Statistics Revision 2 Solutions

1(i) Let the random variables X and Y denote the number of cups of coffee and tea sold in a minute respectively.

Then $X \sim P_o(1.5)$ and $Y \sim P_o(0.5)$

$P(X=1) \bullet P(Y=1) = 0.123$ (shown)

(ii) Let the random variables X' and Y' denote the number of cups of coffee and tea sold in a minute respectively.

Then $X' \sim P_o(4.5)$, $Y' \sim P_o(1.5)$ and $X'+Y' \sim P_o(6)$

$P(X'+Y'<5) = P(X'+Y' \leq 4) = 0.285$ (shown)

(iii) $X+Y \sim P_o(2)$

P(exactly 3 cups are sold in one minute) $= P(X+Y=3) = 0.180$

P (3 cups of coffee and 0 cups of tea are sold in one minute)

$= P(X=3) \bullet P(Y=0) = 0.076$

$\therefore$ P (all drinks sold in one minute are coffee | exactly 3 drinks were sold in one minute)

$= \dfrac{0.076}{0.18} = 0.421$ (shown)

2. $X \sim B(12,\ 0.8)$

$$\frac{P(X=r+1)}{P(X=r)} = \frac{(n-r)}{(r+1)}\frac{p}{q} = \frac{(12-r)(0.8)}{(r+1)(0.2)} = \frac{9.6-0.8r}{0.2r+0.2}$$

$$\frac{P(X=r+1)}{P(X=r)} > 1 \Rightarrow \frac{9.6-0.8r}{0.2r+0.2} > 1, \text{ ie } r < 9.4$$

Hence, $P(X=10) > P(X=9) > P(X=8) \ldots\ldots\ldots\ldots > P(X=0) - - - - - - - (1)$

$$\frac{P(X=r+1)}{P(X=r)} < 1 \Rightarrow \frac{9.6-0.8r}{0.2r+0.2} < 1, \text{ ie } r > 9.4$$

Hence, $P(X=12) < P(X=11) < P(X=10) - - - - - - - (2)$

Reconciling (1) and (2) gives most probable value $=10$ people (shown)

3. $X \sim P_o(3.4)$

$$\frac{P(X=r+1)}{P(X=r)} = \frac{\lambda}{r+1} = \frac{3.4}{r+1}$$

$$\frac{P(X=r+1)}{P(X=r)} > 1 \Rightarrow \frac{3.4}{r+1} > 1, \text{ ie } r < 2.4$$

Hence, $P(X=3) > P(X=2) > P(X=1) > P(X=0) - - - - - - - - - - (1)$

$$\frac{P(X=r+1)}{P(X=r)} < 1 \Rightarrow \frac{3.4}{r+1} < 1, \text{ ie } r > 2.4$$

Hence, $P(X=3) > P(X=4) > P(X=5) > \ldots\ldots\ldots\ldots - - - - - - - - - - - - (2)$

Reconciling (1) and (2) gives most probable value of $X = 3$ (shown)

4(i) Let the random variable X denote the number of broken eggs in a box of 500.

Then $X \sim B(500,\ 0.008)$

$P(X=3) = 0.196$ (shown)

(ii) P (a single box containing no broken eggs) $= P(X=0) = 0.018$

Let the random variable Y denote the number of boxes(out of 100) which contain no broken eggs.

Then $Y \sim B(100,\ 0.018)$

$P(Y=4) = 0.072$ (shown)

5. Let the random variables X and Y denote the mass of sugar in 1kg and 0.5kg labelled bags respectively.

Then $X \sim N(1005,\ 2^2)$ and $Y \sim N(505,\ 2^2)$

(i) $P(X < 1000) = 0.00621$ (shown)

(ii) $Y_1 + Y_2 \sim N(1010,\ 8)$

$P(Y_1 + Y_2 < 1000) = 2.035 \times 10^{-4}$ (shown)

(iii) $Y_1 + Y_2 - X \sim N(5,\ 12)$

$P(Y_1 + Y_2 > X) = P(Y_1 + Y_2 - X > 0) = 0.926$ (shown)

$X_g = X_1 + X_2 + ... + X_{10} \sim N(10050,\ 40)$

$P(X_g > m) = 0.75 \rightarrow P(X_g < m) = 0.25$

$$P\left(Z < \frac{m - 10050}{\sqrt{40}}\right) = 0.25$$

$$\therefore \quad \frac{m - 10050}{\sqrt{40}} = invNorm(0.25) = -0.674 \Rightarrow m = 10045.734 \text{ (shown)}$$

6 (i) Let the random variables X and Y denote the examination marks obtained by boys and girls respectively.

Then $X \sim N(55, 11^2)$ and $Y \sim N(58, 8^2)$

$Y - X \sim N(3,\ 185)$

$P(Y > X) = P(Y - X > 0) = 0.587$ (shown)

(ii) $P(Y - X \geq 20) = 0.106$ (shown)

(iii) $P(|Y - X| > 20) = P(Y - X > 20) + P(Y - X < -20) = 0.151$ (shown)

(iv) $\dfrac{X + Y}{2} \sim N\left(\dfrac{55 + 58}{2} = 56.5, \dfrac{1}{4} \times 11^2 + \dfrac{1}{4} \times 8^2 = 46.25\right)$

$$P\left(\frac{X + Y}{2} > 70\right) = 0.024 \text{ (shown)}$$

(v) $Y - 2X \sim N(58 - 55 \times 2 = -52,\ \ 8^2 + 11^2 \times 4 = 548)$

$P(Y \geq 2X) = P(Y - 2X \geq 0) = 0.013$ (shown)

7(i) Let the random variable X denote the weight of a person.

Then $X \sim N(70, 10^2)$, $\ T = X_1 + X_2 + X_3 + X_4 \sim N(280,\ 400)$

$P(T > 300) = 0.159$ (shown)

(ii) $P(X + 3X > 300) = P(4X > 300) = P(X > 75) = 0.309$ (shown)

8. P(getting a 7 from a single throw of a pair of dice)

$= P$(combination is 6,1) $+ P$(combination is 5, 2) $+ P$(combination is 4, 3)

$$= 2\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right)\left(\frac{1}{6}\right) = \frac{1}{6}$$

Let the random variable X denote the number of 7s obtained in 100 throw of a pair of dice.

Then $X \sim B(100, \frac{1}{6})$

Since $np = \frac{100}{6} > 5$, $nq = \frac{500}{6} > 5$ and $n = 100$ is large,

$X \sim N(\frac{100}{6}, \frac{125}{9})$ approx and $P(X > 25) = P(X > 25.5) = 0.0089$ (shown)

(Note that a direct utilisation of the original binomial distribution **without any approximation** to solve for the required probability is also acceptable.)

Let the number of tosses be n

Then we have $1 - \left(1 - \frac{1}{6}\right)^n \geq 0.9 \rightarrow 1 - \left(\frac{5}{6}\right)^n \geq 0.9$

$$\left(\frac{5}{6}\right)^n \leq 0.1 \Rightarrow n \geq \frac{\ln(0.1)}{\ln\left(\frac{5}{6}\right)} = 12.63, \text{ ie } n = 13 \text{ since it is discrete (shown)}$$

9. Let the random variable X denote the number of forms used in a year.

Then $X \sim B(250, \frac{1}{3})$

Since $np = \frac{250}{3} > 5$, $nq = \frac{500}{3} > 5$ and $n = 250$ is large,

$X \sim N(\frac{250}{3}, \frac{500}{9})$ approx

$$P(X < n) = 0.95 \rightarrow P(X < n - 0.5) = 0.95 \rightarrow P\left(Z < \frac{n - 0.5 - \frac{250}{3}}{\sqrt{\frac{500}{9}}} \right) = 0.95$$

$$\frac{n - 0.5 - \frac{250}{3}}{\sqrt{\frac{500}{9}}} = invNorm(0.95) = 1.645 \Rightarrow n = 96.1$$

Hence, **97** forms must be kept till ensure the 95% probability criteria is met. (shown)

(Note that rounding down to 96 would instead give a probability of less than 95% which is unacceptable)

Let the random variable Y denote the number of unusable forms present in a batch of 250.

Then $Y \sim B(250,\ \frac{1}{100})$

$P(Y \leq 1) = 0.286$ (shown)

10 (i) Let the random variable X denote the number of particles produced via radioactive disintegration in a one second interval

Then $X \sim P_o(69)$

Since $\lambda = 69 > 10$, $X \sim N(69,\ 69)$ approx

$P(X < 60) = P(X < 59.5) = 0.126$

(ii) Redefining X as the number of particles produced in a 2 second interval,

Then $X \sim P_o(69 \times 2 = 138)$

$P(X < 150) = P(X > 150.5) = 0.144$ (shown)

(iii) Redefining X as the number of particles produced in a 10 second interval,

Then $X \sim P_o(69 \times 10 = 690)$

$P(X > 700) = P(X > 700.5) = 0.345$ (shown)

11. Let the random variable $\overline{X}$ denote the sample mean value of size 15.

Then $\overline{X} \sim N\left(60, \frac{4^2}{15} = \frac{16}{15}\right)$

$P(\overline{X} < 58) = 0.0264$ (shown)

(Note that CLT is **not used** because the distribution involved was **originally normal.**)

$P(58 < \overline{X} < 62) = 0.947$

$\therefore$ Expected Number of samples with means lying between 58 and 62

$= 100(0.947) = 94.7 \approx 95$ (shown)

12 (i) Let the random variable X denote the weight of a chocolate.

Then $X \sim N(10,\ 4)$

$P(9.5 < X < 10.5) = 0.197$ (shown)

(ii) $T = X_1 + X_2 + X_3 + ... + X_{25} \sim N(250,\ 100)$

$P(247 < T < 253) = 0.236$ (shown)

(iii) Let the random variable $\overline{X}$ denote the average weight of chocolates in a box.

Then $\overline{X} \sim N\left(10, \frac{4}{25}\right)$ and $P(9.9 < \overline{X} < 10.1) = 0.197$ (shown)

13. $\overline{X} \sim N\left(74, \frac{36}{n}\right)$

$$P\left(\overline{X} > 72\right) = 0.854 \rightarrow P\left(\overline{X} < 72\right) = 0.146 \rightarrow P\left(\overline{X} < \frac{72-74}{\sqrt{\frac{36}{n}}}\right) = 0.146$$

$$\frac{72-74}{\sqrt{\frac{36}{n}}} = InvNorm(0.146) = -1.054 \Rightarrow n = 10 \text{ (shown)}$$

14.(i) Since sample size $n = 30$ is large, by CLT, $\overline{X} \sim N\left(4.5, \frac{4.5}{30}\right)$ approx.

$P(\overline{X} > 5) = 0.0984$ (shown)

(ii) Since sample size $n = 30$ is large, by CLT, $\overline{X} \sim N\left(4.5, \frac{2.25}{30}\right)$ approx.

$P(\overline{X} > 5) = 0.0339$ (shown)

15. Let the random variable X denote the amount in a packet of the first drink.

Since $n = 36$ is large, by CLT,

$T = X_1 + X_2 + X_3 + ... + X_{36} \sim N(36 \times 200 = 7200,\ 36 \times 15^2 = 8100)$ approx.

$P(T > 7000) = 0.987$ (shown)

Let the random variable Y denote the amount in a packet of the second drink.

Since $n = 50$ is large for both batches of drinks, by CLT,

$T_1 = X_1 + X_2 + X_3 + ... + X_{50} \sim N(50 \times 200 = 10000,\ 50 \times 15^2 = 11250)$ approx

$T_2 = Y_1 + Y_2 + Y_3 + ... + Y_{50} \sim N(50 \times 200 = 10000,\ 50 \times 20^2 = 20000)$ approx

and $D = T_1 - T_2 \sim N(0,\ 31250)$

$P(|\ D\ | > 100) = P(D > 100) + P(D < -100) = 0.572$ (shown)

16. Let the random variable X denote the number of sixes obtained in the throwing of 10 dice.

Then $X \sim B(10, \frac{1}{6})$; $E(X) = \frac{5}{3}$ and $Var(X) = \frac{25}{18}$

Since $n = 50$ is large, by CLT,

$\overline{X} \sim N\left(\frac{5}{3}, \frac{\left(\frac{25}{18}\right)}{50} = \frac{1}{36}\right)$ approx and $P(\overline{X} < 2) = 0.977$ (shown)

17 (a)(i) $P\left(\sum_{i=1}^{3} X_i > 10\right)$

$= P(X_1 = 5) \bullet P(X_2 = 5) \bullet P(X_3 = 2) \times \frac{3!}{2!} + P(X_1 = 5) \bullet P(X_2 = 5) \bullet P(X_3 = 5)$

$= (0.6)(0.6)(0.3)(3) + (0.6)^3 = 0.54$ (shown)

(Note that permutation is involved in the computing process)

(ii) Since $n = 100$ is large, by CLT,

$\sum_{i=1}^{100} X_i \sim N(360,\ 324)$ approx and $P\left(\sum_{i=1}^{100} X_i > 350\right) = 0.711$ (shown)

(b) Since $n = 100$ and 200 are large for both separate sets of observations, by CLT,

$\overline{X}_{100} \sim N(3.6, \frac{3.24}{100})$ approx. , $\overline{X}_{200} \sim N(3.6, \frac{3.24}{200})$ approx

and $2\overline{X}_{100} - \overline{X}_{200} \sim N(2 \times 3.6 - 3.6 = 3.6,\ 2^2 \times \frac{3.24}{100} + \frac{3.24}{200} = 0.1458)$

$\therefore P(2\overline{X}_{100} < \overline{X}_{200} + 4.6) = P(2\overline{X}_{100} - \overline{X}_{200} < 4.6) = 0.996$ (shown)

18. Unbiased estimate of population mean $= \frac{\sum x}{n} = 1.14$ (shown)

Unbiased estimate of population variance $= \frac{1}{n-1}\left[\sum x^2 - \frac{(\sum x)^2}{n}\right] = 0.082$ (shown)

19. Unbiased estimate of population mean $= 1000 + \frac{1890}{10} = 1189$ (shown)

Unbiased estimate of population variance

$= \frac{1}{n-1}\left[\sum (x_i - 1000)^2 - \frac{(\sum (x_i - 1000))^2}{n}\right] = 537.8$ (shown)

20. To test: $H_0 : \mu = 1506.5$ $H_1 : \mu > 1506.5$ level of significance: 5%

$\overline{x} = 1506.8,\ n = 11;$ by the GC(Z-test) , $p = 2.52 \times 10^{-10} < 0.05$

Hence, H_0 is rejected and there is sufficient evidence at the 5% level that the machine is providing overweight bags. (shown)

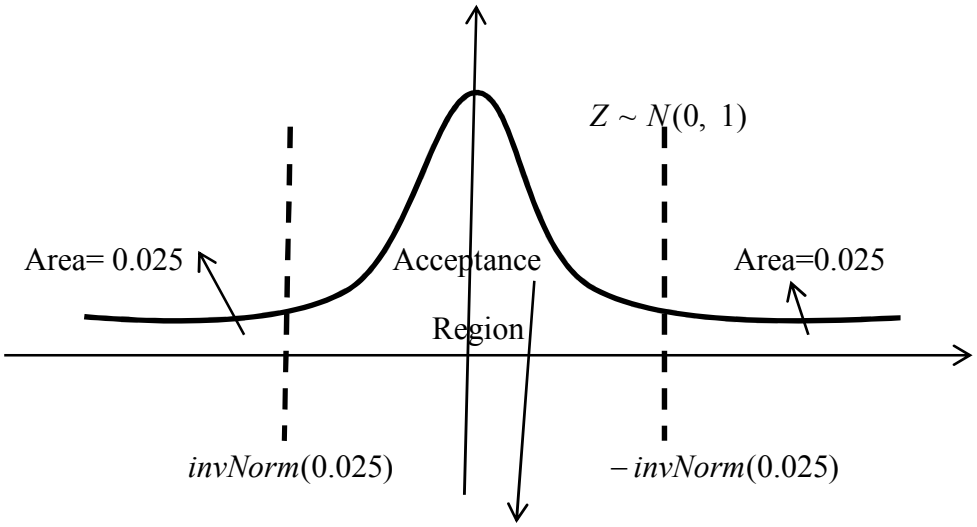
21. To test: $H_0 : \mu = 6$ $H_1 : \mu \neq 6$ level of significance: 5%

If H_0 is to be accepted at the 5% level, then

$$invNorm(0.025) < Z = \frac{\bar{x} - 6}{\left(\frac{0.8}{\sqrt{50}}\right)} < -invNorm(0.025)$$

$$-1.96 < \frac{\bar{x} - 6}{\left(\frac{0.8}{\sqrt{50}}\right)} < 1.96 \Rightarrow 5.778 < \bar{x} < 6.222 \text{ (shown)}$$

(Note: the above inequality is formulated by looking at the bell curve below:)



Test statistic Z must lie in this acceptance region

22. To test: $H_0 : \mu = 2000$ $H_1 : \mu < 2000$ level of significance: 2%

Sample mean $\bar{x} = 2000 - \frac{108}{64} = 1996.75$

Unbiased estimate of population variance

$$= \frac{1}{n-1}\left[\sum (x-2000)^2 - \frac{\left(\sum (x-2000)\right)^2}{n}\right] = 1.111$$

by the GC(Z-test) , $p = 0 < 0.05$

(Note that actual p value is so small that the GC has interpreted it to be approximately zero)

Hence, H_0 is rejected and there is sufficient evidence at the 2% level that the manufacturer is over-estimating the average length of his light bulbs. (shown)

23. To test: $H_0 : \mu = 65$ $H_1 : \mu > 65$

Since H_0 is accepted if $\bar{x} \le 66.5$ and rejected if $\bar{x} > 66.5$, then

$$\alpha = P\left(Z > \frac{66.5 - 65}{\sqrt{\frac{36}{100}}}\right) = P(Z > 2.5) = 0.006$$

$\therefore$ Level of significance is 0.6%. (shown)

24(a) For every year that passes, the child will grow taller by 6cm. (shown)

(b) No, it would not make sense since it implies that an infant at the time of her birth would be 80cm tall. (shown)

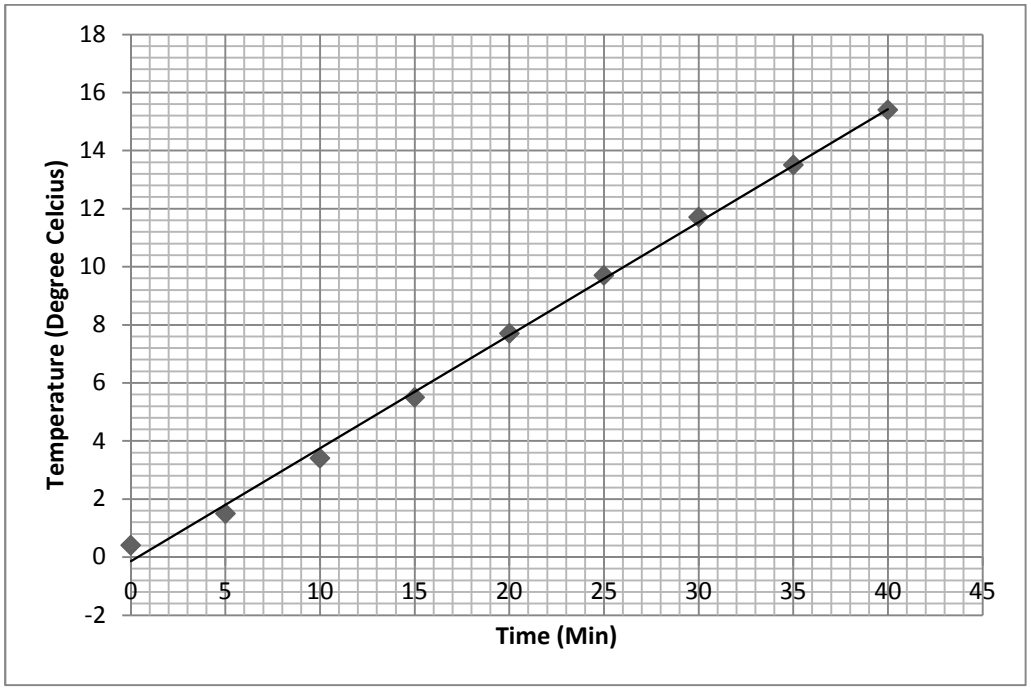
(c) For an American female at 8 years old, $y = 80 + 6(8) = 128$ cm (shown)

Such an interpretation would be **valid** since the age of 8 years falls within the regression analysis range.

For an American female at 25 years old, $y = 80 + 6(25) = 230$ cm (shown)

Such an interpretation would be in**valid** since the age of 25 years is an extrapolated estimation, and in this particular context, it is not reasonable to assume an American female would grow linearly with age indefinitely. (shown)

25 (a)



(b) By the GC, regression line of y on x is $y = 0.389x - 0.142$ (shown)

When $x = 60, \quad y = 0.389(60) - 0.142 = 23.2^{\circ}C$ (shown)

(c) (i) $\frac{x}{60} = t \Rightarrow x = 60t$; $y = 0.389(60t) - 0.142 = 23.36t - 0.142$ (shown)

(ii) $z = 273 + y \Rightarrow y = z - 273; \quad z - 273 = 0.389x - 0.142$

$\therefore \quad z = 272.86 + 0.389x$ (shown)

(d) When temperature is measured at fixed time intervals, it is obvious that **time** is the **dependent variable** while **temperature** assumes the **independent variable** role; hence obtaining the regression line of y on x would be more sensible.

Should time be measured against predetermined temperature levels, then the regression line of x on y would be appropriate since the independent/dependent variable roles are now reversed. (shown)

26 (i) By the GC, regression line of T on V is $T = 0.874\ V + 27.69$ (shown)

(ii) When $V = 60, \ T = 0.874(60) + 27.69 = 80.13$ km/hr (shown)

(iii) Let the random error in T be described by $E \sim N(0, 16)$, then for $V = 60$, required probability

$= P(E > 91 - 80.13) = P(E > 10.87) = 0.0329$ (shown)

27 (i) $\sum u = 56,\ \sum u = 69,\ \sum u^2 = 560,\ \sum v^2 = 887,\ \sum uv = 704$

$$b = \frac{\sum uv - \dfrac{\sum u \sum v}{n}}{\sum u^2 - \dfrac{\left(\sum u\right)^2}{n}} = \frac{704 - \dfrac{(56)(69)}{8}}{560 - \dfrac{56^2}{8}} = 1.315$$

Using $v - \bar{v} = b(u - \bar{u}),$ we have

$$v - \frac{69}{8} = 1.315(u - \frac{56}{8}) \Rightarrow v = 1.315u - 0.583 \text{ (shown)}$$

Hence, equation of regression line of y on x is

$$y - 325 = 1.32(x - 1971) - 0.583 \Rightarrow y = 1.315x - 2268.387 \text{ (shown)}$$

(ii) $$r = \frac{\sum uv - \dfrac{\sum u \sum v}{n}}{\sqrt{\sum u^2 - \dfrac{\left(\sum u\right)^2}{n}}\sqrt{\sum v^2 - \dfrac{\left(\sum v\right)^2}{n}}} = \frac{704 - \dfrac{(56)(69)}{8}}{\sqrt{560 - \dfrac{56^2}{8}}\sqrt{887 - \dfrac{69^2}{8}}} = 0.998 \text{ (shown)}$$

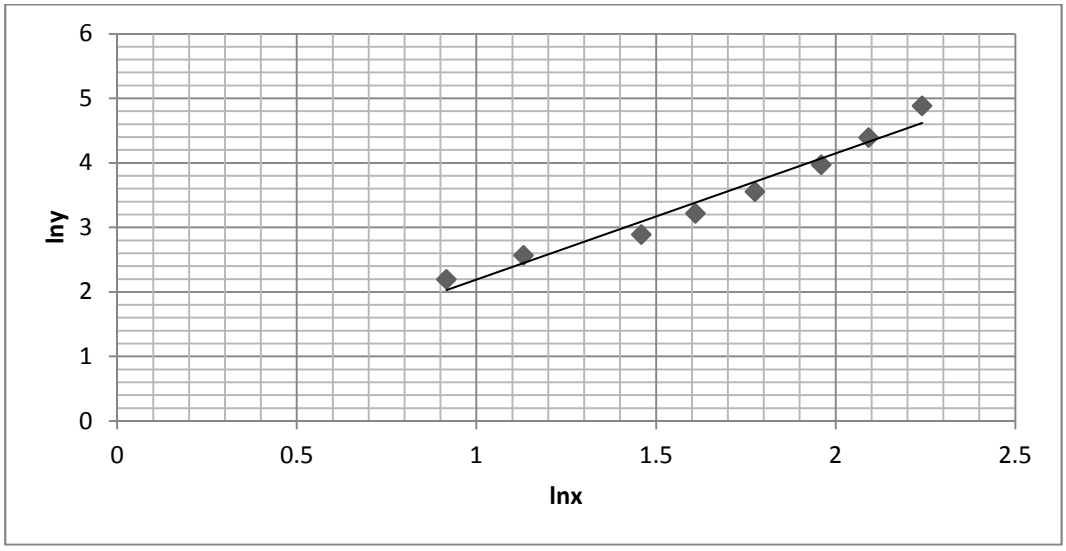
(iii) (a) When $x = 1974,\quad y = 1.315(1974) - 2268.387 = 327$ (shown)

(b) When $x = 1974,\quad y = 1.315(1988) - 2268.387 = 346$ (shown)

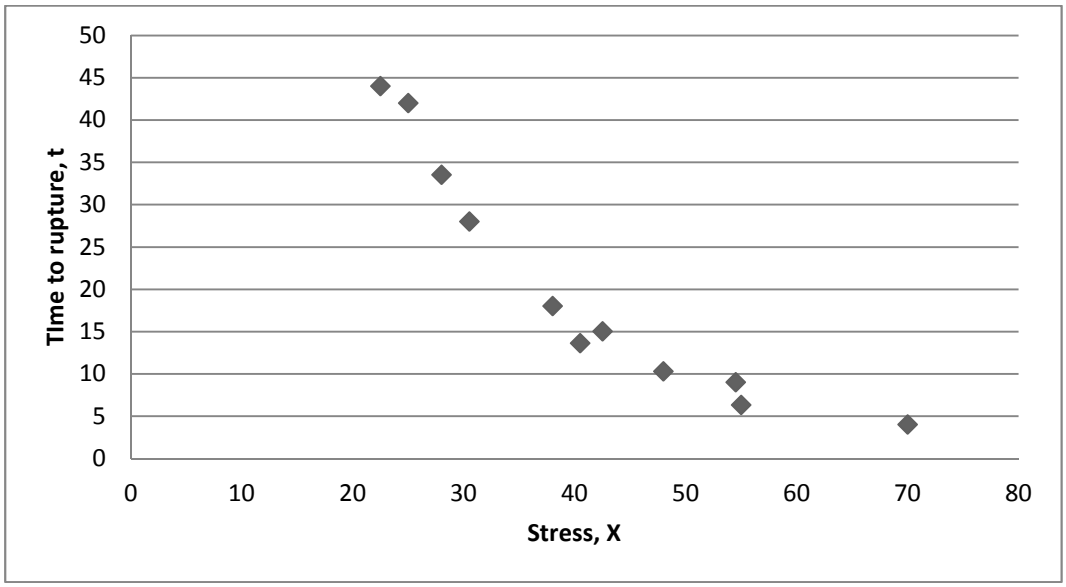
28 (i) The product moment correlation coefficient output as given by the GC based on the data table is **moderately weak** ($r = 0.545$); hence finding the regression line of y on x is not too suitable. (shown)

(ii) The product moment correlation coefficient output as given by the GC based on the data table is very strong ($r = 0.982$); also it is observed that the number of words defined correctly by a child depends on his/her age (ie x is independent variable, y is dependent variable), therefore finding the regression line of y on x is definitely suitable. (shown)

(iii) The equation of the regression line of $\ln y$ on $\ln x$ is given by $\ln y = 0.240 + 1.953 \ln x$ (shown)



29(i)



(ii) Model (a) would be more appropriate as the scatter diagram above reflects a rather obvious **inverse** relationship between the two variables. (shown)