S3 June 2005 (MA)

Q1a) When the population divides naturally into mutually exclusive groups.

b) ADV : enables fieldwork to be done quickly as a representative Sample is easily achieved.

OR : - costs kept to a minimum.

- administering the test is easy.

DISADV : unable to estimate Sampling errors.

OR : - interviewer bias may surface with who is selected.

- non-responses aren't recorded.

Q2) $X \sim N(10, 3^2)$. then $\bar{X} \sim N(10, \frac{3^2}{5})$ by C.L.T.

$$P(\text{required}) = P(7 < \bar{X} < 10) = P\left(\frac{7-10}{\sqrt{\frac{9}{5}}} < Z < \frac{10-10}{\sqrt{\frac{9}{5}}}\right)$$

$$= P(-2.24 < Z < 0)$$

$$= P(Z<0) - P(Z<-2.24)$$

$$= 0.5 - [1 - P(Z < 2.24)]$$

$$= 0.5 - [1 - 0.9875] = \boxed{0.4875}$$

Q3) H_0: There is no association between the treatment of trees and their survival.

H_1: There is an association between the treatment of trees and their survival.

$$\text{Expected no.} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

Expected

	No. action	Remove ...	Spray	
Died within 1 year	7	7	7	21
Survived for 1-4 years	7	7	7	21
survived beyond 4 years	6	6	6	18
	20	20	20	$\boxed{60}$

O_i	E_i	$\frac{(O-E)^2}{E}$
10	7	$\frac{9}{7}$
5	7	$4/7$
6	7	$1/7$
5	7	$4/7$
9	7	$4/7$
7	7	0
5	6	$1/6$
6	6	0
7	6	$1/6$
		3.476

$$\therefore X^2 = \sum \frac{(O-E)^2}{E} = 3.476.$$

$\nu = (3-1)(3-1) = 4$

$\therefore$ critical value $= \chi^2_4(5\%) = 9.488.$

$3.476 < 9.488$ $\therefore$ Result is insignificant (accept H_0)

Evidence suggests no association between treatment of trees and their survival.

Q4a) see mark scheme.

b) The linear relationship between two variables.

c) $S_{xx} = S_{ss}$ (error in question).

$$S_{ss} = 141.51 - \frac{(33.9)^2}{10} = \boxed{26.589} \quad (\Sigma s = 33.9)$$

$$S_{dd} = 1081.74 - \left(\frac{(96.4)^2}{10}\right) = \boxed{152.444} \quad (\Sigma d = 96.4)$$

$$S_{sd} = 386.32 - \frac{(33.9)(96.4)}{10} = \boxed{59.524}$$

d) $$r = \frac{S_{sd}}{\sqrt{S_{ss}\,S_{dd}}} = \frac{59.524}{\sqrt{26.589 \times 152.444}} = \boxed{0.935}$$

c) $H_0: \rho = 0$ critical value : 0.7155
$H_1: \rho > 0$ (1%, 1-tail)

testing for a +ve correlation

0.935 > 0.7155
Result is significant. (reject H_0).
Evidence suggests a positive correlation between s and d.

d) there is (strong) evidence of a positive linear correlation but the scatter diagram does not reflect this.

Q5a) H_0: Poisson distribution is a suitable model.
H_1: Poisson distribution is not a suitable model.

let $X \sim Po(\lambda)$ where X = no. of restarts.

$$\hat{\lambda} = \frac{0(99) + 1(65) + 2(22) + 3(12) + 4(2)}{200}$$

$$= \boxed{0.765}$$

Expected values

recorded for 200 days!

$$E_0 = P(X=0) \times 200 = 200\left[\frac{e^{-0.765}(0.765^0)}{0!}\right]$$

$$= 93.07$$

similarly, $E_1 = 200P(X=1) = 71.20$
$E_2 = 200P(X=2) = 27.23$
$E_3 = 200P(X=3) = 6.94$
$E_4 = 200P(X=4) = 1.56$

need to pool as $E_i > 5$. is required for all cells.

remember, ($E_i > 5$) is required for the test statistic X^2 to be approximated well by the chi-squared distribution(s).

pooling the final two cells to give our final table:

No. of restarts	0	1	2	3 - 4
O_i	99	65	22	14
E_i	93.07	71.20	27.23	8.50
$\frac{(O-E)^2}{E}$	0.378	0.540	1.00	3.56

$$X^2 = \sum \frac{(O-E)^2}{E} = 5.48$$

(anything between 5.4 and 5.5 will be fine)

we estimated λ and pooled two cells together so subtract 2 from total number of cells (4) to give degrees of freedom.

$$\nu = 4 - 1 - 1 = 2$$

$$c.v = \chi^2_2(5\%) = 5.991$$

$5.48 < 5.991$

$\therefore$ Result is insignificant.

Evidence suggests that Poisson is a suitable model for these data.

Q6a) mean $= \frac{\sum x}{n} = \frac{205 + 310 + 405 + 195 + 320}{5}$

$= \boxed{287}$

$s^2 = \frac{1}{4}\left(\sum x^2 - \frac{(\sum x)^2}{5}\right)$

$\left.\begin{array}{l}\sum x = 1435 \\ \sum x^2 = 442575\end{array}\right\}$ $s^2 = \frac{1}{4}\left(442575 - \frac{1435^2}{5}\right)$

$= \boxed{7682.5}$

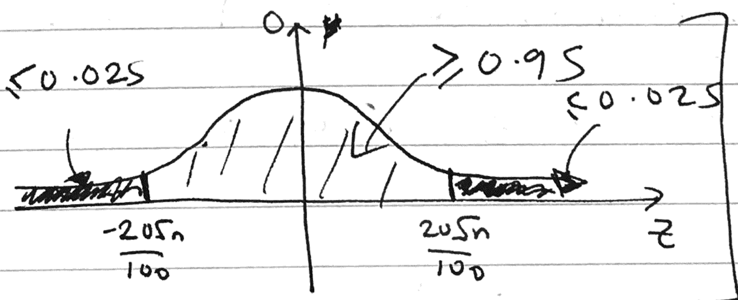
b) $\bar{X}$ = estimate of population mean.

we want: $P(\mu - 20 < \bar{X} < \mu + 20) \geq 0.95$

using C.L.T, $\bar{X} \sim N\left(\mu, \frac{100^2}{n}\right)$

$\Rightarrow P\left(\frac{\mu - 20 - \mu}{\frac{100}{\sqrt{n}}} < Z < \frac{\mu + 20 - \mu}{\frac{100}{\sqrt{n}}}\right) \geq 0.95$

$\Rightarrow P\left(\frac{-20\sqrt{n}}{100} < Z < \frac{20\sqrt{n}}{100}\right) \geq 0.95$



if the middle area ≥ 0.95 then each of the 'side' areas are ≤ 0.025

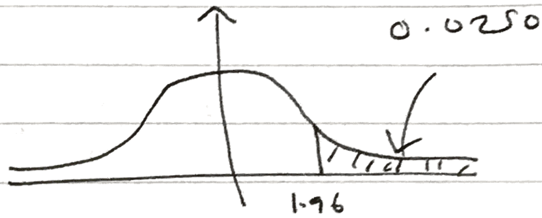
$\Rightarrow$ this means $P\left(Z > \frac{\sqrt{n}}{5}\right) \leq 0.025$

since this region is symmetrical.

$$P\left(Z > \frac{\sqrt{n}}{5}\right) \leq 0.0250$$

but $P(Z > 1.96) = 0.0250$

hence $\frac{\sqrt{n}}{5} \geq 1.96$



$\sqrt{n} \geq 5(1.96)$

$n \geq 96.04\ldots$

$\therefore \boxed{n_{min} = 97}$ (next whole number)

Q7a) $C \sim N(350, 8)$

$P(|C_1 - C_2| > 6)$

$= P(C_1 - C_2 > 6) + P(C_1 - C_2 < -6)$

$= 2P(C_1 - C_2 > 6)$

$\Rightarrow C_1 - C_2 \sim N(0, 16)$

so $2P(C_1 - C_2 > 6) = 2P\left(Z > \frac{6}{4}\right)$

$= 2P(Z > 1.50) = 2(1 - 0.9332) = \boxed{0.1336}$

b) $P(C > L) = P(C - L > 0)$

$C - L \sim N(5, 25)$

$\boxed{\begin{array}{l} E(C-L) = 350 - 345 = 5 \\ Var(C-L) = 17 + 8 = 25 \end{array}}$

$\Rightarrow P(C - L > 0) = P\left(Z > \frac{-5}{\sqrt{25}}\right) = P(Z < 1)$

$= \boxed{0.8413}$

c) $C_1 + \ldots + C_{24} + B = X$

then $E(X) = 24(350) + 100 = 8500$

$Var(X) = 24(8) + 2^2 = 196$

remember $C_1 + \ldots + C_{24}$ is not the same as $24C$ so we do not square the 24. Here we are taking 24 samples of C, not $24 \times$ one sample of C.

$\therefore X \sim N(8500, 196)$

$P(\text{required}) = P(8.510 < X < 8520)$

$= P\left(\frac{8510-8500}{\sqrt{196}} < Z < \frac{8520-8500}{\sqrt{196}}\right)$

$= P(0.714 < Z < 1.429)$

$= P(0.71 < Z < 1.43)$

$= P(Z < 1.43) - P(Z < 0.71)$

$= 0.9236 - 0.7611 = \boxed{0.1625}$

d) All random variables are independent of each other. (ie all cans of cola are independently selected) This is the basis upon which we can use the expectation / variance formulae for combining random variables.