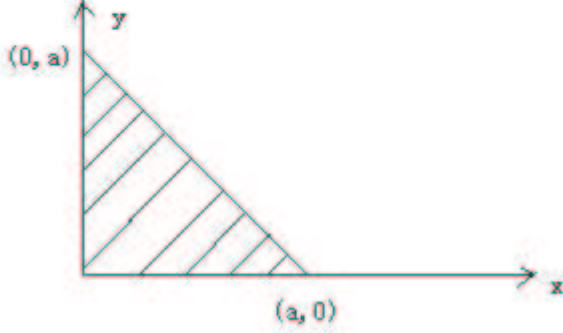
(a) If  $P(A) = P(B) = 1/2$ ,  $P(A|F) = \frac{P(F|A)P(A)}{P(F)} = \frac{P(F|A)P(A)}{P(F|A)P(A) + P(F|B)P(B)} = 9/13$

(b) If  $P(A) = 1/4$ ;  $P(B) = 3/4$ ,  $P(A|F) = 3/7$

Although individually devices from A are more likely to fail, in case(b) more devices are made on line B which results in a higher probability for(b).

7 (i) The distribution is defined in the area shown by the shaded triangle.

(ii)  $f_x(x) = \int_0^{a-x} f(x, y)dy = \frac{2}{a^2} \int_0^{a-x} dy = \frac{2(a-x)}{a^2}$  for  $0 \leq x \leq a$ , and 0 otherwise  
 $F_X(x) = \int_0^x f_X(u)du = \frac{2}{a^2} \int_0^x (a-u)du = \frac{2}{a^2} [au - \frac{1}{2}u^2]_0^x = \frac{2ax-x^2}{a^2}$  for  $0 \leq x \leq a$ .



(iii)  $E[X] = \int_0^a x f_X(x) dx = \frac{2}{a^2} \int_0^a x(a-x) dx = \frac{2}{a^2} [\frac{1}{2}ax^2 - \frac{1}{3}x^3]_0^a = \frac{a}{3}$  by symmetry,  $E[Y] = \frac{a}{3}$  also.

(iv)  $E[x^2] = \int_0^a x^2 f_X(x) dx = \frac{2}{a^2} \int_0^a x^2(a-x) dx = \frac{2}{a^2} [\frac{1}{3}ax^3 - \frac{1}{4}x^4]_0^a = \frac{a^2}{6}$  by symmetry  $E[Y^2] = \frac{a^2}{6}$  also.

$$V[X] = V[Y] = \frac{a^2}{6} - \left(\frac{a}{3}\right)^2 = \frac{a^2}{18}.$$

(V)  $E[XY] = \int_{x=0}^a \int_{y=0}^{a-x} xy f(x,y) dy dx = \int_0^a x \left\{ \int_0^{a-x} \frac{2y}{a^2} dy \right\} dx$   
 $= \int_0^a x \left[ \frac{y^2}{a^2} \right]_0^{a-x} dx = \frac{1}{a^2} \int_0^a x(a-x)^2 dx = \frac{1}{a^2} \int_0^a (a^2x - 2ax^2 + x^3) dx$   
 $= \frac{1}{a^2} [\frac{1}{2}a^2x^2 - \frac{2}{3}ax^3 + \frac{1}{4}x^4]_0^a = \frac{1}{12}a^2.$

$$Cov(X, Y) = E[XY] - E[X]E[Y] = \frac{a^2}{12} - \frac{a^2}{9} = -\frac{a^2}{36},$$

$$\text{Hence } P_{xy} = \frac{cov(X,Y)}{\sqrt{V(X)V(Y)}} = -\frac{a^2/36}{a^2/18} = -1/2.$$

- 8 (i) It is assumed that  $\{e_i : i = 1 \text{ to } n\}$  are independently normally distributed with constant variance  $\sigma^2$ , and mean 0.

$$\frac{\partial S}{\partial \beta_0} = -2(\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})) = 0 \text{ ie. } \sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_i x_{1i} + \beta_2 \sum_i x_{2i}, \text{ or}$$

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 \text{ which gives } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^n x_{1i} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i}) = 0 \text{ when}$$

$$\sum_{i=1}^n x_{1i} y_i = \sum x_{1i} \hat{\beta}_0 + \hat{\beta}_1 \sum_i x_{1i}^2 + \hat{\beta}_2 \sum_i x_{1i} x_{2i} = \sum_i x_{1i} (\bar{y} - \hat{\beta}_1 \bar{x}_1 - \hat{\beta}_2 \bar{x}_2) + \hat{\beta}_1 \sum_i x_{1i}^2 + \hat{\beta}_2 \sum_i x_{1i} x_{2i}.$$

i.e.  $\sum_1^n x_{1i} (y_i - \bar{y}) = \hat{\beta}_1 (\sum_i x_{1i}^2 - \bar{x}_1 \sum_i x_{1i}) + \hat{\beta}_2 (\sum_i x_{1i} x_{2i} - \bar{x}_2 \sum_i x_{1i}) = \hat{\beta}_1 (\sum_i x_{1i}^2 - \frac{\{\sum x_{1i}\}^2}{n}) + \hat{\beta}_2 (\sum x_{1i} x_{2i} - \frac{\{\sum x_{1i}\} \{\sum x_{2i}\}}{n})$  Now  $\sum_i (y_i - \bar{y}) \bar{x}_1 = 0$ , Since  $\bar{x}_1$  is constant over the summation and  $\sum y_i - \bar{y} \bar{x}_1 = 0$  by definition of  $\bar{y}$ .

$$\text{Also } \sum_i x_i^2 - \frac{(\sum x_i)^2}{n} = \sum (x_i - \bar{x})^2, \text{ So this may be written } \sum_i^n (x_{1i} - \bar{x}_1)(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n (x_{1i} - \bar{x}_1)^2 + \hat{\beta}_2 \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2)$$

In exactly the same way,

$$\sum_{i=1}^n (x_{2i} - \bar{x}_2)(y_i - \bar{y}) = \hat{\beta}_1 \sum_{i=1}^n (x_{1i} - \bar{x}_1)(x_{2i} - \bar{x}_2) + \hat{\beta}_2 \sum_{i=1}^n (x_{2i} - \bar{x}_2)^2.$$

(ii) A simple notation for the equations is  $S_{1Y} = S_{11}\hat{\beta}_1 + S_{12}\hat{\beta}_2$ ,  $S_{2Y} = S_{12}\hat{\beta}_1 + S_{22}\hat{\beta}_2$   
 $S_{11} = 28028 - (8 \times 59)^2/8 = 180$ ,  $S_{22} = 29700 - (8 \times 60)^2/8 = 900$ ,  $S_{12} = 28680 - 8 \times 59 \times 60 = 360$ ,  $S_{1Y} = 26860 - 8 \times 55 \times 59 = 900$ ,  $S_{2Y} = 28560 - 8 \times 55 \times 60 = 2160$ .  
Therefore  $900 = 180\hat{\beta}_1 + 360\hat{\beta}_2$   $2160 = 360\hat{\beta}_1 + 900\hat{\beta}_2$ .

These reduce to  $5 = \hat{\beta}_1 + 2\hat{\beta}_2$ , giving  $\hat{\beta}_1 = 1$ ;  $\hat{\beta}_2 = 2$  and  $12 = 2\hat{\beta}_1 + 5\hat{\beta}_2$   $\hat{\beta}_0 = 55 - 59 - (2 \times 60) = -124$ .

The multiple regression equation is  $y = -124 + x_1 + 2x_2$ , The total sum of squares  $\sum(y_i - \bar{y})^2 = 29600 - 8 \times 55^2 = 5400$  The residual (error) s.s. will have  $8 - 3 = 5$ d.f.; hence its value is  $5 \times 36 = 180$

The regression s.s.(2df) is  $5400 - 180 = 520$

$F_{(2,5)} = \frac{520/2}{36} = 72.5$ (0.1% level), very highly significant and so the model is a satisfactory fit to the data.  $\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{1}{\sqrt{1}}$ , *n.s. as  $t_s$* ;  $\frac{\hat{\beta}_2}{SE(\hat{\beta}_2)} = \frac{2}{\sqrt{0.2}} = 4.472$ .

Hence  $\beta_1$  is not significant in the presence of  $\beta_2$ ; but  $\beta_2$  is significant in the presence of  $\beta_1$ . From all these tests we can infer that a simple linear regression of  $y$  on  $x_2$ , arithmetic ability, is adequate as a model.

### *Paper II Statistical Methods*

1 possible approaches would be:

- study the patterns of figures on selected rows,
- consider these figures as proportions of total expenditure,
- plot scatter grams to compare two rows,
- distinguish major from trivial items,
- distinguish luxuries from necessities.

There is not time to do regression or correlation analysis, and the calculation need to be direction and simple. Diagrams might include pie charts, bar charts, simple plots using deciles as horizontal axis, as well as scatter grams.

The choice of low, middle and high income groups could be made using any general definitions known, e.g, for a candidate's particular country;but this would introduce an extra step into the calculations by having to combine deciles. In any case, a "steady trend" shown by 10 points may be visually more convincing.

- (1) We may ignore 'Miscellaneous', being such a small part of the total.
- (2) Tobacco is fairly constant across all deciles; on further comment needed.
- (3) Food is a necessity: % of *thir row ÷ bottom row is*

1	2	3	4	5	6	7	8	9	10	ALL
28.4	22.5	21.8	19.0	18.8	17.4	16.9	16.8	15.3	13.9	17.0

These %'s could be plotted against decile 1-10 (on the x axis) and shows a steady decline from 1 to 10 decile.

- (4) Housing is essential but is provided in a variety of ways, so is with comment in a similar form: % *first row ÷ bottom row is*

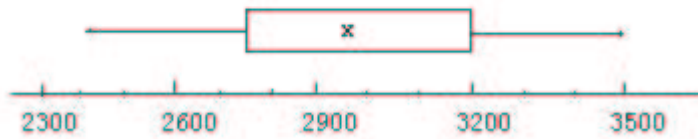
1	2	3	4	5	6	7	8	9	10	ALL
15.0	16.3	16.2	16.0	17.3	15.2	16.6	15.9	15.6	14.5	15.7

As a percentage it stays fairly constant, but an alternative plot would be a scatter gram of actual expenditure on housing against total expenditure—even for a serious news paper this may be less easy to understand than “% constant but total goes up, so housing does too”.

- (5) Somewhere the total expenditure pattern over deciles needs to be shown, especially if it is used as a basis for several different percentages.
- (6) Fuel and power is best plotted on its own, in the Y-direction against decile number as  $x$ ; a jump from 1 to 2, then a steady increase, finally a jump from 9 to 10—could be a function of size of house or of size of household, age structure etc. so some care in interpreting the upper and lower ends should be taken. There are other categories similar to this one.
- (7) Some categories could be combined, e.g. Motoring with Fare etc; although locality obviously affects the balance (city/country) as does place of work, type of work—though, for whatever reason, motoring on its own shows a very marked steady increase, possibly due in part to model and age of car.
- (8) Leisure goods and services could be studied together, although in deciles 9 and 10 they behave rather differently, probably showing what they are made up of different components in different deciles.
- (9) It may be worth drawing pie charts for a selection of deciles, e.g. lowest, middle, highest, showing the %'s of their expenditure in each of the 14 categories. But this needs a lot of arithmetic; it also leads to rather full pie charts unless some categories are combined.
- 2 (i) For the combined data (both sexes), minimum=2412, maximum=3473; quartiles are 2761.5 and 3201.0, median=2951.5.

The median is below the center of the box, so that the central half of the weights show some tendency to have more below “average” than above. The lower whisker is somewhat longer than the upper one, suggesting that relatively small babies can be distinctly small while relatively large ones are not quite so far from the others.

The second smallest is 2596, 184 above the minimum, whereas the second largest is 3462 (and the next 3386), new to the maximum. So the smallest of all may be an “outlier”. Some medical information would be interesting.



(ii) It seems not unreasonable to assume that we have a sample from an approximately normal distribution, and on this basis the 95% confidence interval for  $\mu$  is  $\bar{x} \pm t_{23} \sqrt{s^2/24}$ , where  $\bar{x} = \sum x_i/24$  and  $s^2 = \frac{1}{23}(\sum x_i^2 - (\sum x_i)^2/24)$   $\sum x_i = 71227$ . Therefore  $\bar{x} = 2967.79$  and  $s^2 = 83213.2156$ , so  $s = 288.467$ . The (approximate) interval is  $2967.79 \pm 2.069 \times 288.467/4.899$  or 2846 to 3090.

(iii) Full-term babies:

Males 3279, 3462, 3386, 3199, 3153, 3067.  $n = 6$   $\sum x_i = 19546$ ,  $\bar{x} = 3257.67$ ,  $s^2 = 21885.5$ ,  $s = 147.94$ ;

Females 2967, 3203, 3294, 3473, 2962, 2952.  $n = 6$ ,  $\sum x_i = 18851$ ,  $\bar{x} = 3141.83$ ,  $s^2 = 47102.2$ ,  $s = 217.03$ .

$F_{(5,5)} = 2.15$  for the ratio of variances, which is not significant so it is valid to pool them and use  $s_0^2 = 68987.6335/2 = 34493.82$ . so  $s = 185.725$ .

A two-sample t-test can be used:  $\frac{\bar{x}_M - \bar{x}_F}{S_0 \sqrt{1/t+1/t}}$  that is  $115.84/107.23 = 1.08$  n.s. as  $t_{(10)}$ , giving no evidence against the Null Hypothesis of equal mean weights. Each population must be assumed normally distributed, with the same variance.

3 (a) (i) Type I Error is to reject the Null Hypothesis  $H_0$  when it is true. Level of significance,  $\alpha$ , is the probability of making a Type I Error.

(ii) Type II Error is to "accept", not reject, when it is false. Power is the probability of rejecting  $H_0$  when it is false. So if  $P(\text{Type II Error}) = \beta$ ,  $\text{power} = (1 - \beta)$

(b) (i)  $H_0$  is that the number of accidents is uniformly distributed. The sum of the number of accidents is 65, and so the expected number per day is 65/6. Thus we have the table:

Day	M	Tu	W	Th	F
OBS	17	10	12	11	15
WXP	13	13	13	13	13

and  $\chi_{(4)}^2 = \sum_M \frac{(O-E)^2}{E}$  is a test of the NH against an Alt that the numbers vary from

day to day. Hence  $\chi^2_{(4)} = \frac{1}{13}(4^2 + 3^2 + 1^2 + 2^2 + 2^2) = \frac{34}{13} = 2.615$  n.s. There is no statistical evidence against the Null Hypothesis.

- (ii) We have the same hypotheses and test, with expected frequencies 130, so  $\chi^2_{(4)} = \frac{1}{130}(40^2 + 30^2 + 10^2 + 20^2 + 20^2)$ , is 10 times what it was in (i), so  $\chi^2_{(4)} = 26.15$  giving very strong evidence to reject the NH. The increased amount of data will increase the power of the test to discriminate between the two hypotheses.

- 4 (i)  $H_0$ : cerebral blood flow and senile dementia are independent.  
 $H_1$ : they are not independent.

Calculate expected frequencies in the 2-way table:

<i>OBS Blood Flow</i>	<i>Normal</i>	<i>Reduced</i>		<i>EXP</i>	<i>N</i>	<i>R</i>
<i>Dementia No</i>	92	28	120	<i>No</i>	86	34
<i>YES</i>	80	40	120	<i>YES</i>	86	34
	172	68	240		172	68

$\chi^2_{(1)} = 6^2(\frac{2}{86} + \frac{2}{34}) = 72 \times \frac{120}{86 \times 34} = 2.955$ , n.s., providing no statistical evidence for rejecting  $H_0$ .

- (ii) For a matched case-control study, McNemar's test is required, which has greater power of discrimination because of the matching.

		<i>CONTROL</i>	
		<i>N</i>	<i>R</i>
<i>CASE</i>	<i>N</i>	74	6
	<i>R</i>	18	22

$\chi^2_{(1)} = \frac{(6-18)^2}{6+18} = \frac{144}{24} = 6.00$ , so  $H_0$  is rejected.

- 5 (i) In the model  $y_{ij} = u + \alpha_i + \varepsilon_{ij}$ , the response  $y_{ij}$  on the  $j^M$  unit under "treatment" i contains:

U=a general mean response,

$\alpha_i$ =an effect, or departure from general mean, due to treatment i,

$\varepsilon_{ij}$ =random residual "error" term,  $N \sim (0, \delta^2)$ , with constant  $\delta^2$ , mutually independent for all i,j.

The model is assumed additive. The standard assumptions thus are additivity, normality, constant variance.

- (ii)  $H_0$  is " $u_1 = u_2 = u_3$ ", and the alternative  $H_1$  is that at least one is different from the others.

Total for treatments: 1:161, 2:195, 3:180. Grand total=536  $\sum y_{ij}^2 = 12222$

Corrected total s.s. =  $12222 - 536^2/24 = 251.33$  s.s. for treatments =  $\frac{161^2+195^2+180^2}{8} - \frac{536^2}{24} = 12043.25 - 11970.67 = 72.583$

Analysis of Variance.

SOURCE OF VARIATION	D.F.	SUM OF SQUARES	MEAN SQUARE	
TREATMENTS	2	72.583	36.292	$F_{(2,21)} = 4.26$
RESIDUAL	21	178.750	8.512	
TOTAL	23	251.333		

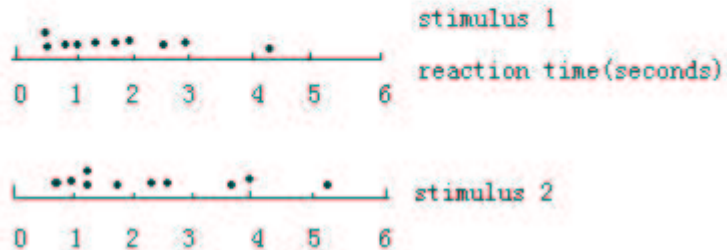
Mean: 1:20.125, 3:22.500, 2:24.375. The significant difference between any two means (since all have 8 replicates) is  $t_{21} \sqrt{\frac{2 \times 8.512}{8}} = 2.08 \times 1.459 = 3.03$  at the 5% level.

The milk yield for treatment 2 (using supplement A) gave the highest average in this trial. But it should be noted that when allowance is made for the variation between individual animals the statistical evidence only indicates that this treatment is better than "no supplement". The position for supplement B remains in doubt.

- 6 (a) (i) When it is reasonable to assume that a set of data come from a population that can be modelled normal distribution, a test for the location "center" of the distribution can be based on the standard normal distribution. If a sample  $\{x_i\}$  is small, we require to know a value for the variance  $\sigma^2$ , Then a test of  $H_0 : "mean = u"$  is to calculate  $z = \frac{(\bar{x}-u)}{\sigma/\sqrt{n}}$ , where n is sample size, as test  $z$  as  $N(0, 1)$ . In large samples,  $\sigma^2$  need not be know exactly but can be estimated by  $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ . This is satisfactory for  $n \geq 30$  in faith symmetrical distributions (even if not exactly normal) but sometimes a value of  $\sigma^2$  from earlier similar work and be used, e.g, in industry for smaller  $n$ , Differences between means from  $N(u_1, \sigma_1^2)$  and  $N(u_2, \sigma_2^2)$  will be  $N(u_1 - u_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$  and then a hypothesis that they have the same mean (location "center") is tested using  $Z = \frac{(\bar{x}_1 - \bar{x}_2) - (u_1 - u_2)}{\sqrt{\delta_1^2/n_1 + \delta_2^2/n_2}}$  Very large sample of data from normal approximations such as Binomial  $(n, p) \sim N(np, np(1-p))$ .
- (ii) For a distribution that is normal with unknown variance but  $n < 30$ ,  $\sigma^2$  is estimated by  $s^2$  and  $t = \frac{\bar{x}-u}{s/\sqrt{n}}$  follows the t-distribution with  $(n-1)$ degrees of freedom. Also, provided two populations have the same variance, the different between their locations can be tested by  $t = \frac{(\bar{x}_1 - \bar{x}_2) - (u_1 - u_2)}{s \sqrt{1/n_1 + 1/n_2}}$ , where sample sizes are  $n_1, n_2$ , the N.H. is " $u_1 = u_2$ " and  $s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$  These test meet the same purposes as those in (i).
- (b) These data are paired, involving the same workers, and therefore a test must be based on the single sample of 12 differences  $d_i = (\text{rate with supervisor present}) - (\text{rate with supervisor absent})$ . We may examine the N.H "mean of D=0". Values  $\{d_i\}$  are  $-5, -3, 0, -3, 1, 2, 6, -3, 0, -3, 1, -3$ .  $\sum d_i = -10$   $\bar{d} = -0.833$ .  $\sum d_i^2 = 112$ ,  $s^2 = \frac{1}{11} (112 - \frac{(-10)^2}{12}) = 9.4242$ . The test is  $t_{(11)} = \frac{-0.833-0}{\sqrt{9.4242/12}} = 0.94$ , n.s., so there is no evidence that, on average, the supervisor's presence made any difference. But the manager may be interested to note

that only one worker(7) showed a substantially higher rate with the supervisor present; usually there was little difference or a small decrease.

- 7 (a) Parametric methods require assumptions about the distribution from which a sample is drawn, usually that it is normal (or approximately so). When samples are small, results can be seriously wrong if this not true, although for large sample the Central Limit Theorem makes it satisfactory to examine means or totals from most distributions. Nonparametric methods require no distributional assumptions, and are generally based on rankings, medians and combinatorial results which allows tables of critical values to be constructed. Skew data can be analyzed without the need for transformations to make them more symmetrical (approximately normal). But in general a nonparametric test is much less powerful than the corresponding parametric one-although an exception is the Mann Whitely test which is almost as good as the corresponding. In general, data that are (approximately) normal, either in the original units or after a suitable transformation, are best analyzed by parametric methods. A dot plot can be a useful guide to th



- (b) For Stimulus 1, the data appear skew,with a possible upper outlier. For Stimulus 2, they are also skew, but more spread out,and again with a possible upper outlier. Because of the skewness, and the possible outliers, parametric joint ranking and stimulus:

0.5	0.5	0.7	0.8	1.0	1.1	1.3	1.3	1.6	1.8	1.8	2.3	2.4	2.6	3.0	3.5	4.0	4.4	5.3	
1	1	1	2	1	2	1	2	2	1	1	2	1	2	2	1	2	2	1	2

Number of times a stimulus 1 time is below a stimulus 2 time  $U = 10 + 10 + 8 + 8 + 6 + 5 + 5 + 3 + 1 = 66$  The number of observations are  $m = n = 10$  for the two stimuli. The rank sum statistic  $T = U + \frac{1}{2} \times 10 \times 11 = 66 + 55 = 121$  A normal approximation to this for 10 or more observations in each set is  $N(105, 175)$ . Standardizing,  $Z = \frac{121 - 105}{\sqrt{175}} = \frac{16}{13.23} = 1.21$  n.s., giving no evidence of difference in location of the two sets of times.

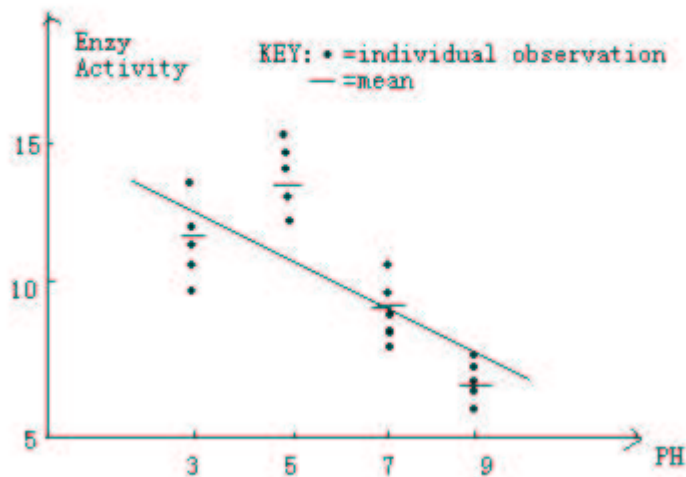
- (a) If two samples are drawn from Normal distributions  $N(u_i, \sigma_i^2)$   $i = 1, 2$  and  $u_i, \sigma_i^2$  not known, an estimate of each variance is  $s_i^2 = \sum_{j=1}^{m_i} (x_{ij} - \bar{x}_i)^2 / (m_i - 1)$  where  $m_i$  are sample sizes. A test of the N.H. " $\sigma_1^2 = \sigma_2^2$ " is given as  $F_{(m_1-1, m_2-1)} = s_1^2 / s_2^2$ , where  $s_1^2 > s_2^2$  to make use of standard tables. (see the example below)

In the Analysis of Variance (see, e.g., question 5), mean squares computed on the Null Hypothesis are all estimates of the same residual variation  $\sigma^2$ , and so the ratio  $\frac{\text{treatments mean square}}{\text{residual mean square}}$  will follow an  $F$ -distribution with the appropriate degrees of freedom.

- (b) Supplier 1:  $m = 16$ ,  $\sum x_i = 30.1$ ,  $\sum x_i^2 = 58.85$ ,  $s^2 = 0.14829$ .  
 Supplier 2:  $m = 16$ ,  $\sum x_i = 30.3$ ,  $\sum x_i^2 = 57.97$ ,  $s^2 = 0.03929$ ,  $F_{(15,15)} = \frac{0.14829}{0.03929} = 3.77$ , leads us to reject a N.H. of equal variances in favor of an A.H.  $\sigma_1^2 > \sigma_2^2$ . The company is well advised to use supplier 1.

*Paper III*  
*Statistical Applications and Practice*

1 PH totals are 56.4; 70.6; 49.7; 31.4;  $G=208.1$ . Between PH's  $ss = \frac{1}{5}(56.4^2 + 70.6^2 + 49.7^2 + 34.1^2) - \frac{1}{20}208.1^2 = 158.99$



Analysis of Variance

<i>SOURCE OF VARIATION</i>	<i>DEGREES OF FREEDOM</i>	<i>SUM OF SQUARES</i>	<i>MEAN SQUARE</i>	
<i>Between PH's</i>	3	158.99	52.997	$F_{(3,16)} = 43.30$
<i>Residual</i>	16	19.58	1.224	
<i>Total</i>	19	178.57		

The analysis shows that the bulk of the variability in the data is due to the difference between the activities at the different PH values. Each observation is assumed to be normally, distributed about its mean, with the same variance for all. It is assumed that the model: *observation = mean for given PH + random residual*  $y_{ij} = m_i + e_{ij}$   $i = 1, 2, 3, 4$ ; explain the data.

The filled live passes through (5,11.4) and (9,7.6), as shown. Although the analysis shows a significant linear component in the regression, the plot clearly indicates the need for a curve. (The deviations from-linearity s.s.is 158.99-91.97=67.02, which is also significantly large). There is no point in studying PH9, nor PH7 because the maximum does not seem to be near there. The maximum will be in the region of 5, probably on the lower side, so values such as 4.25 by steps of 0.25 to 5.25 (omitting 5), or 4.5 by steps of 0.25 to 5.25 including 5.0 in the same experiment, would be appropriate.

2 A  $\chi^2$ -test has to use exact frequencies, not percentages. It can test the Null Hypothesis that the ratios of frequencies between the six categories were the same in both years. On this NH, expected frequencies are given in brackets:

<i>Category</i>	1	2	3	4	5	6	
1996	60(60)	250(205)	160(155)	240(225)	270(320)	20(35)	1000
1998	60(60)	160(205)	150(155)	210(225)	370(320)	50(35)	1000
	120	410	310	450	640	70	2000

$$\chi_{(5)}^2 = 0 + 0 + \frac{(250-205)^2}{205} + \frac{(160-205)^2}{205} + \frac{(160-155)^2}{155} + \frac{(150-155)^2}{155} + \frac{(270-320)^2}{320} + \frac{(370-320)^2}{320} + \frac{(20-35)^2}{35} + \frac{(50-35)^2}{35} = 48.56$$

There is very strong evidence against the NH.

Looking at the percentages through the categories, the numbers opposing have increased over the period, "slight support" having dropped substantially and "great opposition" increased. On the face of it, the magazine was justified.

But much more information could be extracted by attaching a scoring scale, such as -2, -1, 0, 1, 2 to the first five categories, omitting the "don't know" and looking at scores, or some other suitable statistic. If the linear scale is (approximately) valid, this will extract more information than simple categorization. But it may be unwise to assume linearity, and also not easy to

decide how to score very strong views.

$$n_1 = n_2 = 1000, p_1 = 0.51, p_2 = 0.58, p_2 - p_1 = 0.07$$

$$V[p_2 - p_1] = \frac{0.58 \times 0.42}{1000} + \frac{0.51 \times 0.49}{1000} = 0.0004935, SE = 0.0222$$

95% confidence interval for true value of  $p_2 - p_1$  is  $0.07 \pm 1.96 \times 0.0222 = 0.07 \pm 0.0435$  or 0.026 to 0.114 i.e. with 95% probability the difference lies between 2.6% and 11.4%. The evidence is that there has been a shift against, of this extent.

3 (i) (i)(iii)(v) See next sheet.

(ii)  $n = 15$ ,  $\bar{y} = 1000/15 = 66.67$ ,  $\bar{x} = 1231/15 = 82.07$ ,  $\sum y^2 = 88268$ ,  $\sum xy = 77068$ ,  $\sum x^2 = 103411$  Therefore  $\hat{y} = 378.81 - 3.714x$ .

$$\hat{b} = (1706.5 - \frac{12310.00}{15}) \div (103411 - \frac{1231}{15}) = -\frac{4987.67}{2386.93} = -2.094$$

$$\hat{a} = 66.667 + 2.094 \times 82.007 = 238.51 \text{ Therefore } \hat{y} = 238.51 - 2.094x. \text{ Line A.}$$

(iii)  $\hat{y} = 378.81 - 3.714x$  Line B.

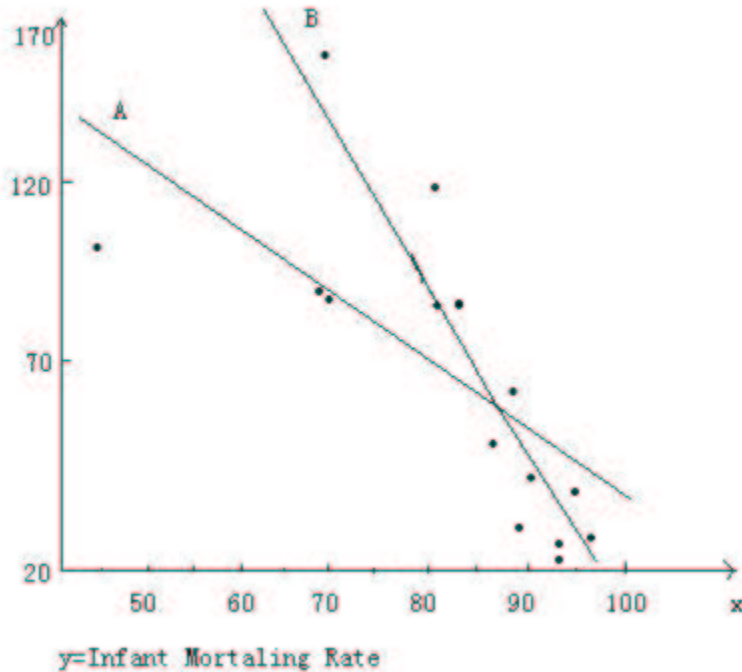
(v) Country C is out of step with the others, and is very influential in determining the fitted line. In fact neither line fits the lower literacy rate values at all well, although line B does rather better than A for the upper values. The variability in y as x decreases suggests that a linear regression assuming constant variance is not likely to do well; a weighted regression might be considered if suitable data is variable.

4 The completed worksheet is attached (see next page). Plot of residuals against quarters:

The 4<sup>M</sup> quarter residual in 1993 is very large. The quarter's delivery was much higher than usual, and was higher than the previous quarter instead of the usual annual pattern for it to be lower. (There was no compensating reduction the following quarter; hence the prediction is also very far from observed. )

5  $F_e^{3+}$ : min=2.25, lower quartile=5.84, median=9.98, upper quartile=18.41, max=39.13.  
 $F_e^{2+}$ : min=4.71, lower quartile=8.32, median=10.35, upper quartile=15.25, max=37.25.

- (i) There does not seem to be a large difference in retention as measured by the middle parts of the data sets (between quartiles).
- (ii) The distributions do not appear at all symmetrical, and  $F_e^{2+}$  has two suspect outliers, well above the other 16 figures. Because of the lack of symmetry and the extreme values in the



data, a mann-whitney test will be more satisfactory than a t-test; the two distributions seem roughly similar shapes. labelling  $F_e^{3+}$  A and  $F_e^{2+}$  B, the joint ranking of the 36 mice is in the order:

AABABAABBABAABABBBABAABBABBBABAABABA

$U$  = number of times an A precedes a B

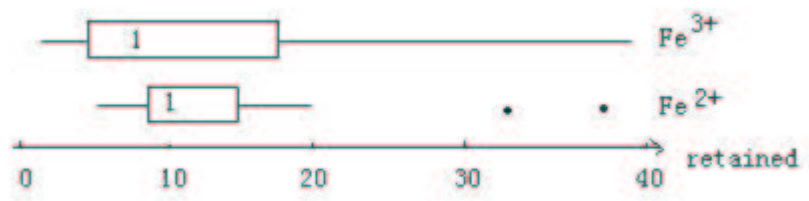
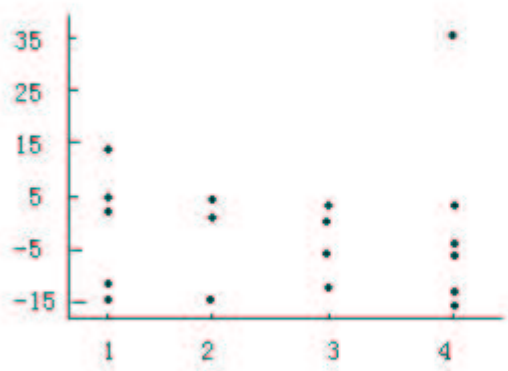
$$= 18 + 18 + 17 + 16 + 16 + 14 + 13 + 13 + 12 + 9 + 8 + 8 + 6 + 3 + 2 + 2 + 1 + 0 = 176$$

The rank-sum statistic  $T = U + \frac{1}{2} \times 18 \times 19 = 176 + 171 = 347$ , and for these sample sizes is approximately normal with mean  $\frac{1}{2} \times 18 \times 37 = 333$  and variance  $\frac{1}{12} \times 18 \times 18 \times 37 = 999$

Therefore  $Z = \frac{347-333}{\sqrt{999}} \sim N(0, 1)$  approximately,  $Z = \frac{14}{31.4} = 0.44$  This provides no evidence against a Null Hypothesis of the same median values of iron retention. The data therefore provide no evidence of difference.

CANDIDATE'S NUMBER

Worksheet for fertilizer deliveries, for use with Question 4:



	Deliveries	M.AV.	Y-M.AV.	Seasonal	Predicted	Residual
1990 2	49.5	*	*	-14.94	*	*
3	94.2	*	*	-0.3475	*	*
4	62.6	69.6750	-7.0750	-2.1675	67.5075	-4.9075
1991 1	72.7	66.3875	6.3125	17.455	83.8425	-11.1425
2	48.9	60.7250	-11.8250	-14.94	45.7850	3.1150
3	68.5	58.8875	9.6125	-0.3475	58.5400	9.9600
4	43.0	58.4125	-15.4125	-2.1675	56.2450	-13.2450
1992 1	77.6	55.6750	21.9250	17.455	73.1300	4.4700
2	40.2	53.400	-13.2000	-14.94	38.4600	1.7400
3	55.3	53.8250	1.4750	-0.3475	53.4775	1.8225
4	38.0	54.6125	-16.6125	-2.1675	52.4450	-14.4450
1993 1	86.0	55.1125	30.8875	17.455	72.5675	13.4325
2	38.1	64.3875	-26.2875	-14.94	49.4475	-11.3475
3	61.4	72.2125	-10.8125	-0.3475	71.8650	-10.4660
4	106.1	72.9750	33.1250	-2.1675	70.8075	35.2925
1994 1	80.5	73.0500	7.4500	17.455	90.5050	-10.0050
2	49.7	63.6625	-13.9625	-14.94	48.7225	0.9775
3	50.4	54.3750	-3.9750	-0.3475	54.0275	-3.6275
4	42.0	52.1375	-10.1375	-2.1675	49.9700	-7.9700
1995 1	70.3	51.6375	18.6625	17.455	69.0925	1.2075

Calculation of seasonal effects

Q1	Q2	Q3	Q4	
6.3125	-11.8250	9.6125	-7.0750	
21.9250	-13.2000	1.4750	-15.4125	
30.8875	-26.2875	-10.8125	-16.6125	
7.4500	-13.9625	-3.9750	33.1250	
18.6625	-11.4625	-0.0750	-10.1375	
	*	*	0.6625	
17.0475	-15.3475	-0.755	-2.575	-1.63
17.455	-14.94	-0.3475	-2.1675	

6 A two-sample t-test would not be appropriate because of the correlation. Paired comparisons can be examined for systolic and for diastolic by examining the two sets of differences, that are given. In each case a one-tail test sets of difference, against a N.H. that the true difference =  $\frac{284}{15} = 18.93$

$$\text{Estimated variance} = \frac{1}{14} \left( 6518 - \frac{284^2}{15} \right) = 81.495$$

$$t_{(14)} = \frac{18.93 - 0}{\sqrt{81.495/15}} = 18.93/2.33 = 8.12$$

very strong evidence against the NH; hence very strong evidence in favor of the drug being effective.

A 95% confidence interval for the true mean change is  $18.93 \pm t_{14,5\%} \times 2.33$  or  $18.93 \pm 2.145 \times 2.33$  which is  $18.93 \pm 5.00$ , i.e. 13.93 to 23.93.

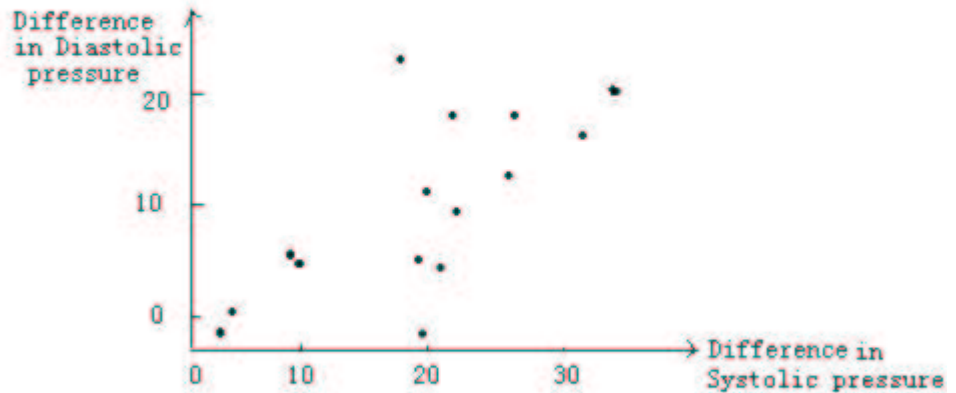
Diastolic Mean difference =  $139/15 = 9.27$

$$s^2 = \frac{1}{14} \left( 2327 - \frac{139^2}{15} \right) = 74.210 \quad t_{(14)} = \frac{9.27 - 0}{\sqrt{74.21/15}} = 9.27/2.22 = 4.17$$

The 95% confidence interval is  $9.27 \pm 2.145 \times 2.22 = 9.27 \pm 4.76$  which is 4.51 to 14.03

Again there is very strong evidence of a real difference, i.e. that the drug is effective, but the reduction in diastolic pressure is less than in systolic.

Generally the two pressures decrease together in the same patient, and roughly speaking the diastolic reduction is about 2/3 as much as the systolic reduction: the slope of a fitted line would be about +2/3.



- (i) (a) Assuming that the list was consolidated, in any order, e.g. alphabetically, by year of entry, by faculty, etc. and that the numbers 0001-8002 could be allocated to the students, then random digits would be used in sets of 4, the corresponding numbered students being included in the sample. 0000 and any set of digits from 8003 to 9999 would be discarded. Only if a student could not be contacted (e.g. was in hospital) would another number be chosen. (Digits which came up a second time, or more, would be discarded—sampling is to be without replacement. )
- (b) The numbers required are:

	<i>M</i>	<i>F</i>	<i>Total</i>
<i>Medicine</i>	77	54	131
<i>Science</i>	116	91	207
<i>Engineering</i>	103	40	143
<i>Social Sciences</i>	77	95	172
<i>Arts</i>	70	77	147
	443	357	800

From the faculty lists, the appropriate numbers of male and female students would be selected by a method similar to that above.

- (c) The quota sample would be carried out to the same specification as in (b), but no lists of students would be required. Instead the survey interviewers would only be asked to interview the correct number of students in each category.
- (ii) A is relatively cheap, quick and simple to administer, and there is no need for trained interviewers.

But the questions must be simple and direct, so that they are easy to answer quickly, and there is minimal scope for misunderstanding. Questions will have to be closed, with boxes to tick or requests for straight forward information, and some pre-testing for possible ambiguities will be essential. Unless the topic of the survey is of general interest, the response rate may be low and follow-up will be needed. Forms may not be returned to a central point: individual collection is much more efficient. Distribution could be through mail (by post for the those not living on campus) for a random sample.

B will generally have a higher response rate; more questions, and more complex questions, can be asked; answers can be clarified if necessary; open questions may be included for a response to be written down or possibly classified by the interviewer; there will be no distribution problems although refusals are still possible for the random sample. But there is a cost for using interviewers which may be substantial; if there are only a few interviewers it may take longer to complete; there is a possibility of interviewer bias, or of the respondent being biased to give a particular type of answer.

Depending on the purpose of the survey and the size of the questionnaire either method is possible but B will usually give better quality data, and if the topic is at all complex the presence of a trained interviewer can be particularly helpful.

8 The total of all 20 measurements is  $5(17.59 + 20.59 + 16.97 + 16.30) = 357.7$  The correction term  $G^2/N$  for calculated sums of squares is  $357.7^2/20 = 6397.4645$  Total for glass fibre are  $5(17.59 + 16.97) = 172.8$  and  $5(20.59 + 16.39) = 184.9$  so Fibre *s.s.* =  $\frac{1}{10}(172.8^2 + 184.9^2) - 6397.4645 = 7.3205$  Total for hardening are  $5(17.59 + 20.59) = 190.9$  and  $5(16.97 + 16.34) = 166.8$  so Hardening *s.s.* =  $\frac{190.9^2 + 166.8^2}{10} - 6397.4645 = 29.0405$

Analysis of Variance

<i>SOURCE</i>	<i>DF</i>	<i>CUM OF SQUARES</i>	<i>MEAN SQUARE</i>	<i>F - ratio</i>	
<i>Fibre</i>	1	7.32		5.80	*
<i>Hardening</i>	1	29.04		22.99	***
<i>Interaction</i>	1	16.02		12.68	**
<i>Residual</i>	16	20.21	1.263		
<i>Total</i>	19	72.59			

The interaction is significant, so the main effects are not studied. Using the pooled variance, the significant difference between any two means at the 5% level is  $t_{(16)}\sqrt{2 \times 1.263/5} = 2.120 \times 0.711 = 1.51$

In the table of means it is clear that glass fibre present and heat hardening leads to great water absorption than any of the other three combinations, whose results are not significantly different from one another.

