The Royal Statistical Society

GRADUATE DIPLOMA IN STATISTICS

May 1998

PaperI : Statistics Theory & Methods

1.In a sample space S, suppose that for the events E_1, E_2,...$P(E_i) > 0$ for all i ;$P(E_i \cap E_j) = 0$ for all i,j,$i \neq j$; $E_1 \cup E_2 \cup \cdots = S$. Let $A \subseteq S$ be any event such that$P(A) > 0$. Then

$$P(E_j|A) = \frac{p(A|E_j)p(E_j)}{\sum\limits_i P(A|E_i)P(E_i)}$$

(i) IF event A is 'policyholder does not make a claim in a year' and E_1, E_2, E_3 are 'policyholder is good, average ,bad risk' respectively then

$$p(E_1|A) = \frac{0.95 \times 0.2}{(0.95 \times 0.2) + (0.85 \times 0.5) + (0.7 \times 0.3)} = \frac{0.19}{0.825} = 0.230$$

(ii)The corresponding probabilities for $E_2, E_3$are $\frac{0.425}{0.825} = 0.515$ and $\frac{0.21}{0.825} = 0.255$. Hence the joint distribution of $x_1,\ x_2,\ x_3$, the number of policyholders of each type among non-claimants,is multinomial with these three probability as P_1, P_2, P_3 In a sample of 4,we require

$$p(x_1 = 2, x_2 = 1, x_3 = 1) + p(2, 2, 0) + p(3, 1, 0)$$

$$= \tfrac{4!}{2!}(0.230)^2(0.515)(0.255) + \tfrac{4!}{2!2!}(0.230)^2(0.515)^2 + \tfrac{4!}{3!}(0.230)^3(0.515)$$

$$= 0.08337 + 0.08418 + 0.02506 = 0.193.$$

(iii) Assume that any individual driver is equally likely to make a claim in any year,and that drivers do or do not make claims independently in different years.
Use Bayes's Theorem with event B"policyholder does not make a claim in 5 years" Then

$$p(E_1|B) = \tfrac{0.95^5 \times 0.2}{(0.95^5 \times 0.2) + (0.85^5 \times 0.5) + (0.7)^5 \times 0.3)}$$

$$= \tfrac{0.15476}{0.15476 + 0.22185 + 00502}$$

$$= \tfrac{0.15476}{0.42703} = 0.362$$

2(a)If Y = No. of arrivals and X = No. turning left, then Y is poisson with mean μ and $X|Y = y$ is Binomial (y,θ),where $0 \le X \le y$

$$\begin{aligned} p(X = x) &= \sum_{y=x}^{\infty} p(X = x|Y = y)p(Y = y) \\ &= \sum_{y=x}^{\infty} \tfrac{y!}{x!(y-x)!}\theta^x(1-\theta)^{y-x}\tfrac{e^{-\mu}\mu^y}{y!} \\ &= \tfrac{\theta^x e^{-\mu}\mu^x}{x!}\sum_{y=x}^{\infty}\tfrac{[(1-\theta)\mu]^{y-x}}{(y-x)!} = \tfrac{(\theta\mu)^x}{x!}e^{-\mu}e^{(1-\theta)\mu} \\ &= \tfrac{(\theta\mu)^x e^{-\theta\mu}}{x!} \qquad \textit{i.e. is poisson mean } \theta\mu \end{aligned}$$

(b)Let z=x+y Then

$$\begin{aligned} p(Z = z) &= \sum_{x=0}^{z} p(X = x)p(Y = z-x) \qquad \textit{since } x,\ y \textit{ independent} \\ &= \sum_{x=0}^{z}\tfrac{e^{-\mu}\mu^x}{x!}\tfrac{e^{-v}v^{z-x}}{(z-x)!} = \tfrac{e^{-(\mu+v)}}{z!}\sum_{x=0}^{z}\tfrac{z!}{x!(z-x)!}\mu^x v^{z-x} \\ &= \tfrac{e^{-(\mu+v)}}{z!}(\mu+v)^z \qquad \textit{i.e. by the biomial theorem} \end{aligned}$$

so that Z is poisson with mean $(\mu + v)$.

3(a)

$$E[u] = \int_0^1 \frac{(m+n+1)!}{(m-1)!(n-1)!}u^m(1-u)^{n-1}du$$

the B(m+1,n) function multiplied by $\frac{(m+n-1)!}{(m-1)(n-1)!}$
Hence

$$E[u] = \frac{(m+n-1)!}{(m-1)!(n-1)!} \times \frac{m!(n-1)!}{(m+n)!} = \frac{m}{m+n}$$

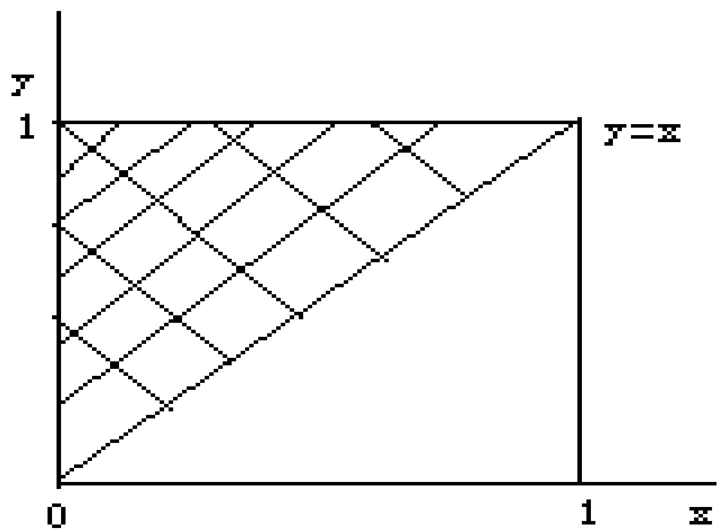
$$E[\mu^2] = \textit{ the same factor } \times B(m+2,n) = \frac{m(m+1)}{(m+n)(m+n+1)}$$

hence

$$v[u] = E[u^2] - (E[u])^2 = f\tfrac{m(m+1)}{(m+n)(m+n+1)} - \tfrac{m^2}{(m+n)^2}$$

$$= \tfrac{m(m+1)(m+n)-m^2(m+n+1)}{(m+n)^2(m+n+1)}$$

$$= \tfrac{mn}{(m+n)^2(m+n+1)}$$



(b)The region of existence for the density is
hence

$$f(x) = \int_{y=x}^{1} 6xdy = [6xy]_{y=x}^{1} = 6x(1-x)\ 0 < x < 1$$

Thus x follows Bera(2,2) $E[x] = \frac{2}{2+2} = \frac{1}{2} \quad v[x] = \frac{2\times 2}{(2+2)^2(2+2+1)} = \frac{1}{20}$

Also $f(y) = \int_{x=0}^{y} 6xdx = [3x^2]_{x=0}^{y} = 3y^2 \quad for\ 0 < y < 1$

Therefore y is Bera(3,1) and $E[y] = \frac{3}{4} \quad v[y] = \frac{3}{16\times 5} = \frac{3}{80}$

$$E[xy] = \int_{y=0}^{1}\int_{x=0}^{y} 6x^2dxdy = \int_0^1 [2x^3y]_{x=0}^{y} = \int_0^1 2y^4dy = [\frac{2y^5}{5}]_0^1 = \frac{2}{5}$$

$$Cov(x,y) = E[xy] - E[x]E[y] = \tfrac{2}{5} - \tfrac{1}{2} \times \tfrac{3}{4} = \tfrac{1}{40}$$

$$\rho_{xy} = \tfrac{cov(x,y)}{\sqrt{v[x]v[y]}} = \tfrac{1}{40} \times \tfrac{1}{\sqrt{\frac{3}{80}\times\frac{1}{20}}} = \tfrac{1}{\sqrt{3}}$$

4: By independence,the joint probability density

$$f(x,y) = \frac{\frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}x^2) y^{\frac{1}{2}k-1} exp(-\frac{1}{2}y)}{2^{\frac{1}{2}k}\Gamma(\frac{1}{2}k)} \qquad -\infty < x < \infty \quad y > 0$$

The given transformation is $x = uv, y = kv^2$

The Jacobian of the transformation is

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial y}{\partial u} \\ \frac{\partial x}{\partial v} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & 0 \\ u & 2kv \end{vmatrix} = 2kv^2$$

Expressing in terms of u,v and multiplying by the Jacobian, the joint probability density is

$$\begin{aligned} f(u,v) &= 2kv^2 \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2 v2} \times \frac{k^{\frac{1}{2}k-1} v^{k-2} e^{-\frac{1}{2}kv^2}}{2^{\frac{1}{2}k}\Gamma(\frac{1}{2}k)} \\ &= \frac{2^{\frac{1}{2}(1-k)} k^{\frac{1}{2}k}}{\sqrt{\pi}\Gamma(\frac{1}{2}k)} v^k e^{-\frac{1}{2}v^2(u^2+k)} \qquad v > 0 \\ f(u) &= \frac{2^{\frac{1}{2}(1-k)} k^{\frac{1}{2}k}}{\sqrt{\pi}\Gamma(\frac{1}{2}k)} \int_0^\infty v^k e^{-\frac{1}{2}v^2(u^2+k)} dv \\ &= \frac{2^{\frac{1}{2}(1-k)} k^{\frac{1}{2}k}}{\sqrt{\pi}\Gamma(\frac{1}{2}k)} 2^{\frac{1}{2}(k-1)} (u^2+k)^{-\frac{1}{2}(1+k)} \Gamma(\frac{k+1}{2}) \\ &= \frac{1}{\sqrt{k}(1+\frac{u^2}{k})^{\frac{1}{2}(k+1)} B(\frac{k}{2},\frac{1}{2})} \end{aligned}$$

using the gamma integral $\Gamma(\frac{1}{2}) = \sqrt{(\pi)}$and the relation between beta and gamma functions.
Thus U is the t-distribution with K degrees of freedom.
The given transformation leads to U, the ratio of a N(0,1) and the squareroot of a $\chi^2_{(K)}$ divided by its d.f., independent of N(0,1). This is the situation when a sample mean from a normal population is divided by an estimate of the standard error of the mean, leading to $t_{(n-1)}$ which is a pivotal quantity in inference.
n=sample size
NOTE. Credit is of course given for reference to interval estimation and /or significance testing.

5.

$$\begin{aligned} M_z &= E[e^{zt}] = \int_{-\infty}^{\infty} e^{zt}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}dz \\ &= \int_{-\infty}^{\infty} e^{\frac{1}{2}t^2}e^{-\frac{1}{2}(z-t)^2}dz = e^{\frac{1}{2}t^2} \end{aligned}$$

since the integral in z is that of a p.d.f. over its whole range and is therefore 1

$$M_x(t) = E[e^{xt}] = \sum_{x=0}^{\infty} e^{xt}\frac{e^{-\mu}\mu^x}{x!} = e^{-\mu}\sum_{x=0}^{\infty}\frac{(\mu e^t)^x}{x!} = e^{-\mu}e^{\mu e^t} = e^{(e^t-1)\mu}$$

Differentiating and setting t=0 gives moments

$$M_x'(t) = \mu e^t e^{(e^t-1)\mu}; \qquad E[x] = M'(0) = \mu$$

$$M_x^{''}(t) = (\mu e^t)^2 e^{(e^t-1)\mu} + \mu e^t e^{(e^t-1)\mu}$$

and

$$M_x^{''}(0) = \mu^2 + \mu = E[x^2]$$

$$v[x] = E[x^2] - (E[x])^2 = \mu^2 + \mu - (\mu)^2 = \mu \quad w = \frac{x}{\sqrt{\mu}} - \sqrt{\mu}$$

and by the usual properties of mgf's

$$M_w(t) = e^{-t\sqrt{\mu}}M_x(\frac{t}{\sqrt{\mu}}) = e^{-t\sqrt{\mu}}e^{\mu(e^{\frac{t}{\sqrt{\mu}}}-1)}$$

Thus

$$\begin{aligned} ln(M_w) &= -t\sqrt{\mu} - \mu + \mu e^{\frac{t}{\sqrt{\mu}}} \\ &= -t\sqrt{\mu} - \mu + \mu(1 + \frac{t}{\sqrt{\mu}} + \frac{t^2}{2\mu} + \frac{t^3}{6\mu^{\frac{3}{2}}}) \\ &= \frac{t^2}{2} + \frac{t^3}{6\sqrt{\mu}} + \cdots \rightarrow \frac{1}{2}t^2 \ as \ \mu \rightarrow \infty \end{aligned}$$

Hence in the limit w has the same mgf as N(0,1),and is therefore distribution as N(0,1).

6

$$f(x_i) = \theta e^{-\theta x_i}, \qquad x > 0 \ \ \theta > 0 \ \ Also \ F(x_i) = 1 - e^{-\theta x_i}$$

$$\begin{aligned} F(u_1, u_n) &= P(U_1 \leq u_1 \cap U_n \leq u_n) \\ &= P(U_n \leq u_n) - P(U_1 \geq u_1 \cap U_n \leq u_n) \\ &= P(all\ x_i\ in\ (0, u_n)) - p(all\ x_i\ in\ (u_1, u_n)) \\ &= \{1 - e^{-\theta u_n}\}^n - \{e^{-\theta u_1} - e^{-\theta u_n}\}^n \qquad 0 < u_1 \leq u_n \end{aligned}$$

The joint pdf is found as $\frac{\partial^2 F}{\partial u_1 \partial u_n}$:

$$f(u_1, u_n) = n(n-1)\theta^2 e^{-u_1} e^{-u_n} \{e^{-\theta u_1} - e^{-\theta u_n}\}^{n-2}, \quad 0 < u_1 \leq u_n$$

The alternative derivation using a multinomial distribution is also acceptable.
Transform to $R = U_n - U_1$ *and* $T = U_1$ (so $U_1 = T, \quad U_n = R + T$):

$$J = \begin{vmatrix} \frac{\partial U_1}{\partial R} & \frac{\partial U_n}{\partial R} \\ \frac{\partial U_1}{\partial T} & \frac{\partial U_n}{\partial T} \end{vmatrix}$$

Which is

$$\begin{vmatrix} 0 & 1 \\ 1 & 1 \end{vmatrix} = 1$$

Thus

$$\begin{aligned} f(r,t) &= n(n-1)\theta^2 e^{-\theta t} e^{-\theta(r+t)} (e^{-\theta t} - e^{-\theta(r+t)})^{n-2} \quad (r, t > 0) \\ &= n(n-1)\theta^2 e^{-n\theta t} e^{-\theta r} (1 - e^{-\theta r})^{n-2} \quad (r, t > 0) \end{aligned}$$

For$f(r)$, we must integrate out the factor $e^{-n\theta t}$ from 0 to ∞, since the rest of the expression does not involve t: $\int_0^\infty e^{-n\theta t} dt = \frac{1}{n\theta}$ and so

$$f(r) = (n-1)\theta e^{-\theta r} (1 - e^{-\theta r})^{n-2} \quad (r > 0)$$

If

$$v = e^{-\theta R}, \quad then \quad R = -\frac{1}{\theta} \log_e v \quad so \quad \frac{dR}{dv} = -\frac{1}{\theta v} \quad and$$

$$f(v) = \frac{1}{\theta v}(n-1)\theta v(1-v)^{n-2} = (n-1)(1-v)^{n-2} \qquad 0 < v < 1$$

so that v is Beta(1,n-1).

7(a) The c.d.f of v is $F_V(v) = p(V \leq v) = p(H^{-1}(u) \leq v) = p(u < H(v))$,which is equal to H(v) because u is uniform (0,1) and so F(u)=u.Hence v has the same distribution as x .

(b) Start at a randomly chose point in the table (and if desired ,read in any direction, not only left-right). Obtain a sequence of 9 digits, e.g. 821 469 344, and takes as the three pseudo-random U(0,1) variate $u_1 = 0.821, \ u_2 = 0.469 \ u_3 = 0.344$

(i)For the $Binomial(4,\frac{1}{4})$,$p(X=x)$ and $p(X\leq x)$ are :

x	0	1	2	3	4
$p(X=x)$	0.3164	0.4219	0.2109	0.0469	0.0039
$p(X\leq x)$	0.3164	0.7383	0.9492	0.9961	1

Since u_1 is between the values $P(x\leq 1)$and $p(x\leq 2)$, take the corresponding binomial observation as 2.For 0.469, take 1 and for 0.344 take 1 again, to give(2,1,1) as the sample of three items "randomly" chosen form the binomial.

(ii) In U(a,b),$F(x)=\frac{x-a}{b-a}$ for $a<x<b$. Set$u=F(x)$ to give $x=a+u(b-a)$; so here $x=-1+2u$. The values corresponding to u_1,u_2,u_3 above are $x_1=0.642,\ x_2=-0.062,\ x_3=-0.312$.

(iii)If $u=\Phi^{-1}(x)$ and$x=\phi^{-1}(u)$ Table 1A shows that $u_1 0.821$ leads to $x_1=0.92;\ u_2=0.469\ to\ x_2=-0.08;\ u_3=0.344\ to\ x_3=-0.40$

8 The transition probability $p_{ij}=p(j\ balls\ in\ uin\ at\ step\ n|i\ balls\ in\ uin\ at\ step\ (n-1))$
Then

$$\begin{array}{lll} p_{01}=1 & P_{0j}=0 & (j\neq 1) \\ p_{i,i-1}=\frac{i}{m} & p_{i,i+1}=\frac{M-i}{M} & (i=1,2,\cdots(M-1)) \\ p_{M,M-1}=1 & p_{Mj}=0 & (j\neq M-1) \end{array}$$

The states refer to one of the uins
If the stationary probabilities are $\Pi=[\Pi_0\Pi_1\cdots\Pi_n]^T$ then $\Pi_j=\sum_{i=0}^{M}\Pi_i P_{i,j}$which leads to

$$\Pi_0=\frac{1}{M}\Pi_1;\quad \Pi_j=\frac{M-j+1}{M}\Pi_{j-1}+\frac{j+1}{M}\Pi_{j+1}\ \ (j=1,2,\cdots,M-1);\ \ \Pi_M=\frac{1}{M}\Pi_{M-1}$$

Using the given probabilities, $\frac{\Pi_0}{\Pi_1}=\begin{pmatrix} M \\ 0 \end{pmatrix}/\begin{pmatrix} M \\ 1 \end{pmatrix}=\frac{1}{M}$
satisfied $\frac{\Pi_M}{\Pi_{M-1}}=\begin{pmatrix} M \\ M \end{pmatrix}/\begin{pmatrix} M \\ M-1 \end{pmatrix}=\frac{1}{M}$
satisfied, For j=1 to (M-1)

$$\frac{M-j+1}{M}\begin{pmatrix} M \\ j-1 \end{pmatrix}+\frac{j+1}{M}\begin{pmatrix} M \\ j+1 \end{pmatrix}=\frac{M-j+1}{M}\frac{M!}{(j-1)!(M-j+1)!}+\frac{j+1}{M}\frac{M!}{(j+1)!(M-j-1)!}$$

$$\frac{(M-1)!}{(j-1)!(M-j)!}+\frac{(M-1)!}{j!(M-j-1)!}=\frac{(M-1)!}{(j-1)!(M-j-1)!}\{\frac{1}{M-j}+\frac{1}{j}\}$$

$$\frac{(M-1)!}{(j-1)!(M-j-1)!}\frac{M}{(M-j)j}=\frac{M!}{j!(M-j)!}$$

satisfying this set of equations since $(\frac{1}{2})^M$ is a common factor
.

1(i) In a uniform distribution over (a,b), the mean is $\frac{1}{2}(a+b)$ and the variance is $\frac{1}{12}(b-a)^2$
Thus for $u(\theta, \theta+1)$, $E^2[u] = \theta + \frac{1}{2}$ and $var[u] = \frac{1}{12}$
Suppose $\{x_i\}$ have this distribution, so that $E[\bar{x}] = \theta + \frac{1}{2}$(because $E[x_i] = \theta + \frac{1}{2}$ for $i = 1, 2, 3, \cdots n$), and $E[\hat{\theta}] = \theta$.

$$v[\hat{\theta}] = v[\bar{x}] == \frac{1}{n^2} V[\sum_{i=1}^{n} x_i] = \frac{1}{n^2} \times n \times \frac{1}{12} = \frac{1}{12n}$$

(ii) The c.f.d.of y is

$$\begin{aligned} p(Y \geq y) &= p\{\max_i(x_i) \leq y\} \\ &= p(x_1, x_2, \cdots, x_n \leq y) \\ &= \prod_{i=1}^{n} p(x_i \leq y) \\ &= (y-\theta)^n \quad \textit{by independence } \theta < y < \theta + 1 \end{aligned}$$

The p.d.f. is the derivative of this,$f(y) = n(y-\theta)^{n-1}$, $\theta < y < \theta + 1$

$$\begin{aligned} E[y] &= \int_{y=\theta}^{\theta+1} ny(y-\theta)^{n-1} dy = \int_{\theta}^{\theta+1} n\{(y-\theta)+\theta\}(y-\theta)^{n-1} dy \\ &= n\int_{\theta}^{\theta+1}(y-\theta)^n dy + n\theta\int_{\theta}^{\theta+1}(y-\theta)^{n-1} dy \\ &= n[\tfrac{(y-\theta)^{n+1}}{n+1}]_{\theta}^{\theta+1} + n\theta[\tfrac{(y-\theta)^n}{n}]_{\theta}^{\theta+1} \\ &= \tfrac{n}{n+1} + \tfrac{n\theta}{n} = \theta + \tfrac{n}{n+1} \end{aligned}$$

$$\begin{aligned} E[y^2] &= \int_{y=\theta}^{\theta+1} ny^2(y-\theta)^{n-1} dy = \int_{\theta}^{\theta+1} n\{(y-\theta)^2 + 2y\theta - \theta^2\}(y-\theta)^{n-1} dy \\ &= n\int_{\theta}^{\theta+1}(y-\theta)^{n+1} dy + 2\theta\int_{\theta}^{\theta+1} ny(y-\theta)^{n-1} dy - n\theta^2\int_{\theta}^{\theta+1}(y-\theta)^{n-1} dy \\ &= n[\tfrac{(y-\theta)^{n+2}}{n+2}]_{\theta}^{\theta+1} + 2\theta(\theta + \tfrac{n}{n+1}) - n\theta^2[\tfrac{(y-\theta)^n}{n}]_{\theta}^{\theta+1} \\ &= n\tfrac{1}{n+2} + 2\theta^2 + \tfrac{2n\theta}{n+1} - \theta^2 = \theta^2 + \tfrac{2n}{n+1}\theta + \tfrac{n}{n+2} \end{aligned}$$

Hence

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

(iii)$\tilde{\theta} = y - (\frac{n}{n+1})$ is an unbiased estimation for θ

$$v[\tilde{\theta}] = v[Y] = \frac{n}{(n+1)^2(n+2)}. \quad \frac{v[\hat{\theta}]}{v[\tilde{\theta}]} = \frac{(n+1)^2(n+2)}{12n^2}$$

Both $\hat{\theta}$ *and* $\tilde{\theta}$ are consistent estimators, since they are unbiased and their variances $\to 0$ as $n \to \infty$.

2 The power of a test is the probability of rejecting the null hypothesis when the alternative hypothesis is correct. If the null and alternative hypotheses are both simple, and the significance level and minimum power are specified, then a lower bound for the required size can be found.

(i)To test $H_0: \ v = v_0$ against $H_1 : v = v_1$, where $v_1 > v_0$, the likelihood ratio:

$$\begin{aligned}\lambda \ &= \tfrac{L(v_0)}{L(v_1)} = \prod_{i=1}^{n} \tfrac{v_0^k}{\Gamma(k)} x_i^{k-1} e^{-v_0 x_i} / \prod_{i=1}^{n} \tfrac{v_1^k}{\Gamma(k)} x_i^{k-1} e^{-v_i x_i} \\ &= (\tfrac{v_0}{v_1})^{nk} e^{-(v_0 - v_1)\sum_{i=1}^{n} x_i}\end{aligned}$$

The neyman-peason lemma gives the most powerful test as the likelihood ratios test with critical region $C = \{\underline{x} : \ \lambda \leq c\}$ for some c, or $C = \{\underline{x} : \ \sum_{i=1}^{n} x_i \leq c^{'}\}$, in which

$$c^{'} = \frac{1}{v_1 - v_0} \ln\{(\frac{v_1}{v_0})^{nk} c\}$$

(ii)The m.g.f of x is $M_x(t) = (1 - \frac{t}{v})^{-k} \quad t < v$
so that of $sum x_i$ is $\prod_{i=1}^{n} M_{x_i}(t) = (1 - \frac{t}{v})^{-nk}, \quad t < v$
Because of the uniqueness theorem for generating functions this implies that $y = \sum_{i=1}^{n} x_i$ is Gamma(nk,v).

(iii)For $k = \frac{1}{n}$ y is Gamma(1,v), which is exponential(v). If $H_0$is true, $Y \sim exonential(v_0)$, and $c^{'}$ must satisfy

$$\alpha = \int_0^{c^{'}} v_0 e^{-v_0 y} dy = [-e^{-v_0 y}]_0^{c^{'}}, \quad \text{ı.e. } c^{'} = \frac{\ln(1-\alpha)}{v_c}$$

The required critical region is then

$$C = \{\underline{x} : \sum_{i=1}^{n} x_i \leq \frac{-\ln(1-\alpha)}{v_0}\}$$

(iv)power is

$$p(\sum_{i=1}^{n} x_i \leq \frac{-\ln(1-\alpha)}{v_0} | v = v_1) = \int_0^{\frac{-\ln(1-\alpha)}{v_0}} v_1 e^{-v_1 y} dy$$
$$e^{-v_1 y}|_0^{\frac{-\ln(1-\alpha)}{v_0}} = 1 - (1-\alpha)^{\frac{v_1}{v_0}}$$

3 suppose that $x = (x_1, \cdots, x_n)$ is a set of data form a population in which θ is an unknown parameter. A statistic $V(x;\theta)$ is a pivotal quantity of:

(i) $q(x;\theta)$ involves θ but no other unknown parameters;

(ii) The distribution of q does not depend on θ, or on any other unknown parameters. To find a 100z% confidence set for θ, find a set c such that $p\{q(x;\theta) \in c\} = z$ since the distribution of q does not involve θ, c is independent of θ. Then if x take the observed value x, the confidence set for θ is $\{\theta : \ q(x,\theta) \in c\}$

(i) The distribution function of y is

$$F_Y(y) = P(Y \leq y) = p(-\ln x \leq y) = p(x \geq e^{-y}) = 1 - F_x(e^{-y})$$

The probability density function of y is therefore

$$f_Y(y) = e^{-y} f_x(e^{-y}) = \lambda e^{-\lambda y}, \quad y > 0$$

which is exponential with parameter λ

(ii)Let$w = y\lambda$. The w has distribution function

$$F_W(w) = P(W \leq w) p(y \leq \frac{w}{\lambda}) = 1 - F_x(e^{-\frac{w}{\lambda}})$$

and its density function is

$$f_W(w) = \frac{1}{\lambda} e^{-\frac{w}{\lambda}} f_x(e^{-\frac{w}{\lambda}}) = e^{-w} \quad w > 0$$

Therefore $w = y\lambda$ is exponential with parameter 1.Thus $y\lambda$ is a function of λ whole distribution does not depend on λ. Hence it is a pivotal quantity .

(iii) A 95% confidence interval for λ is $\{\lambda : \ R^1 < y\lambda < R^2\}$ where R_1,R_2 are the lower and upper $2\frac{1}{2}$% points of exp(1); so $\int_0^{R_1} e^{-w}dw = 0.025$, i.e. $[-e^{-w}]_0^{R_1} = 0.025$, so $e^{-R_1} = 0.975$, $\ R_1 = 0.025$ so $\int_0^{R_2} e^{-w} = 0.975$ requires $e^{-R_2} = 0.025$, so that $R_2 = 3.689$ Hence a 95% confidence interval forθ is (0.025/y; 3.689/y).

4 The likelihood function based on observations $(x_1, x_2, \cdots, x_n)$ i s

$$L_{(n)}(p) = \begin{pmatrix} n \\ \sum\limits_{i=1}^{n} x_i \end{pmatrix} p^{(\sum\limits_{i=1}^{n} x_i)} (1-p)^{n-\sum\limits_{i=1}^{n} x_i} \quad 0 \le p \le 1$$

The likelihood ratio is

$$\lambda_{(n)} = \frac{L_n(0.35)}{L_n(0.70)} = (\frac{0.35}{0.70})^{\sum\limits_{i=1}^{n} x_i} (\frac{(0.65)}{0.30})^{n-\sum\limits_{i=1}^{n} x_i} = (0.5)^{\sum\limits_{i=1}^{n} x_i} (2.167)^{n-\sum\limits_{i=1}^{n} x_i}$$

(i) In a sequential probability ratio test,

$$\begin{array}{ccc} \textit{continue sampling} & if & A < \lambda_n < B \\ \textit{accept } H_0 & if & \lambda_n \ge B \\ \textit{accept } H_1 & if & \lambda_n \le A \end{array}$$

where $A = \frac{\alpha}{1-\beta} = \frac{0.01}{0.98} = \frac{1}{98}$; $\ B = \frac{1-\alpha}{\beta} = \frac{0.90}{0.02} = 49.5$
Therefore continue sampling if

$$\begin{array}{lc} & \ln A < \sum x_i \ln(0.5) + (n - \sum x_i) \ln(2.167) < \ln B \\ i.e. & -4.585 < -0.6931 \sum x_i + 0.7732(n - \sum x_i) < 3.902 \\ or & -4.585 < -1.4663 \sum x_i + 0.7732n < 3.902 \\ i.e. & 3.127 > \sum x_i - 0.527n > -2.661 \\ so\ that & 0.527n + 3.127 > \sum x_i > 0.527n - 2.661 \end{array}$$

Also, stop and accept H_0 if $\sum x_i \le 0.527n - 2.661$
and, stop and accept H_1 if $\sum x_i \ge 0.527n - 3.127$

(ii)

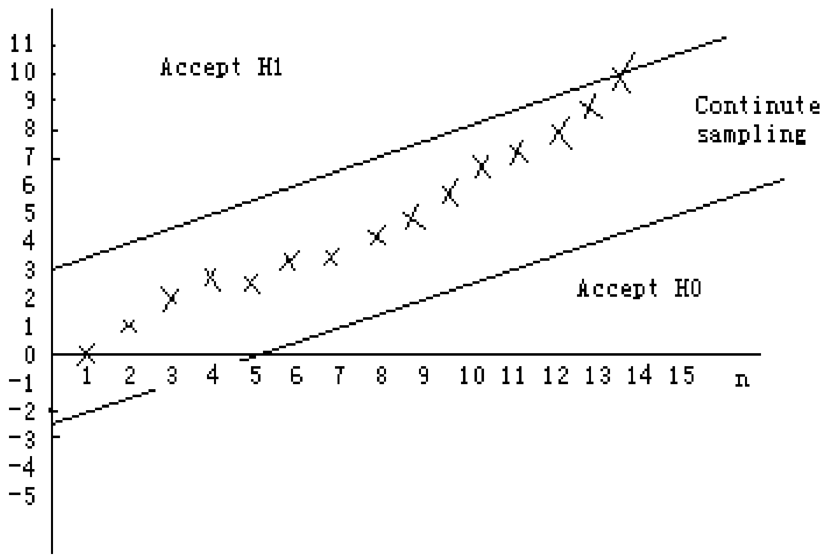
$$z_i = \ln(\frac{p_0(x_i)}{p_1(x_i)}) = x_i \ln(0.5) + (1 - x_i) \ln(2.167) \quad [i = 1, 2, \cdots, n]$$

Expect sample size when H_1 is true is

$$E_1(n) = \frac{(1-\beta)\ln A+\beta\ln B}{E_1[z_i]} = \frac{0.98\ln(1/98)+0.02\ln(49.5)}{-0.2532}$$

$$= \frac{-4.41523}{-0.2532} = 17.44 \quad (say\ approx\ 17.5)$$

(iii)Plot $\sum_{i=1}^{n} x_i$ against n, and stop sampling as soon as the sample path goes outside the 'continue sampling ' region between the two parallel lines.
For the given data, stop after patient 15,and accept H_1



5 In a bayesian analysis, if the prior and posterior distributions belong to the same family, then this family is said to be conjugate to the distribution yielding the observations

(i)The prior distribution of θ is

$$\Pi(\theta) x \theta^2 e^{-\theta/3} \quad \theta > 0$$

and the likelihood function is

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x_i}}{\Pi(x_i!)} \quad \theta > 0$$

The posterior distribution of θ is

$$\Pi(\theta|x) \propto x\theta^2 e^{-\theta/3} e^{-n\theta} \theta^{\sum_{i=1}^{n} x_i} = \theta^{2+\sum_{i=1}^{n} x_i} e^{-(n+\frac{1}{3})\theta} \quad \theta > 0$$

This is gamma with $v = n + \frac{1}{3}$ and $k = 3 + \sum_{i=1}^{n} x_i$

(ii)with a squared-error loss function,the bayes estimation $\tilde{\theta}$ is the expect value of θ in the posterior distribution. For $y \sim \Gamma(k, v)$, the m.g.f. is

$$M_y(t) = (1 - \frac{t}{v})^{-k} = 1 + \frac{kt}{v} + \cdots \qquad (k < v)$$

so that the mean is k/v.
The bayes estimator $\tilde{\theta}$ is thus.

$$E(\theta|x) = \frac{3 + \sum_{i=1}^{n} x_i}{\frac{1}{3} + n}$$

(iii)The posterior distribution is now $\Gamma(29, 13/3)$. Using the given result, $26\theta/3 \sim \chi^2_{(58)}$, and so $p(39.67 < \frac{266}{3} < 82.11) = 0.95$ *or* $p(\frac{3}{26} < \theta < \frac{3}{26} \times 82.11) = 0.95$. The interval is (4.58; 9.47).

(iv)$p(0) = e^{-\theta}$. A bayes estimate of $e^{-\theta}$ is the expected value of $e^{-\theta}$ on the posterior distribution. When $y \sim \Gamma(29, 13/3)$,

$$E[e^{-\theta}|x] = \int_0^{\infty} e^{-\theta} \Pi(\theta|x) d\theta = M_y(t) \quad \textit{evaluated at } t = -1$$

This is $(\frac{13/3}{13/3+1})^{29} = (\frac{13}{16})^{29}$

6 (i)Let n_i denote the number of values in category i (i=0,1,2,3). The probabilities of an observation falling into categories of 0,1,2,3 are $(1-p)^3,\ 3p(1-p)^2,\ 3p^2(1-p)\ p^3$. The distribution of $\{n_i\}$ is multinomal with these probabilities:

$$p(n_0, n_1, n_2, n_3) = \frac{216!}{\prod_{i=0}^{3}(n_i!)} (1-p)^{3n_0} \{3p(1-p^2)\}^{n_1} 3p^2(1-p)^{n_2} p^{3n^3}$$

and

$$L(p) \propto (1-p)^{3n_0+2n_1+n_2} p^{n_1+2n_2+3n_3} \quad \textit{for } 0 < p < 1$$

$n_0 = 110,\ n_1 = 85,\ n_2 = 20,\ n_3 = 1$; hence

$$L(p) \propto (1-p)^{520} p^{128} \quad 0 < p < 1$$

(ii)

$$\ln L = const + 520 \ln(1-p) + 128 \ln p$$

and

$$\frac{d(\ln L)}{dp} = -\frac{520}{1-p} + \frac{128}{p}$$

also

$$\frac{d^2(\ln L)}{dp^2} = \frac{-520}{(1-p)^2} - \frac{128}{p^2} < 0$$

The maximum likelihood estimate of p is found from $\frac{d}{dp}(\ln L) = 0$; $\quad \hat{p} = \frac{128}{648} = 0.198$
We need the probabilities in a binomial distribution (3,0.198);

$$p_0 = 0.5168 \quad p_1 = 0.3816 \quad p_2 = 0.0939 \quad p_3 = 0.0077$$

giving $E_i = 216 p_i, \quad i = 0, 1, 2, 3$ Therefore

$$E_0 = 111.63, \quad E_1 = 82.43 \quad E_2 = 20.28 \quad E_3 = 1.66$$

combine the last two categories:

	0	1	2 *and* 3
Observed	111	85	21
Expected	111.63	82.43	21.94

three items in table one parameter estimated: df=1.

$$\chi^2_{(1)} = \frac{0.63^2}{111.63} + \frac{2.57^2}{82.43} + \frac{0.94^2}{21.94} = 0.12$$

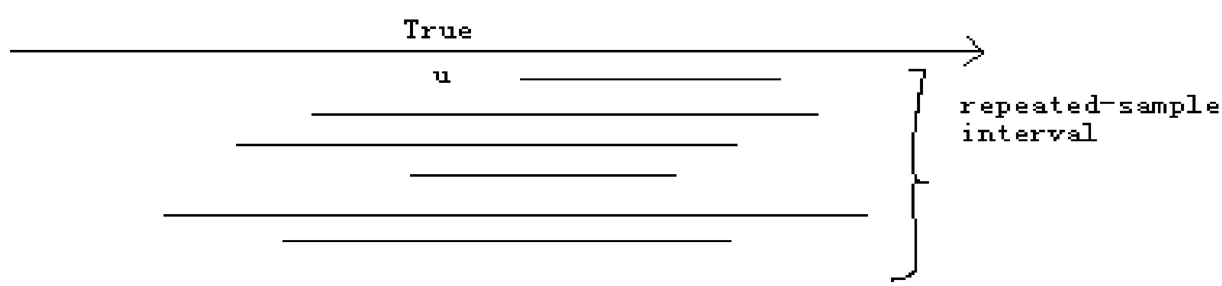
not significant. No evidence of lack of fit.

(iii)Applying the centra Limit Theorem, an approximate 90% confidence interval is $\hat{p} \pm 1.645\sqrt{\frac{\hat{p}(1-\hat{p})}{3n}}$ The variance arises in the following way :

$$\begin{aligned} E(-\frac{d^2 \ln L}{dp^2}) &= E[\frac{3n_0+2n_1+n_2}{(1-p)^2} + \frac{n_1+2n_2+3n_3}{p^2}] \\ &= n[\frac{3p_0+2p_1+p_2}{(1-p)^2} + \frac{p_1+2p_2+3p_3}{p^2}] \\ &= 3n[\frac{(1-p)^3+2p(1-p)^2+(1-p)p^2}{(1-p)^2} + \frac{p(1-p)^2+6p(1-p)+3p^2}{p^2}] \\ &= 3n(\frac{1}{1-p} + \frac{1}{p}) = \frac{3n}{p(1-p)} \end{aligned}$$

and so the variance is $p(1-p)/3n$.
For the given data, $\hat{p} = 0.1975$ and $\sqrt{\frac{\hat{p}(1-\hat{p})}{648}} = 0.01564$, giving the confidence interval 0.1975 ± 0.0257 *or* $(0.172$ *to* $0.223)$

7. Interval estimation .
Classical methods are "frequentist". A confidence interval is a random interval, determined from sample data each time a new sample is selected from the same population, which has a specified probability of containing the true(population) value of the parameter being studied. This probability has to be understood in the sense of repeated sampling from a population, and so it is not entirely clear how the idea applies to a single sample of data, e.g. from an experiment thus when a 95% confidence interval for μ is found, in a normal distribution with σ^2 unknown, using $\bar{x} \pm t_{(n-1)}s/\sqrt{n}$



as the limits, repeated sampling could yield the intervals shown, with centers depending on $\bar{x}$ and width with depending on s^2. In the long run, only 5% of these would be expected not to include the true μ. A Bayesian interval is an interval within which the parameter falls with specified probability. Because we do not assume the parameters to have a "true" value but only a posterior distribution (depending on an assumed prior distribution and on the available data), this gives a clear definition of the concept without involving hypothetical "repeated sampling". There can be argument about the assumptions of the prior distribution and the derivation of posterior. If a uniform distribution is used as prior, the Bayesian approach is similar to the likelihood approach Because the data have considerable influence, although it is not exactly the same.

The likelihood approach is to include in the interval all values of parameters which give a log likelihood that is within a certain distance of the maximum likelihood given by $\hat{\theta}$. Its logical basis is that the likelihoods function represents the plausibility of the different values of θ, and there can be some argument about this, as well as about choice of "certain distance ". In large samples, the log likelihood is approximately quadratic,the data assume major importance in the Bayesian approach, and so the these approaches give similar results.

8. The Kolmogorov-Smirnov test is a goodness-of-fit test for data when the null hypothesis states that they are drawn from the distribution $F_0(x)$. This hypothesis c.d.f. can be calculated for each observed sample value of x, and the sample c.d.f. $\{F_{(n)}(x)\}$ is then compared with it. As in the following example, the test uses the set $\{D_{(k)}\}$ of differences between these two c.d.f.'s at the points $x_1, \cdots, x_n$.
If the sample values are ranked so that $x_1 \leq x_2 \leq \cdots \leq x_n$ then

$$F_{(k)}(x) = \left\{ \begin{array}{ll} 0 & for x < x_{(1)} \\ \frac{k}{n} & for x_{(k)} \leq x \leq < x_{(k+1)} \quad k = 1, 2, \cdots, n-1 \\ 1 & for x \geq x_{(n)} \end{array} \right.$$

If H_0 is $F(x) = F_0(x)$ and H_1 is $F(x) \neq F_0(x)$, and we define $D_{(k)} = |F_{(k)}(x) - F_0(x)|$, the test statistic is $D_n = \max\limits_{x_{(1)} \cdots x_{(n)}} (D_{(k)})$ and D_n is referred to the tables using sample size n.
H_0 is rejected when D_n is above the critical value in the tables. Merits of this test are: (1) small sample sizes can be used (unlike the χ^2 goodness-of-fit test) because $D_n$has a known distribution which can be tabulated; (2) a one-sided alternative hypothesis can be used (again unlike the χ^2 test);(3) a "confidence band" for an unknown $F(x)$ can be constructed using this test statistic.
If H_0 specifies an exponential distribution with mean 40, then

$$f(x) = \frac{1}{40} e^{-\frac{x}{40}}, x > 0$$

and so

$$F(\xi) = \int_0^{\xi} \frac{1}{40} e^{-\frac{x}{40}} = [-e - \frac{x}{40}]_0^{\xi} = 1 - e^{-\frac{\xi}{40}}, \xi > 0.$$

Ranked data are 1, 6, 12, 18, 23, 32, 58, 68, 101, 116. n=10.

$k:$	1	2	3	4	5
$k/n:$	0.1	0.2	0.3	0.4	0.5
$x_{(k)}:$	1	6	12	18	23
$F_0(x_{(k)}):$	0.0247	0.1393	0.2592	0.3624	0.4373
$D_{(k)}:$	0.0763	0.0607	0.0408	0.0376	0.0627

$k:$	6	7	8	9	10
$k/n:$	0.6	0.7	0.8	0.9	1.0
$x_{(k)}:$	32	58	68	101	116
$F_0(x_{(k)}):$	0.6607	0.7654	0.8173	0.9199	0.9460
$D_{(k)}:$	0.0493	0.0664	0.0173	0.0199	0.0660

$D_{1c} = 0.0753$, the maximum value of $\{D_{(k)}\}$; and since this is much less than the 5% critical value in the table, 0.409, we do not reject H_0. The data seem to be consistent with the proposed distribution exponential with mean 40.

Applied Statistics I

1. (a) Linear models usually assume that there is random natural variation component ε which follows a normal distribution; i.e. the variance of the observation which is represented by this component, is the parameter σ^2 in a normal distribution.
If the data are known to be from another distribution, a transformation can help to normalize them and to make σ^2 constant: e.g. $\ln y$ is useful when y is skew to the right and approximately lognormal, $\sqrt{y}$ is useful for binomial data. These will all allow standard analysis of variance methods to be used on the transformed data.

If it is known, or discovered from a study of the data such as a plot of residuals against fitted values, that there is a relation between the magnitude of y_i and $var(y_i)$, e.g.σ or σ^2 is proportional to μ, then an appropriate transformation can make the variance of the data roughly constant.

Finally, a model may not be linear (in its parameters) in the original units y, but can be made so by transformation. A multiplicative model $y = \alpha x_1^{\beta} x_2^{\gamma} x_3^{\delta}$ is made "linear" by taking $\ln y = \ln \alpha + \beta \ln x_1 + \gamma \ln x_2 + \delta \ln x_3$; a term ε will be added to the right hand side which will be assume $N(0, \sigma^2)$

(b)Both data sets are skew to the right, and for both the mean and standard deviation are roughly equal .

(i)Hence a log transformation should be useful. After this, normal-theorem methods, and t-test, will be valid way of comparing average percentage level.

(ii) After transformation, using natural logs:

$$\begin{array}{llllllll} A: & 0.9163 & -0.2231 & 0.0000 & 2.7279 & 1.6292 & 1.3863 & 1.0728\ (mean) \\ B: & 2.0149 & 2.8736 & 2.9601 & 3.9279 & 1.6864 & 0.7419 & 2.3675 (mean) \end{array}$$

variance are $\sigma_A^2 = 1.2004$, $\sigma_B^2 = 1.2546$; $n = 6$ $\hat{\sigma}^2 = 1.2275$, the pooled variance for all the data (clearly σ_A^2 and σ_B^2 do not differ significantly)

$$E[\ln B - \ln A] = 2.3675 - 1.0728 = 1.2947$$

This is $E[\ln \frac{B}{A}]$; we will thus find limits for the ratio, rather than the difference in impurities.
In logarithmic units a 95% confidence interval is :

$$1.294 \pm T_{(10)}\sqrt{1.2275(\frac{1}{6}+\frac{1}{6})},$$

since $\hat{\sigma}^2$ has 10 d.f. this is $1.2947 \pm 2.228 \times 0.6397$, i.e 1.2947 ± 1.4252 *or* $(-0.131; 2.720)$
Taking exponentials, the 95% limits for $\frac{B}{A}$ are 0.878 to 15.2

2(i)If both the judges and the piece of beef have been selected at random from a larger number that were available, then a "random effect" model is appropriate rather than "fixed effects"

(ii)The grand total of x is $G = 1103$ $N = 27$ $s = \sum x^2 = 57217$. The corrected total sum of squares= $57217 - 1103^2/27 = 12157.41$. The ss for judges= $\frac{1}{9}(515^2 + 267^2 + 321^2) - \frac{G^2}{N} = 48839.4 - 45059.59 = 3779.85$
The ss for beef pieces= $\frac{1}{3}(121^2 + \cdots + 153^2) - \frac{G^2}{N} = 3975.41$

Souce of Variation	*D.F.*	*s.s.*	*M.s.*	$E[M.S.]$	
judge	2	3779.85	1889.93	$\sigma^2 + 9\sigma_\sigma^2$	$F(2,16) = 6.78*$
Pieces	8	3975.41	496.93	$\sigma^2 + 3\sigma_p^2$	$F(8,16) = 1.81$ *n.s.*
PResidual	16	4402.15	275.13	σ^2	
Total	26	12157.41			

$$\hat{\sigma}_\sigma^2 = 179.42, \quad \sigma_p^2 = 73.93, \quad \hat{}\sigma^2 = 275.13$$

which is the basic "random variation " of the process. The additional component $\hat{\sigma}_p^2$ is the repeat differences between beef pieces, which is relative small; $\hat{\sigma}_\sigma^2$ is additional variation between judges, which is larger. These are the variance components. The test of hypotheses"$\sigma_\sigma^2 = 0$" and "$\sigma_p^2 = 0$" are made using the F values given above : there is no real evidence of difference due to pieces but there is evidence of a judge difference.

(iii) The variance of each measurement is quite large, suggesting that the judging process is not very reliable. Besides this, the extra variance of the judges is considerable; people may be finding it hard to carry out the task in a reliable way. There is not much suggestion of difference among the beef pieces used.

(iv) $P\{\frac{16\hat{\sigma}^2}{\chi_u^2} < \sigma^2 < \frac{16\hat{\sigma}^2}{\chi_L^2}\} = 0.95$; 16 are the d.f. of the estimate, and χ_L^2, χ_U^2. the lower and upper $2\frac{1}{2}\%$ points of $\chi^2_{(16)}$, i.e. 6.91, 28.85. The 95% limits for σ^2 are therefore 152.6 to 637.1.

3(a)(i) A stationary time series has the joint distribution of $x(t_1) \ \cdots \ x(t_n)$ the same as that of $x(t_1+\tau) \ \cdots \ x(t_n+\tau)$ for all $\{t_i\}$ and all τ. In particular, the distributions of all members of the series are identical (consider n=1),so $E(x_t) = \mu$ and $var(x_t) = \sigma^2$ for all t .
Any pair of x's has autocovariance $E[(x_t - \mu)(x_{t+j} - \mu)] = z_j$ and autocorrelation $\rho_j = z_j/\sigma^2$, which is the same as ρ_{-j}
The general autocovariance function consists of collection of autocovariance coefficients at lag τ,$z(\tau) = E[(x_t - \mu)(x_{t+\tau} - \mu)]$ and the corresponding autocorrelation function is $z(\tau)/z(0)$.
When fitting an autoregressive process, the highest order coefficient, say α_p, measures the excels correlation at at lag P which is not accounted for by a model going only as far as (p-1). It is called the p^{th} partial autocorrelation coefficient; plotting it against p gives the partial autocorrelation function.
A partial autocorrelation function becomes effectively zero at lag p, if an AR(p) process is an appropriate model.
For a first-order process the theoretical autocorrelation decrease exponentially, but for higher orders there is no simple shape to identify.

(ii)$\nabla = x_t - x_{t-1} = a_t - \theta a_{t-1}$ which is a first-order MA process, so long as $\{a_t\}$ is "white noise". Now∇_t is stationary .

(b) The pattern of $\hat{r^k}$ for x_t suggests non-stationary, while $\hat{r_k}$ for ∇_t suggests an MA(1) process for the differences; also $\hat{\Phi_{kk}}$ for ∇_t is consistence with a first-order process. Hence we may propose

$$\nabla_t = a_t - \theta a_{t-1}$$
$$or \quad x_t = x_{t-1} + a_t - \theta a_{t-1} \qquad |\theta| < 1$$

After fitting the model, residuals may be examined; any pattern in them can indicate amendments needed to the model .
The estimate of θ, and its standard error, will show how reliable the model may be
There are general methods (e.g. Box& Pierce) of examining, in large samples, the autocorrelation coefficients.

4(i)

$$\begin{array}{r} y_1 = \beta_1 + \epsilon_1 \\ y_2 = \beta_2 + \epsilon_2 \\ y_3 = \beta_1 - \beta_2 + \epsilon_3 \end{array} \qquad x = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$$

$$x^T x = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \quad and \quad (x^{'}x)^{-1} = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$x^T y = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} y_1 + y_3 \\ y_2 - y_3 \end{bmatrix}$$

$$\hat{\beta} = (x^T x)^{-1} x^T y = \frac{1}{3}\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} \begin{bmatrix} y_1 + y_3 \\ y_2 - y_3 \end{bmatrix} = \frac{1}{3}\begin{bmatrix} 2y_1 + y_2 + y_3 \\ y_1 + 2y_2 - y_3 \end{bmatrix}$$

(ii)$\hat{\beta}_1 = \frac{2995}{3} = 998.33$ and $\hat{\beta}_2 = \frac{920}{3} = 306.67$ In order to find $\sigma^2 = var(\epsilon)$ we require residuals. Comparing y_1, y_2, y_3 with their estimates using $\hat{\beta}_1, \quad \hat{\beta}_2$, we find

$$\epsilon_1 = 1.6667, \quad \epsilon_2 = -1.6667 \quad \epsilon_3 = 1.6667 \quad \epsilon_1^2 + \epsilon_2^2 + \epsilon_3^2 = 8.3$$

Residual s.s.=8.3333=residual m.s with 1 degree of freedom, Liquid remaining $=\beta_1 - \beta_2$.

$$\begin{aligned} var(\hat{\beta}_1 - \hat{\beta}_2) &= v(\hat{\beta}_1) + v(\hat{\beta}_2) - 2cov(\hat{\beta}_1, \hat{\beta}_2) \\ &= \tfrac{2}{3}\sigma^2 + \tfrac{2}{3}\sigma^2 - 2\tfrac{1}{3}\sigma^2 \\ &= \tfrac{2}{3}(8.3) = 5.5556 \end{aligned}$$

The estimated $\hat{\beta}_1 - \hat{\beta}_2 = 691.6667; \quad t_{(1;5\%)} = 12.706$; hence a 95% confidence interval is

$$691.67 \pm 12.706\sqrt{5.5556} = 691.67 \pm 29.95 = 661.7 \ to \ 721.6$$

(iii) Incorporating information on the different variances,

$$v = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad v^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix}$$

and

$$x^T v^{-1} x = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{pmatrix}$$

so that

$$x^T v^{-1} x = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}$$

$$(x^T v^{-1} x)^{-1} = \frac{4}{5} \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$

Also

$$x^T v^{-1} y = \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} y_1 + \frac{1}{2} y_3 \\ y_2 - \frac{1}{2} y_3 \end{pmatrix}$$

So the weighted β are found from

$$\begin{aligned} \hat{\beta}_w &= \tfrac{4}{5} \begin{pmatrix} \frac{3}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{2} y_1 + \frac{1}{2} y_3 \\ y_2 - \frac{1}{2} y_3 \end{pmatrix} = \tfrac{1}{5} \begin{pmatrix} 3y_1 + 2y_2 + 2y_3 \\ y_1 + 4y_2 - y_3 \end{pmatrix} \\ &= \tfrac{1}{50} \begin{pmatrix} 4990 \\ 1530 \end{pmatrix} = \begin{pmatrix} 998 \\ 306 \end{pmatrix} \end{aligned}$$

5(i)Because the variables are correlated, all the coefficients of variables already in the model will change every time a new combination, or a new variable, is introduced. Part of the combination of x_1, for example, will become "explained" by its relation to x_2 and to x_3.

(ii)If a particular item of data (a particular subject) has "high influence" then estimates of parameters in a linear model will alter substantially if that point is omitted from the data set. A "high influence "diagnosis could therefore be a warning that the parameter estimates are unreliable because they depend heavily on certain of the data items.
A large standardized residual at a data point indicates that the fitted model does not go very near to the observed value there (standardized means that assessing fit) This can give information on how the model might be improved by including extra terms.

(iii)If we use forward selection, begin with x_1:

souce	*S.S.*	*DF*	*M.S.*	*F ratio*
x_1	2461.8	1	2461.8	225.2
Residual	2787.2	255	10.93	
	5249.0	256		

Then it is better to add x_2 than x_3.

x_2 *after* x_1	136.5	1	136.5
x_1	2461.8	1	
Residual	2650.7	254	10.44
	5249.0	256	

13.07 sig. at 0.1% (critical value appx. 6.74). Adding x_3 to x_2 and x_1 does not significantly improve fit

x_3 *after* x_1, x_2	31.4	1	31.4
x_1 *and* x_2	2598.3	2	
Residual	2619.3	253	10.35
	5249.0	256	

3.03 n.s(at 1%). The appropriate model is that containing x_1 and x_2.
Note:The same result is found by backward selection, beginning with the full model $x_1,\ x_2,\ x_3$ Omitting x_3 does not have any significant effect. After that omitting x_1 will make the fit significant effect worse, and so for x_2 although the effect is not so strong.

6(i)

$$logit\Pi = \log \frac{\Pi}{1-\Pi}$$

log to base e.

(ii)The advantage is that we do not need to make any assumption about the way in which the proportion changes from one age-group to the next; (0,1,2)would assume the same difference between young and middle ages as between middle and old ages, which is likely not be true. But the disadvantage is that it uses up an extra degree of freedom in fitting, which is lost from residual.

(iii)Add terms in x_1x_2 *and* x_1x_3:

$$logit\Pi = \beta_0 + \beta_1x_1 + \beta_2x_2 + \beta_3x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_{13}$$

(iv)The total column contains only two items of data. We are therefore fitting a straight line to just two points, and so there is no residual left with which to test the goodness of fit.

(v)$x_1 = 1,\ x_2 = 0,\ x_3 = 1$ identifies females over 60. The fitted value of logit Π is $0.0655 + 1 \times 0.884 + 0 \times (-0.1354) + 1 \times (-0.1953) = -0.0414$ so

$$\frac{\Pi}{1-\Pi} = e^{-0.0414} = 0.9594 \quad giving \;\; \hat{\Pi} = 0.9594(1-\hat{\Pi})$$

Or $1.9594\hat{\Pi} = 0.9594$ so that $\hat{\Pi} = 0.4897$. About 49%are prepared to take part.

7(i)Fisher's linear discriminant function finds $y = a^T x$ which will maximize $(\frac{\mu_H - \mu_E}{\sigma})^2$.

(ii) $\sum$ has determinant $(98 \times 92) - 57^2 = 5767$
Hence

$$\Sigma^{-1} = \frac{1}{5767}\begin{pmatrix} 92 & -57 \\ -57 & 98 \end{pmatrix}$$

$$a^T = (\mu_E - \mu_H)^T \Sigma^{-1} = (9 \;\; 8)\begin{pmatrix} 92 & -57 \\ -57 & 98 \end{pmatrix} \times \frac{1}{5767} = \frac{1}{5767}(372 \;\; 271) = (0.0645 \;\; 0.0470)$$

i.e. $y = 0.0645x_1 + 0.0470x_2$ is the discriminant function.

(iii)

$$\mu_{y(E)} = 20 \times 0.0645 + 19 \times 0.0470 = 2.183$$

$$\mu_{y(H)} = 11 \times 0.0645 + 11 \times 0.0470 = 1.227$$

and with E,H equally probable the decision rule uses $\frac{1}{2}(2.183+1.227) = 1.705$ as dividing point in classification. Allocate "honest"if $y < 1.075$.

$$\sigma_y^2 = 98(0.0645)^2 + 2(57)(0.0645)(0.0470) + 92(0.0470)^2 = 0.9565$$

The value of y has a normal distribution with mean 1.227 and variance 0.9565, so $z = \frac{1.705-1.227}{\sqrt{0.9565}} = 0.489$ is the cut-off value in a N(0,1) above which an incorrect allocation is made. $1 - \Phi(0.489) = 0.312$ is therefore the probability of incorrect classification. It is only necessary to consider 'honest' as there are only two possible classification to be used.

(iv)

$$Honest: \quad z = \frac{2-1.227}{\sqrt{0.9565}} = 0.790 \quad 1 - \phi(0.790) = 0.215$$

$$Exaggerator: \quad z = \frac{2-2.183}{\sqrt{0.9565}} = -0.187 \quad \phi(-0.187) = 0.426$$

Assuming P(honest)=0.9, P(Exaggerator)=0.1. we now have P(incorrect|H)=0.215, p(incorrect|E)=0.426. Overall probability of incorrect classification is $(0.9 \times 0.215) + (0.1 \times 0.426) = 0.236$

8(i)The model is $x_{ijk} = \mu + S_i + E_j + (SE)_{ij} + \epsilon_{ijk} \quad (i, j, k = 1, 2, 3)$ where μ is a grand mean consumption level,S_i is a fixed effect of speed,E_j a fixed effect of engine size, and $(SE)_{ij}$ an interaction between speed and size. The random terms ϵ are mutually independent, all with mean 0 and variance σ^2, drawn from a normal distribution.

(ii)There are three complete replicates of the size-speed combinations. The grand total $G = 1006.1$. $N = 27$ $G^2/N = 34790.267$. The total $\sum x^2 = 38253.85$; hence total $s.s. = 763.58$ speed $s.s. = \frac{1}{9}(355.7^2 + 374.3^2 + 276.1^2) - \frac{G^2}{N} = 604.64$, and that for $Engine = \frac{1}{9}(358.8^2 + 324.1^2 + 323.2^2) = 91.57$
Total for engine /speed combinations are

	1100	1500	1800
30 :	133.9	114.0	107.8
50 :	128.5	123.2	122.6
70 :	96.4	86.9	92.8

ss speeds +engines+interaction= $\frac{1}{3}(133.9^2 + \cdots + 92.8^2) - \frac{G^2}{N} = 750.97$ Analysis of variance.

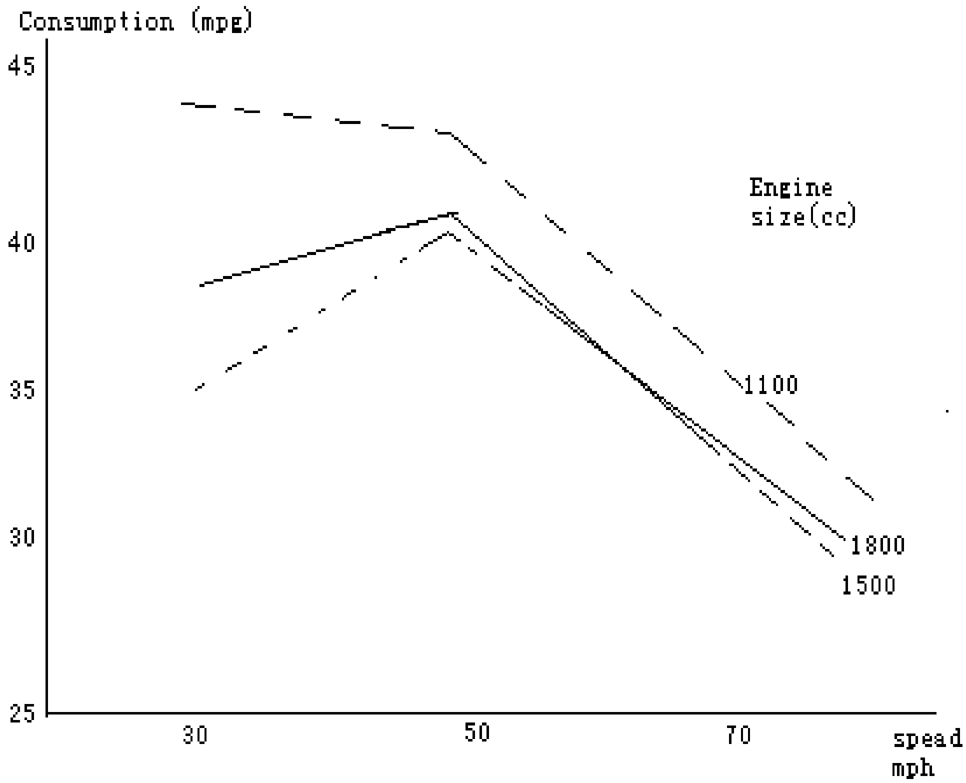
Source	*D.F*	*Sum of squares*	*M.s.*
speeds	2	604.64	302.32
Engines	2	91.57	45.79
Interaction	4	54.76	13.69
Residual	18	12.61	0.7006
Total	26	763.58	

$F_{(4,18)} = 19.53^{**}$
The interaction is very highly significant. Result must therefore be interpreted in terms of the interaction.A graph of mean is useful.

Means *E*	1100	1500	1800
$S : 30$	44.63	38.00	35.93
50	42.83	41.07	40.87
70	32.13	28.97	30.93

The stand error of difference between two means is $\sqrt{\frac{2}{3} \times 0.7006} = 0.683$
"Least significant differences" are

$$t_{(18)} \times 0.683 = \begin{cases} 1.43 & 5\% \\ 1.97 & 1\% \\ 2.68 & 0.1\% \end{cases}$$



Consumption of 1100cc engine is always significantly above that of the other two sizes, at any speed.
speeds three sizes show a sharp drop form 50mph to 70mph, and at All three speeds 1500 and 1800 engine do not differ from one another.
At 30mph, all size of engine differ significantly from one another. 1100cc is higher in consumption at 30mph than at 50, whereas both other sizes are lower at 30 than at 50mph.

Applied Statistics II

1(a)(i) In a linear model, terms are added together, and there is among them a residual term to explain natural variation which is assumed to follows a normal distribution whose mean is 0 and variance σ^2, which is constant over all the observations made all terms and all residuals are mutually independent. The model includes term for all the source of variation present in the observations made.

(ii) If any systematic variations among observed residuals is detected, a further term may be required in the model. If there is evidence of non-constant variance (e.g/ larger observations have larger residuals) a variance-stablishing transformation such as log or square root may be appropriate. A complete transformation of a model sometimes makes it linear in its parameters, e.g a log transformation of a multiplicative model. When an individual contrast is studied in a block design, calculating the value of the contrast in each block can overcome non-constant variance.

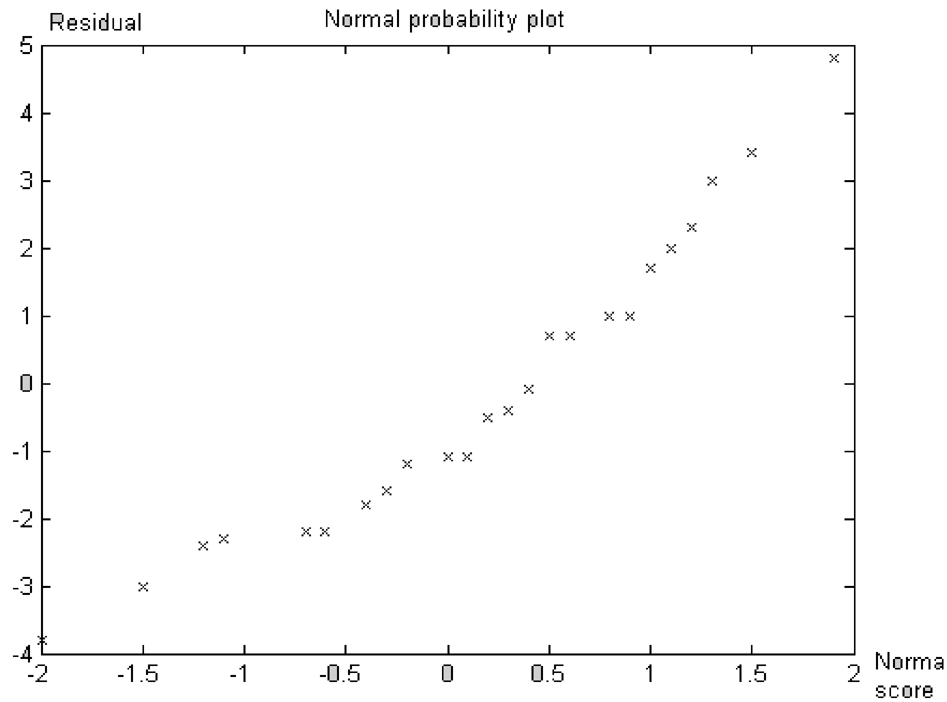
(b)(i)

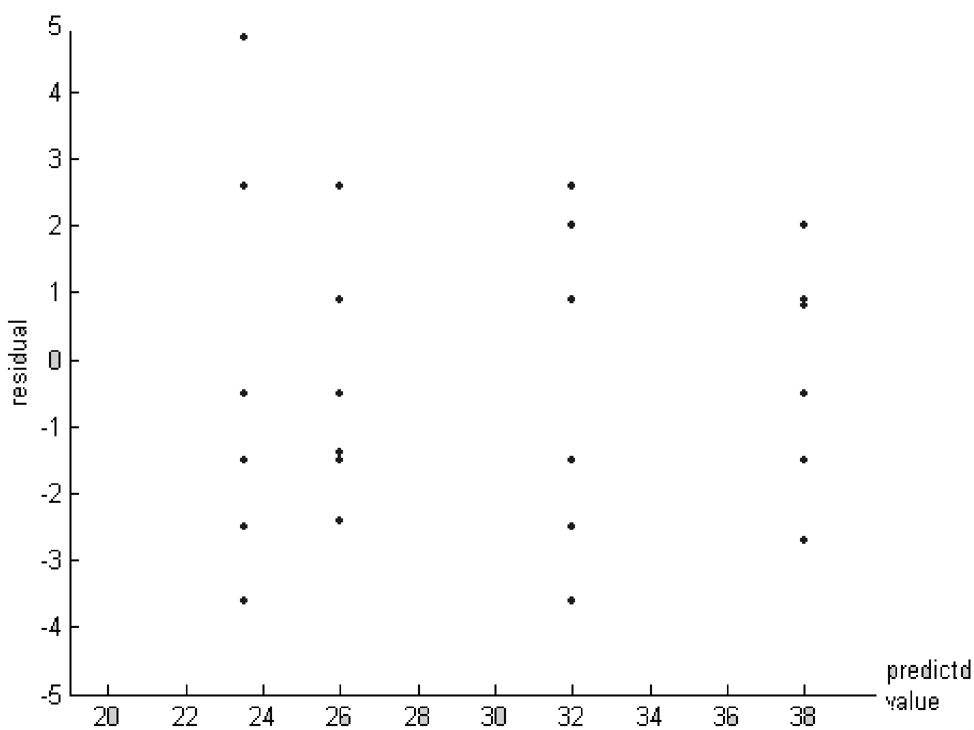
$$y_{ijk} = \alpha + \tau_i + k_j + \Phi_{ij} + \epsilon_{ijk} \quad i = 1, 2; \; j = 1, 2 \; k = 1 \; to \; 6$$

y_{ijk} is an observation, hα the overall mean, τ_i an effect due to time, k_j an effect due to culture medium, ϕ_{ij} an interaction of medium and time, all of these terms being fixed-effect terms. Finally ϵ_{ijk} are a set of i.i.d N(0,σ^2) residual terms, There are 6 replicates (assumed "completely randomize")of the four treatments $T_{12}c_1,\ total$ 140, $T_{18}c_1,\ total$ 223; $T_{12}c_2,\ total$ 156 $T_{18}c_2,\ total$ 192 $G = 771$ $\hat{\alpha} = \frac{711}{24} = 29.625$ Mean of $T_{12}c_1$ is 23.333, Which is the predicted value for each observation there in; similarly we have for $T_{18}c_1$ the prediction 31.167; for $t_{12}c_2$, 26.000; for $T_{18}c_2$, 32.000. Residuals found as observed minus predicted value, are:

$T_{12}c_1$:	−2.333	−1.333	−0.333	4.667	−3.333	2.667
$T_{18}c_1$:	0.167	1.833	0.833	0.833	−2.167	−1.167
$T_{12}c_2$:	−1.000	0.000	−2.000	−1.000	3.000	1.000
$T_{13}c_2$	−1.000	2.000	−3.000	1.000	−2.000	3.000

(ii)From the plots on the following page, we use that although the normal probability plot gives an apparently adequate straight line there is some evidence from the plot of residuals against fitted values that variance may not be constant. $T_{12}c_1$ is more variable, and $T_{18}c_1$ less variable than the c_2 combinations. with only 6 replications no firm conclusion can be drawn, however.


Residual
Normal probability plot
Normal score


residual
predictde value

2(a)

D	A	D	C	A	D	C	B
C	B	A	B	C	B	A	D

Pen *I* *II* *III* *IV*

Additives are A,B,C,D *square* = 1 animal.

Source	*DF*
$Pens \equiv litters$	3
Additives	3
Residual	9
	15

pens and litters are confounded.

(II)

A_1	A_1	B_4	B_4	C_2	C_2	D_3	D_3
A_1	A_1	B_4	B_4	C_2	C_2	D_3	D_3

Pen *I* *II* *III* *IV*

1,2,3,4 are litter numbers.

Source	*DF*
$Pens \equiv litters \equiv additives$	3
Residual(betwen animals within pens)	12
	15

pens and additives and litters are confounded.

(III)

A_1	C_4	C_3	D_1	C_2	A_3	A_4	D_2
D_3	B_2	A_2	B_4	B_1	D_4	B_3	C_1

Pen *I* *II* *III* *IV*

Source	*DF*
Pens	3
litters	3
Additives	3
Residual	6
	15

(b)(A)This applies to II. A pen is the unit rather than an animal, and the residual is only within pen variation. For other designs the unit is an animal, and the residual is a measure of the overall variation.

(B)This applies to II: see the table given above. There is confounding of all three source of variation so that only 3 d.f. are used by them.

(c)This applies to II: the suggestion is :

	A_1	A_2	B_1	B_2	C_1	C_2	D_1	D_2
	A_3	A_4	B_3	B_4	C_3	C_4	D_3	D_4
Pen	*I*		*II*		*III*		*IV*	

pens≡ Additives in analysis

Source	*DF*
Pens ≡ Additives	3
Litters	3
Residual	9
	15

The comment is true but still needs to make a serious assumption of no effect of pens.

(D) This applies to III:only here are pens, Additives and Litters capable of separate estimation. It is a 'Latin square' type of analysis in this sense.

(E)This applies to II: see the table in part(a). Litters and additives are not confounded in any other design.

(F) This applies t III, because it is the only design in which each pen contains an animal from each litters.

(G)This applies to III,because we can take out all the three effects,pens litters and additives, each of which uses up 3 d.f..

3(i) The test of 6 treatments uses all combinations of the 2-level factor W(watering) and the 3-lever factor F(fertilizer). Total are $G^2/N = 903^2/12 = 67950.75$

Fertilizer :	*O*	*AS*	*MP*	*TOTAL*
W Heavy :	154	199	173	526
Light	101	110	166	377
	255	309	339	903

The arrangement was completely randomize.
The corrected total

$$s.s. = (72^2 + \cdots + 81^2) - G^2/N = 72065 - G^2/N = 4114.25$$

The treatment

$$s.s. = \frac{1}{2}(154^2 + \cdots\ 166^2) - G^2/N = 3600.75$$

Source of variation	*DF*	*S.S*	*M.S*	
Between treatments	5	3600.75	720.15	$F_{(5,6)} = 8.41*$
Within treatments	6	513.50	85.583	
Total	11	4114.25		

The value of $F_{(5,6)}$ is almost significant at 1%, so there is evidence of difference among treatments.

(iii) The 5 d.f. for treatments can be divided into 5 orthogonal contrasts, each with 1 d.f., as specified. Denoting waterig levels as H,L:

Treatment	*HO*	*HAS*	*HMP*	*LO*	*LAS*	*LMP*	*Value*	*Divisor*	*S.S.*
Total	154	199	173	101	110	166			
Contrast(*a*)	1	1	1	-1	-1	-1	149	12	$1850.083**$
(*b*)	0	1	-1	0	1	-1	-30	8	112.500 *n.s.*
(*c*)	0	1	-1	0	-1	1	82	8	$840.500*$
(*d*)	2	-1	-1	2	-1	-1	-138	24	$793.500*$
(*e*)	2	-1	-1	-2	1	1	10	24	4.167 *n.s.*
									3600.750

Each contrast can be tested as $F_{(1,6)}$ against the residual mean square, with the result shown in the final column. Hence watering lever has a very important effect(see(a)), fertilizing also has an effect (see(d)) and the comparison between the two fertilizers is different at the two watering levers(see(c)). Heavy watering gives heavies plant roots.
Contrast (d):

$$mean\ \ non-fertilized = \tfrac{255}{4} = 63.75$$

$$mean\ \ fertilized = \tfrac{309+339}{8} = 81.00$$

so fertilizing gives heavier plant root,

$$\begin{array}{llll} Contrast: & means & H & L \\ & AS & 99.5 & 55.0 \\ & Mp & 86.5 & 83.0 \end{array}$$

Heavy watering is beneficial with AS, but not with MP; AS is better than MP under H but the opposite is true with h.
and (e)give no further information when (c)and(d)have been examined. Since the contrasts each have 1 d.f., no further t-test are required as they would be equivalent to F.

(ii)To give the variances of the contrasts expressed in terms of treatment means, note that each observation has varianceσ^2; if the positive and negative terms in the contrast are based on m and n observations the variance will be$\sigma^2(\frac{1}{m}+\frac{1}{n})$
(a)Compares 6 observations on H with 6 on L,so $var[(a)] = \frac{\sigma^2}{6}+\frac{\sigma^2}{6}=\frac{\sigma^2}{3}$ This is estimated by (85.583)/3 so that its SE is 5.34
(b)compares 4 observations on AS with 4 on MP, so $var[(b)] = \frac{\sigma^2}{4}+\frac{\sigma^2}{4}=\frac{\sigma^2}{2}$ This is estimated by (85.583)/2 so that SE is 6.54
(c)compares the 4 observations HAS,LMP with the 4 HMP,LAS,and so has the same variance as (b)and the SE=6.54
(d)compares the 8 observations AS,MP with the 4 control observations and so has variance$\frac{\sigma^2}{8}+\frac{\sigma^2}{4}=\frac{3\sigma^2}{8}$ so that its SE is estimated as$\sqrt{\frac{3}{8}\times 85.583}=5.67$
(e)compares the 6 observations HO,LAS,LMP with the 6 observations LO,HAS,HMP and so has the same variance as(a),i.e.SE is 5.34

4(i) I is a fractional factorial design which can be used to fit a linear relation between the response and $x_A,\cdots,x_E$ II gives (k-1)d.f. towards estimating residual, so that there can be a lack-of -fit test of the fitted model.III are the "axial"points which allows quadratic and interaction terms to be fitted, so that maximum peak height may be estimated.

(ii)Using I=ABCDE as defining relation, the aliases are:

$$\begin{array}{lll} A=BCDE & AB=CDE & BD=ACE \\ B=ACDE & AC=BDE & BE=ACD \\ C=ABDE & AD=BCE & CD=ABE \\ D=ABCE & AE=BCD & CE=ABD \\ E=ABCD & BC=ADE & DE=ABC \end{array}$$

This allows all the required terms to be fitted
If a $\frac{1}{4}$ replicate were to be used, a defining relation could be I=ABD=ACE=BCDE;This is the best type available Aliases now are

$$A = BD = CE = ABCDE$$
$$B = AD = ABCE = CDE \quad and\ similarly\ for C, D, E$$
$$BC = ACD = ABE = DE \quad and\ similarly\ for B, E$$

Thus we alias each main effect with at least one two-factor interaction and so will not be able to fit all the items needed in the model. The two-factor interaction not in these alias sets are aliased with other two-factor interactions, causing more difficulty in fitting the model.

(iii)Completed table is :

Source	*DF*	*Mean square*	
Constant term	1	108.17	
First Order	5	84.15	$F_{(5,10)} = 2.52\ ns$
Interaction	10	131.80	$F_{(10,10)} = 3.95*$
Second order	5	70.91	$F_{(5,10)} = 2.12\ ns$
Lack of fit	6 } 10	33.396	
Pure error	4 }		
Total	31		

An initial test shows lake of fit is not significant different from pure error ($F_{6,4} = 145$) so there is no evidence of lack of fit.Also we may pool these two estimates of σ^2 to have 10 d.f. The only significant part of the model is second order, i.e. quadratic terms.
The fitted second-order model may be used to locate, the maximum or minimum responses, and the levels of the factors which correspond to these.
Canonical analyzes can also locate ridges, saddle points and other types of interaction.
Contour diagrams, if suitable graphical facilities are available, will allow detailed study of the patterns of responses rates of change as factor-levers change, experimental regions for any follow-up work. with 5 factors, They can only be studied three at a time with suitable choice of fixed values for the other two.
Because the linear terms were not significant, there is likely to be a maximum (or minimum) near the center(00000)

5(a)Fertility relates to the number of live births a woman has had ,i.e. is the "opposite" of childlessness .
Period analysis considers all births occurring in a specified period of time, usually one year.
Cohort analysis considers all births occurring to a specific group of women, usually to all those born in a particular year, or all those married in a particular year.

1.

$$\text{Birth rate} \quad = \quad 1000 \times \frac{\text{number of live births}}{\text{total population}}$$

$$= \frac{1000 \times 10122}{315000 + 285000} = \frac{10122}{600} = 16.87 \text{ per 1000 per year}$$

2.

$$\begin{aligned} \text{General fertility rate} &= 1000 \times \frac{\text{number of live births}}{\text{no.of females aged 15-44}} \\ &= \frac{1000 \times 10122}{129000} = 78.47 \text{ per 1000 females of childbearing age.} \end{aligned}$$

3.

$$\begin{aligned} \text{Fertility rate of ages 20 to 24} &= 1000 \times \frac{\text{no. of live births to females aged 20-24}}{\text{number of females aged 20-24}} \\ &= \frac{1000 \times 3008}{20000} = 150.40 \text{ per 1000} \end{aligned}$$

4.

$$\begin{aligned} \text{infant morality rate} &= 1000 \times \frac{\text{number of deaths under 1 year age}}{\text{number of live births}} \\ &= \frac{1000 \times 210}{10122} = 20.75 \text{ per 1000 live births} \end{aligned}$$

5.

$$\begin{aligned} \text{neonatal morality rate} &= 1000 \times \frac{\text{deaths aged under 28 days}}{\text{number of live births}} \\ &= \frac{1000 \times 126}{10122} = 12.45 \text{ per 1000 live births} \end{aligned}$$

6.

$$\begin{aligned} \text{postneonatal morality rate} &= 1000 \times \frac{\text{deaths aged under 28 days and 1 year}}{\text{number of live births}} \\ &= \frac{1000 \times (210 - 126)}{10122} = 8.3012 \text{ per 1000 live births} \end{aligned}$$

7.

$$\begin{aligned} \text{stillbirth rate} &= 1000 \times \frac{\text{number of stillbirths}}{\text{total live births +stillbirths}} \\ &= \frac{1000 \times 200}{200 + 10122} = 19.38 \text{deaths per 1000 births} \end{aligned}$$

8.

$$\begin{aligned} \text{perinatal mortality rate} &= 1000 \times \frac{\text{number of still births+deaths under 1 week}}{\text{total births live births +still births}} \\ &= \frac{1000 \times (200 + 106)}{200 + 10122} = 29.65 \text{deaths per 1000 births} \end{aligned}$$

9.

$$\text{Material mortality rate} = 1000 \times \frac{\text{number of maternal deaths}}{\text{total live births +still births}}$$
$$= \frac{3000}{200 + 10122} = 0.29\text{deaths per 1000 births}$$

6. Ratio estimators are appropriate
(i)Estimate of total sugar content is $N\bar{y}$,where $\bar{y}$ is the mean sugar content in the random sample of n oranges. A measurement of the sugar content of each sampled orange is required.

(b) With the given assumptions ,$\hat{\tau}_y = \frac{\bar{y}}{\bar{x}}\tau_x$ where $\bar{x}$is the mean weight of the sample oranges whose mean sugar content is $\bar{y}$. For each sampled orange, its sugar content y and weight x must be measured.

(ii) $var(\gamma) = E[(\gamma - R)^2]$ where R is the population value of γ,i.e.$\frac{\mu_y}{\mu_x}$ writing f=n/N, the estimated variance of a mean of any variate say z,from a finite population is$(1-f)s_z^2/n$
Now

$$\gamma - R = \frac{\bar{y}}{\bar{x}} - R = \frac{\bar{y} - R\bar{x}}{\bar{x}} \approx \frac{\bar{y} - R\bar{x}}{\mu_x}$$

if n is reasonably large; i.e. $\gamma - R \doteq \frac{1}{n\mu_x}\sum_{i=1}^{n}(y_i - Rx_i)$ where (x_i, y_i) are measured on the i^{th} sample member.Also

$$Var(\gamma) \doteq \frac{1}{\mu_x^2}E[(\bar{y} - R\bar{x})^2] = \frac{1-f}{n\mu_x^2}\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1}$$

(iii)

$$v[\hat{\gamma_y}] = \tau_x^2 v[\frac{\bar{y}}{\bar{x}}] = \tau_x^2 v[\gamma] = \frac{\tau_x^2(1-f)}{\mu_x^2 n(N-1)}\sum_{i=1}^{N}(y_i - Rx_i)^2$$

Usually the sum is taken as $\sum_{i=1}^{n}$ over the sample values, which is a further approximation that is acceptable in reasonable sample sizes. However in the present example we are given extra information and need not make this approximation.
If there is a good positive correlation between x and y as we are told here, then ratio estimation are more precise than estimates based on y alone. strictly, $p > \frac{1}{2}$ is needed for $v[\bar{y}_R]$ to be $< v[\bar{y}]$
(iv) $\tau_x = 1800$ $\sum_{i=1}^{N}\frac{(y_i - Rx_i)^2}{N-1} = (0.0030)^2$ $\tilde{x} = 0.4$ we require $v[\hat{\tau_y}] \leq 3^2$ (approximately).Assume f negligible Hence $9 \geq \frac{1800^2}{(0.4)^2}\frac{1}{n}(0.0030)^2$ or $3\sqrt{n} \geq \frac{1800 \times 0.0030}{0.4} = 13.5$

giving $\sqrt{n} = 4.5$ and so n=20.25 sample about 20 or 21

7(a)When a population is divided into groups or strata, and a (simple) random sample is taken independently from each stratum, the process is called stratified random sampling. Proportional allocation is when the sampling fraction f, the same for all strata,and optimal allocation is when the stratum sample sizes $\{n_i\}$ are chosen to satisfy conditions such as minimizing the variance of the estimator $\bar{y}_{st}$ for total cost fixed at C, or minimizing total cost for a given target value of variance.

(b)$v_{ran} = (1-f)\frac{s^2}{n}$ In general, the variance is stratified sampling is

$$v(\bar{y}_{sr}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h(N_h - n_h)\frac{s_h^2}{n_h} = \sum(1-f_h)w_h^2\frac{s_h^2}{n_h}$$

with proportional allocation $\frac{n_h}{N_h} = \frac{n}{N}$, i.e. $w_h = \frac{N_h}{N} = \frac{n_h}{n}$; $f_h = f$ Thus

$$v_{prop} = (1-f)\sum_{i=1}^{L}\frac{w_h s_h^2}{n} = \sum_{i=1}^{L}\frac{w_h s_h^2}{n} - \sum_{i=1}^{L}\frac{w_h s_h^2}{N}$$

Using the subdivision of sum of squares as in analysis of variance

$$\begin{aligned}(N-1)s^2 &= \textstyle\sum_h \sum_i N_h(y_{h_i} - \bar{y})^2 \\ &= \textstyle\sum_h \sum_i N_h(y_{h_i} - \bar{y_h})^2 + \sum_h N_h(\bar{y_h} - \bar{y})^2 \\ &= \textstyle\sum_h (N_h - 1)s_h^2 + \sum_h N_h(\bar{y_h} - \bar{y})^2\end{aligned}$$

Therefore

$$\begin{aligned}v_{ran} &= \tfrac{(1-f)}{n(N-1)}[\textstyle\sum_h (N_h - 1)s_h^2 + \sum_h N_h(\bar{y_h} - \bar{y})^2] \\ &= v_{prop} - \tfrac{1-f}{n}\textstyle\sum_h w_h s_h^2 + \tfrac{1-f}{n}[\frac{\sum_h (N_h-1)s_h^2}{N-1} + \frac{\sum_h N_h(\bar{y_h}-\bar{y})^2}{N-1}] \\ &= v_{prop} + \tfrac{1-f}{n(N-1)}\{\textstyle\sum_h N_h(\bar{y_h} - \bar{y})^2 + \sum_h[(N_h - 1) - \frac{N-1}{N}N_h]s_h^2\} \\ &= v_{prop} + \tfrac{1-f}{n(1-N)}[\textstyle\sum_h N_h(\bar{y_h} - \bar{y})^2 - \frac{1}{N}\sum_h (N - N_h)s_h^2]\end{aligned}$$

(c)(i)Optimum allocation uses $n_h = n\frac{N_h s_h}{\sum_{h=1}^{L} N_h s_h}$ $\quad n = 100$
Values of N_h s_h are 3270.2 6131.3 5904.1 6613.2 4140.5 2938.0 and 5209.6,so $\sum N_h s_h = 34206.9$. The values of n_h, to the nearest integer, are 10 18 17 19 12 9 15.

(ii)Proportional allocation has $n_h = 100N_h/2010$,and these values are 20 23 19 17 8 6 7. The minimum variance is the value of $v(\bar{y}_{st})$ for

$$n_h = \frac{nN_h s_h}{\sum N_h s_h} = \sum_{h=1}^{L} w_h^2 s_h^2/n_h - \frac{1}{N}\sum_h w_h s_h^2$$

This reduces to

$$\sum_{h=1}^{L} \frac{(w_h s_h)^2}{n w_h s_h}(\sum w_h s_h) - \frac{1}{N}\sum_{h=1}^{L} w_h s_h^2 = \frac{1}{n}(\sum w_h s_h)^2 - \frac{1}{N} w_h s_h^2$$

The value of this is $\frac{17.0183^2}{100} - \frac{343.2788}{2010} = 2.725$.
For proportional allocation, variance is $\frac{1-f}{n}\sum w_h s_h^2$
which is $\frac{1910}{2010} \times \frac{343.2788}{100} = 3.262$

$$\begin{aligned}
v_{ran} = \quad & 3.262 + \tfrac{1910}{100\times 2010\times 2009}[(394\times 20.9^2) + (461\times 10.0^2) + (391\times 2.0^2) \\
& +(334\times 8.2^2) + (169\times 15.8^2) + (113\times 23.8^2) + (148\times 37.5^2) \\
& -\tfrac{1}{2010}\{(1616\times 8.3^2) + (1549\times 13.3^2) + (1619\times 15.1^2) \\
& +(1676\times 19.8^2) + (1841\times 24.5^2) + (1897\times 26.0^2) + (1862\times 35.2^2)\}] \\
= \quad & 3.262 + 4.73\times 10^{-6}[556547.18 - 6106060.81/2010] \\
= \quad & 3.262 + 2.618 = 5.880
\end{aligned}$$

Relative efficiencies for optimum and proportional compared with random are $\frac{5.880}{2.725} = 216\%$ and $\frac{5.880}{3.262} = 180\%$ respectively.

8(i)

$$\sum_{i,j} y_{ij} = 165.06, \quad n = 40, \quad \bar{y} = 4.1265kg/plot.$$

Hence an estimate of wheat yield perhectare is $\hat{y} = 16 \times 4.1265 = 66.0kg$
$s_b^2, s_w^2$are variances between and within fields, f_1 is the sampling fraction for fields and f_2 for plots within frields Also $n = 10$ and $m = 4$; $N = 100$ and $M = 16$.

$$s_b^2 = \frac{1}{n-1}\sum_i (\bar{y}_i - \bar{y})^2$$

and

$$s_w^2 = \frac{1}{n(m-1)}\sum_i\sum_j (y_{ij} - \bar{y}_i)^2$$

i=1 to n, j=1 to m

$Field$	$\sum y_{ij}$	$\bar{y}_i$	$\sum y_{ij}^2$	s_i^2
1	15.16	3.790	58.7184	0.420667
2	17.50	4.375	77.6052	0.347567
3	17.14	4.285	73.8004	0.118500
4	18.20	4.550	83.9504	0.389133
5	14.54	3.635	53.5108	0.219300
6	16.28	4.070	67.0088	0.249733
7	16.88	4.220	71.9264	0.230933
8	15.18	3.795	59.7332	0.708367
9	17.02	4.255	73.5652	0.381700
10	17.16	4.290	74.0912	0.158267
	165.06		693.910	

$$s^2 = \frac{1}{39}(693.91 - \frac{165.06^2}{40}) = 0.327946$$

$$v_b^2 = \frac{1}{9}\sum_{i=1}^{10}(\bar{y}_i - 4.1256)^2 = 0.087345$$

$$s_w^2 = \frac{1}{30}\sum_{i=1}^{10}\sum_{j=1}^{4}(y_{ij} - \bar{y}_i)^2 = \frac{1}{10}\sum_{i=1}^{10} s_i^2 = 0.321517$$

Hence

$$\hat{v}(\bar{y}) = \frac{0.9}{10} \times 0.087345 + 0.1 \times \frac{0.75}{40} \times 0.321517 = 0.0084639$$

so that $\hat{SE}(\bar{y}) = 0.0920$, SE of estimate= $SE(16\bar{y}) = 1.472$

(ii)For random sampling

$$v(\bar{y}) = (1 - \frac{400}{1600})(\frac{1}{40})(0.327946) = 0.007994$$

The ratio of variance multistage: $random = \frac{0.008464}{0.007994} = 1.0588$ giving relative efficiency $\frac{1}{1.0588} = 0.9445$ or 94.45%

(iii)Total cost $c = 4n + mn$, which must be $\leq 100 units$. The theoretical variance of $\bar{y}$ (whose unbiased estimate is as given in (i)) is

$$v = \frac{(1-f_1)s_B^2}{n} + \frac{(1-f_2)s_w^2}{mn}$$

where s_B^2 and s_w^2 are the population values of s_b^2 and s_w^2.
A lagrange method will minimize $v + \lambda(100 - 4n - mn) \equiv L$ say

$$L = s_B^2(\frac{1}{n} - \frac{1}{N}) + s_w^2(\frac{1}{mn} - \frac{1}{Mn}) = \lambda(100 - 4n - mn)$$

$$\frac{\partial L}{\partial n} = -\frac{s_B^2}{n^2} - \frac{s_w^2}{mn^2} + \frac{s_w^2}{Mn^2} - 4\lambda - m\lambda = 0$$

for max or in

$$\frac{\partial L}{\partial m} = \frac{s_w^2}{m^2 n} - \lambda n = 0 \quad when \quad -\lambda = \frac{s_w^2}{m^2 n^2}$$

First equation becomes

$$\frac{(4+m)s_w^2}{m^2 n^2} = \frac{s_B^2}{n^2} + \frac{s_w^2}{mn^2} - \frac{s_w^2}{Mn^2}$$

or

$$\frac{s_w^2(4+m-m)}{m^2} + \frac{s_w^2}{M} = s_B^2 \quad giving \quad m^2 = \frac{4}{\frac{s_B^2}{s_w^2} - \frac{1}{M}}$$

setting $c = 100$, i.e. $(4+m)n = 100$ will give the value of n.

An unbiased estimator of s_B^2 is $s_b^2 - \frac{(1-f_2)}{m}s_w^2$, whose value is $0.087345 - \frac{0.75}{4} \times 0.321517 = 0.027061$ Hence

$$m^2 = 4/(\frac{0.027061}{0.321517} - \frac{1}{16}) = 184.627 \quad i.e \; m = 13.59 \quad n = 5.69$$

Rounding to the nearest integer,the choice is between :

	$n=5$	$n=6$
$m=13$	85	102
$m=14$	90	108

Since $cost = (4+m)n$ can not be > 100, we must use $m = 14 \quad n = 5$, that is use 5 fields and select 14 plots from each