S3 June 2009 (MA)

Q1a) $k = \frac{50000}{100} = 500.$

So by random numbers pick the first name between 1 and 500. Then pick every 500^{th} name afterwords.

bi) ADV : Costs kept to a minimum
OR :- administering the test is easy.

DISADV : not possible to estimate sampling errors
OR : - not a random process
- non-responses aren't recorded.

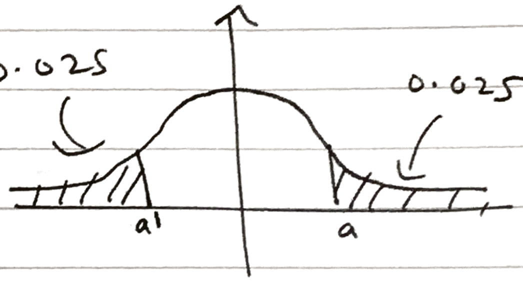
ii) ADV : simple to use
OR : suitable for large samples.

DISADV : - only random if the ordered list is truly random.
OR : - requires a list of entire population.

[NOT TO BE CONFUSED WITH CONFIDENCE INTERVALS!]

Q2a) heights are normally distributed...

$H \sim N(20.1, 0.5^2)$



$P(H > a) = 0.025$

$P\left(Z > \frac{a - 20.1}{\sqrt{0.5^2}}\right) = 0.025$

from tables, $\frac{a - 20.1}{0.5} = 1.96$

$\therefore \; a = 1.96(0.5) + 20.1 = 21.08$ = upper limit

lower limit(a') $= 20.1 - (21.08 - 20.1) = 19.12$ //

so limits are : $[19.12, 21.08]$

b) 98% C.I $= \left[\bar{x} \pm 2.3263\left(\frac{\sigma}{\sqrt{n}}\right)\right]$

$\Rightarrow \left[20.1 \pm 2.3263\left(\frac{0.5}{\sqrt{10}}\right)\right]$

$\Rightarrow [19.7, 20.5]$

c) Claim is not correct as 19.5 is outside the confidence interval.

(Position is already a rank)

Q3a)

Individual	BMI	Rank	Position	d	d^2
A	17.4	1	3	2	4
B	21.4	6	5	1	1
C	18.9	3	1	2	4
D	24.4	8	9	1	1
E	19.4	4	6	2	4
F	20.1	5	4	1	1
G	22.6	7	10	3	9
H	18.4	2	2	0	6
I	25.8	9	7	2	4
J	28.1	10	8	2	4
					32

$\sum d^2 = 32$

$\therefore r_s = 1 - \frac{6(32)}{10(99)} = \boxed{0.806}$

b) $H_0 : p = 0$

$H_1 : p > 0$

critical value $= \pm 0.5636$ (5%, 1-tail)

not $p < 0$ because that would mean the higher the BMI, the better the finishing position.

$0.806 > 0.5636$

$\therefore$ Result is significant (reject H_0).

Evidence suggests that the doctor's belief is correct.

c) Position is given as a rank and is not normally distributed.

Q4) $X \sim N(55, 3^2)$

By C.L.T, $\bar{X} \sim N(55, \frac{3^2}{8})$

$P(\text{required}) = P(\bar{X} > 57) = P\left(Z > \frac{57-55}{\frac{3}{\sqrt{8}}}\right)$

$= P(Z > 1.89) = 1 - P(Z < 1.89)$

$= 1 - 0.9706$

$= \boxed{0.0294}$

Q5a) $\text{mean} = \dfrac{0(40) + 1(33) + 2(14) + 3(8) + 4(5)}{100}$

$= \boxed{1.05}$ $\left(= \dfrac{\text{total goals}}{\text{total games}} \right)$

b) $r = 100 \times P(X=1) = 100\left[\dfrac{e^{-1.05}(1.05^1)}{1!}\right]$

$\boxed{X \sim Po(1.05)}$

$= \boxed{36.743}$

$s = 100 \times P(X=2) = 100\left[\dfrac{e^{-1.05}(1.05^2)}{2!}\right]$

$= \boxed{19.290}$

We multiply by 100 as there are 100 games played.

c) $\underline{H_0}$: Poisson ($\lambda = 1.05$) is a suitable model for these data.

$\underline{H_1}$: Poisson ($\lambda = 1.05$) is not a suitable model for these data.

No of goals	0	1	2	≥ 3
O_i	40	33	14	13
E_i	34.994	36.743	19.290	8.972
$\frac{(O-E)^2}{E}$	0.7161	0.3813	1.4507	1.8084

Pool last 2 cells as expected frequency must be greater than 5 for the test statistic X^2 to be approximated well by the chi-squared distribution(s).

$$X^2 = \sum \frac{(O-E)^2}{E} = 4.36$$

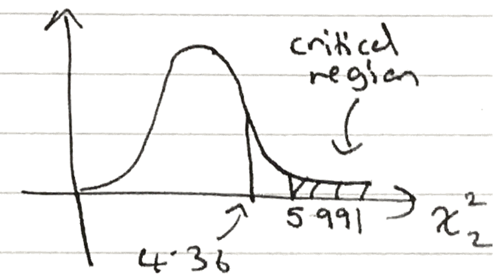
degrees of freedom $= \nu = 4 - 1 - 1 = 2$

−2 as there are 4 cells after pooling and we calculated λ.

critical value $= \chi^2_2(5\%) = 5.991$

$4.36 < 5.991$

$\therefore$ Result is insignificant. Evidence suggests that Poisson ($\lambda = 2$) is a suitable model for these data.



let μ_1 indicate mean for upper shore limpets and μ_2 indicate mean for lower shore limpets.

(Q6a) $H_0: \mu_1 = \mu_2$ critical value: ± 1.6449 (5%, 1-tail)

$H_1: \mu_1 < \mu_2$

$$\text{Test stat} = \frac{\bar{x} - \bar{y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}}$$

$$= \frac{5.05 - 4.97 - (0)}{\sqrt{\frac{0.42^2}{120} + \frac{0.67^2}{150}}}$$

$$= 1.1975$$

$1.1975 < 1.6449$

$\therefore$ Result is insignificant (accept H_0)
Evidence suggests there is a difference in mean length of limpets from upper / lower shores.

b) - Sample variance = population variance.
(by definition, test statistic uses population variance

- Samples are independent of each other.

$$\begin{bmatrix} \sum x = 600.9 \\ \sum x^2 = 72216.31 \end{bmatrix}$$

Q7a) $\text{mean} = \dfrac{\sum x}{n} = \dfrac{600.9}{5} = \boxed{120.18}$

variance estimate $= s^2 = \dfrac{1}{n-1}\left(\sum x^2 - \dfrac{(\sum x)^2}{n}\right)$

$$= \frac{1}{4}\left(72216.31 - \frac{(600.9)^2}{5}\right)$$

$$= \boxed{0.037}$$

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

$$P(\text{required}) = \left[P(\mu - 0.05 < \bar{L} < \mu + 0.05) > 0.90\right]$$

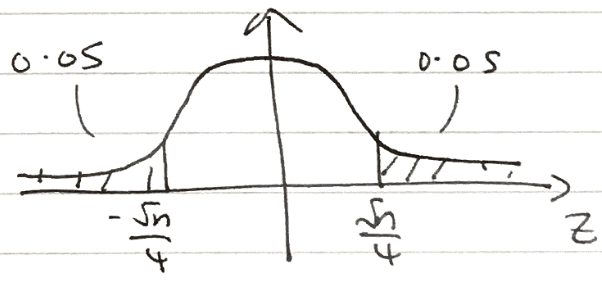
this is what we want.

now by C.L.T, $\bar{L} \sim N\left(\mu, \dfrac{0.2^2}{n}\right)$ (for a sample of n)

$$\Rightarrow P\left(\frac{\mu - 0.05 - \mu}{\frac{0.2}{\sqrt{n}}} < Z < \frac{\mu + 0.05 - \mu}{\frac{0.2}{\sqrt{n}}}\right) > 0.90$$

$$\Rightarrow P\left(-\frac{\sqrt{n}}{4} < Z < \frac{\sqrt{n}}{4}\right) > 0.90$$

Looking at the bell curve, we can see that:



$$P\left(Z > \frac{\sqrt{n}}{4}\right) \le 0.05$$ //

$P\left(Z > \frac{\sqrt{n}}{4}\right) \leq 0.05$

[as the unshaded area increases, the shaded area decreases hence $P(Z > \frac{\sqrt{n}}{4})$ will be less than (or equal to) 0.05]

this must mean: $\frac{\sqrt{n}}{4} \geq 1.6449$

(the z-value corresponding to $p = 0.05$ is 1.6449)

$\sqrt{n} \geq 1.6449(4)$

hence $n \geq (4 \times 1.6449)^2$

$n \geq 43.29\ldots$

$\boxed{n_{min} = 44}$ (next integer).

Q8a) $E(A) = 4E(X) - 3E(Y)$

$= 4(30) - 3(20) = \boxed{60}$

b) $Var(A) = 4^2 Var(X) + 9 Var(Y)$

$= 16 Var(X) + 9 Var(Y)$

$= 16(9) + 9(4) = \boxed{180}$

c) $B = Y_1 + Y_2 + Y_3 + Y_4$

$P(B > A) = P(Y_1 + Y_2 + Y_3 + Y_4 > A)$

$\Rightarrow P(Y_1 + Y_2 + Y_3 + Y_4 - A > 0)$ //

let $C = Y_1 + Y_2 + Y_3 + Y_4 - A$,

$Y \sim N(20, 2^2)$
$A \sim N(60, 180)$

$E(C) = 4(20) - 60 = 20$

$Var(C) = 4(4) + 180 = 196$

$\therefore P(\text{required}) = P(C > 0) = P\left(Z > \frac{-20}{\sqrt{196}}\right)$

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

$= \boxed{0.923}$