THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE
(3 papers)

SOLUTIONS 1997

I. STATISTICAL THEORY

1 (a i) Label the 4 men A, B, C, D. Then A may have B or C or D as partner; the other two are the opposing pair. There are 3 ways. [Alternatively $\frac{\binom{4}{2}}{2} = 3$].
Label the players M_1, M_2, W_1, W_2. Then M_1 may play with W_1 or W_2 as partner; M_2 then has the other woman as partner. There are 2 ways.

(a ii) Suppose $n = 4m$. (If it is not, then 1 or 2 or 3 players cannot take part). Groups of 4 can be chosen in $\binom{4m}{4}$ ways, and each group can be matched in 3 ways (as above) giving a total of $3\binom{4m}{4}$, where $4m$ is the multiple of 4 that is as near to n (below) as possible.
Suppose $n_1 = 2m_1$ (otherwise leave out one man) and $n_2 = 2m_2$ (otherwise leave out one woman).
Two of each sex may be chosen in $\binom{2m_1}{2} \cdot \binom{2m_2}{2}$ ways and the total number of matches is then $2\binom{2m_1}{2} \cdot \binom{2m_2}{2}$.

(b i) $\binom{10}{5} = 252$, as once the first 5 are chosen the teams are chosen completely.

(b ii) Teams each consist of a goalkeeper and 4 others. Team 1 can be completed in $\binom{8}{4}$ ways $=70$ ways, which defines the selection completely.

(b iii) Label the goalkeepers G_1, G_2, the strikers S_1, S_2, S_3. Then these are 5 other players.
We may have, in one team, G_1 with S_1 or S_2 or S_3 and choose the other 3 from 5: there are $3 \times \binom{5}{3} = 30$ ways for this. Also G_1 may have two of the strikers and two others in the same team. There are $\binom{3}{2} = 3$ ways for strikers and $\binom{5}{2} = 10$ ways for others, making 30 ways in all. The total number of ways is thus $30 + 30 = 60$. Choosing one team fixes the other.

2. (i) $Y = pX_1 + (1-p)X_2$ is $N(1750p + 2000(1-p), p^2 \cdot 300^2 + (1-p)^2 \cdot 400^2)$
i.e. $N(2000 - 250p, 10^4\{9p^2 + 16(1-p)^2\})$
or $N(2000 - 250p, 10000(25p^2 - 32p + 16))$.

(ii) $E[Y] = 2000 - 250p$ and has maximum value (2000) for $p_1 = 0$.

(iii) $V[Y]$ is minimized when $\frac{d}{dp}(25p^2 - 32p + 16) = 0$
i.e. $50p - 32 = 0$ or $p_2 = 16/25$.
The second deviation is > 0, indicating a minimum.

(iv) $E[Y|p_1 = 0] = £2000$. $E[Y|p_2 = 16/25] = 2000 - \frac{250\times16}{25} = £1840$.

(v) (a) On $p = p_1 = 0$, $Y \sim N(2000, 400^2)$.

$P(Y < 1480) = P(Z < \frac{1480-2000}{400}) = P(Z < -1.30) = 0.0968.$

(b) On $p = p_2 = 16/25$, $Y \sim N(1840, 10^4\{\frac{16^2}{25} - \frac{32\times 16}{25} + 16\})$.

i.e. $N(1840, 16 \times 10^4\{1 - \frac{16}{25}\}) = N(1840, 240^2)$.

$P(Y < 1480) = P(Z < \frac{1480-1840}{240}) = P(Z < -1.50) = 0.0668.$

Z stands for the standardized variate $N(0,1)$. Use mixed strategy (b) because its probability of ruin is only two-thirds of that on (a). His expectation is lower, but variability is also lower, on (b).

3. $p(B) = p = \frac{1}{4}$. Family size $n = 5$. The distribution of r, the number with blue eyes is binomial $(n = 5, p = 1/4)$.

(i) $P(0) = (\frac{3}{4})^5 = \frac{243}{1024}$, so $P(\text{at least 1 with blue eyes}) = 1 - P(0) = \frac{781}{1024} = 0.7627$.

(ii) $P(\text{at least 3 B} \mid \text{at least 1 B}) {=} P(r \geq 3)/P(r \geq 1) = \frac{P(r\geq 3)}{781/1024}$.

$$\begin{aligned} P(3) + P(4) + P(5) &= \binom{5}{3}(\tfrac{1}{4})^3(\tfrac{3}{4})^2 + \binom{5}{4}(\tfrac{1}{4})^4(\tfrac{3}{4}) + (\tfrac{1}{4})^5 \\ &= \tfrac{1}{1024}\{10 \times 9 + 5 \times 3 + 1\} = \tfrac{106}{1024}. \end{aligned}$$

So required answer is $\frac{106}{1024} \div \frac{781}{1024} = \frac{106}{781} = 0.1357$.

(iii) Given that a particular one - the youngest - has blue eyes means that of the other four, at least two have blue eyes. This is found as $P(2)+P(3)+P(4)$ in binomial $(4, 1/4)$: $\binom{4}{2}(\frac{1}{4})^2(\frac{3}{4})^2 + \binom{4}{3}(\frac{1}{4})^3(\frac{3}{4}) + (\frac{1}{4})^4 = \frac{1}{256}\{6\times 9 + 4 \times 3 + 1\}$ $= \frac{67}{256} = 0.2617$.

(iv) (a) Using binomial $(5, 1/4)$ and excluding $r = 0$, the expected number is $\frac{1}{1-P(0)}\sum_{r=1}^{5} rP(r) = \dfrac{1024}{781}\{1\times 5 \times (\dfrac{1}{4})(\dfrac{3}{4})^4 + 2 \times 10 \times (\dfrac{1}{4})^2 (\dfrac{3}{4})^3 + 3 \times 10 \times (\dfrac{1}{4})^3 (\dfrac{3}{4})^2 +$ $4 \times 5 \times (\dfrac{1}{4})^4 (\dfrac{3}{4}) + 5 \times (\dfrac{1}{4})^5\} = \dfrac{1}{781}(5 \times 81 + 20 \times 27 + 30 \times 9 + 20 \times 3 + 5) =$ $\dfrac{1280}{781} = 1.64$.

(b) In binomial $(4, 1/4)$, $E[r] = np = 1$.

So expected number is 1(youngest) + 1(others) = 2.

(v) Specific information about one child reduces the "subspace" in which we have to search for the values of r concerning the others.

4. (i) To fit a bolt with $X = 9.98$, we must have Y between 10.00 and 10.18.

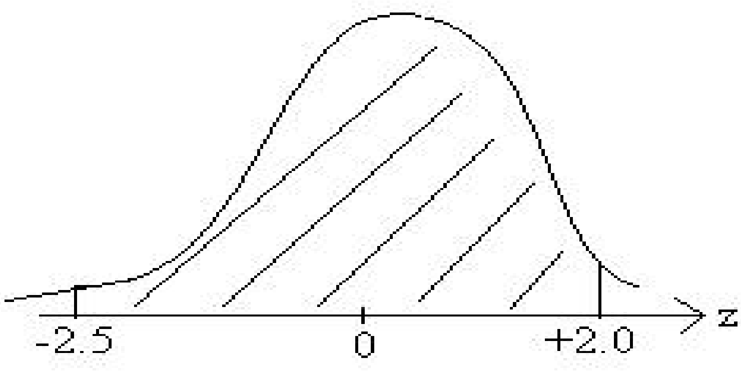
$Y \sim N(10.10, 0.0016)$. $z = \frac{Y - 10.1}{0.04} \sim N(0, 1)$.

For $Y = 10.0$, $z = \frac{-0.1}{0.04} = -2.5$,

For $Y = 10.18$, $z = \frac{+0.08}{0.04} = +2.0$.

$P(z < 2.0) = 0.97725$. $P(z < -2.5) = 0.00621$.

We require the difference of these, which is 0.97104.



(ii) $Y - X \sim N(10.1 - 10.0, 0.0016 + 0.0009) \sim N(0.1, 0.0025)$.

P(fit satisfactorily) $= P(0.02 \leq Y - X \leq 0.2)$. Corresponding z values for 0.02, 0.2 are $z = \frac{0.02-0.1}{0.05} = \frac{-0.08}{0.05} = -1.60$; $z = \frac{0.2-0.1}{0.05} = +2.00$.

$P(z < -1.60) = 0.05480$, $P(z < +2.00) = 0.97725$, difference is 0.92245.

(iii) $Z \sim N(10.3, 0.0144)$. $P(Z > 10.06) = P(z > \frac{10.06-10.3}{0.12})$, where $z \sim N(0,1)$, i.e. $= P(z > -\frac{0.24}{0.12}) = P(z > -2.0) = 0.97725$.

(a) Plates are independent, so required probability is $(0.97725)^2 = 0.95502$.

(b) We require $10.08 \leq Y \leq 10.26$ for nut and bolt to fit. Corresponding z values are $\frac{10.08-10.10}{0.04} = -\frac{0.02}{0.04} = -0.5$ and $\frac{10.26-10.10}{0.04} = \frac{0.16}{0.04} = +4.0$, above which we may ignore the probability (strictly it is 0.00003). $P(z < -0.5) = 0.30854$, and the required probability is $1 - 0.30854 = 0.69146$. (strictly 0.69143).

Nut and bolt must fit and bolt go through the holes. Given random choice, and hence independence, this has probability $0.69146 \times 0.95502 = 0.66036$ (or 0.66033).

(iv) $n = 25$, $\bar{X} \sim N(10.0, \frac{0.0009}{25}) \sim N(10.0, (0.006)^2)$.

The permitted deviation of $\bar{X}$ from 10.0 is only 0.01, corresponding to $z = \pm\frac{0.01}{0.006} = \pm 1.667$. $P(z > 1.667) = 0.04779 = P(z < -1.667)$.

Hence the probability is $2 \times 0.04779 = 0.09558$ of stopping.

5. P(positive)$= p$. P(no positive in k)$= (1-p)^k$,

(i) and so the probability of a pooled-sample positive is $1 - (1-p)^k$.

(ii) $S = m$(one for each group) + k individual tests if the group was positive, taken over each of the m groups $= m + kX$, where m is the number of groups and each group has probability $1 - (1-p)^k$ of requiring k tests.

Therefore X is binomial $(m, 1 - (1-p)^k)$.

(iii) $E[S] = E[m] + kE[X] = \frac{N}{k} + k \cdot \frac{N}{k} \cdot \{1-(1-p)^k\} = N\{\frac{1}{k} + 1 - (1-p)^k\}$.

$V[S] = k^2V[X] = k^2m\{1-(1-p)^k\}(1-p)^k = Nk(1-p)^k\{1-(1-p)^k\}$.

(iv) $\frac{dE}{dk} = -\frac{N}{k^2} - N\frac{d}{dk}(1-p)^k = -\frac{N}{k^2} - N(1-p)^k \ln(1-p)$.

(using the result that $\frac{d}{dx}(a^x) = a^x \ln a$).

For minimum, set $\frac{dE}{dk} = 0$, giving $1 + k^2(1-p)^k \ln(1-p) = 0$.

(v) $1 + k^2(0.99)^k \ln(0.99) = 0$, so that $k^2 = \frac{-1}{(0.99)^k \ln(0.99)}$.

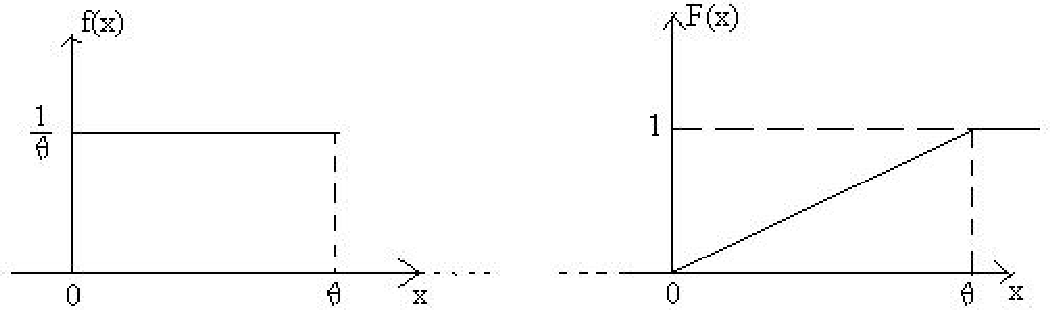
Find $E[S]$ for $k = 10$ and 11 (since it must be an integer).

$E[S|k = 10] = 9900(0.1 + 1 - 0.99^{10}) = 1936.62$.

$E[S|k = 11] = 9900(\frac{1}{11} + 1 - 0.99^{11}) = 1936.15$.

Take $k = 11$.

6. (i) $P(X \leq \xi) = \int_0^\xi \frac{dx}{\theta} = \frac{\xi}{\theta}$, for $0 \leq \xi \leq \theta$; $= 0$ for $\xi < 0$; $= 1$ for $\xi > \theta$.



$E[X] = \int_0^\theta \frac{xdx}{\theta} = \frac{1}{2\theta}[x^2]_0^\theta = \frac{\theta}{2}$.

$E[x^2] = \int_0^\theta \frac{x^2dx}{\theta} = \frac{1}{3\theta}[x^3]_0^\theta = \frac{\theta^2}{3}$; $V[X] = E[x^2] - (E[X])^2 = \frac{\theta^2}{3} - \frac{\theta^2}{4} = \frac{\theta^2}{12}$.

(ii) Sampled items chosen at random, hence $\{X_i\}$ are independent. $P(X \leq x) = \frac{x}{\theta}$ for each item, all are required to be $\leq x$, so probability for n items is $(\frac{x}{\theta})^n$, when $0 \leq x \leq \theta$. This is $F(X_{(n)})$, where $X_{(n)}$ is the sample maximum, and so $f(x_{(n)}) = F^{'}(x_{(n)}) = \frac{nx^{n-1}}{\theta^n}$, when $0 \leq x \leq \theta$, =0 otherwise.

$E[X_{(n)}] = \int_0^\theta \frac{nx^n}{\theta^n}dx = [\frac{nx^{n+1}}{(n+1)\theta^n}]_0^\theta = \frac{n\theta}{n+1}$,

$E[X_{(n)}^2] = \int_0^\theta \frac{nx^{n+1}}{\theta^n}dx = [\frac{nx^{n+2}}{(n+2)\theta^n}]_0^\theta = \frac{n\theta^2}{n+2}$.

Hence

$$\begin{aligned} V[X_{(n)}] &= \tfrac{n\theta^2}{n+2} - (\tfrac{n\theta}{n+1})^2 = \theta^2\{\tfrac{n}{n+2} - \tfrac{n^2}{(n+1)^2}\} \\ &= \tfrac{\theta^2}{(n+1)^2(n+2)}\{n(n+1)^2 - n^2(n+2)\} = \tfrac{n\theta^2}{(n+1)^2(n+2)}. \end{aligned}$$

$\frac{n+1}{n}X_{(n)}$ is unbiased for estimating θ.

$Var[\frac{n+1}{n}X_{(n)}] = (\frac{n+1}{n})^2\frac{n\theta^2}{(n+1)^2(n+2)} = \frac{\theta^2}{n(n+2)}$.

The likelihood of a sample $\{x_1, \cdots, x_n\}$ is $\frac{1}{\theta^n}$, $(0 \leq x \leq \theta)$.

Setting $\hat{\theta} = X_{(n)}$, where is the lowest value of θ possible on the evidence of the sample values, gives the largest possible value of the likelihood. Hence $X_{(n)}$ is the m.l. estimator.

(iii) Method of moments estimator $\tilde{\theta}$ is found from setting $\bar{x} = E[x]$, i.e., $\bar{x} = \frac{1}{2}\tilde{\theta}$ so that $\tilde{\theta} = 2\bar{x}$. $V[\tilde{\theta}] = 4V[\bar{x}] = \frac{4}{n}V[x] = \frac{4}{n}\cdot\frac{\theta^2}{12} = \theta^2/3n$.

(iv) $\frac{n+1}{n}X_{(n)}$ is unbiased, and $X_{(n)}$ very nearly so if n is at all large; their variances are much smaller than that for $\tilde{\theta}$, the estimator based on the mean. Hence, use $X_{(n)}$ if there are a reasonable number of offcuts; if only few, multiply by the factor $\frac{n+1}{n}$.

7. (i) $Y_i = \beta x_i + e_i$, $i = 1, 2, \cdots, n$, $\{e_i\}$ i.i.d. $N(0, \sigma^2)$.

Likelihood $L = \prod_{i=1}^{n}\{\frac{1}{\sigma\sqrt{2\pi}}\exp[-\frac{(y_i - \beta x_i)^2}{2\sigma^2}]\}$,

$\ln L = \Lambda = -n\ln(\sigma\sqrt{2\pi}) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)^2$.

$\frac{\partial\Lambda}{\partial\beta} = 0 + \frac{1}{2\sigma^2}\cdot 2\sum_{i=1}^{n}(y_i - \beta x_i)x_i$ and is zero when $\sum(y_i - \hat{\beta}x_i)x_i = 0$ i.e. $\sum y_i x_i = \hat{\beta}\sum {x_i}^2$ or $\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2}$.

$\frac{\partial^2\Lambda}{\partial\beta^2} = -\frac{\sum x_i^2}{\sigma^2}$, confirming maximum.

(ii) If now $\{e_i\}$ are $N(0, \sigma^2 x_i)$,

$L = \prod_{i=1}^{n}\{\frac{1}{\sigma\sqrt{2\pi x_i}}\exp[-\frac{(y_i - \beta x_i)^2}{2x_i\sigma^2}]\}$

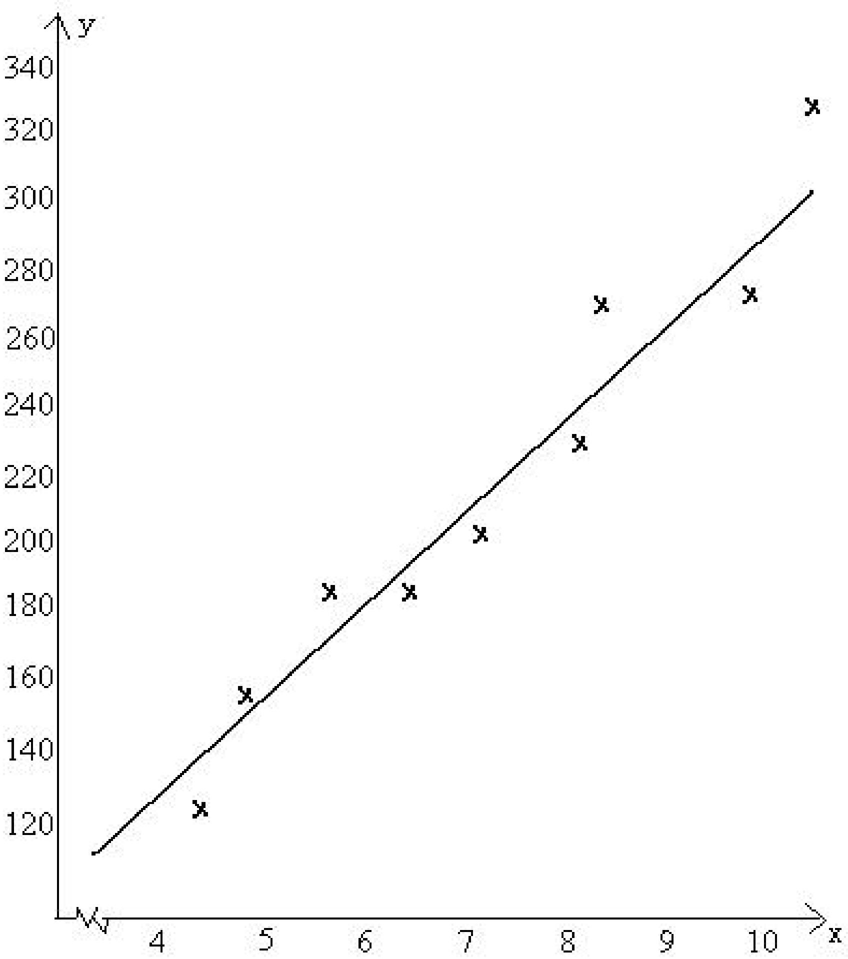
and $\Lambda = -n\ln(\sigma\sqrt{2\pi}) - \frac{n}{2}\sum_{i=1}^{n}\ln x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\frac{(y_i - \beta x_i)^2}{x_i}$.

$\frac{\partial\Lambda}{\partial\beta} = 0 + 0 + \frac{1}{2\sigma^2}\cdot\sum_{i=1}^{n}\frac{1}{x_i}2(y_i - \beta x_i)x_i = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \beta x_i)$.

This is zero when $\sum(y_i - \hat{\beta}x_i) = 0$ i.e. $\sum y_i = \hat{\beta}\sum x_i$ or $\hat{\beta}_2 = \frac{\sum y_i}{\sum x_i}$.

$\frac{\partial^2\Lambda}{\partial\beta^2} = -\frac{\sum x_i}{\sigma^2}$, confirming maximum.

The first case (i) has $L = \text{Constant} - \frac{1}{2\sigma^2}\sum e_i^2$, considered as a function of β; similarly case (ii) has $L = \text{Constant} - \frac{1}{2\sigma^2}\sum \frac{e_i^2}{x_i}$. Considering as a function of β. Thus L is maximized when (i) $\sum e_i^2$ or (ii) $\sum e_i^2/x_i$ is minimized (note the - sign).



(iii)

										SUM
X	4.3	4.9	6.5	5.7	7.2	8.3	8.4	9.6	10.1	65.0
Y	123	156	183	183	204	234	270	273	324	1950
XY	528.9	764.4	1043.1	1189.5	1468.8	1942.2	2268.0	2620.8	3272.4	15098.1
X_2	18.49	24.01	32.49	42.25	51.84	68.89	70.56	92.16	102.01	502.70

$n = 9$. $\hat{\beta}_1 = \frac{15098.1}{502.7} = 30.034$. $\hat{\beta}_2 = \frac{1950}{65} = 30.000$

The regression lines are indistinguishable between the two models. However, the residuals (difference between y and the value on the line at the same x - value - i.e. the vertical differences) show a definite tendency to increase as x increases. For this reason, model (ii) is likely to be better.

8. (a) (i) A binomial distribution with large n and very small p may be approximated by a Poisson with $\mu = np$. It is desirable that np should be at least 5, but in addition n should be ≥ 20 and $p \leq 0.1$. When all these conditions are met the approximation will be a good one.

(ii) A Poisson with large mean can be approximated by $N(\mu, \mu)$. The approximation will be good for $\mu \geq 10$, but adequate down to $\mu = 5$.

(b) (i) Poisson, mean 5:

$$P(4) + P(5) + P(6) = e^{-5}(\tfrac{5^4}{4!} + \tfrac{5^5}{5!} + \tfrac{5^6}{6!})$$
$$= 625e^{-5}(\tfrac{1}{24} + \tfrac{5}{120} + \tfrac{25}{720}) = 625e^{-5}(\tfrac{1}{12} + \tfrac{5}{144}) = 0.49716.$$

(ii) $N(5, 5)$ with a continuity correction is required: find $P(3\frac{1}{2} < X < 6\frac{1}{2})$ in $N(5, 5)$. Corresponding r-values are $\frac{3\frac{1}{2}-5}{\sqrt{5}} = -0.6708$ and $\frac{6\frac{1}{2}-5}{\sqrt{5}} = +0.6708$. $P(z < -0.6708) = P(z > +0.6708) = 0.25117$, and so the required probability is $1 - 2 \times 0.25117 = 0.49766$. The error is 0.0005, and % error $\frac{0.0005}{0.49716} \times 100 = 0.1\%$.

Using the continuity correction with $\mu = 5$, and calculating values which we near to the mean, leads to a very good approximation.

(c) $P(0) = e^{-\lambda t} = P(T > t)$ for the first event observed $= 1 - F(t)$. Hence $F(t) = 1 - e^{-\lambda t}$ and $g(t) = F^{'}(t) = \lambda e^{-\lambda t}$. $(t \geq 0; \lambda > 0)$.

[$g(t) = 0$ unless $t \geq 0, \lambda > 0$ since neither time of events nor rate of events occurring can be negative.] Use integration by parts.

$$\begin{aligned} E[T] &= \int_0^\infty \lambda t e^{-\lambda t} dt = \int_0^\infty t d(-e^{-\lambda t}) = [-te^{-\lambda t}]_0^\infty + \int_0^\infty e^{-\lambda t} dt \\ &= [-\frac{1}{\lambda} e^{-\lambda t}]_0^\infty = 1/\lambda. \\ E[T^2] &= \int_0^\infty \lambda t^2 e^{-\lambda t} dt = \int_0^\infty t^2 d(-e^{-\lambda t}) = [-t^2 e^{-\lambda t}]_0^\infty + \int_0^\infty 2te^{-\lambda t} dt \\ &= \frac{2}{\lambda} E[T] = 2/\lambda^2. \end{aligned}$$

Hence $V[T] = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = 1/\lambda^2$.

(d) $\lambda = 5$, so that $E[T] = 0.2$ and $V[T] = 0.04$. For $n = 100$, a sample mean $\bar{T}$ is approximately $N(0.2, \frac{0.04}{100})$, and the range required is from 0.18 to 0.22, within 10% of 0.2. The corresponding values of r are $\frac{0.18-0.2}{\sqrt{0.0004}} = \frac{-0.02}{0.02} = -1$, and the other $= +1$. $P(r > 1) = 0.1587 = P(r < -1)$ and so the probability between these values is $1 - 2 \times 0.1587 = 0.6826$.

II. STATISTICAL METHODS

1. (i) For 1988, $\bar{x}_1 = 53.4$, $s_1 = 19.7$; also $n = 750$;

   for 1990, $\bar{x}_2 = 55.3$, $s_2 = 19.5$; also $n = 633$.

   If μ_1, μ_2 are the corresponding population means, H_0 is $\mu_1 = \mu_2$ (or, strictly, $\mu_1 \geq \mu_2$) and H_1, to be tested, is $\mu_2 > \mu_1$.

   $V(\bar{x}_2 - \bar{x}_1) = \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} = \frac{19.7^2}{750} + \frac{19.5^2}{633} = 1.11816$, $SE = 1.057$.

   As these are large samples of date we use a normal (r) test:

   $r = \frac{55.3-53.4}{1.057} = \frac{1.9}{1.057} = 1.798$.

   The form of H_1 requires a one-tail test, with critical value 1.645 at 5%. Hence we reject H_0.

   A 95% confidence interval for the increase is $1.9 \pm 1.96 \times 1.057 = 1.9 \pm 2.07$, or (-0.17; 3.97).

   If we are certain that there must have been an increase we may prefer to quote this result as $(0, 3.97)$.

   (ii) For 1988, $p_M = \frac{349}{750} = 0.4653$ and $p_F = 0.5347$; $n = 750$.

   For 1990, $p_M = \frac{321}{633} = 0.5071$ and $p_F = 0.4929$; $n = 633$.

   The hypotheses H_0: $p_{M,1988} = p_{M,1990}$ and H_1 : p_M has changed can be examined in a 2 × 2 table of 'observed' frequencies and 'those expected on H_0'.

OBSERVED(EXPECTED)	1988	1990	TOTAL
MALE	349(363.34)	321(306.66)	670
FEMALE	401(386.66)	312(326.34)	713
	750	633	1383

$$\begin{aligned}\chi^2_{(1)} &= \frac{(349-363.34)^2}{363.34} + \cdots + \frac{(312-326.34)^2}{326.34} \\ &= (14.34)^2\{\frac{1}{363.34} + \frac{1}{306.66} + \frac{1}{386.66} + \frac{1}{326.34}\} \\ &= 205.6356 \times 0.011664 = 2.40 n.s.\end{aligned}$$

   There is no evidence of change.

   [An alternative method is to use normal approximations for p_M: $N(p, \frac{p(1-p)}{n})$ in each year and consider the difference. This would be needed if confidence intervals had been required.]

2. (a) $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where y_{ij} is the observation measured as the j^{th} of the items receiving treatment i; μ is a grand (overall) mean term; τ_i is an effect

(deviation from mean) due to treatment i; ϵ_{ij} are i.i.d. $N(0, \sigma^2)$ residual terms. There are $i = 1$ to v treatments, r_i replicates of each, and $\sum_{i=1}^{v} r_i = N$, the total number of items in the experiment.

(b)

"Treatment"	r_i	$\sum y_{ij}$	$\sum y_{ij}^2$	$\bar{y}_i$
1	6	128.0	2792.00	21.33
2	4	79.4	1582.06	19.85
3	4	90.7	2064.51	22.68
4	3	60.5	1243.25	20.17
	17	358.6	7681.82	

Although results 1 are rounded to whole numbers, analysis of variance will have to assume that all observations on all treatments have the same variance σ^2. Also the material used in the trial should have been selected at random from what was available, and the samples examined under identical conditions in random order.

(i) Total corrected $S.S. = 7681.82 - G^2/N = 7681.82 - 7564.35 = 117.47$. "Treatments" $S.S. = \frac{128^2}{6} + \frac{79.4^2 + 90.7^2}{4} + \frac{60.5^2}{3} - \frac{G^2}{N} = 7583.4625 - 7564.35 = 19.11$.

Analysis of Variance	$D.F.$	$S.S.$	$M.S.$	
Treatments(Storages)	3	19.11	6.371	$F < 1$
Residual	13	98.36	$7.566 = \hat{\sigma}^2$	
TOTAL	16	117.47		

We are not given any specific contrasts among storages to be tested, but even if the whole Treatments $S.S.$ were due to one contrast this would still not be significant as $F_{(1,13)}$ ($\frac{19.11}{7.566} = 2.52$, less than the 5% point 4.67). We may say confidently, that there are no significant differences among these "Treatments".

(ii) Given the result in (i), there could be no change to the inference. [In a borderline case, some intelligence in looking at individual differences may be called for, as σ^2 may be slightly overestimated.]

3.

Pair	A	B	C	D	E	F	G	H	I	J	
Sign (Gp.2 - Gp.1)	$+$	$+$	$+$	$-$	$-$	$+$	$-$	$+$	$+$	$+$	$7+; 3-$
Difference	$+3$	$+8$	$+5$	-1	-1	$+25$	-1	$+3$	$+19$	$+10$	
Rank	$4\frac{1}{2}$	7	6	2	2	10	2	$4\frac{1}{2}$	9	8	

(i) The number of + signs should be binomial ($n = 10, p = 1/2$) on the Null Hypotheses of no difference between groups (i.e. training methods). Using a continuity correction, find $P(r \geq 7)$ in $N(5, 5/2)$:

$r = \frac{6\frac{1}{2}-5}{\sqrt{2.5}} = \frac{1.5}{1.581} = 0.949$, n.s., so no evidence of difference.

The exact probability $P(7)+P(8)+P(9)+P(10)$ in $B(10, 1/2) = \frac{1}{2^{10}}(\binom{10}{7}+\binom{10}{8}+\binom{10}{9}+\binom{10}{10}) = \frac{1}{2^{10}}(120 + 45 + 10 + 1) = \frac{176}{1024} = 0.172$, and so the probability of the given result in a 2-tail test (A. H. "there is a difference between groups", direction not specified) is 0.344. Again no evidence of any difference.

(ii) The sum of the positive ranks is 49, and of negative 6. The value 6 is (approximately) $N(\frac{1}{4}n(n+1), \frac{1}{24}n(n+1)(n+2))$, making no allowance for the ties in the ranks (3 of -1 and 2 of +3). $n = 10$, the number of non-zero differences, so $\frac{n(n+1)}{4} = 27.5$ and $\frac{1}{24}n(n+1)(n+2) = 96.25$. Using a continuity correction, $r = \frac{6.5-27.5}{\sqrt{96.25}} = \frac{-21.0}{9.81} = -2.14^*$.

At the 5% level, there is significant evidence against the N. H. [Using the Wilcoxon table, the critical number is 8, and 6, being less than this, is significant at 5%.]

This test uses the information on numerical sizes of differences, whereas the sign test does not. All the negative ones were very small.

If the differences had appeared to be normally distributed, a t-test (paired version) would have been appropriate. This seems very unliablely, since there is no clustering around a mean, and there are several large values.

4. (A)

	1	2				More extreme tables are								
R	11	13	:	24	|	10	14	:	24	and	9	15	:	24
NR	7	2	:	9		8	1	:	9		9	0	:	9
	18	15		33		18	15		33		18	15		33

Together, these form the "tail" of the distribution when margins are fixed. probability are

$$\frac{18!\ 15!\ 24!\ 9!}{33!\ 11!\ 7!\ 13!\ 2!};\ \frac{18!\ 15!\ 24!\ 9!}{33!\ 10!\ 8!\ 14!\ 1!};\ \frac{18!\ 15!\ 24!\ 9!}{33!\ 9!\ 9!\ 15!\ 0!}.$$

i.e. $\frac{18\times17\times14}{55\times31\times29} = 0.08664$; $\frac{9\times17}{31\times290} = 0.01703$; $\frac{17}{31\times29\times15} = 0.00126$.

For a 2-tail test of the Null Hypothesis of no difference, the probability is $2(0.08664+0.01702+0.00126) = 0.2098$. There is no significant evidence for any difference between the two drugs.

(B) The $\chi^2_{(1)}$ test also tests the N. H. that the proportions recovering are the same on each drug. 'Expected' frequencies are those given by this N. H. with the same marginal totals as the 'Observed'.

OBSERVED (EXPECTED)	Drug1	Drug2	
Recovered	11(12.65)	13(11.35)	24
Not recovered	18(16.35)	13(14.65)	31
	29	26	55

$\chi^2_{(1)} = (1.65)^2(\frac{1}{12.65} + \frac{1}{11.35} + \frac{1}{16.35} + \frac{1}{14.65}) = 2.7225 \times 0.296578 = 0.807$ n.s. Again there is no significant evidence against the N. H.

(ii) The inference is the same in this, rather small, trial whether or not the drop-outs are included. There were 11 drop-outs on each drug, which is a considerable proportion of patients beginning the trial; however, no particular reasons for drop-out are known.

5. (i) The Central Limit Theorem says that if $\{x_i\}$ are n independent observations, all taken from the same distribution with (finite) mean and variance μ and σ^2, then the limiting distribution as $n \to \infty$ of the mean $\bar{X}$ is $N(\mu, \sigma^2/n)$. Therefore if μ, σ^2 are known, the actual form of the distribution of $\{x_i\}$ is not important provided n is large. However, if the X distribution is skew, n in practice needs to be very large, 500+, whereas if this distribution is symmetrical, even though not itself normal, a sample of less than 50 observations may be adequate for the the approximation to be used. It is the basis for "large sample" tests of means and differences of means; and can also be applied to discrete distributions, e.g. in testing proportions.

(ii) If there are systematic differences between "blocks" or groups of items which have to be used in the same experiment, as well as random variation and systematic effects of treatments, a linear model (as in 2(a)) needs to contain a term for blocks: $y_{ij} = \mu + \tau_i + \beta_j + e_{ij}$, $(i = 1$ to $v, j = 1$ to $r)$, where every treatment, 1 to v, appears once in every block, 1 to r. A "two-way" analysis of variance, to remove blocks as well as treatments from the total sum of squares, is required:

SOURCE OF VARIATION	D.F.
Blocks	$r-1$
Treatments	$v-1$
Residual	$(r-1)(v-1)$
TOTAL	$rv-1$

The $\{e_{ij}\}$ are required to be i.i.d. $N(0, \sigma^2)$, and an estimate of σ^2 is provided by the residual mean square in this analysis.

(iii) Degrees of freedom is a parameter in a χ^2-distribution. The square of a $N(0, 1)$ variate is $X^2_{(1)}$, and the sum of the squares of n independent $N(0, 1)^{'}$s is $\chi^2_{(n)}$. Thus for $X = N(\mu, \sigma^2)$, so that $\frac{X-\mu}{\sigma} = r$ is $N(0, 1)$ the sum of n independent observations $\sum_{i=1}^{n}(\dfrac{x_i - \mu}{\sigma})^2$ is $\chi^2_{(n)}$. When μ is not known and must be replaced by an estimate $\bar{x}$ from a sample, $\sum_{i=1}^{n}(\dfrac{x_i - \bar{x}}{\sigma})^2$ is $\chi^2_{(n-1)}$ because the n values $(x_i - \bar{x})$ used in the calculation have one constraint placed on them, namely $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ by definition of $\bar{x}$. In other applications of χ^2, such as tests in contingency tables or tests of goodness of fit, the rule for degrees of freedom is "number of independent items of information minus number of constraints placed on them". Constraints include fixing the marginal totals and the grand total in tables when calculating expected values, and having to estimate parameters from the observed data (e.g. mean in a Poisson distribution) before expected values can be calculated. In the analysis of variance, d.f. for total are $N - 1$ when N observations are available, $v - 1$ for v treatments etc. The t-statistic always has the same number of d.f. as the estimate s^2 used in it, e.g. the residual d.f. in the analysis of variance.

(iv) Confidence intervals often give more information than significance tests based on the same sample of data. When a parameter (e.g. μ) has been estimated (by, e.g. $\bar{x}$) we may set up a Null Hypothesis that μ takes a certain value and then test whether $\bar{x}$ is close enough to this for the NH not to be rejected. However, there is a whole range of values of μ which would be consistent with the observed value of $\bar{x}$, and it is this range which forms a confidence interval. Fox example, if $\{x_i\}$ are taken from $N(\mu, \sigma^2)$, with σ^2 known, and their mean is $\bar{x}$, then we know that if $i = 1$ to n then $\bar{x}$ has the distribution $N(\mu, \sigma^2/n)$, so that $P(-1.96 \leq \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq +1.96) = 0.95$, which can be written as $P(\bar{x} - 1.96\sigma/\sqrt{n} \leq \mu \leq \bar{x} + 1.96\sigma/\sqrt{n}) = 0.95$, giving a 95% confidence interval for the true value of μ, based on the sample mean $\bar{x}$. With probability 0.95, the interval contains μ. To alter the probability, or confidence level, to e.g. 90% or 99% the corresponding r-values $(N(0, 1))$ must be used instead of 1.96, e.g. 1.645 and 2.576. Confidence intervals can be set up whenever we know the distribution of a parameter estimate, e.g.

s^2 for σ^2, b for β in linear regression. If an interval is wide relative to the size of the estimate, the lack of precision is immediately clear; this information is hidden in a significance test.

6. If we assume the data follow a normal distribution with mean μ and variance σ^2, then we can use the data (16 observations) to test the Null Hypotheses that (i) $\mu \geq 400$, (ii) $\sigma^2 \leq 64$ against the Alternatives $\mu < 400$ and $\sigma^2 > 64$. The mean of the sample is $\bar{x} = 396.125$ and variance $s^2 = 77.9833$.

(i) For the mean, $t_{(15)} = \frac{396.125-400}{\sqrt{77.9833/16}} = -\frac{3.875}{2.208} = -1.76$, which is just on the borderline of significance at 5% in a 1-tail test. The evidence from these data is that the mean is not likely to be ≥ 400.

(ii) For the variance, $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{(n-1)}$ i.e. $\frac{15\times 77.9833}{64}$ is $\chi^2_{(15)} = 18.28$, which is less than the 5% (upper) point of $\chi^2_{(15)}$, so there is no evidence to reject the hypothesis that $\sigma^2 \leq 64$, even though the observed value is above this.

With a sample four times as large, i.e. $n = 64$, assuming the same estimates $\bar{x}$ and s^2, $t_{(63)}$ would be $\sqrt{4}$, i.e. 2, times as large, providing very strong evidence against the Null Hypothesis for μ. The variance would be based on 63 d.f., and the χ^2 statistic would be $\frac{63\times 77.9833}{64} = 76.76$, which is not significant and so there is still no evidence against the Null Hypothesis for σ^2.

7. (i) If the process is producing individual rejects "at random", i.e. singly and at unpredictable instants of time, but at a constant rate over the period of study, then the number of rejects during a fixed time of observation will follow a Poisson distribution.

(ii) The mean must be estimated from the data:

$$\bar{x} = \frac{1}{160}(0 + 49 + 86 + 51 + 44 + 10) = \frac{240}{160} = 1.5$$

Expected frequencies are $160e^{-1.5}(1.5)^r/r!$ for $r = 0, 1, \cdots$.

r :	0	1	2	3	4	≥ 5	TOTAL
Obs:	38	49	43	17	11	2	160
Exp:	35.70	53.55	40.16	20.08	7.53	2.98	160

(The last two cells may be combined, but this is not really necessary.)

χ^2 has 4 d.f., since 1 parameter had to be estimated and the totals of Obs and Exp have to be the same.

$$\begin{aligned}\chi^2_{(4)} &= \tfrac{(38-35.70)^2}{35.70} + \tfrac{(49-53.55)^2}{53.55} + \tfrac{(43-40.16)^2}{40.16} + \tfrac{(17-20.08)^2}{20.08} + \tfrac{(11-7.53)^2}{7.53} + \tfrac{(2-2.98)^2}{2.98} \\ &= 3.13, \text{not significance.}\end{aligned}$$

There is no reason to reject the hypothesis that the data follow a Poisson distribution. Therefore the number of rejects per unit time is likely to remain reasonably constant and they do not arise in any regular or predictable way.

8. (a). The F distribution with (v_1, v_2) degrees of freedom is the distribution of the ratio of two χ^2 distributions - specifically $F(v_1, v_2) = \frac{\chi_{(v_1)}}{v_1} / \frac{\chi_{(v_2)}}{v_2}$. Therefore, two independent estimates of variance from the same population, based respectively on $(v_1 + 1)$ and $(v_2 + 1)$ observations may be compared in an F distribution. An example of this is in (b), assuming observations normally distributed.

   Also, two sums of squares of normally distributed variables can be compared. An examples is in the analysis of designed experiments (see Question 2), where the residual sum of squares provides an estimate of natural variation, σ^2, and the treatment sum of squares also provides an estimate of this on the Null Hypothesis of no treatment differences: the actual estimates are the sums of squares divided by their degrees of freedom, i.e. the "mean squares". Hence these mean squares can be compared in an F-test, as long as observations are normally distributed.

   Similarly, in linear regression, the sum of squares of deviations from the fitted "live" provides a test of fit: sums of squares for regression is $\chi^2_{(p-1)}$ when p x-variables are used, and the residual sum of squares is $\chi^2_{(n-p)}$. Hence two mean squares can be found whose ratio will be $F(p-1, n-p)$ if a linear fit is adequate.

   (b). For men, $v_1 = 12$ and for women, $v_2 = 10$. Calculate $s^2 = \frac{1}{v-1} \sum (x_i - \bar{x})^2$ for each sample. For men, $s_1^2 = 30.3333$ and for women $s_2^2 = 7.3778$, $F_{(11,9)} = \frac{s_1^2}{s_2^2} = 4.1114^*$. Since this is significant at 5% on the Null Hypothesis that the two variances are equal, we must reject that hypothesis.. Each set of data is assumed to be from a normal population, there is some suggestion that this may not be true for the men, but rather there are two sub-populations.

   $\frac{(v_2-1)s_2^2}{\sigma_2^2} \div \frac{(v_1-1)s_1^2}{\sigma_1^2}$ is the ratio $\frac{\chi_{(v_2-1)^2}}{\chi_{(v_1-1)^2}}$ i.e. $F(v_2-1, v_1-1)$. Therefore $\frac{v_2-1}{v_1-1} \cdot \frac{s_2^2}{s_1^2} \cdot \frac{\sigma_1^2}{\sigma_2^2} \sim F(v_2-1, v_1-1)$, or $\frac{\sigma_1^2}{\sigma_2^2} = \frac{(v_1-1)s_1^2}{(v_2-1)s_2^2} F(v_2-1, v_1-1)$. Limits for $\frac{\sigma_1^2}{\sigma_2^2}$ are $\frac{11}{9} \cdot 4.1114 \cdot F(9, 11)$, where the upper and lower $2\frac{1}{2}$% points of $F(9, 11)$ are to be used. The upper point is 3.59. For the lower point, use the fact that $P(F > F^*) = P(\frac{1}{F} < \frac{1}{F^*})$, where F^* is the critical value. But $\frac{1}{F}$ is also an F-variable, with upper and lower degrees of freedom interchanged. The upper $2\frac{1}{2}\%$ point of $F(11, 9)$ is 3.92. Hence the required lower $2\frac{1}{2}\%$ points $1/3.92 = 0.255$. The 95% confidence interval is 5.025×0.255 to 5.025×3.59 i.e. (1.2814 to 18.04).

III. STATISTICAL APPLICATIONS & PRACTICE

1. (i) Missing entries are ($*$ indicates cannot be calculated):
   Moving Average $*$, $*$, 108.250, 141.913, $*$, $*$.
   Difference $*$, $*$, 14.90000, -7.4250, 0.7500, 13.8875, $*$, $*$.

   (ii) see next sheet for graph.

   (iii) Seasonal effects:

Quarter	1	2	3	4	
	-7.4250	0.4375	14.9000	-5.7625	
	-5.2375	0.7500	16.7125	-10.6250	
	-7.3875	1.3000	8.0500	-3.0375	
	-8.1750	4.9875	13.8875	-6.8000	
MEAN	-7.05625	1.86875	13.3875	-6.55625	: 1.64375
Correction	-0.41094	-0.41094	-0.41094	-0.41094	(≈ 0)
SEASONAL	-7.4672	1.4578	12.9766	-6.9672	

   (iv) Since the data are given to 1 decimal place, 7.5 should be added to each Q1 item, 7.0 to each Q4 item, 1.5 subtracted from each Q2 item and 13.0 from each Q3 item, to "deseasonalise".

   (v)

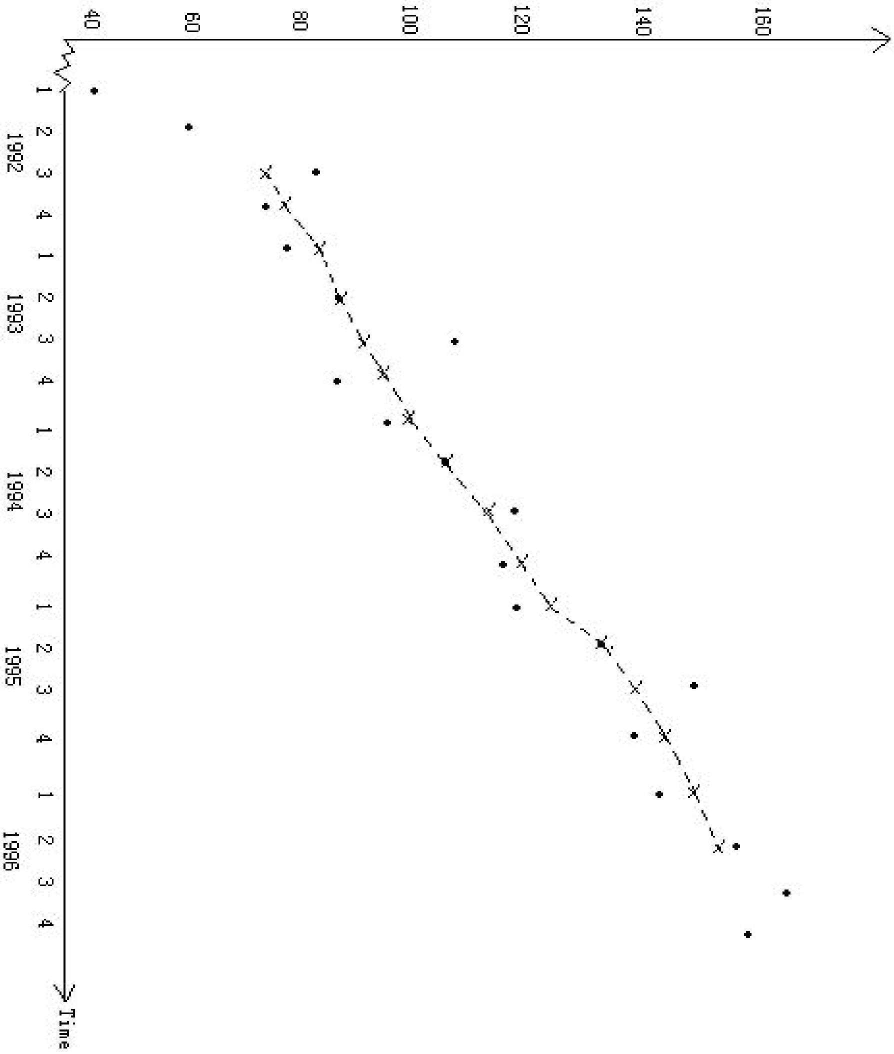
1997 :	$Q1$	$Q2$	$Q3$	$Q4$
$50 + 6t$:	176	182	188	194
Seasonalised:	168.5	183.5	201.0	187.0

   This assumes same general trend and seasonal effects continue.

2. (i) The calculations of sums of squares and products are heavily influenced by the two point $(3.48, 6.05)$ and $(3.49, 6.29)$. These lead to $\sum(x-\bar{x})(y-\bar{y})$ being negative, and hence the slope is negative.

   (ii) Removing these two points gives completely different summary line for the remainder. The new values of sums etc. are: $\sum x = 100.04 - 3.48 - 3.49 = 93.07$; $\sum y = 117.12 - 6.05 - 6.29 = 104.78$; $\sum x^2 = 436.9760 - 3.48^2 - 3.49^2 = 412.6855$; $\sum xy = 508.0134 - (3.48 \times 6.05) - (3.49 \times 6.29) = 465.0073$. $\sum(x-\bar{x})(y-\bar{y}) = 465.0073 - \frac{1}{21}(104.78 \times 93.07) = 0.63232$. $\sum(x-\bar{x})^2 = 412.6855 - 93.07^2/21 = 0.20812$. $\hat{b} = 3.038$.

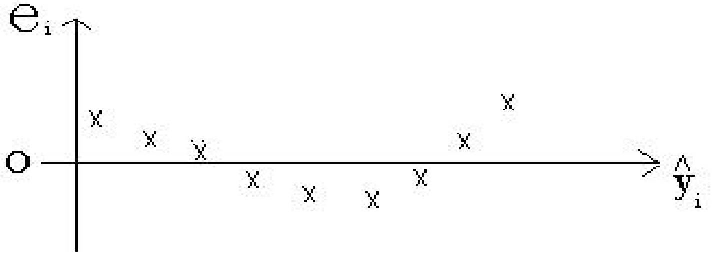
   (iii) A regression line has to go through the mean $(\bar{x}, \bar{y})$ of all the data. If there are two (or more) parts to the population or set of data, as here, then it



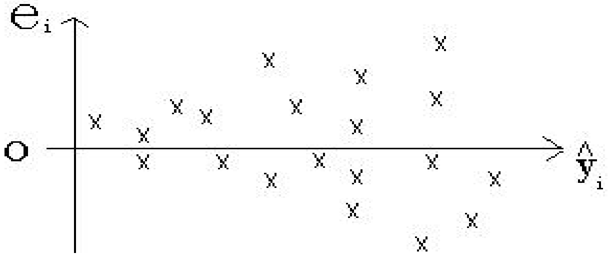
does not explain the data well. The sub-populations have to be separated. In this case there are only two points that are well away from the rest. All the remaining 21 have their x values (log surface temperature) between 4.2 and 4.6 approximately. For temperatures in this range therefore we can use the regression line with slope +3.038 as a summary of the relationship. We do not have enough information to propose relationships outside this range of x-values (which correspond to $y^{'}$s between about 4.3 and 5.6).

3. A set of data may be fitted by a statistical model, e.g. a linear regression $y_i = a + bx_i + e_i$ or an experimental design model such as $y_{ij} = \mu + t_i + b_j + e_{ij}$ for randomized complete blocks. The terms $\{e_i\}$ or $\{e_{ij}\}$ are generally assumed $N(0, \sigma^2)$, independently of one another. After the parameters a, b or μ, $\{t_i\}$,

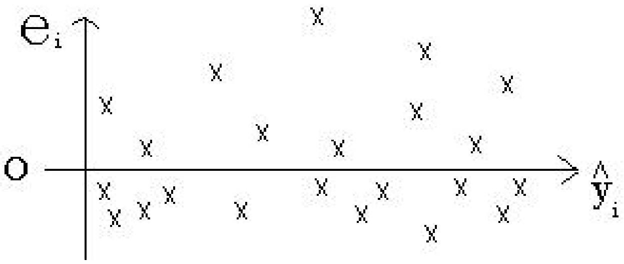
$\{b_j\}$ have been estimated the fitted values $\hat{y}_i = \hat{a}+\hat{b}x_i$ or $\hat{y}_{ij} = \hat{\mu}+\hat{t}_i+\hat{b}_j$ can be found. Then the differences $y_i - \hat{y}_i$ or $y_{ij} - \hat{y}_{ij}$, observed minus expected (fitted), are the residuals. These residuals should be from the same $N(0, \sigma^2)$ population. Their sizes should bear no systematic relationship to the sizes of the corresponding $\hat{y}_i$, $\hat{y}_{ij}$, or (for example) to x_i if x represents time in a set of time-series data or if x is any variable on a quantitative scale such as a level of fertilized application. Clearly they should cluster around 0 and be symmetrical. We may examine several of these properties in diagrams. A useful one is to plot the residuals e_i against corresponding fitted values y_i. If the wrong model has been fitted, e.g. a linear regression which should be a curve, the residuals will show a regular patten, e.g.,



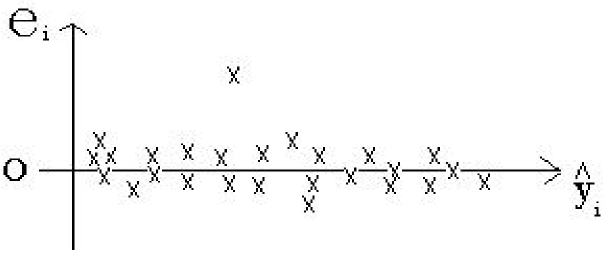
If the values of σ^2 is not constant, but increases as y increases, a 'fan' shape may appear.



A skew distribution, rather than normal, will have all the largest residuals on the same side of 0:



There may be outliers in the data, which will show up as isolated large values (positive or negative) of e_i:



However, this does not always happen (cf. Qu 2 where the two "odd" points can be fitted quite well by the first line with negative slope). Normal probability plotting can also be used to check the assumption of normality. The residuals, ordered by size, are plotted against the expected values of normal order statistics. Noticeable non-linearity is a warning that the assumption may be valid.

4. 2×2 factorial experiment in 5 replicates, completely randomized.

TOTALS.

Time:	*H*	*L*		
1210	73.88	68.93	:	142.81
1240	71.25	71.03	:	142.28
	145.13	139.96		285.09

(i) Correction term $G^2/N = 285.09^2/20 = 4063.8154$.

S.S. for Times $= \frac{1}{10}(145.13^2 + 139.96^2) - \frac{G^2}{N} = 1.33645$

S.S. for Temperatures $= \frac{1}{10}(142.81^2 + 142.28^2) - \frac{G^2}{N} = 0.01405$

S.S. for all "treatments" $= \frac{1}{5}(73.88^2{+}68.93^2{+}71.25^2{+}71.03^2){-}\frac{G^2}{N} = 2.46914$.

Corrected total S.S. $= 4067.00 - 4063.8154 = 3.1846$.

Analysis of Variance.

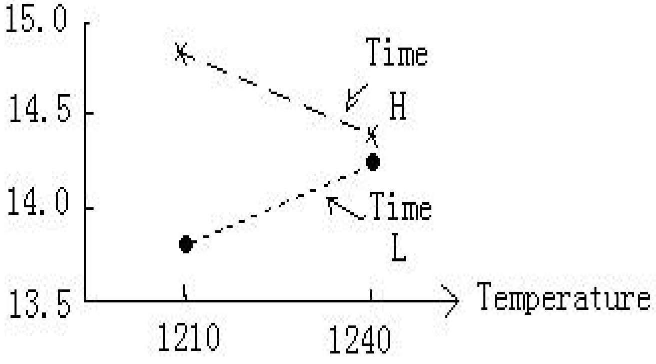
	D.F.	*S.S.*	*M.S.*	
Temperatures	1	0.01405	0.0141	
Times	1	1.33645	1.3365	
Interaction	1	1.11864	1.1186	$F_{(1,16)} = 25.03^{***}$
	3	2.46914		
Residual	16	0.71546	$0.0447 = s^2$.	
TOTAL	19	3.18460		

Since there is a very strong interaction of time with temperature, main effects should not be quoted.

(ii) Means are:

	Time:	*H*	*L*
Temperature	1210	14.78	14.79
	1240	14.25	14.21

A graph shows the results clearly:



The standard error of a single mean is $\sqrt{s^2/5} = 0.095$. Hence at 1240^0 C, time has no effect, while at 1210^0 C time H gives a thicker layer.

(iii) Report should make the point that neither time nor temperature alone determines the thickness of the layer; also for a thicker layer we should use the lower temperature and longer time, while the lower temperature and shorter time gives a relatively thin layer. At the higher temperature, with either time, the thickness of the layer is between these other two, and apparently not affected by time.

5. (i) We need to find out whether there are systematic trends along the rows, and / or whether one row is likely to do better than the other.

We also want to know whether all the grow-bags came from the same source, contain the same compost mixture, are the same size, have equally good drainage, the same thickness of wall so that temperature is likely to be the same.

Reasons for blocking would be: difference between rows, trend along rows, different sorts of bag.

(ii) The experimental unit is a bag of 4 plants. We would analyse the total (or mean) yield of plants per bag. If any plants died, we would need to adjust for this, so it should be recorded.

(iii) If there is no known or suspected systematic variation revealed in the answers to (i), a completely randomized design may be used, with a fully random choice of 16 bags for each of the four nutrient solutions. This could be achieved by using a random number table, reading digits in pairs, discarding pairs 00, 65 - 99, taking the first 16 positions for treatment A, the next 16 for B, the next 16 for C and the others for D, 01 - 64 represent the two rows with 32 bags in each.

If the answers to (i) indicate likely differences in the positions, make up 16 blocks each of which is as homogeneous as possible. Number the bags 1,2,3,4 in each block and permute these numbers at random to determine the order in which the 4 nutrients will be allocated to bags.

(iv) For the completely randomized design, the analysis is:

Source of Variation	D.F.
Nutrients	3
Residual	60
TOTAL	63

Using blocks of 4 in a randomized complete block design gives:

Source of Variation	D.F.
Blocks	15
Nutrients	3
Residual	45
TOTAL	63

6. Since we have a table of frequencies in various categories, an appropriate Null Hypothesis is that the ratio Good:Fair:Poor is the same in each area. A $\chi^2_{(8)}$ test is suitable. “Expected” frequencies are calculated from margin totals as usual, e.g. Ruthven / Good is $\frac{680\times 3689}{5842} = 429.39$.

OBS(EXP)	Good	Fair	Poor		
Area R	459(429.39)	178(210.56)	43(40.04)	:	680
M	926(969.29)	506(475.32)	103(90.39)	:	1535
W	954(930.77)	442(456.43)	78(86.79)	:	1474
D	985(995.82)	507(488.32)	85(92.86)	:	1577
A	365(363.72)	176(178.36)	35(33.92)	:	576
	3689	1809	344		5842

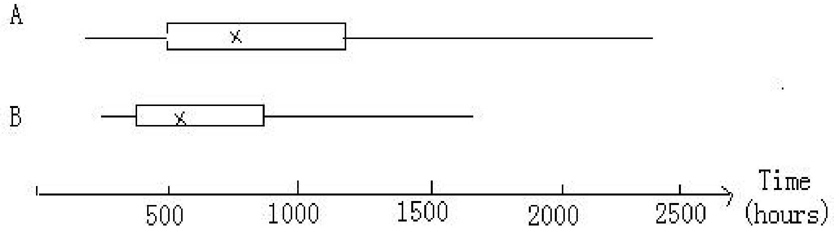
(Rounding expected frequencies is to the nearest 0.01).

$$\begin{aligned}\chi^2_{(8)} = \sum \tfrac{(0-E)^2}{E} = \; & 2.04185 + 1.93340 + 0.57977 + 0.11756 + 0.00450 + 5.03492 \\ & +1.98027 + 0.45620 + 0.71458 + 0.03123 + 0.21882 + 1.75918 \\ & +0.89024 + 0.66530 + 0.03439 = 16.46^*.\end{aligned}$$

This indicates that there are departures from a constant ratio in some areas. Comparing observed and corresponding expected frequencies shows that Ruthven has more 'Good' and less 'Fair' than expected; Mossmont has less 'Good' and more 'Fair' or 'Poor'; Windgyle has more 'Good' and less 'Fair' or 'Poor'; Dundonan has more 'Fair' and less 'Good' or 'Poor'.

(ii) Since respondents rate their own health this is very subjective and unlikely to produce the same ratings for the same condition in different people or areas. Also we obtain relatively little information per person and so require a large number of observations.
Actual measurements on a smaller number of people could provide data on blood pressure, cholesterol, weight and many other objective ways of assessing health, as well as observing the presence or absence of infections, and the general environmental conditions such as air quality.

7. (i) For type A, min = 171; lower quartile, $q = 396.5$; median, $M = \frac{1}{2}(568+795) = 681.5$; upper quartile, $Q = 1158$; max = 2415.
For B, min = 212; $q = 298.5$; $M = 447.5$; $Q = 823.5$; max = 1678.



The two distributions are distinctly skew, since the medians are not in the middle of the boxes made by the quartiles, and also the upper whiskers are very long. The variability in the distributions appears not to be the same either.

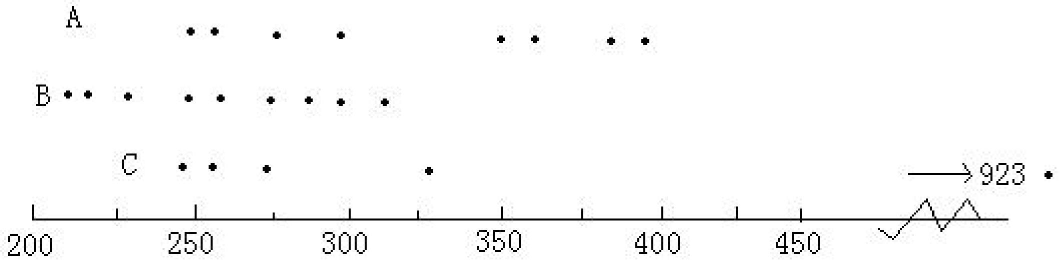
(ii) The t-test requires symmetry (strictly normality) of sets of data and, at least approximately, the same variance. Since neither of these seems very likely in the populations from which the samples were drawn, a Mann - Whitney test is preferred. This requires data to be of similar shape, but that is more reasonable. The Null Hypothesis will be that the populations have the same median values. The ranks of Type A are: 1, 4, 8, 9, 13, 14, 16, 17, 20, 21, 25, 28, 30, 32, 33, 34, 35, 37, 38, 40; and of type B: 2, 3, 5, 6, 7, 10, 11, 12, 15, 18, 19, 22, 23, 24, 26, 27, 29, 31, 36, 39.

Rank sums are: A, 455; B, 365. [check: sum $= 820 = \frac{1}{2} \cdot 40 \cdot 41$]. The mean of all ranks is 410, and the normal approximation to rank sum has variance $\frac{1}{12} \cdot 20 \cdot 20 \cdot 41$, i.e. s.d.$=36.97$.

(This form of the test is usually called Wilcoxan's Rank Sum test.)

Hence $r = \frac{455-410}{36.97} = 1.22$ is approximately $N(0,1)$; the value is not significant, so there is no evidence that medians differ.

8. A dot - plot for each set of data, on the same scale, is useful.



(i) 923 for C seems highly unlikely. This level may be physically impossible, to judge from all the other observations. Or it may be a recording error for 293 (or even 329).

(ii) Residual S.S. $= 19 \times 19797 = 376143$; hence treatment S.S. $= 70262$ and M.S. $= 35131$. The variance ratio is then $\frac{35131}{19797} = 1.77$.

The analysis, by itself, suggests that there are no significant treatment differences, and also that the standard deviation of an observation is very large ($\sqrt{19797} = 140.7$).

(iii) Revised sums and S.S. are:

	A	B	C	TOTAL
Sum	2533	2308	1097	5938
n	8	9	4	21
Sum of squares	826145	602898	305129	1734172

Treatments $SS = \frac{2533^2}{8} + \frac{2308^2}{9} + \frac{1097^2}{4} - \frac{5938^2}{21} = 1694737 - 1679040 = 15697.$

and Total $SS = 1734172 - 5938^2/21 = 55132$.

Source of variation	DF	Sum of Squares	$M.S.$	
Treatments	2	15697	7849	$F_{(2,18)} = 3.58^*$
Residual	18	39435	2191	
TOTAL	20	55132		

F is just significant at 5%. The estimated variance of an observation is 2191, S.D. = 46.8. Means are: A, 316.6; B, 256.4; C, 274.3. $\text{Var}[\bar{x}_A - \bar{x}_B] = s^2(\frac{1}{8} + \frac{1}{9}) = 517.32$, S.D. = 22.7, $t_{(18)} = \frac{60.2}{22.7} = 2.65^*$, so A and B appear to differ. $\text{Var}[\bar{x}_A - \bar{x}_C] = s^2(\frac{1}{8} + \frac{1}{4}) = 821.63$, S.D. = 28.7, $t_{(18)} = \frac{42.3}{28.7}$, not significant. A and C do not differ; nor will B and C.

(iv) The one doubtful observation greatly increased the variance estimate.