THE ROYAL STATISTICAL SOCIETY

HIGHER CERTIFICATE
(3 papers)

SOLUTIONS 1998

Royal Statistical Society

HIGHER CERTIFICATE IN STATISTICS

May 1998

PAPER I : Statistical Theory

1.(i)The probabilities of reading the three newspapers are

$P(A) = \frac{1}{3}$, P(B)=$\frac{1}{2}$, P(C)=$\frac{1}{6}$.

Let E be the event that an early is made.

P(E|A)=0.02, P(E|B)=0.01, P(E|C)=0.05 .

$$\begin{aligned} \text{P(E)} &= P(E|A)P(A) + P(E|B)P(B) + P(E|C)P(C) \\ &= (0.02 \times \tfrac{1}{3}) + (0.01 \times \tfrac{1}{2}) + (0.05 \times \tfrac{1}{6}) \\ &= \tfrac{1}{6}(0.04 + 0.03 + 0.05) \\ &= \tfrac{0.12}{6} = 0.02. \end{aligned}$$

Since all entries are correct, the probability of winning is P(A|E),P(B|E),P(C|E) respectively for readers of A, B, C.

P(A|E)P(E)=P(E|A)P(A) or P(A|E)=$\frac{0.02\times 1/3}{0.02} = \frac{1}{3}$;

P(B|E)=P(E|B)P(B)/P(E)=$\frac{0.01\times 1/2}{0.02} = \frac{1}{4}$;

P(C|E)=P(E|C)P(c)/P(E)=$\frac{0.05\times 1/6}{0.02} = \frac{5}{12}$.

[Check: these must sum to 1 . $\surd$]

(ii)Since the readerships are large, we may ignore the need for a finite population correction, and so the required probability will be $(\frac{1}{3})^2 = \frac{1}{9}$.

(iii)Similarly, for any two different newspapers, in either order for first, second, the probability will be: $2\{(\frac{1}{3} \times \frac{1}{4}) + (\frac{1}{3} \times \frac{5}{12}) + (\frac{1}{4} \times \frac{5}{12})\} = \frac{1}{6} + \frac{5}{18} + \frac{5}{24} = \frac{4}{9} + \frac{5}{24} = \frac{47}{72} = 0.653$.

There are 6 possible orders for one each of A, B, C, so the probability is $6\times\frac{1}{3}\times\frac{1}{4}\times\frac{5}{12} = \frac{5}{24} = 0.208$.

2.(i)For a non-sufferer, X∼N(7,9), so

$$\begin{aligned} \text{P(x}\geq 10) &= P(z = \tfrac{x-7}{3} \geq \tfrac{10-7}{3}) \qquad \text{where Z}\sim\text{N(0,1)} \\ &= 1 - P(Z \leq \tfrac{10-7}{3} = 1) \\ &= 1 - 0.8413 \\ &= 0.1587. \end{aligned}$$

(ii)For a sufferer, X∼N(19,36), so

$$\begin{aligned} \text{P(x}<10) &= P(Z = \tfrac{x-19}{6} < \tfrac{10-19}{6}) \qquad \text{where Z}\sim\text{N(0,1)} \\ &= P(Z < -\tfrac{3}{2}) = 0.0668. \end{aligned}$$

[Calculate as 1-P(Z<+3/2) if using tables.]

(iii)If critical level is x_0, P(X≥x_0|x ∼N(7,9))=0.05 .

In N(0,1) the upper 5% point is Z_0=1.645 ; hence $Z_0 = \frac{x_0-7}{3} = 1.645$ or $x_0 = 7+(3\times 1.645) = 11.935$.

$$\begin{aligned} \text{(iv)P(X}\geq 10) &= P(X \geq x_0|sufferer)P(has\ disease) + P(X \geq 10|non-sufferer)P(does\ not\ have\ disease) \\ &= (1 - 0.0668) \times 0.1 + 0.1587 \times 0.9 \\ &= 0.2362. \end{aligned}$$

[use answers(i),(ii) and information that 10% if population affected.]

(v)P(disease|$x \geq 10$)=(0.9332×0.1)/(0.2362)=0.3951 .

3. Sample size n=10. In the first scheme, suppose r_1, r_2 are the numbers of defects classified 'major' or 'minor' respectively.

The batch is accepted only if (i)$r_1 = r_2 = 0$ or (ii)$r_1 = 0, r_2 = 1$ followed by $r_1 = r_2 = 0$ in the second sample.

The probability is $(1-p_1)^{10}(1-p_2)^{10} + (1-p_1)^{10} \cdot 10P_2(1-p_2)^9(1-p_1)^{10}(1-p_2)^{10}$

$= (1-p_1)^{10}(1-p_2)^{10}\{1+10p_2(1-p_2)^9(1-p_1)^{10}\}$.

and so the probability of rejection is

$1-(1-p_1)^{10}(1-p_2)^{10}\{1+10p_2(1-p_2)^9(1-p_1)^{10}\}$,

In the second scheme, accept only if $r_1 = 0, r_2 = 0$ or 1 .

The probability of this is $(1-p_1)^{20}(1-p_2)^{20} + (1-p_1)^{20} \cdot 20P_2(1-p_2)^{19}$ and the probability of rejection is therefore $1-(1-p_1)^{20}(1-p_2)^{19}(1+19p_2)$.

Probabilities are:

	$p_1 = 0.01$	0.01	$p_1 = 0.02$	0.02
	$p_2 = 0.02$	0.05	$p_2 = 0.02$	0.05
$Scheme1$	0.150	0.304	0.241	0.385
$Scheme2$	0.231	0.398	0.372	0.509

Scheme 2 gives higher probabilities of rejection for all these values of p_1, p_2.

4.(a)(i)P(≥1 replacement)=1-P(0)=1-P(T>1)=1-$\int_1^\infty \frac{1}{5}e^{-t/5}dt$

$=1-[-e^{-t/5}]_1^\infty = 1-e^{-1/5} = 0.1813.$

(ii)E[net profit]=profit on sale - replacement cost × P(replacement)

=£[100-70×0.1813] if there is only one replacement

=£87.31 .

(b)(i)

x	10	11	12	13	14	
$P(X=x)$	0.2	0.4	0.2	0.1	0.1	$=P(Y=y)$
y	+4000	+2000	0	−2000	−4000	

Distribution of Y given by second and third rows above.

P(loss)=P(Y<0)=0.2 .

(ii)E[X]=(10×0.2)+(11×0.4)+(12×0.2)+(13×0.1)+(14×0.1)=11.5 .

V[X] $= E[X^2] - (E[X])^2$.

$= (100 \times 0.2) + (121 \times 0.4) + (144 \times 0.2) + (169 \times 0.1) + (196 \times 0.1) - 11.5^2$

$= 133.7 - 132.25 = 1.45$

E[Y]=2000(12-E[X])=£1000 .

V[Y]=4×10^6 . V[X]=5.8×10^6 .

5.(i)E[X] $= \sum_{x=0}^{\infty} xe^{-\lambda}\lambda^x/x!$.

$= \sum_{x=0}^{\infty} e^{-\lambda}\lambda^x/(x-1)!$

$= \lambda \sum_{x=0}^{\infty} e^{-\lambda}\lambda^{x-1}/(x-1)!$ *(in which the term for $x=0$ is o)*

$= \lambda$

(ii)If P(X=k)=P(X=k+1), $\frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-\lambda}\lambda^{k+1}}{(k+1)!}$, *i.e.* $\frac{\lambda}{k+1} = 1$ *so that* $\lambda = k+1$.

(iii)Since the mode has maximum probability, it is unique as in (ii) if λ is an integer but otherwise satisfies $\frac{P(X=m)}{P(X=m-1)} > 1$ and $\frac{P(X=m+1)}{P(X=m)} < 1$, where m is the modal value.

If $\frac{e^{-\lambda}\lambda^m}{m!} \cdot \frac{(m-1)!}{e^{-\lambda}\lambda^{m-1}}$, then $\frac{\lambda}{m} > 1$; i.e. $m < \lambda$;

also if $\frac{e^{-\lambda}\lambda^{m+1}}{(m+1)!} \cdot \frac{m!}{e^{-\lambda}\lambda^m} < 1$, then $\frac{\lambda}{m+1} < 1$; i.e. $\lambda < m+1$ or $\lambda - 1 < m$;

hence $\lambda - 1 < m < \lambda$.

(iv)(a) $\lambda = 1$, so P(0)=$e^{-\lambda}$=$\frac{1}{e}$=0.3679 .

(b) P(0)+P(1)+P(2)=e^{-1}(1+1+$\frac{1}{2}$)=0.9197 .

(v)Number of faults in $20m^2$ will follow Poisson with mean 4.

(a)$P(\geq 3) = 1 - P(0) - P(1) - P(2)$.
$= 1 - e^{-4}(1 + 4 + \frac{4^2}{2!})$
$= 1 - 13e^{-4}$
$= 1 - 0.2381 = 0.7619$

(b)Number of rooms with ≥ 3 faults is Binomial(50,0.7619) which can be approximated as $N(50 \times 0.7619, 50 \times 0.7619 \times 0.2381)$ or $N(38.095, 9.0704)$. The probability of being >40 is the value corresponding to 40.5(with continuity correction) in this distribution:

$Z = \frac{40.5-38.095}{\sqrt{9.0704}} = \frac{2.405}{3.0117} = 0.7986$

$P(Z > 0.7986) = 0.2123$

[The answer without a continuity correction would be 0.2635.]

6.(a)If the residual(error) terms are i.i.d. $N(0, \sigma^2)$, then the least squares estimates are also maximum likelihood. $Y = y - y_0$, $X = k(x - x_0)$ transforms $\hat{Y} = \hat{A} + \hat{B}X$ into $\hat{y} - y_0 = \hat{A} + \hat{B}k(x - x_0)$ or $\hat{y} = (\hat{A} - \hat{B}kx_0) + y_0 + \hat{B}kx$, giving in the usual notation $\hat{a} = y_0 + \hat{A} - \hat{B}kx_0$ and $\hat{b} = \hat{B}k$.

Since the scale of Y is not changed, the estimate of σ^2 will not be changed: $s^2 = S^2$.

(b)$\sum t$=15, $\sum w$=588, n=6, $\sum w^2$=71360, $\sum t^2$=55, using t=(age-84)/7, w=weihgt-500. $\sum wt$=1960.

$w - \bar{w} = \hat{b}(t - \bar{t})$ where $\hat{b} = \frac{\sum(w-\bar{w})(t-\bar{t})}{\sum(t-\bar{t})^2}$.
$= \frac{1960-15\times588/6}{55-15^2/6}$
$= \frac{490}{17.5} = 28$

Hence w-98=28(t-2.5)=28t-70, or w=28t+28 .

This transforms back to (weight-500)=$\frac{28}{7}$(age-84)+28

or weight=500+4(age)-336+28 or weight=4(age)+192 .

The fitted values and residuals are:

Age	84	91	98	105	112	119
Weight	528	556	584	612	640	668
Residual	−1	−1	1	3	0	−2

sum of squares of residuals =16, hence residual mean square with 4 degrees of freedom is 16/4 =4.

The residuals go rather systematically up and then down again, which suggests the need for a curvilinear model, such as adding a $(time)^2$ term, or plotting log(weight) against log(age).

7.(a)The total sum of probabilities must be 1. Hence k=1/900.

(b)Summing in rows :

x	1	2	3	$TOTAL$
$P(x)$	7/90	26/90	57/90	: 1

and in columns :

y	1	2	3	4	$TOTAL$
$P(y)$	17/450	31/225	3/10	118/225	: 1

These are the marginal distributions of X and Y.

(c)E[X]=(7+52+171)/90=23/9 . (=2.56).

E[Y]=$\frac{1}{450}((1 \times 17) + (2 \times 62) + (3 \times 135) + (4 \times 236)) = \frac{149}{45}$. (=3.31)

E[X^2]=$(\frac{1}{90})(1 \times 7 + 4 \times 26 + 9 \times 57) = \frac{624}{90} = \frac{104}{15}$

V[X]=$E[X^2] - (E[X])^2 = \frac{104}{15} - \frac{23^2}{81} = \frac{8424-7935}{15\times81} = \frac{163}{405}$. (= 0.4025)

E[Y^2]=$\frac{1}{450}(1 \times 17 + 62 \times 4 + 135 \times 9 + 236 \times 16) = \frac{5256}{450}$

$V[Y] = E[Y^2] - (E[Y])^2 = \frac{5256}{450} - (\frac{149}{45})^2 = \frac{5256\times45-149^2\times4}{45\times450}$
$= \frac{14510}{45\times450} = \frac{1451}{2025}$. (=0.7165)

xy :	1	2	3	4	6	8	9	12
probability :	3/900	20/900	42/900	72/900	156/900	136/900	171/900	300/900

E[XY]= $\frac{1}{900}(3+40+126+288+936+1088+1539+3600) = \frac{7620}{900} = \frac{127}{15}$.

Cov[X,Y]=E[XY]-E[X]E[Y]=$\frac{127}{15} - \frac{23}{9} \cdot \frac{149}{45} = \frac{27\times127-23\times149}{9\times45} = \frac{2}{405}$.

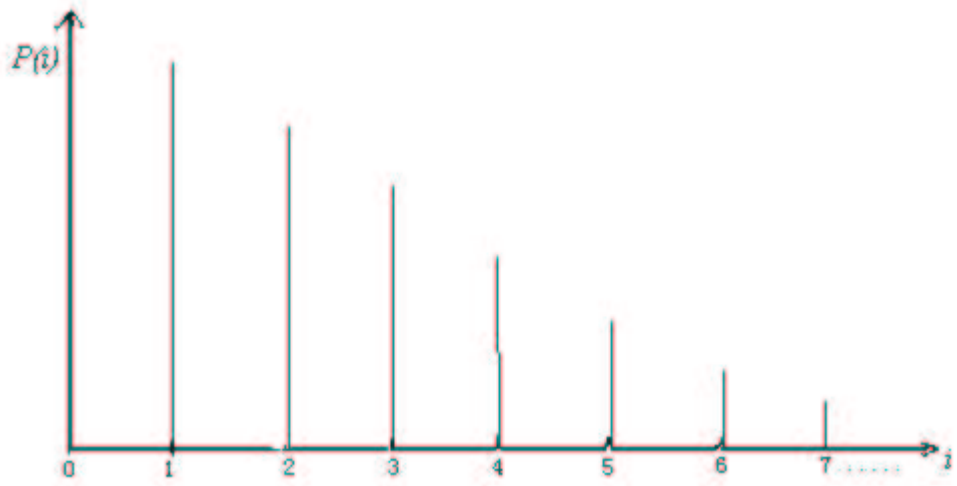
$P_{XY} = \frac{2/405}{\sqrt{\frac{163}{405}\times\frac{1451}{2025}}} = \frac{2\sqrt{5}}{\sqrt{163\times1451}} = 0.00920$.

(d)When Y=1, the conditional distribution of X is the first column of the table, scaled to sum to 1:

x	1	2	3
Probability	3/34	10/34	21/34

(e)E[X|Y=1]=$\frac{1}{34}(3\times1+10\times2+21\times3)$ =86/34=43/17=2.529 .

8.



E[X]=$\sum_{x=1}^{\infty} xP(X=x) = \sum_{x=1}^{\infty} xq^{x-1}p = p(1+2q+3q^2+\cdots)$

$= p/(1-q)^2 = p/p^2 = 1/p$.

F(x)=$P(X\le x) = \sum_{n=1}^{x} pq^{n-1} = p\frac{1-q^x}{1-q} = 1-q^x \quad (x=1,2,\cdots)$.

(Also F(x)=0 for x<1)

Strictly, F is a step function, changing value for each integer value of x and holding the value 1-$q^{[x]}$ until the next change. ([x] is the integral part of x.)

For the median M, F(M)=$\frac{1}{2}$.

Hence 1-$q^x = \frac{1}{2}$ or $q^x = \frac{1}{2}$ so that $x\ln q = -\ln 2$ or $x = -\ln 2/\ln q$. M is the smallest integer not less than this.

P(Y=X)=$\sum_{x=1}^{\infty} p^2q^{2(x-1)} = p^2\sum_{x=1}^{\infty} q^{2(x-1)} = p^2/(1-q^2) = \frac{p^2}{(1-q)(1+q)} = \frac{p}{1+q}$

1.(a) $Y_{ij} = \mu + \alpha_i + \beta + j + \epsilon_{ij}$

Y_{ij}: observation; μ: general mean; α_i: effect due to treatments; $\beta + j$: effect of being in block j

α_i and β_j are deviations from the general mean, due to which treatment has been given and which block the unit(plot) is in; these are independent of one another.

$\{\epsilon_{ij}\}$ are mutually independent random residual terms, representing natural variation between experimental units, each distributed normally with mean C and (constant) variance σ^2.

(b)Location totals: (1)31; (2)47; (3)42; (4)34. G=154. N=12.

Treatment totals: A,47; B,59; C,48. $\sum y^2$=2060.

Total ss=2060-$154^2/12$=83.667 .

Location ss=$\frac{1}{3}(31^2 + 47^2 + 42^2 + 34^2) - 154^2/12 = 53.667$.

Treatment ss=$\frac{1}{4}(47^2 + 59^2 + 48^2) - 154^2/12 = 22.167$.

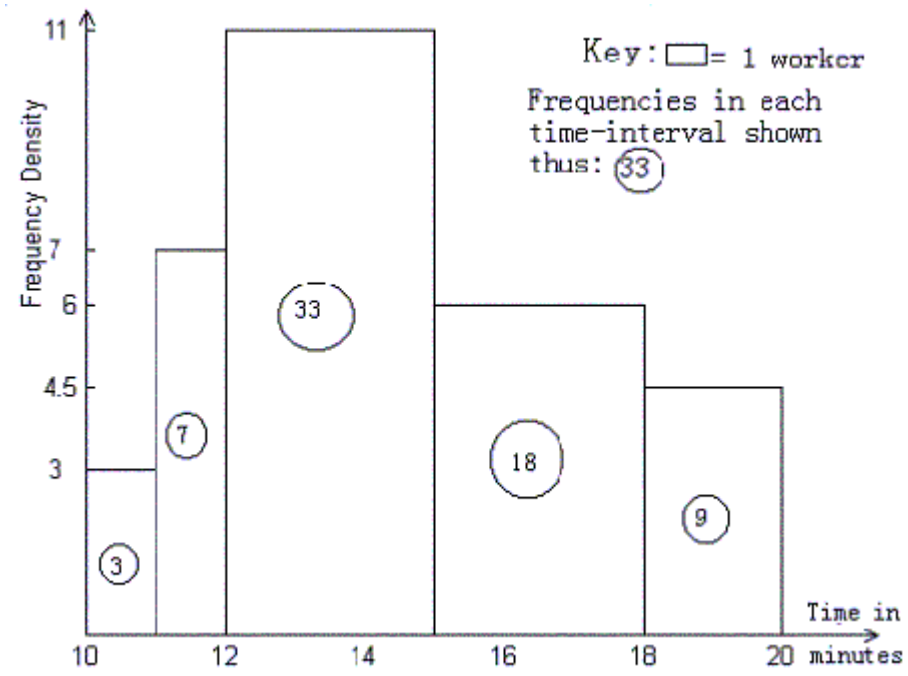
Analysis of Variance:

SOURCE	*D.F.*	*SUM OF SQUARES*	*M.S.*	
Treatments	2	22.167	11.084	$F(2,6) = 8.49^*$
Locations	3	53.667	17.889	$F(3,6) = 13.70^{**}$
Residuals	6	7.883	1.306	
TOTAL	11	83.667		

The locations differed significantly, and therefore it was useful to use the randomized block scheme with locations as blocks. We assume there is no blocks × treatments interaction.

For treatments, means are: *A* 11.75 . We are not told which comparisons(contrasts) among *C* 12.00
B 14.75
treatments are important, but it is clear that the significance of F(2,6) must be due to the difference between B and the other two.

2.(i)Histogram of time taken to complete a standard task.



Mid-point of *time interval* (*mins.*)(t)	*Frequency* (f)	ft	ft^2	*Cumulative frequency*
10.5	3	31.5	330.75	3
11.5	7	80.5	925.75	10
13.5	33	445.5	6014.25	43
16.5	18	297.0	4900.50	61
19.0	9	171.0	3249.00	70
	70	1025.5	15420.25	

Mean=$\frac{1025.5}{70} = 14.65$.

Median=$11.95 + \frac{25}{33} \times 3 = 14.22$.

(assuming records to nearest 0.05min).

Variance=$(15420.25 - 1025.5^2/70) \div 69 = 5.7489$. SD=2.40 .

With few intervals, the histogram alone is not very informative, but the mean and median are roughly the same, and near to the middle of the range of the data. Therefore we may treat the data as approximately normal, and certainly as sufficiently symmetrical to use large-sample tests.

(ii)An approximate 95% confidence interval is $14.65 \pm 1.96\sqrt{\frac{5.7489}{70}}$ i.e. 14.65 ± 0.56, which is (14.09 to 15.21).

3.(a)If we can assume that the lifetime distribution for the bulbs is normal with variance σ^2, and all observations are independent of one another, then $(n-1)s^2/\sigma^2$ will be distributed χ^2_{n-1}. Here n=10, and on H_0 we take $\sigma^2 = 150^2$. Then effectively we test $H_0 : \sigma^2 \leq 150^2$ against $H_1 : \sigma^2 > 150^2$.

For the data,$(n-1)s^2 = 9 \times 35410.99$. Hence $\chi^2_{(9)} = 14.16$, which is not significant at the 5% level. Therefore we do not have enough evidence to reject H_0 which says $\sigma \leq 150$.

(b)Since twelve randomly selected batches were used from each process we have independent estimates of variances σ_1^2, σ_2^2. The Null Hypothesis will be $\sigma_1^2 = \sigma_2^2$, and AH $\sigma_1^2 > \sigma_2^2$.

From the data, $s_1^2 = 0.012536$ and $s_2^2 = 0.003590$.

Assuming that the distributions of impurity levels are normal, s_1^2/s_2^2 is distributed as F(11,11). $\frac{s_1^2}{s_2^2} = 3.49^*$, significant at the 5% level so that H_0 is rejected in favor of H_1: there is evidence of a reduction in process variability.

4.(a)Because the measurements are taken on the same volunteers, the paired t-test is appropriate.

Differences (B-A) are: -5, -2, -8, 1, -3, 0, -6, 2, -1, -5, 0, -4.

Assuming that these are normally distributed, the N.H. that the mean difference is 0 uses $t_{(11)} = \frac{\bar{d}-0}{s/\sqrt{12}}$.

The observed mean difference is $\bar{d} = -\frac{31}{12} = -2.583$. $s^2 = (3.088)^2$.

Hence $t_{(11)} = -\frac{2.583}{3.088/3.464} = -2.898^*$.

Reject the N.H. There is evidence of a change in blood pressure.

The estimated mean increase is 2.583 units. A 95% confidence interval for this is $2.583 \pm 2.201 \times 3.088/3.464 = 2.583 \pm 1.962$ or (0.62 to 4.55) units.

(b)On the Null Hypothesis of no difference in improvement under the two treatments, expected numbers are calculated:

$OBS(EXP)$	$Improved$	$Not\ Improved$	$TOTAL$
A	45(54)	55(46)	: 100
B	63(54)	37(46)	: 100
	108	92	200

$\chi^2_{(1)} = \sum \frac{(O-E)^2}{E} = 9^2(\frac{2}{54} + \frac{2}{46}) = 162(\frac{1}{54} + \frac{1}{46}) = 6.52^*$.
($\chi^2_{(1)} = 5.80$ if Yates' correction is used: not essential).
We have evidence to reject the Null Hypothesis at the 5% significance level. This is an indication of treatment difference.

5.(a)If p is the probability of success at any attempt, and the rat does not 'learn' which routes are failures, so that each result is independent of others, then the geometric distribution explains the number of trials needed to gain one success.
(b)The value of p must be estimated from the data.
$\hat{p} = \frac{1}{\bar{x}}$, $\bar{x} = [(1 \times 56) + (2 \times 27) + (3 \times 13) + (4 \times 3) + (6 \times 1)]/100 = 1.67$.
Hence $\hat{\beta} = 0.5988$. Calculate P(1) etc. on geometric distribution.
P(1)=0.5988 . P(2)=0.5988×0.4012=0.2402 .
P(3)=0.5988 × $(0.4012)^2$=0.0964 . P(4)=0.0387 etc.

$x:$	1	2	3	≥ 4	$TOTAL$
$OBS:$	56	27	13	4	100
$EXP.\ ON\ GEOMETRIC:$	59.88	24.02	9.64	6.46	100

Combine "≥ 4" into one class to avoid very small expected frequencies. One parameter was estimated, so χ^2 has 2 d.f. for testing fit to the geometric.

$$\chi^2_{(2)} = \frac{(56-59.88)^2}{59.88} + \frac{(27-24.02)^2}{24.02} + \frac{(13-9.64)^2}{9.64} + \frac{(4-6.46)^2}{6.46} = 2.73 \quad n.s.$$

There is no evidence against the hypothesis of fit to a geometric distribution, nor therefore against the conditions stated in (a).

6.(a)When a large sample of data is available from any population, not necessarily normal, and including discrete data as well as continuous, the sample mean or total follows a distribution that is approximately normal. Therefore with large samples of data significance tests of, and confidence intervals for, a population mean may be found, at least approximately, without knowing what distribution the population has. This extends, for example to proportions in a binomial.

In practice, when distributions are reasonably symmetrical, even when not normal, samples may be as small as about 30, while if data are very skew then very large samples-several hundred-may be required to give acceptable results. An examination of data, possibly by graphical methods, is a useful guide when applying the approximation.

We may treat $\bar{x}$ as $N(\mu, \sigma^2/n)$ when n is sample size and μ, σ^2 are the (known or estimated) mean and variance of the population distribution. The only theoretical restriction is that μ and σ^2 must be finite.

Many estimates of practically important items are the sum of several independent components, e.g. crop yields of many plants forming a plot, and so their total tends to be normally distributed, $N(n\mu,\ n\sigma^2)$.
(b)(i)For difference in means, a 95% confidence interval is approximately
$(\bar{x_1} - \bar{x_2}) \pm 1.96\sqrt{s_1^2/n_1 + s_2^2/n_2}$.
For the given data, this is $(3.75 - 2.10) \pm 1.96\sqrt{\frac{2.74^2}{125} + \frac{1.40^2}{108}}$, i.e. $1.65 \pm 1.96 \times 0.2797$

giving 1.65 ± 0.55 or (1.10 to 2.20).

Since this interval does not contain zero, it is likely that there will be more calls each shift in district 1 than in 2, the mean difference being between 1.1 and 2.2(with probability 0.95).

(ii)For difference in proportions, $\frac{\hat{p_1}-\hat{p_2}}{\sqrt{\frac{p_1(1-p_1)}{n_1}+\frac{p_2(1-p_2)}{n_2}}} \sim N(0,1)$.

The NH is "$p_1 = p_2$".

Hence $\frac{26}{125} - \frac{15}{108} = 0.208 - 0.139 = 0.069$ is the estimated difference.

Its variance is $\frac{0.208\times0.792}{125} + \frac{0.139\times0.861}{108} = 2.426 \times 10^{-3}$; S.E.=0.049.

Hence $\frac{0.069}{0.049} = 1.40$ n.s. as N(0,1)and there is no significant evidence of a difference in proportions.

7.Given $\mu = 1.81,\ \sigma^2 = (0.025)^2$. n=10.

(i)For A, $\bar{x}$=1.80 and s^2=0.001977.

For B, $\bar{x}$=1.85 and s^2=0.000689.

$\frac{(n-1)s_A^2}{\sigma^2} = \frac{9\times0.001977}{0.025^2} = 28.47^{***} \sim \chi^2_{(9)}$, giving very strong evidence to reject an NH that A's variability is the same as the laboratory standard, and to accept an AH that it is greater.

$\frac{(n-1)s_B^2}{\sigma^2} = \frac{9\times0.000689}{0.025^2} = 9.92,\ \ n.s.\ as\ \chi^2_{(9)}$, so there is no statistical evidence that B's variability is unacceptable.

(ii)For A, $\frac{\bar{x}-1.81}{\sqrt{0.001977/10}} = \frac{-0.01}{0.014} = -0.71\ n.s.\ as\ t_{(9)}$.

No evidence that A's results are biased.

For B, $\frac{\bar{x}-1.81}{\sqrt{0.000689/10}} = \frac{0.04}{0.0083} = 4.82^{***}\ as\ t_{(9)}$.

B's results do seem to be biased, because this value of $t_{(9)}$ leads us to reject the N.H. "mean=1.81"

Hence worker A produces results which are unbiased but very variable, while B is biased but precise.

8.(a)Parametric methods require a distribution (often the normal) to be specified as a model for the observations. If this is not correct, inferences can be seriously affected. Non-parametric methods rely on such things as rank ordering of data, and require no distributional assumptions. They allow more general types of analysis, not based on means and variances(parameters) but are less powerful than parametric methods when there are available for the corresponding problem.

(b)These data are very skew, even after differences have been taken. Even if a logarithmic transformation is taken, the resulting data cannot be taken as anywhere near symmetrical.

(i)A sign test uses only +/-, and there are 3 +'s, 7 -'s in 10 pairs. If the populations are the same, H_0 says that the proportion of + signs is $\frac{1}{2}$, so the number of +'s is Binomial(10,1/2).

$P(\leq 3 +'s) = P(0) + P(1) + P(2) + P(3) = \frac{1}{2^{10}}(1 + 10 + 45 + 120) = 0.172$.

This is >0.5, so does not provide evidence of any difference.

(ii)Wilcoxon's signed ranks test uses magnitudes also.

H_0 is "opulations the same", H_1"different". Ranked absolute values

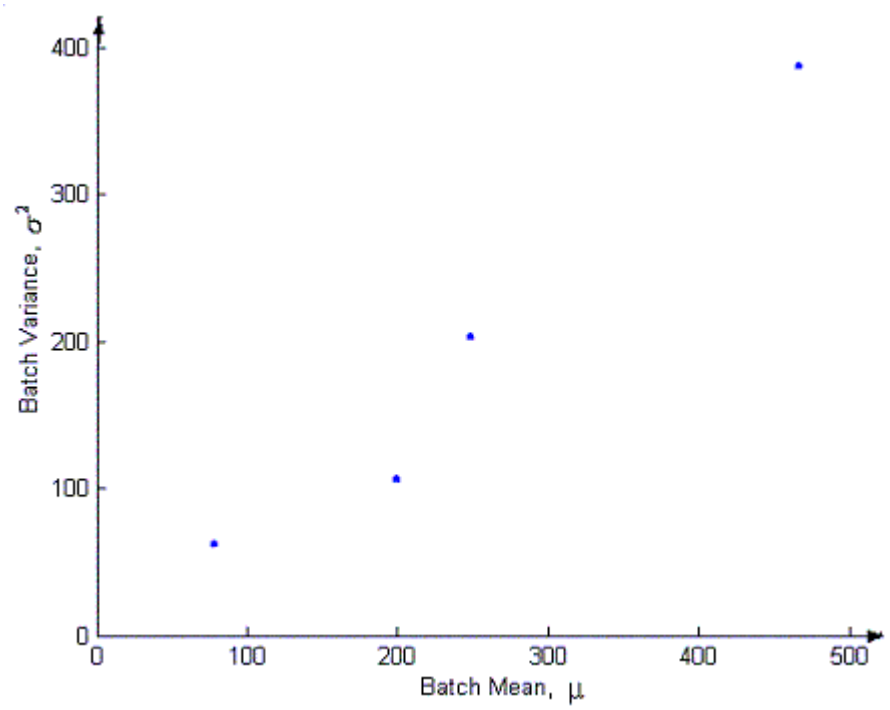
are	2	3	4	12	12	13	73	120	125	147
rank	1	2	3	$4\frac{1}{2}$	$4\frac{1}{2}$	6	7	8	9	10
sign	+	+	−	+	−	−	−	−	−	−

The sum of the + ve ranks is $7\frac{1}{2}$ and is below the(2-sided) critical table value 8 for n=10, and so the NH is rejected. Using this test, there is evidence of different location of population values.

By using the values, more information is available, since the + signs are attached to values that are numerically quite small. The Wilcoxon test thus gains greater power to discriminate between the two sets of data that produced the differences given.

PAPER III : Statistical Applications & Practice

1.(i)



The variance σ^2 appears to be proportional to the mean μ, whereas for Analysis of Variance we must assume that all observations have the same variance. The square root transformation will stabilize the variance when $\sigma^2 \propto \mu$, and has often been found suitable for counts.

(ii)After the square root transformation the variance is quite similar for all batches, and so can reasonably be pooled in an analysis of variance.

(iii)The total sum of squares of $\sqrt{y_{ij}}$ is the sum of all the y_{ij}, most easily calculated as 6 $\times$ the sum of those means,it is 5955.

The sum of all $\sqrt{y_{ij}}$ is 6$\times$ the sum of those means, 361.73 .

N=24, so $G^2/N = 361.73^2/24 = 5452.0247$.

Corrected total SS=5955-5452.0247=502.975 .

Transformed batch totals are A, 129.63; B,84.65; C,52.94; D,94.51 .

Batch SS= $\frac{1}{6}(129.63^2 + \cdots + 94.67^2) - G^2/N = 498.699$.

Analysis of variance:

SOURCE	*D.F.*	*SUM OF SQUARES*	*MEAN SQUARE*	
Batches	3	498.699	166.233	$F(3,20)^{***\cdots}$ Extremely highly significant.
Residual	20	4.276	0.2138	
TOTAL	23	502.975		

Clearly there are large differences between batches. This can be explored using least significant differences, since no further information is available to set up definite comparisons(contrasts) between batches for testing.

The l.s.d. between two mean$=t_{(20)}\sqrt{\frac{2\times 0.2138}{6}} = 0.267 \times \begin{cases} 2.086(5\%) \\ 2.845(1\%) \\ 3.850(0.1\%) \end{cases} = \begin{cases} 0.557(5\%) \\ 0.759(10\%) \\ 1.028(0.1\%) \end{cases}$

Batch means are:

C	*B*	*D*	*A*
8.8233	14.1083	15.7517	21.6050

.

Hence all batches differ very significantly from are another.

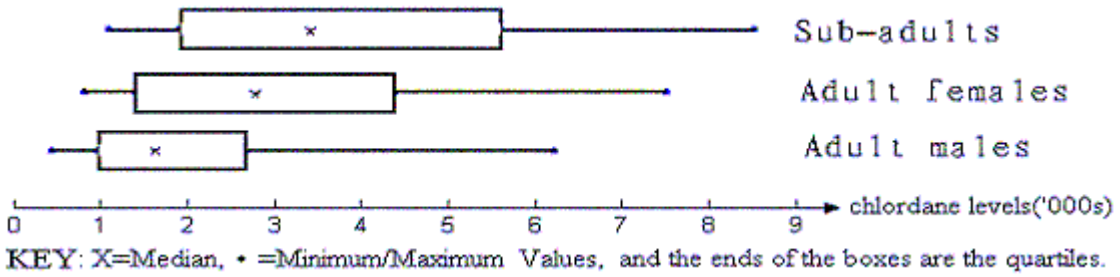
2.For sub-adults, $n_S = 15$; for females, $n_F = 23$; for males, $n_M = 20$. Medians and quartiles are: $M_S = 3329(13^{M}_{\cdot}\ item)$; $M_F = 2859(12^{M}_{\cdot}\ item)$; $M_M = 1693(averrage\ of\ 10^{M}_{\cdot}\ and\ 11^{M}_{\cdot}\ items)$.

Transformed: $M_S = 3.52$, $M_F = 3.46$, $M_M = 3.23$.

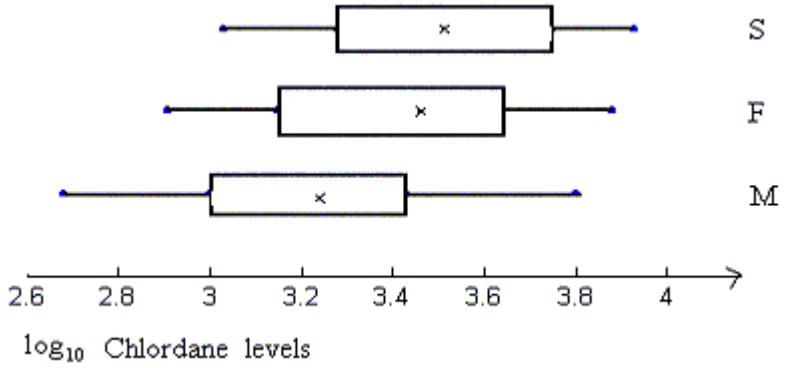
Lower quartiles: $q_S = \frac{1}{2}(1846 + 1960) = 1903$ *or* 3.28; $q_F = 1409$ *or* 3.15; $q_M = \frac{1}{2}(916 + 1089) = 1002.5$ *or* 3.00.

Upper quartiles: $Q_S = 5602$ *or* 3.75; $Q_F = 4397$ *or* 3.64; $Q_M = \frac{1}{2}(2520 + 2888) = 2704$ *or* 3.43.

(i) (ii)



These three distributions are all distinctly skew to the right.



The diagrams for $log_{10}(data)$ show much more symmetry, and much more constant variability. The basic assumptions for analysis of variance are therefore much more reasonable in these units.

(iii)Analysis of Variance

SOURCE	*DF*	*SUM OF SQUARES*	*MEAN SQUARE*	
Between groups	2	0.8930	0.4465	$F_{(2,65)} = 5.29^{**}$
Residual	65	5.4844	0.0844	
Total	67	6.3774		

There are substantial differences between the three groups of animals.

(iv)Using the log data, we may compare the rations of chlordane levels in the different groups. The 95% limits in log_{10} units are:

$$\begin{aligned}
&\textit{Sub adults}: \quad 3.51 \pm 2.00\sqrt{\tfrac{0.0844}{25}} \quad \textit{or} \quad 3.51 \pm 0.12, \quad \textit{i.e.} \quad 3.39 \textit{ to } 3.63\\
&\textit{Females}: \quad 3.42 \pm 2.00\sqrt{\tfrac{0.0844}{23}} \quad \textit{or} \quad 3.42 \pm 0.12, \quad \textit{i.e.} \quad 3.30 \textit{ to } 3.54\\
&\textit{Males}: \quad 3.23 \pm 2.00\sqrt{\tfrac{0.0844}{20}} \quad \textit{or} \quad 3.23 \pm 0.13, \quad \textit{i.e.} \quad 3.10 \textit{ to } 3.36
\end{aligned}$$

To answer the questions the investigators had in mind, the sub-adults could be compared with adults of each sex in significant tests. The confidence intervals in this particular experiment indicate what the results of these comparisons would be: sub-adults and adult females show no real difference(intervals overlap considerably) but adult males and sub-adults do show significant difference(no overlap of intervals).

[NOTE that males and females differ at the 5% significance level; but it not clear that we need to make this comparison, and the confidence intervals alone do not tell us this.]

(v)Anti-logs to base 10 give the intervals as follows:

Males 1260 to 2290; Females 2000 to 3470;

Sub-adults 2450 to 4270 (to nearest 10 ng/g).

3.(a)(i)In a sample survey, people are not compelled to respond and will only do so if the topic of the enquiry interests them, if they think it is important, and if there is nothing in the approach or the questionnaire that annoys them or puts them off. Questions of a private or sensitive native will not be answered. In a postal questionnaire, or if they come early in an interview, this will usually result in total non-response; at best be some missing data.

There is always some non-response in a postal questionnaire survey because people simply do not complete and mail it back. Interviews of a selected random sample of people cannot always be carried out 100% because some individuals refuse or are not available for interview.

(ii)Since non-response is often concentrated in certain parts of the target/study population, it is important to minimize it to avoid serious bias.

Care over wording of questions, fore-testing of them, reminders is a mail survey, repeated visits for interviews, keeping a questionnaire or interview as short as possible, even offering a reward or prize to those who do reply, can sometimes help reduce non-response.

If people cannot be interviewed because they have moved, then can be replaced by 'reserve' randomly selected people. But unless there is a good reason replacements should not be made as they can lead to bias.

(b)(i)$\hat{p_E} = \frac{67}{125} = 0.536$. $\hat{p_B} = \frac{126}{200} = 0.630$. For comparing these, with the null hypothesis "ture π_E=true π_B", a 2×2 table is :

OBSERVED	*Improve*	*Not*		*EXPECTED*	
Eng.	67	58	: 125	74.23	50.77
Bankg.	126	74	: 200	118.77	81.23
	193	132	325		

$\chi^2_{(1)} = \sum_{all\ cells\ of\ table} \frac{(O-E)^2}{E}$ $= (7.23)^2(\frac{1}{74.23} + \frac{1}{50.77} + \frac{1}{118.77} + \frac{1}{81.23})$
$= 52.2729 \times 0.0539$
$= 2.82\ n.s.$

This gives no evidence of difference between the population values of the proportion.

(ii)$\hat{p_B} - \hat{p_E} = 0.094$. $\frac{p_E(1-p_E)}{n_E} = \frac{0.536 \times 0.464}{125} = 0.0019896$

$\frac{p_B(1-p_B)}{n_B} = \frac{0.63 \times 0.37}{200} = 0.0011655$. Variance of difference=0.0031551.

$0.094 \pm 2\sqrt{0.0031551} = 0.094 \pm 0.112$ or -0.018 to 0.206.

With probability 0.95, $\pi_B - \pi_E$ lies between -0.018 and +0.206.

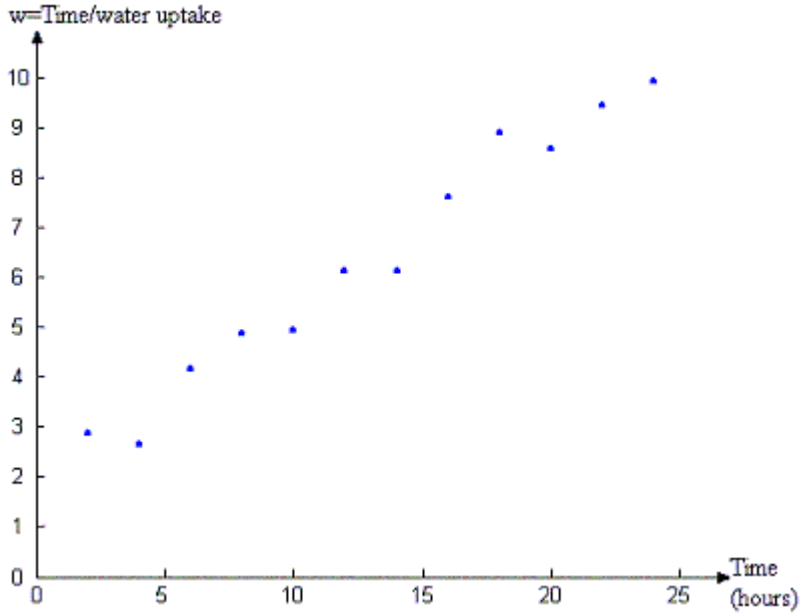
(iii)When $\pi = 0.6$, $2\sqrt{\frac{0.6 \times 0.4}{n}}$ is required to be 0.04.

Hence $4 \times \frac{0.24}{n} = (0.04)^2$ *or* $n = \frac{4 \times 0.24}{(0.04)^2} = 600$.

4.(i)$y = \frac{ct}{d+t}$ *or* $yd + yt = ct$ *or* $d + t = c\omega$.

This can be written as $\omega = \alpha + \beta t$, *where* $\alpha = \frac{d}{c}$; $\beta = \frac{1}{c}$.

(ii)



(iii)The fitted line is $\omega - \overline{\omega} = \hat{\beta}(t - \bar{t})$ where

$\hat{\beta} = \frac{\sum(\omega-\bar{\omega})(t-\bar{t})}{\sum(t-\bar{t})^2} = \frac{\sum \omega t - \sum \omega \sum t/n}{\sum t^2 - (\sum t)^2/n}$.

$= \frac{1190.46022 - 156 \times 76.39073/12}{2600 - 156^2/12}$

$= \frac{197.38073}{572} = 0.34507$

Hence $\hat{\alpha} = \frac{76.39073}{12} - \hat{\beta} \times 13 = 1.87997$.

(iv)$\hat{c} = 1/\hat{\beta} = 0.898 \quad and \quad \hat{d} = \hat{c}\ \hat{\alpha} = 5.448$.

(v)Using $y = \frac{2.898t}{5.448+t}$, when t=16, gives y=2.16.

(vi)The parameters $\hat{c}$ and $\hat{d}$ are non-linear functions of $\hat{\alpha}$ and $\hat{\beta}$ so there are no simple formula for the relations between standard errors.

5.(a)Writing p for price, q for volume, o for January 1995 and l for January 1997, the Paosehe Price index for 1997 is

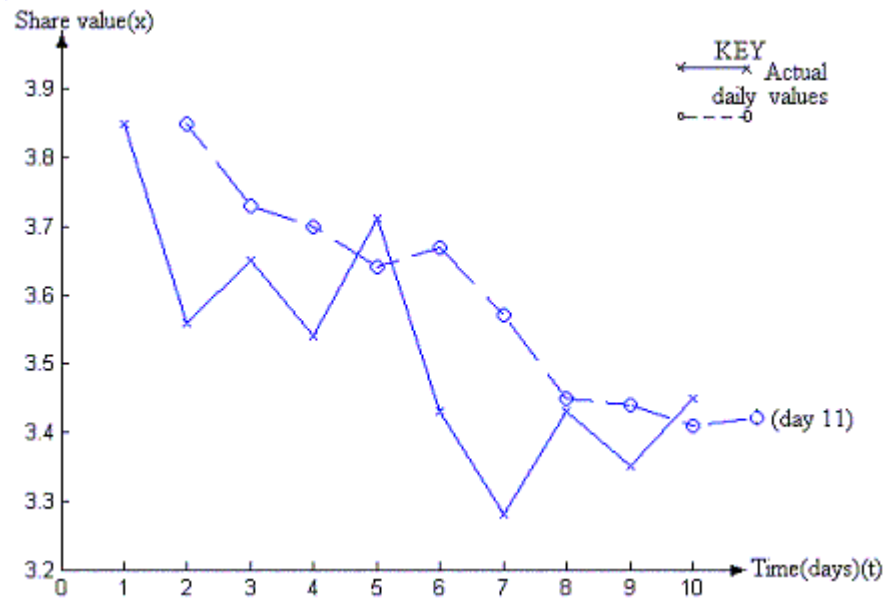
(i)$\frac{\sum p_1 q_1}{\sum p_0 q_1} = \frac{2.45\times167.3+6.43\times777.3+2.66\times165.2+15.15\times3193.7}{2.98\times167.3+4.40\times777.3+3.85\times165.2+11.41\times3193.7} = \frac{54231.911}{40994.811} = 1.3229$.

The index is thus 132.29.

(ii)$\frac{15.15}{11.41}$=1.3278, i.e. 132.78%.

The weight(q) for D is so large that the price change in D dominates the index.

(b)(i) (ii)



Share values and predictions by exponential smoothing

(ii)Exponential smoothing relates forecasts F at successive times t and actual figures as:

$$F_t = \alpha x_{t-1} + (1-\alpha)F_{t-1}$$

.

Calculations for $\alpha = 0.4$ are on the next page.

(iii)If $\alpha = 0.9$, predictions will track actual prices(x) much more closely.
or $\alpha = 0.4$, $F_t = 0.4 \times x_{t-1} + 0.6F_{t-1}$, for t=2,$\cdots$,11.

Day	1	2	3	4	5	6	7	8	9	10	11
x	3.85	3.56	3.65	3.54	3.71	3.43	3.28	3.43	3.35	3.45	–
F	(3.85)	3.85	3.73	3.70	3.64	3.67	3.57	3.45	3.44	3.41	3.42

6.(i)Given that $\mu = 0.52$ *and* $\sigma = 23.43$, the cumulative frequencies in a normal distribution are given by 520 $\Phi(Z)$;

for -67.5, $\Phi(\frac{-67.5-0.52}{23.43}) = \Phi(-2.903) = 0.00185$; $CF = 0.962$;

for -52.5, $\Phi(\frac{-52.5-0.52}{23.43}) = \Phi(-2.263) = 0.01182$; $CF = 6.146$.

Because of the very small expected frequency in the first group we shall combine it with the second, to give Obs.=7, Exp.=6.146. Continuing down the table, the cumulative frequency to -7.5 is 6.146+21.063+57.512+105.632=190.353.

For +7.5, $\Phi(\frac{7.5-0.52}{23.43}) = \Phi(0.298) = 0.61715$; $CF = 320.918$.

Now 320.918-190.353=130.565, which is the frequency in (-7.5,+7.5).

For +37.5, $\Phi(\frac{37.5-0.52}{23.43}) = \Phi(1.578) = 0.94272$. Hence the frequency in (22.5,37.5)=490.214-320.918-108.573=60.723 .

Now check that 490.214+(22.873+5.789+1.106)=519.982$\doteq$n within acceptable rounding error.

For testing normality the last two groups are combined: O=3, E=6.895.

There are 9 groups, two parameters were estimated, so χ^2 has 6 d f.

$$\chi^2_{(6)} = \frac{(7-6.146)^2}{6.146} + \frac{(26-21.063)^2}{21.063} + \frac{(54-57.512)^2}{57.512} + \frac{(90-105.632)^2}{105.632} + \frac{(147-130.565)^2}{130.565} + \frac{(102-108.573)^2}{108.573} + \frac{(66-60.723)^2}{60.723} + \frac{(25-22.873)^2}{22.873} + \frac{(3-6.895)^2}{6.895} = 9.20 \quad n.s.$$

This provides no evidence against the fit to a normal distribution.

(ii)Unless there are enough observations to combine into several groups, the χ^2 is a very poor approximation, and also the pattern in the data is difficult to detect. Power against an alternative of non-normality is very low.

(iii)There will be 30 residuals, which should be i.i.d. $N(0, \sigma^2)$. If it is only normality that is tested, and not only of the other assumptions made in analysis of a randomized block, a normal probability plot is suitable. The residuals are ranked in order, from largest negative to largest positive. Normal probability paper allows these to be plotted against the order-statistics for a normal distribution with a sample of 30 items; the i^{th} observed value is plotted against $\Phi^{-1}(i/31)$. This should give roughly a straight line. Further information comes from identifying which blocks and treatments give the largest residuals, positive or negative. If, for example, one treatment seems to have mostly large residuals it may be indicating that variances differ from one treatment to another.

7.(i)$P(T \geq t_0) = \int_{t_0}^{\infty} \lambda e^{-\lambda t} dt = [-e^{-\lambda t}]_{t_0}^{\infty} = e^{-\lambda t_0}$,

(ii)For 12 observations, $L = \prod_{i=1}^{12} \lambda e^{-\lambda t_i}$, so the log likelihood is $(\ln L) = n \ln \lambda - \lambda \sum_{i=1}^{12} t_i = 12 \ln \lambda - 6028\lambda$.

$\frac{d}{dt}(\ln L) = \frac{12}{\lambda} - 6028 = 0$ when $\hat{\lambda} = \frac{12}{6028} = 0.001991$.

(iii) $\frac{d^2}{d\lambda^2}(\ln L) = -\frac{12}{\lambda^2}$, and $Var(\hat{\lambda}) \approx -1/(\frac{-12}{\lambda^2}) = \lambda^2/12$.

(iv)The likelihood function is now

$$\begin{aligned} \mathrm{L} &= \lambda^{12} e^{-6028\lambda} \cdot e^{-\lambda(641+234+87)} \qquad (using\ (i)) \ . \\ &= \lambda^{12} e^{-6990\lambda} \end{aligned}$$

The same analysis now gives $\hat{\lambda} = \frac{12}{6990} = 0.001717$.

8.(i)The treatment combinations used in a factorial design are made up of all possible combinations of levels, or amounts, of several factors that may affect the response being measured. For example,

an industrial process may depend on the time for which it runs, T, and the temperature at which it is operated, U. If several values of T and of U are used, say T_1, T_2, T_3, T_4 and U_1, U_2, U_3 a factorial design requires all twelve combinations T_1U_1 to T_4U_3 to be used, usually in two or more complete replicates.

The response to one factor may take a different pattern at different levels of the other: e.g. at U_1 there may be a linear change from T_1 to T_4 whereas it is quadratic at U_2, and irregular at U_3.

This is interaction between T and U and is not discovered unless a factorial design is used.

Totals of responses are required for analysis, i.e. means$\times$ 5.

	Goats	*Red Deer*	*Camelids*
Sown grasses	4.535	10.165	5.685
Natural grasses	6.740	12.185	9.865
Heathers	3.230	9.440	4.850

Grand total G=66.695.[MISPRINT ON PAPER]

(ii)Correction term $G^2/N = 66.695^2/45 = 98.8494$.

Total S.S.=114.85-G^2/N=16.0006 .

Total for Animals are 14.505, 31.790, 20.400; for plants 20.385, 28.790, 17.520.

S.S. Animals= $\frac{1}{15}(14.505^2 + 31.790^2 + 20.400^2) - G^2/N = 10.2945$.

S.S. Plants= $\frac{1}{15}(20.385^2 + 28.790^2 + 17.520^2) - G^2/N = 4.5748$.

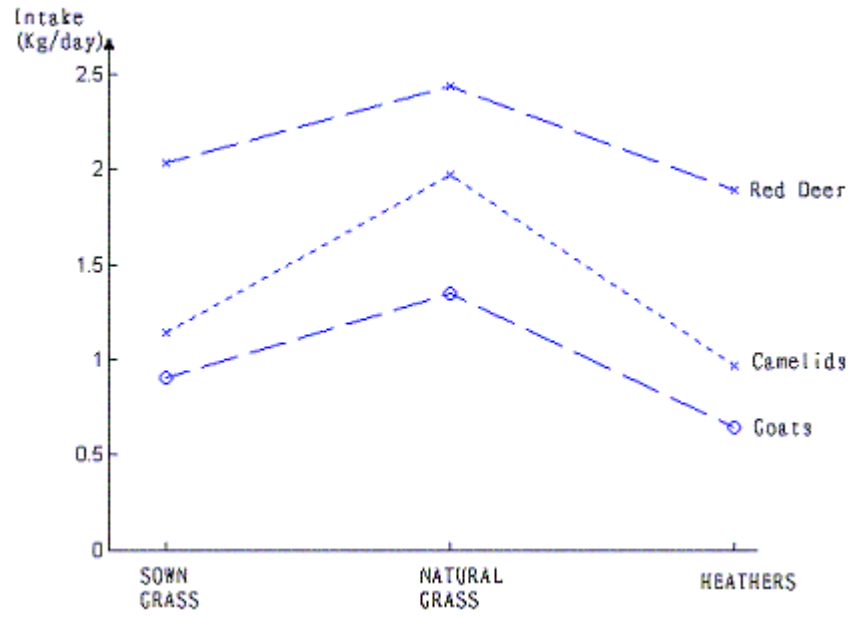
S.S. Animals+S.S. plants +S.S. Interaction= $\frac{1}{5}(4.535^2 + \cdots + 4.850^2) - \frac{G^2}{N} = 15.2509$.

Hence the Analysis of Variance:

SOURCE OF VARIATION	*D.F.*	*SUM OF SQUARES*	*MEAN SQUARE*	*VARIANCE RATIO*
Animal Types	2	10.2945	5.1473	
Plant Types	2	4.5748	2.2874	
Interaction	4	0.3816	0.0954	$F(4,36) = 4.58^{**}$
Residual	36	0.7497	0.02083	
TOTAL	44	16.0006		

There is strong evidence of interaction between Animals and Plants. Graphs of the means for the nine combinations are required.

In the presence of interaction, the main effects of the factors have little meaning. We may note that levels for Red Deer are consistently above those for Camelids, which are also above those



Camelids show a different pattern from Goats and Red Deer, in that their intake from natural grasses is relatively higher than from sown grass or heathers; the figure 1.973 is the odd one out in the table of means.