Semester 1	Solution to Tutorial Set 13 and Revision Problems	2013

1. Ans: (b) since r is negative and close to -1.

2. Ans: (d) $r^2 = (-0.7761)^2 = 0.60238$=60.238%.

3. (a) $r^2 = (-0.953)^2 = 0.908209 = 90.1\%$. Hence the regression line of Y on X can explain 90.1% of the variation in Y.

   (b) When $x_1 = 30$, $\hat{y}_1 = a + bx_1 = 18.23810 - 0.0940476 \times 30 = 15.41667$ and hence $r_1 = y_1 - \hat{y}_1 = 16 - 15.41667 = 0.583328$.
   When $x_2 = 40$, $\hat{y}_2 = a + bx_2 = 18.23810 - 0.0940476 \times 40 = 14.47620$ and hence $r_2 = y_2 - \hat{y}_2 = 14 - 14.47620 = -0.476196$.

   (c) The t-test for the significance of the regression model is:

   1. Hypotheses: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
   2. Test statistic: $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}} = \dfrac{-0.0940476}{0.788585/\sqrt{4200}} = -7.72901$ where

      $$s^2 = \frac{L_{yy} - bL_{xy}}{n-2} = \frac{40.88 - (-0.0940476)(-395)}{8-2} = 0.788585^2$$
   3. P-value: $2P(t_6 < -7.72901) < 0.002$ (df=8-2)
   4. Conclusion: Since P-value< 0.05, there is strong evidence in the data against H_0. The regression model is significant.

   (d) For an increase of 1gm of dose, the breathing rate is decreased by 0.094 in breath per min.

4. Ans: (e) Note that $\overline{X}_1 - \overline{X}_2 \sim N\left(\mu_1 - \mu_2, \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)\right)$ under the assumption of equality of variance. When σ^2 is unknown, it is estimated by $s_p^2 = \dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$. Hence the mean estimate of $\overline{X}_1 - \overline{X}_2$ is

   $$\bar{x}_1 - \bar{x}_2 = 10 - 12 = -2$$

   and the SE estimate is

   $$s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 4.0139 \times \sqrt{\frac{1}{9} + \frac{1}{11}} = 1.8041$$

   where $s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} = \sqrt{\dfrac{8(5^2) + 10(3^2)}{8+10}} = 4.0139.$

5. Ans: (b) We have $n = 200$ and $p = 0.6$ (one sample of binary data). Since n is large, the sample proportion $\hat{p} \sim N\left(p, \dfrac{p(1-p)}{n}\right)$, i.e. $N\left(0.6, \dfrac{0.6(1-0.6)}{200}\right)$ by CLT.

$$P(\hat{p} > 0.58) = P\left(Z > \frac{0.58 - 0.6}{\sqrt{0.6 \times 0.4/200}}\right) = P(Z > -0.58) = P(Z < 0.58) = 0.7190$$

6. Ans: (e) $Z_{\text{obs}} = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} = \dfrac{0.76 - 0.68}{\sqrt{0.72(1-0.72)(\frac{1}{100} + \frac{1}{100})}} = 1.2599$ where the pooled proportion is $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} = \dfrac{76 + 68}{100 + 100} = 0.72$.

7. Given $X_1, X_2, \ldots, X_9 \sim N(5, 3^2)$, $\sum_{i=1}^{9} X_i \sim N(9(5), 9(3^2))$, i.e. $N(45, 9^2)$ (Read P.23 of the lecture note on Week 6). Hence

$$P\left(\sum_{i=1}^{9} X_i > 54\right) = P\left(Z > \frac{54 - 45}{9}\right) = P(Z > 1) = 1 - 0.8413 = 0.1587.$$

8. We have $n = 10$, $\bar{x} = 36$ and $s = 0.034$. The 80% CI for the population mean μ in Celsius is

$$\bar{x} \mp t_{10-1,0.1} \frac{s}{\sqrt{n}} = 36 \mp 1.383 \frac{0.034}{\sqrt{10}} = (35.98513,\ 36.01487)$$

Since $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ in Celsius and $F = \frac{9}{5}C + 32$, the sample mean in Fahrenheit

$$\bar{Y} = \frac{9}{5}\bar{X} + 32 \sim N\left(\frac{9}{5}\mu + 32, \frac{\left(\frac{9}{5}\right)^2 \sigma^2}{n}\right).$$

Based on the sample mean of $\bar{x} = 36$ in Celsius, the sample mean $\bar{y}$ in Fahrenheit is

$$\bar{y} = \frac{9}{5}\bar{x} + 32 = \frac{9}{5}36 + 32 = 96.8$$

and hence the 80% CI for the population mean μ in Fahrenheit is

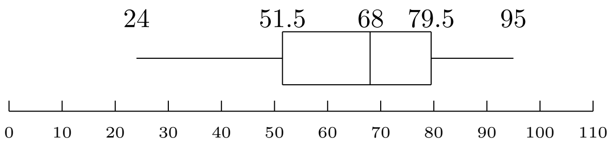
$$\bar{y} \mp t_{10-1,0.1} \frac{\frac{9}{5}s}{\sqrt{n}} = 96.8 \mp 1.383 \frac{\frac{9}{5}0.034}{\sqrt{10}} = (96.77323,\ 96.82677)$$

Solutions to Additional Problems and Revision for Week 13

1. (i) The correct median is $\frac{X_{(20)}+X_{(21)}}{2}$ where $X_{(k)}$ denotes the k-th ordered observation in ascending order. However the leaves are not ordered. 64 and 69 are not the 20^{th} and 21^{st} ordered observations.

(ii) The correct median is $\frac{X_{(20)}+X_{(21)}}{2} = \frac{67+69}{2} = 68$.
To draw boxplot, we also need
Min $= 24.0$;
$Q_1 = \frac{X_{(10)}+X_{(11)}}{2} = \frac{50+53}{2} = 51.5$;

$Q_2 = 68$;
$Q_3 = \frac{X_{(30)}+X_{(31)}}{2} = \frac{79+80}{2} = 79.5$;
Max $= 95.0$.
IQR $= Q_3 - Q_1 = 79.5 - 51.5 = 28$.
To check if there are outliers, we calculate
LT=$Q_1 - 1.5 \times IQR = 51.5 - 1.5(28) = 9.5$ and
UT=$Q_3 + 1.5 \times IQR = 79.5 + 1.5(28) = 121.5$.
Since all observations lie within (LT,UT), there is no outlier.
The boxplot is



The distribution is slightly left skewed.

(iii) 7/40 or 17.5%.

2. The expected frequencies are

Observed frequencies, O_i	38	43	10	5	96
(a) Expected frequencies, E_i	$96(\frac{1}{4}) = 24.0$	$96(\frac{1}{4}) = 24.0$	$96(\frac{1}{4}) = 24.0$	$96(\frac{1}{4}) = 24.0$	96
(b) Expected frequencies, E_i	$96(\frac{4}{10}) = 38.4$	$96(\frac{4}{10}) = 38.4$	$96(\frac{1}{10}) = 9.6$	$96(\frac{1}{10}) = 9.6$	96

(a) 1. The hypotheses are $H_0 : p_1 = p_2 = p_3 = p_4 = \frac{1}{4}$ vs H_1: not all equalities hold.
2. The goodness of fit statistic is

$$\chi^2_{\text{obs}} = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{14^2}{24} + \frac{19^2}{24} + \frac{14^2}{24} + \frac{19^2}{24} = 46.42$$

3. P-value $= P(\chi^2_3 \geq 46.42) < 0.01$

4. Conclusion: We have very strong evidence against H_0.

(b) 1. The hypotheses are $H_0 : p_1 = p_2 = \frac{4}{10}$; $p_3 = p_4 = \frac{1}{10}$ vs H_1: not all equalities hold.
2. The goodness of fit statistic is

$$\chi^2_{\text{obs}} = \sum_i \frac{(O_i - E_i)^2}{E_i} = \frac{0.4^2}{38.4} + \frac{4.6^2}{38.4} + \frac{0.4^2}{9.6} + \frac{4.6^2}{9.6} = 2.78$$

3. P-value $= P(\chi^2_3 \geq 2.78) > 0.10$

4. Conclusion: The data are consistent with H_0.

3. (i) P(pass or better)=0.85 and therefore, P(not qualify for a grade)=0.15
Hence the required probability=$2 \times 0.15 \times 0.06 = 0.0180$.

(ii) Each trial has a probability of success $p = 0.6$ and there are n (fixed) number of independent trials. Thus, $X \sim B(n, 0.60)$.

(a) Since $X \sim B(10, 0.60)$, $P(X \leq 6) = 0.6177$ from the binomial table.
Hence $P(X \geq 7) = 1 - 0.6177 = 0.3823$.

(b) Since $X \sim B(25, 0.60)$, $P(X = 12) = \binom{25}{12}(0.60)^{12}(0.40^{13} = 0.0760$.

(iii) $P(X > 63.5) = P(Z > \dfrac{63.5 - 48}{6}) = P(Z > 2.58) = 0.0049$.

Let p be the true proportion of students live in apartments.
The hypotheses are $H_0 : p = 0.25$ vs $H_1 : p > 0.25$

Let X be the number of students live in apartments in a sample of 192. Under H_0,

$$\hat{p} = \frac{X}{n} \sim N\left(p, \frac{p(1-p)}{n}\right) \quad \text{or} \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1).$$

Since $\hat{p} = \dfrac{64}{192} = \dfrac{1}{3}$, the test statistic is $z_{\text{obs}} = \dfrac{1/3 - 1/4}{\sqrt{\frac{0.25(0.75)}{192}}} = 2.67$.

The P-value= $P(Z \geq 2.67) = 0.0038$
This probability is too small. Therefore we have very strong evidence against H_0.

4. $$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = 24625 - \frac{475^2}{10} = 2062.5$$
$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = 9547.56 - \frac{307.6^2}{10} = 85.784 \text{ and}$$
$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = 14234.5 - \frac{475(307.6)}{10} = -376.5$$

(a) $r^2 = \dfrac{L_{xy}^2}{L_{xx}L_{yy}} = \dfrac{(-376.5)^2}{(2062.5)(85.784)} = 0.801$

Since r is negative and r^2 is large, there is a strong negative linear relationship between x and y variables.

(b) $b = \dfrac{L_{xy}}{L_{xx}} = \dfrac{-376.5}{2062.5} = -0.183$ and $a = 30.76 - (-0.183 \times 47.5) = 39.431$.

Therefore, $y = 39.431 - 0.183x$.

(c) $x = 125$ is very far away from the given range (25-70) for x. Thus it is not reasonable to use this model to estimate the value of y when $x = 125$.

5. Q11.21-Q11.23 (P.227)
$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = 46689410 - \frac{23670^2}{12} = 335$$
$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = 4033.83 - \frac{214.9^2}{12} = 185.3292 \text{ and}$$
$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = 423643.3 - \frac{23670(214.9)}{12} = -246.95$$

(a) $b = \dfrac{L_{xy}}{L_{xx}} = \dfrac{-246.95}{335} = -0.7371642$ and

$$a = \bar{y} - b\bar{x} = \frac{214.9}{12} - (-0.7371642)\frac{23670}{12} = 1471.9646766.$$

Therefore, Infant-mortality rate $= 1471.9646766 - 0.7371642\times$ Chronological year.

(b) The t-test for the significance of the regression model is:

1. Hypotheses: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
2. Test statistic: $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}} = \dfrac{-0.7371642}{0.5465988/\sqrt{335}} = -24.6841376$ where
$$s^2 = \frac{L_{yy} - bL_{xy}}{n-2} = \frac{185.3292 - (-0.7371642)(-246.95)}{12-2} = 0.5465988^2$$
3. P-value: $2P(t_{10} < -24.6841376) < 0.000$ (df=8-2)
4. Conclusion: Since P-value< 0.05, there is strong evidence in the data against H_0. The regression model is significant.

(c) When the year is $x = 1989$, the infant-mortality rate is $\hat{y} = 1471.9646766 - 0.7371642 \times 1989 = 5.750$.

6. Q11.36 (P.228) and Q11.38 to Q11.41 (P.229)

$$L_{xx} = \sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n} = 1.31 - \frac{2.38^2}{8} = 0.60195$$
$$L_{yy} = \sum_i y_i^2 - \frac{(\sum_i y_i)^2}{n} = 30.708 - \frac{(-15.55)^2}{8} = 0.4826875 \text{ and}$$
$$L_{xy} = \sum_i x_i y_i - \frac{(\sum_i x_i)(\sum_i y_i)}{n} = -4.125 - \frac{2.38(-15.55)}{8} = 0.501125.$$

(a) $r^2 = \dfrac{L_{xy}}{\sqrt{L_{xx}L_{yy}}} = \dfrac{0.501125}{\sqrt{(0.60195)(0.4826875)}} = 0.9296786$

(b) $b = \dfrac{L_{xy}}{L_{xx}} = \dfrac{0.501125}{0.60195} = 0.8325027$ and

$$a = \bar{y} - b\bar{x} = \frac{-15.55}{8} - (0.8325027)\frac{2.38}{8} = -2.1914196.$$

Therefore, Log mortality $= -2.1914196 + 0.8325027\times$ Log annual cigarette consumption.

(c) The t-test for the significance of the regression model is:

1. Hypotheses: $H_0 : \beta = 0$ vs $H_1 : \beta \neq 0$
2. Test statistic: $t_{\text{obs}} = \dfrac{b}{s/\sqrt{L_{xx}}} = \dfrac{0.8325027}{0.0967320191/\sqrt{0.60195}} = 6.6772187887$ where
$$s^2 = \frac{L_{yy} - bL_{xy}}{n-2} = \frac{0.4826875 - (0.8325027)(0.501125)}{8-2} = 0.0967320191^2$$
3. P-value: $2P(t_{10} > 6.6772187887) < 0.000$ (df=8-2; 0.0005464681 from R)
4. Conclusion: Since P-value< 0.05, there is strong evidence in the data against H_0. The regression model is significant.

(d) When the log annual cigarette consumption is $x = \log(1) = 0$, the log mortality rate is $\hat{y} = -2.1914196 + 0.8325027 \times 0 = -2.1914196$. Hence the mortality rate $10^{-2.1914196} = 0.006435472$.

(e) The linear relationship only exists on log scale.