The Royal Statistical Society

HIGHER CERTIFICATE

IN STATISTICS 2000

SOLUTIONS

Paper I : *Statistics Theory*

1(i) For one leg, there are 4 possible colors in each of the three positions, making $4 \times 4 \times 4 = 64$ combinnations. For the other leg there will also be 64, and of the 64 on one leg may be combined with any of the 64 on the other to make $64 \times 64 = 4096$ combinations in all.
ALTERNATIVE,there are 6 positions altogether, and four possible colors may appear on each,making $4^6 = 4096$

(ii)Consider the middle position on one leg: it can hold any one of 4 colors, there are only 3 possibilities for each of upper and lower (they could be the same);so each leg has $3 \times 4 \times 3$ possible patterns ,i.e.36.Any of these patterns can also appear on the leg, making $36 \times 36 = 1296$ possible combinations in all.

(iii)In this case the possible number of combinations is $4 \times 3 \times 2$on each leg, i.e.$24 \times 24 = 576$ altogether.

(iv)In(i),since there are 64 combinations for one leg. there are only 63 for the other: $64 \times 63 = 4032$in all.
In(ii),similarly ,there are $36 \times 35 = 1260$.
In(iii), similarly, there are $24 \times 23 = 552$.

2(a)(i)
$$P(x \ irregular \ in \ sample) = \binom{10}{x} \binom{40}{5-x} / \binom{50}{5}$$

For x=0,1,2,3,4,5
The sampling leads to the hypergeometric distribution:

$$\begin{array}{llll}
Account: & Ok & Irrigular & Total \\
Sampled & 10-x & x & 10 \\
Notsampled & 35+x & 5-x & 40 \\
Total & 45 & 5 & 50
\end{array}$$

So

$$P(x=0) = \frac{\binom{10}{0}\binom{40}{5}}{\binom{50}{5}} = \frac{1 \times 40! \times 5! \times 45!}{5! \times 35! \times 50!} = \frac{40 \times 39 \times 38 \times 37 \times 36}{50 \times 49 \times 48 \times 47 \times 46} = 0.3106$$

(ii)

$$P(x \geq 2) = 1 - P(0) - P(1).$$

$$P(x=1) = \frac{\binom{10}{1}\binom{40}{4}}{\binom{50}{5}} = \frac{10 \times 40! \times 5! \times 45!}{4! \times 36! \times 50!} = \frac{10 \times 5 \times 40 \times 39 \times 38 \times 37}{50 \times 49 \times 48 \times 47 \times 46} = 0.4313$$

$$and \;\; P(x \geq 2) = 0.2581.$$

(b)P(server wins from deuce)=

$$P(ww) + P(wlww) + P(lwww) + p(wlwlww) + p(wllwww)$$
$$+P(lwwlww) + P(lwlwww) + P(wlwlwlww) + \cdots$$

where the number of possible sequence doubles each deuce.this is

$$(\tfrac{2}{3})^2 + 2(\tfrac{2}{3})^2(\tfrac{1}{3} \times \tfrac{2}{3}) + 4(\tfrac{2}{3})^2(\tfrac{1}{3} \times \tfrac{2}{3})^2 + 8(\tfrac{2}{3})^2(\tfrac{1}{3} \times \tfrac{2}{3})^3 + \cdots$$

$$= (\tfrac{2}{3})^2[1 + (\tfrac{2}{3})^2 + (\tfrac{2}{3})^4 + (\tfrac{2}{3})^6 + \cdots]$$

$$= \tfrac{9}{4}[1 + \tfrac{9}{4} + (\tfrac{9}{4})^2 + (\tfrac{9}{4})^3 + \cdots] = \tfrac{4}{9} \times \tfrac{1}{1-\tfrac{4}{9}} = \tfrac{4}{5}$$

$$P(score\ does\ not\ change\ after\ 2\ points) = (\tfrac{2}{3} \times \tfrac{1}{3}) + (\tfrac{1}{3} \times \tfrac{2}{3}) = \tfrac{4}{9}.$$

$$P(game\ ends\ after\ 2\ points) = \left(\tfrac{2}{3} \times \tfrac{2}{3}\right) + \left(\tfrac{1}{3} \times \tfrac{1}{3}\right) = \tfrac{5}{9}.$$

$$P(N = 2k) = \tfrac{5}{9} \times P(k-1\ sequences\ of\ 2\ points\ which\ do\ not\ change\ score)$$

$$= \tfrac{5}{9} \times \left(\tfrac{4}{9}\right)^{k-1} \quad for\ k = 1, 2, 3, \cdots$$

so that N=2,4,6,$\cdots$

writing 2k=n , $P(N = n) = \tfrac{5}{9} \times \tfrac{9}{4} \times \left(\tfrac{4}{9}\right)^{n/2} = \tfrac{5}{4} \times \left(\tfrac{2}{3}\right)^n$

(c)

$$P(C) = 0.3 \qquad P(V|C) = 0.8. \qquad P(L) = 0.4 \qquad P(V|L) = 0.6$$
$$P(D) = 0.2 \qquad P(V|D) = 0.9 \qquad P(O) = 0.1 \qquad P(V|O) = 0$$
$$P(NV|C) = 0.2 \quad P(NV|L) = 0.4 \quad P(NV|D) = 0.1, \quad P(NV|O) = 1$$

$$P(L|NV) = P(NV|L)P(L) \ / \sum_{x=C,L,D,O} P(NV|x)P(x)$$

$$= \tfrac{0.4 \times 0.4}{(0.4 \times 0.4) + (0.3 \times 0.2) + (0.1 \times 0.2) + (1 \times 0.1)} = \tfrac{0.16}{0.34} = 0.4706.$$

$$P(both\ L|NV) = 0.4706^2 = 0.2215.$$

3.(a) Linear combinations of normal variables remain normal.

$$y \sim N\left(\sum_{i=1}^{n} a_i\mu_i,\ \sum_{i=1}^{n} a_i^2\sigma_i^2\right)$$

Hence $x_1 + x_2 \sim N(\mu_1 + \mu_2,\ \sigma_1^2 + \sigma_2^2);\ \ x_1 - x_2 \sim N(\mu_1 - \mu_2,\ \sigma_1^2 + \sigma_2^2)$

(b) Let M,S be travelling times of manager and secretary.

$$M \sim N(35, 16) \quad and \quad S \sim N(33, 9)$$

(i)Assuming journey times are independent ,

$$M - S \sim N(2, 25) \ \ and \ \ P(M - s < 0) = P\left(z < \frac{0-2}{5}\right)$$

$$where\ \ z \sim N(0., 1); \quad this\ \ is\ \ P(z < -0.4) = 0.3446$$

3

(ii) If secretary leaves t minutes earlier, the difference in arrival times will be N(2+t,25),assuming that the journey times still have the same distributions as before.

$$P(secretary\ arrives\ first)\ =P(M-S+t>0)$$

$$=1-\Phi(\tfrac{-1(2+t)}{5})\geq 0.9\ \ if$$

$$\frac{t+2}{5}\geq 1.2826\ \ fromN(0,1)\ table\ i.e.\ t=4.408min.$$

(iii)
$$P(M<30)=\phi(\tfrac{30-35}{4})=\phi(-1.25)=0.01056$$

$$P(S<30)=\phi(\tfrac{30-33}{3})=\phi(-1)=0.1587$$

$$and\ require\ probability\ is\ 0.1056\times 0.1587=0.0168.$$

(c) The number of breakdowns per week will follow a poisson distribution with mean $20\times 0.02=0.4$

(i)
$$P(no\ breakdowns)=e^{-0.4}$$

$$P(none\ in\ 4\ weeks)=(e^{-0.4})^4=e^{-1.6}=0.2019$$

(ii)
$$P(1\ or\ more\ in\ a\ week)=1-e^{-0.4}=0.39297$$

$$required\ probability=(0.3297)^4=0.0118$$

In 52 weeks, number of breakdowns in poisson with mean $0.4\times 52=20.8$ A normal approximation N(20.8,20.8) may be used, and using a continuity correction we require $P(number>26.5)$

$$z=\frac{26.5-20.8}{\sqrt{20.8}}=\frac{5.7}{4.5607}=1.2498$$

$$P(z>1.2498)=P(z<-1.2498)=0.1056\ \ from\ tables$$

4

4(i)

$$P(x = y) = P(x = y = 1) + P(x = y = 2) + \cdots + P(x = y = 6) = \frac{6}{36} = \frac{1}{6}$$

$$P(x > y \ or \ y > x) = 1 - P(x = y) = \frac{5}{6}$$

By symmetry of the joint distribution, $P(x > y) = P(y > x)$, so each of these must be $\frac{5}{12}$

ALTERNATIVELY enumerate all possibilities

(ii)

$$P(z > \xi) = P(neither \ A \ or \ B \ throws \ a \ 6 \ in \ frist \ \xi - 1 \ attempts)$$

$$= (\frac{5}{6} \times \frac{5}{6})^{\xi-1} = (\frac{25}{36})^{\xi-1}, \quad \xi = 1, 2, 3, \cdots$$

$$P(z = \xi) = P(z \geq \xi) - P(z \geq \xi + 1)$$

$$= (\frac{25}{36})^{\xi-1}(1 - \frac{25}{36}) = \frac{11}{36}(\frac{25}{36})^{\xi-1} \quad \xi = 1, 2, 3, \cdots$$

(iii)(a)
$$P(z \leq 4) = 1 - P(z \geq 5) = 1 - (\frac{25}{36})^4 = 0.233$$

(b)
$$E[z] = \sum_{\xi=1}^{\infty} \xi P(z = \xi) = \frac{11}{36} \sum_{\xi=1}^{\infty} \xi(\frac{25}{36})^{\xi-1}$$

Now $\sum_{\xi=1}^{\infty} \xi x^{\xi-1}$ is derivative of $\sum_{\xi=1}^{\infty} x^{\xi} = \frac{1}{1-x}$ by the used results for a geometric series.

Thus $\sum_{\xi=1}^{\infty} \xi x^{\xi-1} = \frac{d}{dx}(\frac{1}{1-x}) = \frac{1}{(1-x)^2}$, we have $x = \frac{25}{36}$;

Therefore
$$E[z] = \frac{11}{36} \times \frac{1}{(1 - \frac{25}{36})^2} = \frac{36}{11} = 3.27$$

Alternatively consider E[z] and $-\frac{25}{36}E[z]$, and add to give $\frac{11}{36}E[z]$ which is 1 as an infinite geometric series

5(i)

$$P(group\ test\ +\ ve) = 1 - P(group test\ -\ ve)$$

$$= 1 - P(all\ k\ individuals\ -\ ve)$$

$$= 1 - (1 - P)^k$$

(ii) The N persons form m independent groups ,each group having a group test: also if the group test is positive there are k individual tests, so this happens with probability $(1 - (1-p)^k)$

Hence the number of test is $s_k = m + k \times x$,where x is the number out of the m groups where individual tests have to be made; so x is $Binomial(m, 1 - (1-p)^k)$.

(iii)The mean and variance of $Binomial(n, \pi)$ are $n\pi$, $n\pi(1 - \pi)$. hence

$$E[x] = m(1 - (1-p)^k), \quad V[x] = m(1-p)^k(1 - (1-p)^k)$$

since N=mk

$$E[s_k] = m + kE[x] = \frac{N}{k} + N(1 - (1-p)^k) = N[\frac{1}{k} + 1 - (1-p)^k]$$

$$V[s_k] = k^2 V[x] = Nk(1-p)^k[1 - (1-p)^k]$$

(iv)When P=0.01

$$E(s_{10}) - E(s_{11}) = N(\frac{1}{10} - \frac{1}{11} - 0.99^{10} + 0.99^{11})$$

$$= N(\frac{1}{10} - \frac{1}{11} - 0.01 \times 0.99^{10}) = 4.7089 \times 10^{-5}N$$

$$E(s_{12}) - E(s_{11}) = N(\frac{1}{12} - \frac{1}{11} - 0.99^{12} + 0.99^{11})$$

$$= N(\frac{1}{12} - \frac{1}{11} + 0.01 \times 0.99^{11}) = 1.3776 \times 10^{-3}N$$

Both of these are positive, so $E(s_{11})$ is less than $E(s_{10})$ and $E(s_{12})$.

ALTERNATIVELY by directly calculation as below.

(v)When p=0.05,

$$E[s_4] = N[1.25 - 0.95^4] = 0.435494N$$

$$E[s_5] = N[1.2 - 0.95^5] = 0.426219N$$

$$E[s_6] = N[\tfrac{7}{6} - 0.95^6] = 0.431575N$$

$$E[s_7] = N[\tfrac{8}{7} - 0.95^7] = 0.444520N$$

k=5 minimizes $E[s_k]$ in this range.

(vi)When k=1

$$E[s] = N(2 - (1 - p)) \quad = 1.01N \qquad for \ P = 0.01$$

$$= 1.05N \qquad for \ P = 0.05$$

this suggests that when P is very small the total number of tests $s_k$can be very substantially reduced. Even for P=0.05 it is (more than) halved. Also the group size may be larger the smaller P is: the above results suggest optima of k=11 or 5 for p=0.01 or 0.05. As P increase there is less scope for economy of testing.

6. Note that

$$\int_0^\infty \lambda^2 x e^{-\lambda x} dx = 1$$

(i)(a)
$$E[x] = \lambda^2 \int_0^\infty x^2 e^{-\lambda x} dx \quad = \lambda^2[-\tfrac{1}{\lambda}x^2 e^{-\lambda x}]_0^\infty + \lambda^2 \int_0^\infty \tfrac{1}{\lambda}e^{-\lambda x} \times 2x dx$$

$$= 0 + \tfrac{2}{\lambda} \int_0^\infty \lambda^2 x e^{-\lambda x} dx = \tfrac{2}{\lambda}$$

(b)
$$E[x^2] = \lambda^2 \int_0^\infty x^3 e^{-\lambda x} dx \quad = \lambda^2[-\tfrac{1}{\lambda}x^3 e^{-\lambda x}]_0^\infty + \lambda^2 \int_0^\infty \tfrac{1}{\lambda}e^{-\lambda x} \times 3x^2 dx$$

$$= 0 + \tfrac{3}{\lambda}E[x] = \tfrac{6}{\lambda^2}$$

Therefore

$$v[x] = \frac{6}{\lambda^2} - (\frac{2}{\lambda})^2 = \frac{2}{\lambda^2}$$

(c)

$$P(X > x) = \int_x^\infty \lambda^2 u e^{-\lambda u} du \quad = [-\lambda u e^{\lambda u}]_x^\infty + \int_x^\infty \lambda e^{-\lambda u} du$$

$$= \lambda x e^{-\lambda x} + [-e^{-\lambda u}]_x^\infty = e^{-\lambda x}(1 + \lambda x).$$

(ii) $\lambda = 0.01,\ \ x = 500\ in(c)$, $so\ \ P(x > 500) = e^{-5}(1+5) = 0.04043$

(iii)Assume $X \sim N(\frac{2}{\lambda}, \frac{2}{\lambda^2})$ $i.e.\ N(200, 20000)$

Now
$$P(x > 500) \quad = 1 - \phi(\frac{500-200}{100\sqrt{2}}) = 1 - \phi(\frac{3}{\sqrt{2}})$$

$$= 1 - \phi(2.1213) = 1 - 0.9835 = 0.0165$$

(iv)Using the correct distribution (which is positive skewed),with $\lambda = 0.01\ and\ x = 450,\ P(twin) = e^{-4.5}(1 + 4.5) = 0.0611$. The skewness raises the right-hand tail probability considerably.

7(i)$E[x] = \sum\limits_{i=1}^{k} \frac{i}{k} = \frac{1}{k}\frac{1}{2}k(k + 1) = \frac{k+1}{2}$

(ii)(a)The moment estimator $\hat{k_1}$ is found from setting $E[x] = \bar{x}$ i.e. $\bar{x} = \frac{1}{2}(\hat{k_1} + 1)$ or $\hat{k_1} = 2\bar{x} - 1$

(b)$Likelihood = \prod\limits_{i=1}^{4} f(x_i) = \frac{1}{k^4}$ provided all x_i lie between 1 and k inclusive.
The maximum of this occurs when $\hat{k_2}$ is chosen to be as small as possible given the data values; i.e. $\hat{k_2} = x_{(4)}$, the sample maximum.

(c)For$\{x_i\} = \{1, 10, 3, 2\},\ \ \hat{k_1} = 2(\frac{16}{4}) - 1 = 7 \quad \hat{k_2} = x_{(4)} = 10$
$\hat{k_1}$ is impossible, since there is a data value above 7.$\hat{k_2}$ is consistent with the data.

8(i) The t-value is given as 2.70, and residual d.f.=7, so p-value is $2P(t_7 > 2.70)$.
From tables,$P(t_7 > 2.70) \doteq 0.0153$.The (2-tail) p-value therefore is about 0.031 for intercept(interpolation between p=0.05 and 0.01 in a suitable table would give the same answer).
For the slope, the t-value is 10.48,so $P < 0.001$, $corr(x, y) = \sqrt{R^2} = \sqrt{0.940} = 0.97$
From plot1A,the linear relation may be breaking down above about x=8. Plot 1B supports the inadequacy of a linear model for the full set of data, because the residuals do

not appear to be a normally distribution set with mean zero.

(ii)When x^2 is included, x ceases to be significant. The information in x^2 is clearly taking up that previously given by x (with which it is strongly correlated).Now the model explain 98.1% of the total variation among the y-values.
Residual variance may increase with x;whereas it should be constant .There appears to be a non-random pattern of residual.

(iii)Plot 3A shows that the logarithmic trend explains data better. Plot 3B is more like a random scatter of residual(though still a bit suspect)The amount of total variation in the explain by regression 3 is 99.1% the test of any of these models.
The total sum of $squares = (n-1) \times variance$ of y;the units of y in regression 3 are logarithmic, whereas in 1 and 2 they were natural.

(iv)When x=10

$$1 \quad gives \quad y = 78.33 + 540 \doteq 618$$
$$2 \quad gives \quad y = 170 + 40 + 500 \doteq 710$$
$$3 \quad gives \quad log_{10}y = 2.16 + 0.689 = 2.849, \ i.e. \ y \doteq 706$$

However data do not go so far as x=10 and it is only in regression 3 that we have a model that seems at all safe to use for extrapolation beyond the range of available data. On all counts (see (iii) and(iv)) Regression 3 appears best, and also it requires only two parameters.

Paper II: Statistical Methods

1(i) Use stems of 5 units, rather than 10,to give clearer results.

Division1

```
1 | 4                        Median = ½(21 + 22) = 21.5
· | 6 6 7 8 9 9              Lower quartile = 19
2 | 0 0 0 1 2 3 4 4          Upper quartile = 26
· | 5 6 7 7 8 9              Interquartile range = 7
3 | 4
```

Division2

```
1 |
  . |  5  5  8
  2 |  1  3  4  4
  . |  5  6  8  8  9  9  9
  3 |  1  2  3  4  4
  . |  7  8  9
```

$Median = \frac{1}{2}(28 + 29) = 28.5$
$Lower\ quartile = 24$
$Upper\ quartile = 33$
$Interquartile\ range = 9$

The means and variances are required later.

$$Division 1: \quad \tilde{x}_1 = 22.23; \quad s_1^2 = 24.7554; \quad s_1 = 498.$$
$$Division 2: \quad \tilde{x}_2 = 27.82; \quad s_1^2 = 46.8225; \quad s_1 = 684.$$

(ii)The average values are considerably higher in Division 2;So is the variability especially because of the three high values 37,38,39.The distributions are otherwise reasonably symmetrical.

(iii)assuming the normal distribution can be used as a model, $(\bar{x}_1 - \bar{x}_2) \pm t_{(42)}\sqrt{2s^2/22}$ is the confidence interval, where s^2 is the pooled variance $\frac{21s_1^2 + 20s_2^2}{42}$, with 42 d.f. $\bar{x}_1 - \bar{x}_2 = -5.59$; $t_{(42)}$ at 5% is (approx.) 2.019;$s^2 = 35.7890$ so the interval is $-5.59 \pm 2.019\sqrt{\frac{s^2}{11}}$ i.e. $-5.59 \pm 2.019 \times 1.804.giving \quad -5.59 \pm 3.642$, i.e. $(-9.23\ to\ -1.95)$.The variance in the two populations are assumed to be the same.

2 N=48 mice, Table of results is:

survival :	No	Yes	
control :	14(10.5)	10(13.5)	24
Drug	7(10.5)	17(13.5)	24
	21	27	48

values in brackets are the frequencies expected on the null hypothesis that survival rates are the same in both groups of mice.

$$x_{(1)}^2 = \sum \frac{(O-E)^2}{E} = (3.5)^2(\frac{1}{10.5} + \frac{1}{10.5} + \frac{1}{13.5} + \frac{1}{13.5}) = 4.148$$

with Yates correction, each (O-E) is reduced by 0.5 before squaring, which gives $3^2(\frac{2}{10.5} + \frac{2}{13.5}) = 3.408$ n.s. as $X_{(1)}^2$.

Yates' correction can sometimes over-correct for continuity (discreteness). so the "exact" probability of obtain the results on the N.H.-which could be found by Fisher's Exact Test-is found about the 5% value(3.84). this leaves the significance or otherwise

of the result in doubt, but indicates that further data would be needed before a clear decision could be made. 48 is a very small number of units upon which to compare two proportions.

For the second trial, N=144, and the results are:

$$
\begin{array}{lccc}
survival: & No & Yes & \\
control: & 42(31.5) & 30(40.5) & 72 \\
Drug & 21(31.5) & 51(40.5) & 72 \\
& 63 & 81 & 144
\end{array}
$$

$$x_2^{(1)} = (10.5)^2(\frac{2}{31.5} + \frac{2}{40.5}) = 12.44$$

and with Yates' correction,the value of $X_{(1)}^2$. is $10^2(\frac{2}{31.5} + \frac{2}{40.5}) = 11.29$

Both values are significant at the 0.1% lever, leaving little doubt that there is a difference between Control and Drug.

since we are only asked to test "the effect" of the drug a 1-tail test may not be valid; but the data from the second experiment show a firm indication in favor of the drug.

3.N.H. "$\mu = 10$", Also for s.d., N.H. is "$\sigma = 0.04$"
for the sample,

$$n = 10 \quad \sum x = 100.17 \quad \sum x^2 = 1003.4242, \quad s^2 = 2.35667 \times 10^{-3}, \quad s = 0.04865$$

$$t_{(9)} = \frac{\bar{x} - \mu}{s\sqrt{10}} = \frac{10.017 - 10.000}{\sqrt{2.235667 \times 10^{-4}}} = \frac{0.017}{0.0150} = 1.137 \ \ n.s.$$

The N.H. "$\mu = 10$" is not rejected on this evidence.

$$X_{(9)}^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{9 \times 2.35667 \times 10^{-3}}{(0.04)^2} = 13.256$$

less than the value for 5% significance, so the N.H. is not rejected.

We have assumed the determinations of potency are independently normally distributed with the same variance.

with n=30,$t_{(9)}$ has $\sqrt{10}$ replace by $\sqrt{30}$, and it is now $t_{(29)}$. the value of $t_{(29)}$ *will be* $\frac{\sqrt{30}}{\sqrt{10}} \times$ 1.137 = 1.97, which is approaching the 5% significance point, so now the N.H. is open to doubt (although technically not rejected at 5%)

For the variance,$x_{(29)}^2 = \frac{29s^2}{\sigma^2}$, so in the previous calculation 9 is replaced by 29 and $x_{(29)}^2 = \frac{29}{9} \times 13.256 = 42.714$.. This again is very near the 5% point, but this time is just significant and we may reject the N.H. for σ^2

In both cases the large amount of data has given a more powerful test

4(i)A suitable N.H. is that the population distributions are the same with A.H.that they are different. On the N.H. the number of instance where the new types will give a higher figure is $Binomial(10, p = y_2)$.
there are 8 differences in favor of the original,and 2 for the new.
In $B(10, \frac{1}{2})$, $P(2 \ or \ less) = (1 + 10 + 45)(\frac{1}{2})^{10} = 0.0547$ not significant. so the N.H. is not rejected.

(ii)For a Wilcoxon test the N.H. is that population distributions are identical,and the A.H is that they differ in location.
The actual difference(original-new) for each car are: +0.4,+0.3,-0.6,+0.8,+0.2,-0.1,+0.3+0.4,+1.1, +0.2 and the ranks are $6\frac{1}{2}, 4\frac{1}{2}, 8, 9, 2\frac{1}{2}, 1, 4\frac{1}{2}, 6\frac{1}{2}, 10, 2\frac{1}{2}$. The sums of ranks are $T_- = 8 + 1 = 9$; $T_+ = 46$ the smaller of these is compared with the critical value for n=10, which is $T = 8$; $T_- > 8$ so the N.H. is not rejected.
(However, the result in (ii) is nearer to significance than that in(i)).
If we could assume that the differences for the 10 cars followed a normal distribution, then a paired t-test could be applied. It would, if valid, be were powerful than either of those above.

5(i) The 1-way analysis compares the N.H. "all μ_i are the same ", where $\{\mu_i\}$ denote the means of the sores using the 4 points i=1,2,3,4; The A.H. is that there are differences among the means
Paint total are 315,335,350,370,grand total=1370, N=16 observations. Sum of squares of observations=118290.Total corrected sum of squares $= 118290 - \frac{1370^2}{16} = 983.75$. s.s. for $paint = \frac{1}{4}(315^2 + 335^2 + 350^2 + 370^2) - \frac{1370^2}{16} = 406.25$
Analysis Of Variance

Source of Variation	Degrees of freedom	Sum of squre	Mean square
Paints	3	406.25	135.42
Residual	12	577.50	48.125
Total	15	983.75	

$$F(3, 12) = 2.81 \ n.s.$$

Assuming the scores can be modelled by a normal distribution, the linear model underlining this analysis is $y = \mu_i + \varepsilon_{ij}$, $\{\varepsilon_{ij}\}$ all $N(0, \sigma^2)$. we do not reject the N.H. that all μ_i are the same.
If the columns repeat geographical differences, we should remove these in a 2-way analysis.The area total are 343,321,336,370 and give a sum of square $\frac{1}{4}(343^2 + 321^2 + 336^2 + 370^2) - 1370^2/16 = 315.25$.

ANALYSIS OF VARIANCE

Source	DF	SS	MS	
Paint	3	406.25	136.42	$F(3,9) = 4.65$
Areas	3	315.25	105.08	$F(3,9) = 3.61 n.s.$
Residual	9	262.25	$29.139 = \hat{\sigma}^2$	
Total	15	983.75		

Making allowance for area differences reduce the residual variation to that which is purely random natural variation, so make the paints comparison more precise. Now the N.H.should be rejected. there is evidence of a different among the paints.

6:As shown in figure 1
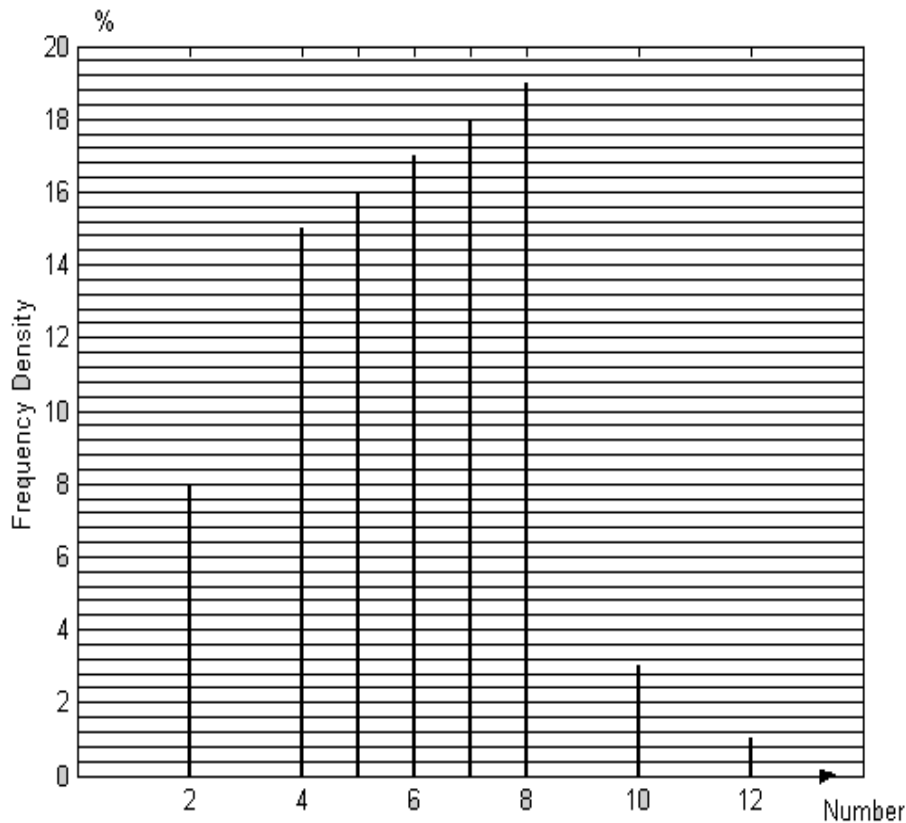
The 6th question

The model class interval is "7 but less than 8" For mean, we use mid-points of interval as x:

$$
\begin{array}{cccc}
x & f & F & fx \\
1 & 3 & 3 & 3 \\
4.5 & 8 & 11 & 24 \\
4.5 & 15 & 26 & 67.5 \\
5.5 & 16 & 42 & 88 \\
6.5 & 17 & 59 & 110.5 \\
7.5 & 18 & 77 & 135 \\
9 & 19 & 96 & 171 \\
11 & 3 & 99 & 33 \\
13 & 1 & 100 & 13 \\
& & & 645
\end{array}
$$

Estimation of mean $\bar{x} = \frac{645}{100} = 6.45$

Median is midway between 50^M and 51^{sr} in rank orders. There are 42 up to 6, and 17 in the interval 6 to 7.

$M = 6 + \frac{50.5-42}{17} \times 1 = 6.5$

By using mid-points to calculate $\bar{x}$, we assume the observations in the interval are uniformly distributed. Because we have a slight overestimation,it seems that there was a slight tendency for them to occur nearer the lower limits of intervals. we do not know what to use for x in the final interval; 13 may be too high but as the frequency is only 1 the effect is small. the median being overestimated indicates that in "'6 to 7"most of the observations actually fall in the lower part of the interval.

7(i)The mean of the data is $\frac{1}{100}(0 + 25 + 36 + 36 + 28 + 45 + 30) = 2.00$. In a poisson distribution with mean 2, $P(0) = e^{-2} = 0.1353$,so the expected frequency of zeros is 13.53.similarly $P(1) = 2e^{-2} = 0.2707$, $P(2) = \frac{4e^{-2}}{2} = 0.2707$ $P(3) = \frac{8e^{-2}}{3} = 0.1804$ $P(4) = \frac{16e^{-2}}{4} = 0.0902$ $P(5) = 0.0361$, $P(\geq 6) = 0.0166$. The table of observed and corresponding expected frequencies is :

$$
\begin{array}{ccccccccc}
count & 0 & 1 & 2 & 3 & 4 & \geq 5 & & Total \\
Obs.freq & 24 & 25 & 18 & 12 & 7 & 14 & : & 100 \\
Exp.freq & 13.53 & 27.07 & 27.07 & 18.04 & 9.02 & 5.27 & : & 100
\end{array}
$$

Comparing these in a χ^2 test, there will be 4 degrees o freedom since we are using an estimate of the mean.

$$
\chi_4^2 = \sum \frac{O-E}{E} = \frac{10.47^2}{13.53} + \frac{2.07^2 + 9.07^2}{27.07} + \frac{6.04^2}{18.04} + \frac{2.02^2}{9.02} + \frac{8.73^2}{5.27} = 28.236
$$

There is strong evidence against the N.H. of a poisson distribution, which is therefore rejected.

(ii)kolmogorov-smirnor uses cumulative probabilities:

	0	1	2	3	4	5	6	(∞)
OBS	0.24	0.49	0.67	0.79	0.86	0.95	1.00	$(-)$
EXP	0.1353	0.4060	0.6767	0.8571	0.9473	0.9843	0.9945	(1.0000)
$\|O-E\|$	0.1047	0.0840	0.0067	0.0671	0.0873	0.0334	0.0046	$(-)$

The maximum modulus of difference in cumulative probabilities is 0.1047; for n=100 observations, the upper 5% tail starts at $\frac{1.36}{\sqrt{n}}$ i.e. 0.1360, so the observed difference is not significant and there is no evidence for rejecting the poisson N.H. Chi-squared tests the pattern of frequencies, which had too heavy a tail that was balanced by too many zeros for a mean=2; Kolmogrorov-smirnov by using cumulative probabilities is not so affected by the upper-tail.

The gross value added by all manufacturing fell in the period 1991/2/3, so any study of individual component must allow for this. Also the data are based on 1995 price, so change in costs of raw materials, labor and manufacture over the nine years will reflect this: not all price can be automatically increase to compensate for increase in expenditure. The weightings are 1995,so if there were substantial change over the period. This would influence index movements.

One approach is to compare trends, and illustrate these on a graph with time on the horizontal axis. The extent to which individuals reflected the general trend (bottom row) should be commented on. with 14 individual rows, there is a lot of information, and perhaps the main components as given by 1995 weight are enough to illustrate.

An alternative is to recalculate indices relative to the overall index for each year, e.g. Food 1989:95.6/97.9=97.7(see table below).

	1989	1990	1991	1992	1993	1994	1995	1996	1997
Food.etc	97.7	99.5	104.5	106.4	105.2	103.0	100.0	100.6	102.0
Pulp, paper, etc	96.2	98.7	99.1	100.3	102.1	100.0	100.0	97.6	96.9
chemical	85.4	85.5	92.5	95.4	96.1	96.5	100.0	100.3	100.3
Basic Metals, etc	114.5	113.8	108.8	103.4	101.0	98.8	100.0	99.3	99.9
Electrical	81.7	82.7	83.6	85.0	88.4	94.7	100.0	103.6	103.8
Transport	113.7	111.4	109.7	107.5	104.3	102.2	100.0	105.3	108.9

Thus relative to the picture for total manufacturing, both chemicals and electrical have steadily increased their value added over time,metals have steadily reduced, food rose until 1993 and has since dropped back, pulp and paper rose and fell again while transport fell and then rose again.

Transport was never below the level for total.

although most other categories formed small parts of the total (by 1995 weight things), some points stand out; coke etc. is nearly always well below. Total except for 1995; textiles show a steady fall,rubber and plastic a steady rise.

Higher value added will often reflect greater efficiency, but will also be affected by the

15

factors mentioned above.

Paper III

Statistical Applications & Practice

1(I)A χ^2 test for 2×2 table may be used. Expected value on the Null Hypothesis of no difference between proportion are shown in brackets,$(109.14 = \frac{200 \times 191}{350}, etc)$

	K	L	Total
+	127(109.14)	64(81.86)	191
−	73(90.86)	86(68.14)	159
	200	150	350

$$\chi^2_{(1)} = \frac{127 - 109.14^2}{109.14} + \cdots + \frac{86 - 68.14^2}{68.14} = 15.01$$

(yate's correction may be used but is not necessary since the result is in no doubt). This is clear evidence to reject the N.H.

(ii)The two proportions are:$P_k = \frac{127}{200} = 0.635, \quad P_L = \frac{64}{150} = 0.427$. Using a normal approximation to the binomial distribution gives the distribution $N(P_K - P_L; \frac{P_K(1-P_K)}{200} + \frac{P_L(1-P_L)}{150})$ for the true difference $\Pi_K - \Pi_L$ between the population proportions. A 95% confidence interval for $\Pi_K - \Pi_L$ is

$$(P_K - P_L) \pm 1.96\sqrt{\frac{0.635 \times 0.365}{200} + \frac{0.427 \times 0.573}{150}}$$

i.e. $0.208 \pm 1.96 \times 0.0528$ or 0.208 ± 0.104 which is $(0.104$ to $0.312)$

(iii) The proportion of positive response among those showing the K reaction is, with 95% probability, between 10.4% and 31.2% greater than for those showing the L reaction.

(iv) The point estimation is 0.208, or 20.8%. we know that if we took another sample we would not get exactly the same estimate of the difference, but without the confidence interval we would not know how much to expect repeated estimation to vary

(v)Interval width depends on the square root of sample size,though the square root of the variance of $(P_K - P_L)$ and so to narrow the interval by a factor $\frac{1}{4}$ we increase sample size by $4^2 = 16$. The necessary size is then about $16 \times 350 = 5600$.

2(i)

(ii)
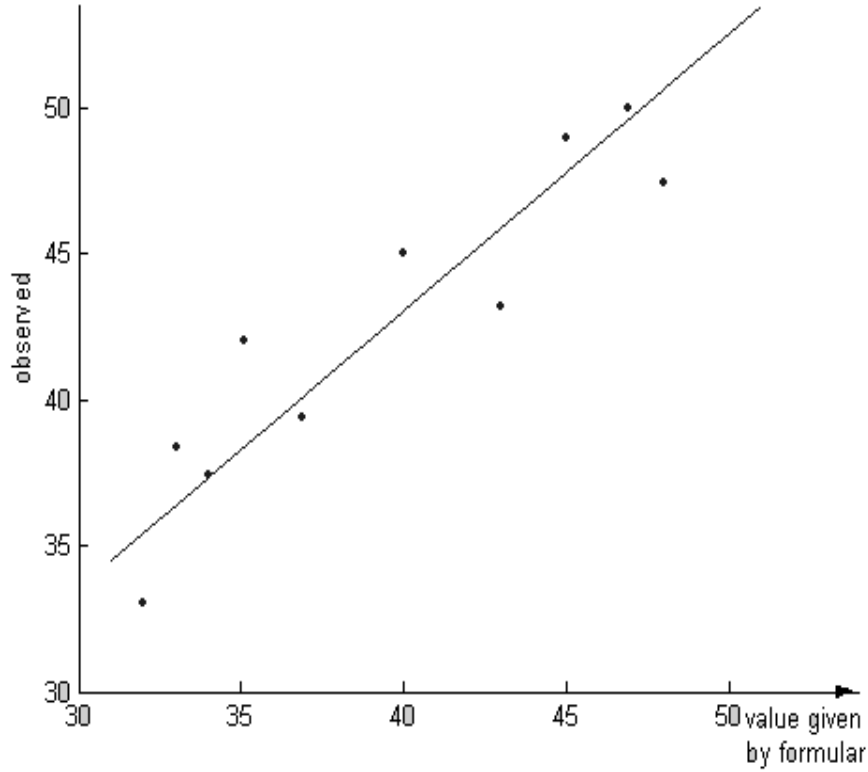
$$\sum y = 424.8 \quad s_{xy} = 17005.66 - (424.8 \times 393.9)/10 = 272.788$$
$$\sum x = 393.9 \qquad s_{xx} = 15840.23 - 393.9^2/10 = 324.509$$

Hence the slope

$$\hat{b} = \frac{272.788}{324.509} = 0.8406$$

The fitted line is $y - \bar{y} = \hat{b}(x - \bar{x})$ or $y - 42.48 = 0.8406(x - 39.39)$ i.e. $y = 0.8406x + 9.3681$ or $y = 9.37 + 0.84x$

(iii)The proportion of total variation (sum of square), that is explained by the relation of observation to formula is to 0.843.[$(229.31/272.18) \approx 0.8425$] this is reasonably good.

(iv)The calculation is based on the theoretical model $y = \alpha + \beta x + \varepsilon$,where $var(\varepsilon) = \sigma^2$ is estimated by $s^2 = (2.315)^2 = 5.36$ It has 8 d.f. $var(\hat{b}) = \sigma^2/s_{xx}$ estimated as $\frac{5.36}{324.509} = 0.016517$, so $\hat{SE} = 0.1285$

17

A 95% interval for β is :

$$\hat{b} \pm 2.306 \times 0.1285 = 0.8406 \pm 0.2963 \quad i.e.(0.544 \ to \ 1.137)$$

$[t_{(8,5\%)} = 2.306]$. A 95% interval for α is :

$$\hat{a} \pm 2.306\sqrt{s^2(\frac{1}{10} + \frac{\bar{x}^2}{s_{xx}})} = 9.368 \pm 2.306 \times \sqrt{5.36 \times 4.8813} = 9.368 \pm 11.795$$

This give(-2.427 to +21.16)
Note that this is very imprecisely determined(and is a little use since there are no data near to zero to confirm whether a linear relation still holds)

3(i)N=18 observations. $Total \ (corrected) \ ss = 286327 - 2225^2/18 = 11292.28$

$$ss \ for \ times = \frac{1}{3}(415^2 + \cdots + 267^2) - \frac{2225^2}{18} = \frac{840655}{3} - \frac{2225^2}{18} = 5183.61$$

Analysis of variance:

Source of variation	DF	Sum of squares	Meansquare	
Times	5	5183.61	1036.72	
Residual	12	6108.67	509.06	$F(5, 12) = 2.04n.s.$
Total	17	11292.28		

This provides no evidence in favor of any time effect.

(ii)Omitting the given sample,grand total is now 2161,N=17, $G^2/N = 274701.2353$; sum of all $squres = 286327 - 64^2 = 282231$ total for time 36(2 observations)=286,and ss for times is

$$\frac{286^2}{2} + \frac{1}{3}(415^2 + 409^2 + 396^2 + 388^2 + 267^2) - \frac{G^2}{N} = 5581.76$$

Residual now has 11 d.f. and total 16. residual ss=1948.00

	DF	SS	MS	
Times	5	5581.76	116.35	$F(5, 11) = 6.30$
Residual	11	1948.00	$177.09 = s^2$	
Total	16			

We should now reject the hypothesis that the mean results at each time are all the same.

	(0)	(24)	(36)	(48)	(72)	(120)
Means	138.3	136.3	134.0	132.0	129.3	89.0

18

The reason for the different conclusion is that now 120 stands out from the others. the SE of difference between two means (not including (36)) is $\sqrt{\frac{2}{3}} \times 177.09 = 10.87$,and between (36) and any other is $\sqrt{(\frac{1}{2} + \frac{1}{3})(177.09)} = 12.15$ so t test (11 d.f.) would confirm this conclusion.

The value at 36 hours which has been omitted looked suspiciously low, and could perhaps have been a measure for recording error. Now it can not be checked, but if there are laboratory records available. that may help to decide whether or not to induce it in the analysis. It makes a serious difference to the conclusions.

4 (i) N=32.total G=70.6. week totals:(9),5.0;(12)19.3;(15)25.0;(18)21.3. Hence

$$SS \ Time = \frac{1}{8}(5.0^2 + 19.3^2 + 25.0^2 + 21.3^2) - 70.6^2/32 = 28.7613.$$
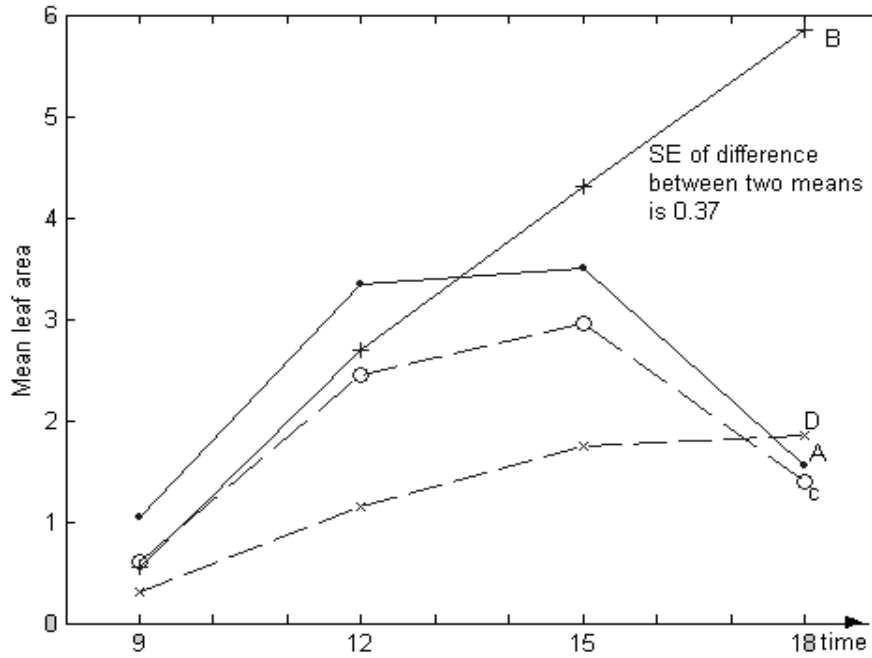
Analysis of variance

Source	DF	SS	MS	
Species	3	18.8012	6.2671	
Time	3	28.7613	9.5871	$F(9,16) = 17.44$
Species × Time	9	21.0862	2.3429	
Residual	16	2.1500	0.1344	
Total	31	70.7987		

There is clear evidence of an interaction, i.e. species behave differently over time. Main effects of species and time are not therefore relevant.

(ii)Means:

time	(9)	(12)	(15)	(18)
A	1.05	3.35	3.50	1.55
B	0.55	2.70	4.30	5.85
C	0.60	2.45	2.95	1.40
D	0.30	1.15	1.75	1.85

species

(iii)A increase up to week (15), then decrease; C has a similar pattern at a lower level of area, B goes on increasing; so does D,slowly.

The figure shows a graph with "Mean leaf area" on the y-axis (0 to 6) and "time" on the x-axis (9, 12, 15, 18). Curves labelled B, D, A, C. A note reads "SE of difference between two means is 0.37".

5(i)

$$Mean = \frac{1}{120}(0 + 30 + 64 + 60 + 52 + 45 = 30 + 14 + 8) = \frac{303}{120} = 2.525$$

$$P(r) = e^{-2.525}(2.525)^r/r! \quad for \ r = 0, 1, 2 \cdots$$

Expected frequencies are 120P(r).

τ =	0	1	2	3	4	5	6	≥ 7
E_i =	9.607	24.258	30.625	25.776	16.271	8.217	3.458	1.788

7 is 1.247; 8 is 0.394; $\geq$ 9 is 0.147.

$$o_i = 8 \quad 30 \quad 32 \quad 20 \quad 13 \quad 9 \quad 8$$

.

(ii)

$$x_5^2 = \frac{1.607^2}{9.607} + \frac{(-5.742)^2}{24.258} + \frac{(-1.375)^2}{30.625} + \frac{5.776^2}{25.776} + \frac{3.271^2}{16.271} + \frac{(-0.783)^2}{8.217} + \frac{(-2.754)^2}{6.246} = 5.162 \, n.s.$$

20

So a Null Hypothesis that the poisson model holds is not rejected.

(iii) Assuming that the Poisson model is valid we use 2.525 as the variance, so $2.525 \pm 1.96\sqrt{\frac{2.525}{120}}$ is an approximate 95% confidence interval for the true mean.

(a) This is $2.525 \pm 1.96 \times 0.145 = 2.525 \pm 0.284$ or $(2.24$ to $2.81)$.
Note. it is often specified that this approximation require a mean of at least 5 to be satisfactory.

(b)$P(\geq 1) = 1 - P(0) = 1 - 0.080 = 0.92$. This is an estimate of the proportion of non-zero minutes, and variance is $\frac{0.92 \times 0.08}{120}$ which is 0.0006133, whose square root is 0.0248.
An approximate 95% interval for the true proportion is $0.92 \pm 1.96 \times 0.0248$ or 0.92 ± 0.049, i.e. 0.87 to 0.97.

6 (i)
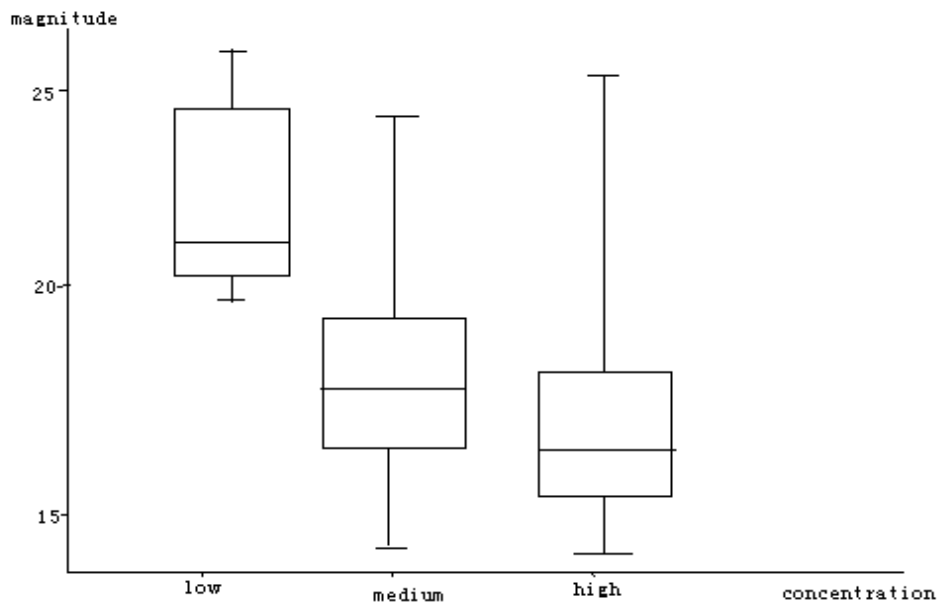
O outside W F and P

(ii)The whiskers run from minimum value to lower quartile and form maximum value to upper quartile; the median is marked inside the box. A symmetrical distribution would have the median at approximately the middle of the box, and the two whiskers of about the same length.

21

All of these patterns show skewness, positive in direction. For low concentration, the average magnitude is considerably more than for higher concentration, but there is skewness; the range of this set of data is less than for the other. The median brightness appears to reduce as concentration rises. For median concentration the distribution is nearer to symmetry except for the upper whisker.
For high concentration there is again high skewness(which may include outliers at the upper end if we had the original data.).

(iii) An analysis of variance assumes (approximate) normality, and the same variance in each set of data; neither of these seem to hold here. Because of the skewness, the mean will not be a good central measure either.
(A transformation such as logarithmic may improve matters, but if there were outliers these would still affect the analysis).

7:

Means are:A,10.76;B,6.75(Marked x)
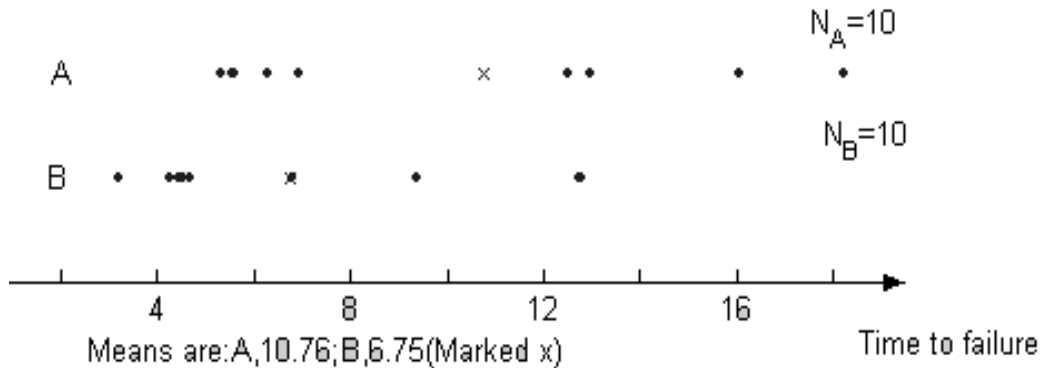
Because of the very irregular distribution and total table lake of concentration about the mean, the assumption of normality can not be made and so a t test is not valid. The summary measure of location, if one were needed, should be the median (9.71 for A and 4.68 for B). The Maun Whitney U test, will examine whether the two distribution of times differ; so will the Wilcoxon rank same test.
Jocut ranking:

| (1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
3.19	4.26	4.47	4.53	4.67	4.69	5.30	5.53	5.60	6.30
B	B	B	B	B	B	A	A	A	A
(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
6.79	6.92	9.37	12.51	12.75	12.78	12.95	16.04	18.21	18.24
B	A	B	A	B	B	A	A	A	A

Sums of ranks are: A,134 ;B,76 ($check$: $134 + 76 = 210 = \frac{1}{2} \times 20 \times 21$)

$$U_A = 134 - \frac{1}{2} \times 10 \times 11 = 79 \quad and \quad U_B = 76 - \frac{1}{2} \times 10 \times 11 = 21 \quad U_{min} = 21$$

From table this is significant at 5%.

ALTERNATIVELY, count the number of times member of A come before B's in the ranking :U=4+4+4+4+3+2=21.(B's before A's gives U=79). A normal approximation, strictly only for N_A, N_B both>10, is that U is $N(\frac{1}{2} \times 10 \times 10; \frac{1}{12} \times 10 \times 10 \times 21)$,so $\frac{21-50}{\sqrt{175}} = -\frac{29}{13.23}$ is N(0,1); this is 2.19 and so gives the same significance
we may conclude that there is evidence of difference between the time distributions. (The data suggest B tends to fail earlier)

8
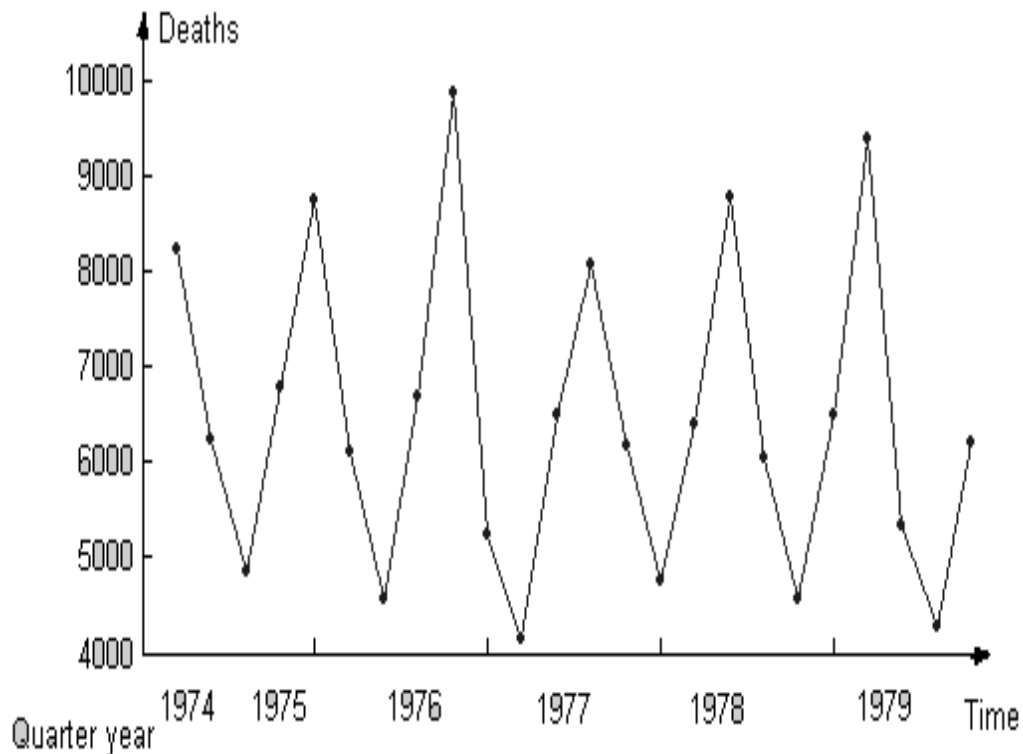
O outside W F and P

There is a clear seasonal pattern of deaths, highest in quarter 1 and lower in quarter 3 each year. Trend is small, and an additive model should be appropriate.

The centred 4-point average to go in 1974(3) is found by taking the first four item's average and setting it at $2\frac{1}{2}$, then the average of items 2-3-4-5 set at $3\frac{1}{2}$,finally averaging these two to go at 3. So we have $\frac{1}{8}(8291 + 2(6223 + 4841 + 6785) + 8760) = 6593.625$ etc. Average the figures for each quarter in the final column gives seasonal means which total 19.82. Effect are mean $-\frac{1}{4}(19.82)$ for the four quarters.

The trend at t=25,26,27,28 of y=68.31-31.8t is 6036.0, 6004.2,5972 and 5940.6. Predicted values are trend +seasonal effect, giving for

1980 (1)8542.83; (2)5445.41; (3)4050.41; (4)6014.55.

CALCULATIONS:

		Y	M.AV.	Y-M.AV
1974	1	8291		
	2	6223		
	3	4841	6593.63	-1752.63
	4	6785	6636.00	149.00
1975	1	8760	6583.13	2176.88
	2	6093	6535.88	-442.8
	3	4548	6662.38	-2114.38
	4	6700	6691.25	8.75
1976	1	9875	6532.5	3324.38
	2	5227	6455.88	-1228.88
	3	4145	6204.75	-2059.75
	4	6489	6096.00	393.00
1977	1	8059	6289.63	1769.83
	2	6155	6355.25	-200.25
	3	4766	6433.25	-1667.25
	4	6393	6509.63	-116.63
1978	1	8779	6469.25	2309.75
	2	6046	6454.88	-408.88
	3	4552	6543.13	-1991.13
	4	6492	6531.63	-39.63
1979	1	9386	6407.63	2978.38
	2	5347	6335.25	-988.25
	3	4259		
	4	6206		

Calculation of seasonal effects:

	Q1	Q2	Q3	Q4	
	2176.9	-442.88	-1752.63	149.000	
	3324.4	-1228.88	-2114.38	8.750	
detrended data	1769.4	-200.25	-2059.75	393.00	
	2309.8	-408.88	-1667.25	-116.625	
	2978.4	-998.25	-1991.13	-39.625	
seasonal means	2511.78	-653.83	-1917.03	78.9	19.82
seasonal effects	2506.83	-658.79	-1921.99	73.95	