Semester 2, 2012 (Last adjustments: November 12, 2012)

Lecture Notes

MATH1905 Statistics (Advanced)

Lecturer

Dr. John T. Ormerod

School of Mathematics & Statistics F07

University of Sydney

(w) 02 9351 5883

(e) john.ormerod (at) sydney.edu.au

Semester 1, 2012 (Last adjustments: November 12, 2012)

Monday, 24th October 2011

Lecture 15 - Content

- **Extended Answer Section of 2010 Exam**
- **Extended Answer Section of 2011 Exam**

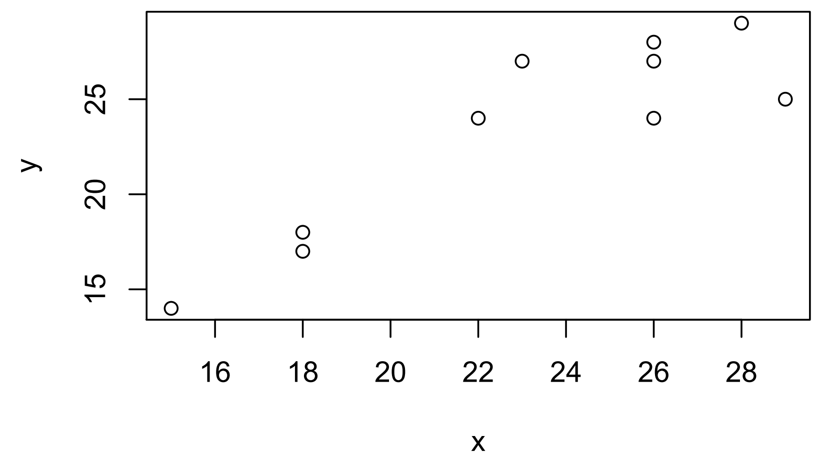
Question 1 (16 marks in all)

The following R-output gives the maximum temperatures x_i and y_i, in $^\circ$C, on two successive days (day 1 and day 2) in 10 Australian weather observation stations, $i = 1, \ldots, 10$:

```
> x = c(22,18,26,26,29,15,23,18,28,26)
> y = c(24,17,28,24,25,14,27,18,29,27)
```

Additionally you might find the following R output of use:

```
> length(x)        > sum(x^2)
[1] 10             [1] 5539
> sum(x)           > sum(y^2)
[1] 231            [1] 5669
> sum(y)           > sum(x*y)
[1] 233            [1] 5580
```



(a) (3 marks) *Calculate the coefficient of correlation.*

$$r_{xy} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{5580 - 10 \times 23.1 \times 23.3}{\sqrt{(5539 - 231^2/10) \times (5669 - 233^2/10)}}$$
$$= \frac{197.7}{\sqrt{202.9 \times 240.1}} = 0.8957.$$

(b) (4 marks) *Find the equation of the least squares line of temperature on day 2 (y) on temperature of day 1 (x).*

$$b = \frac{S_{xy}}{S_{xx}} = \frac{197.7}{202.9} = 0.9744$$

and

$$a = \overline{y} - b\overline{x} = 23.3 - b \times 23.1 = 0.7920.$$

Thus, the regression line is $y = 0.7920 + 0.9744 \times x$. [2 marks for writing down the regression line explicitly, 1 mark if a correct, 1 mark if b correct].

(c) (2 marks) *The simplest weather forecast is 'tomorrow will be like today'. Use the regression line to give an improved forecast in one of those weather stations, if today's temperature there is 15°C*

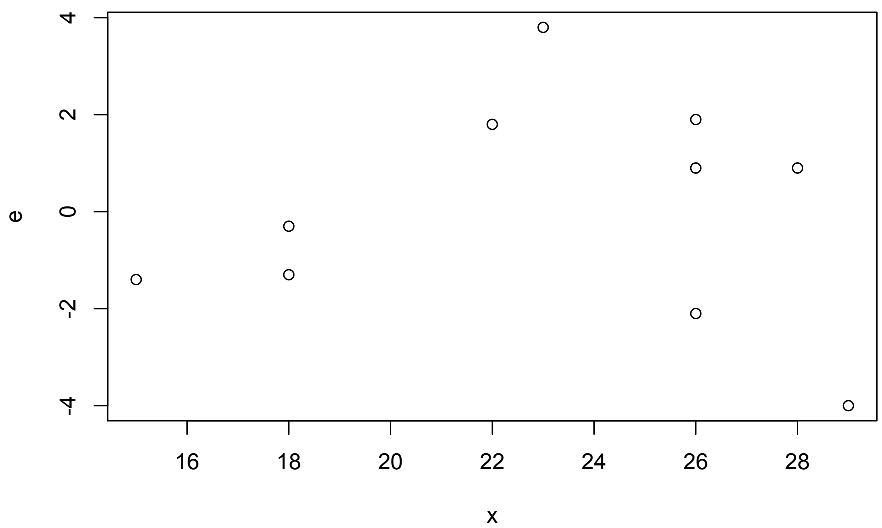
Let $x = 15$, then $y = 0.7920 + 0.9744 \times 15 = 15.408$ degrees.

(d) (6 marks) *The residuals* $e_1, \ldots, e_{10}$ *from the least squares fit are as follows (to 1dp):*

```
> round(lm(y~x)$res,1)
   1    2    3    4    5    6    7    8    9   10
 1.8 -1.3  1.9 -2.1 -4.0 -1.4  3.8 -0.3  0.9  0.9
```

Draw a residual plot and hence comment on the appropriateness of the linear regression.

The residual plot should look as follows (deduct marks if axes are not labeled and if the e is not drawn on the y-axis etc) [3 marks]



The sample size is only 10 and therefore it is difficult to justify with certainty the appropriateness of the linear regression [1 mark]. There seems to be some 'quadratic' curvature left [1 mark], also the variability of the residuals seem to be larger for larger x values [1 mark].

(e) (1 mark) *About what percentage of the variability of y's is explained by the regression line?*

About 80%, since $r^2 = r_{xy}^2 = 0.8957^2 = 0.8022$. [1 mark]

Question 2 (15 marks in all)

The following data are measurement of weight gain (in gm) after 10 male rats and 10 female rats were given the same diet over the same period of time. The 10 male rats and 10 female rats were chosen independently.

Male (x)	2.6	4.8	12.5	8.7	9.7	8.2	9.4	8.7	9.2	10.0
Female (y)	8.1	7.6	10.5	8.9	11.2	6.9	11.7	12.6	10.3	7.1

(a) (8 marks) *Assume that both samples can be modelled by a normal distribution with the same population variance. Given that* $\sum_{i=1}^{10} x_i = 83.8$, $\sum_{i=1}^{10} y_i = 94.9$, $\sum_{i=1}^{10} x_i^2 = 771.76$, $\sum_{i=1}^{10} y_i^2 = 938.03$, *is there evidence of a difference in weight gains between male rats and female rats?*

Here, a two-sample t-test is the appropriate test to be used because male rats and female rats were chosen independently [1 mark].

Our null and alternative hypotheses are [1 mark]:

$$H_0\colon \mu_x = \mu_y \quad \text{vs} \quad H_1\colon \mu_x \neq \mu_y$$

Thus, the pooled variance is [1 mark]

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1) \times s_y^2}{18} = \frac{1}{2}(s_x^2 + s_y^2)$$
$$= (7.724 + 4.158778)/2 = 5.941389$$

$$t_{(n_x+n_y-2)} \sim \frac{\overline{X} - \overline{Y}}{s_p\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}} = \frac{8.38 - 9.49}{\sqrt{5.941389}\sqrt{1/5}} = -1.018272, \quad \text{[2 marks]}$$

The degrees of freedom is $\nu = 18$ [1 mark].

From the tables we find the critical value at the 5% level to be between -2.228 and -2.086. Alternatively the P-value is $P = 2 \times P(t_{18} > 1.018272)$. From tables $P(t_{10} > 0.700) = 0.25$ and $P(t_{20} > 1.064) = 0.15$ so $0.3 < P < 0.5$ [1 mark].

Thus we accept the null hypothesis that there is no difference in weight gains between male rats and female rats. [1 mark]

(b) (7 marks) Let $X \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu = 1$ and $\sigma^2 = 1/2$. *Use Chebyshev's inequality to bound* $\mathrm{P}(X > 2)$.

Chebyshev's inequality is

$$\mathrm{P}(|X - \mu| \geq c\sigma) \leq \frac{1}{c^2} \quad \text{[1 mark].}$$

Here, $\mu = 1$ and $\sigma = \frac{1}{\sqrt{2}}$ [1 mark].

Thus,

$$\mathrm{P}(X > 2) \leq \mathrm{P}(|X - 1| > 1) \leq \mathrm{P}(|X - 1| > \sqrt{2}\frac{1}{\sqrt{2}}) < 1/2. \quad \text{[3 marks].}$$

Is Chebyshev's bound obtained in (i) sharp in this case, i.e. are the two probabilities the same? (Justify your answer)

No, using normal tables, we obtain that $\mathrm{P}(X > 2) = \mathrm{P}((X-\mu)/\sigma > (2-\mu)/\sigma)$ [1 mark], and $\mathrm{P}(Z > \sqrt{2}) = 1 - 0.9207 < 1/2$ [2 marks]

Question 3 (17 marks in all)

The number of radioactive counts in 100 one minute intervals for a particular machine were

No. of counts (x):	0	1	2	3	4	5
Observed frequency:	15	28	23	16	12	6

Suppose that we can model the number of counts in one minute by a Poisson random variable X, where

$$\mathrm{P}(X = i) = e^{-\lambda}\lambda^i / i!, \quad i = 0, 1, 2, \ldots .$$

(a) (4 marks) *Prove that the probability generating function of the random variable* X *is* $\pi(s) = e^{\lambda(s-1)}$. For $X \sim \mathcal{P}(\lambda)$,

$$\begin{aligned}
\pi(s) &= \sum_{i=0}^{\infty} s^i \,\mathrm{P}(X = i) \\
&= \sum_{i=0}^{\infty} e^{-\lambda} \frac{\lambda^i}{i!} s^i \\
&= e^{-\lambda} \sum_{i=0}^{\infty} \frac{e^{\lambda s}}{e^{\lambda s}} \frac{(\lambda s)^i}{i!} \\
&= e^{-\lambda} e^{\lambda s} \\
&= e^{\lambda(s-1)}.
\end{aligned}$$

(b) (2 marks) *Use the probability generating function in part (i) to prove that* $\mathrm{E}(X) = \lambda$.

Because, $\pi'(s) = \lambda e^{\lambda(s-1)}$ it follows that $\pi'(1) = \mathrm{E}\,X = \lambda$.

(c) (2 marks) *It is known that Var*$(X) = \mathrm{E}(X) = \lambda$, *determine* $\mathrm{E}(X^2)$.

From the definition of the $\mathrm{Var}(X) = \mathrm{E}(X^2) - (\mathrm{E}(X))^2$.

It follows that $\mathrm{E}(X^2) = \lambda + \lambda^2$.

(d) (2 marks) *Calculate the expected number of 0's in a sample of size 100 from a Poisson random variable with mean 2.*

$$\mathrm{P}(X = 0) = e^{-2} \quad \Rightarrow 100 \times e^{-2} = 13.53353.$$

(e) (7 marks) *You are given that the above sample average is $\overline{x} = 2$. Complete the table of expected frequencies below and test the goodness of fit of the Poisson distribution as a model for the number of radioactive counts.*

No. of counts (x):	*0*	*1*	*2*	*3*	*4*	*5*
Expected frequency:	*?*	*27.07*	*27.07*	*18.04*	*9.02*	*?*

The expected frequencies have to add to the sample size, i.e $E_5 = 100 - 9.02 - \ldots - 13.53353 = 5.26647$. [1 mark]

Thus the test statistic becomes,

$$X^2 = \sum O_i^2/E_i - n = \frac{15^2}{13.53353} + \ldots + \frac{6^2}{5.26647} - 100 = 2.12. \text{ [3 marks]}$$

Since here $g = 6$ and $k = 1$ the degrees of freedom is $\nu = g - k - 1 = 6 - 1 - 1 = 4$, we have that the P-value is larger than 0.25 since from the tables $\mathrm{P}(\chi_4^2 > 5.385) = 0.25$. [2 marks]

Thus there is not sufficient statistical evidence to say that a Poisson distribution is not a good model. [1 mark]

Question 4 (17 marks in all)

When cancerous tumours are removed from the colon it is not always possible to remove all cancerous cells without removing too much of the patient's vital organs. Consider the following data:

		Was the cancer controlled?	
		Yes	No
Was cancer present	Yes	8	182
at the edge of surgery?	No	11	58

(a) (8 marks) *Is there any evidence that cancer at the edge of surgery affects the chance of the cancer being controlled?*

This a 2×2 contingency table, i.e. with entries x_{ij} the expected frequencies under independence are

$$E_{ij} = n \times \frac{x_{i\bullet}}{n} \times \frac{x_{\bullet j}}{n} = \frac{x_{i\bullet} \times x_{\bullet j}}{n},$$

where $x_{i\bullet}$ denotes the sum of the ith row and $x_{\bullet j}$ denotes the sum of the jth column. Here $n = 259$ and the row and column totals are:

		Was the cancer controlled?		
		Yes	No	Total
Was cancer present	Yes	8	182	190
at the edge of surgery?	No	11	58	69
	Total	19	240	259

Thus $E_{11} = \frac{190 \times 19}{259} = 13.93822$, $E_{12} = \frac{190 \times 240}{259} = 176.0618$, $E_{21} = \frac{69 \times 19}{259} = 5.061776$ and $E_{22} = \frac{69 \times 240}{259} = 63.93822$. [3 marks]

		Was the cancer controlled?	
		Yes	No
Was cancer present	Yes	8	182
at the edge of surgery?	No	11	58

From previous page $E_{11} = 13.93822$, $E_{12} = 176.0618$, $E_{21} = 5.061776$ and $E_{22} = 63.93822$.

Thus the test statistic becomes,

$$X^2 = \sum O_{ij}^2/E_{ij} - n = \frac{8^2}{13.94} + \ldots + \frac{58^2}{63.94} - 259 = 10.24814. \text{ [2 marks]}$$

The degrees of freedom here uses $g = 4$ and $k = 2$ so $g - k - 1 = 1$ or $(r-1)(c-1) = 1$ [1 mark].

The P-value is less than 0.01 since from the tables $\mathrm{P}(\chi_1^2 > 6.635) = 0.01$. [1 mark] Thus there is statistical evidence that cancer at the edge of surgery affects the chance of the cancer being controlled. [1 mark]

(b) (5 marks) Consider the 69 patients who had 'no cancer present at the edge of surgery' only:

Provide a conservative 95% confidence interval for the proportion of having controlled cancer.

The formula for a conservative CI is $\widehat{p} \pm z_{\alpha/2}\sqrt{1/(4n)}$ [1 mark].

Here

$$\frac{11}{69} \pm 1.96/\sqrt{4 \times 69} = (\, 0.0414 \,,\, 0.2774 \,) \text{ [2 marks]}$$

Determine how much smaller the length of the conservative confidence interval in (i) is when the sample size was 100 instead of 69 patients.

This is only determined by the $\sqrt{n}$ term, thus increasing the sample size from 69 to 100, makes the CI smaller by a factor of $\sqrt{100/69} = 1.2$. [2 marks]

(c) (4 marks) *Let* A *and* B *be two independent events. Show that* $\mathrm{P}(A \cap B^c) = \mathrm{P}(A) \times \mathrm{P}(B^c)$, *where* B^c *denotes the complement of B.*

Note that $A = (A \cap B^c) \cup (A \cap B)$ and because those two events are mutually exclusive it follows from Axiom 3 that $\mathrm{P}(A) = \mathrm{P}(A \cap B^c) + \mathrm{P}(A \cap B)$ [1 mark].

Thus, $\mathrm{P}(A \cap B^c) = \mathrm{P}(A) - \mathrm{P}(A \cap B)$ [1 mark]

and using that A and B are independent we obtain

$$\begin{aligned} \mathrm{P}(A \cap B^c) &= \mathrm{P}(A) - \mathrm{P}(A)\,\mathrm{P}(B) \text{ [1 mark]} \\ &= \mathrm{P}(A)(1 - \mathrm{P}(B)) \\ &= \mathrm{P}(A)\,\mathrm{P}(B^c) \text{ [1 mark]} \end{aligned}$$

Extended Answer Section of 2011 Exam

Question 1 (14 marks in all)

The following R-output gives daily temperature, x_i in degrees Fahrenheit, and Ozone level, y_i in parts per billion in New York over 16 successive days.

```
> x = c(61,61,67,81,79,76,82,90,87,82,77,72,65,73,76,84)
> y = c(4,32,23,45,115,37,29,71,39,23,21,37,20,12,13,135)
```

Additionally you might find the following R output of use:

```
> sum(x)
[1] 1213
> sum(y)
[1] 656
> sum(x^2)
[1] 93125
> sum(y^2)
[1] 46868
> sum(x*y)
[1] 52111
> sort(round(lm(y~x)$resid,0))
 10  15   7   9  14  11   1   4   6   3   8  13  12   2   5  16
-31 -28 -25 -25 -23 -22  -7  -7  -4   0   1   1   4  21  67  77
```

From these we can calculate the summary statistics:

$$S_{xx} = (\sum x_i^2) - (\sum x_i)^2/n = 93125 - 1213^2/16 = 1164.4375$$

$$S_{yy} = (\sum y_i^2) - (\sum y_i)^2/n = 46868 - 656^2/16 = 19.972$$

$$S_{xy} = (\sum x_i y_i) - (\sum x_i)(\sum y_i)^2/n = 52111 - 1213 \times 656/16 = 2378$$

Part a (8 Marks)

Subpart i

Calculate the correlation coefficient. How would the correlation coefficient change if the daily temperatures were measured in degrees Celsius instead of degrees Fahrenheit?

[1 mark for r, 1 mark for comment] The correlation coefficient is:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \frac{2378}{\sqrt{1164.4375 \times 19.972}} = 0.4931091$$

If the measurements where converted to Celsius from Fahrenheit then the correlation coefficient r would remain unchanged.

Subpart ii

Calculate the least squares regression fit and the proportion variability of y's is explained by the regression line.

[1 mark for b, 1 mark for a, 1 mark for r^2]

$$b = \frac{S_{xy}}{S_{xx}} = \frac{2378}{1164.4375} = 2.042187752$$

$$\begin{aligned} a &= \overline{y} - b\overline{x} \\ &= 656/16 - 2.042187752 \times 1213/16 \\ &= -113.8233589. \end{aligned}$$

The proportion ov variability explained by the regression fit is given by $r^2 = 0.4931091^2 = 0.243156584$.

Subpart iii

Use the R *output above to calculate a 5 number summary. Use the 5 number summary to comment on the distribution of the residuals.*

[1/2 for min, 1/2 for max, 1/2 for median, 1/2 for Q_1, 1/2 for Q_2, 1/2 for Q_3 and 1/2 for comment] From the R output we have

- ☐ $\min = -31$
- ☐ $Q_1 = \frac{-25-24}{2} = -24$
- ☐ $Q_2 = \frac{-7-4}{2} = -5.5$
- ☐ $Q_3 = \frac{1+4}{2} = 2.5$
- ☐ $\max = 77$

From the 5 number summary we can see that the data is left skewed. (The distance between Q_1 and Q_2 is larger than between Q_2 and Q_3. Note that we can't really use the minimum and maximum values to determine skewness because they can be made arbitrarily large without effecting the values of the quartiles.)

Part b (6 Marks) – Subpart i

Consider the paired data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

For a least squares regression fit show that the residuals $\widehat{e}_i = y_i - a - bx_i$ *satisfy*

$$\sum_{i=1}^{n} \widehat{e}_i = 0 \qquad \text{and} \qquad \sum_{i=1}^{n} \widehat{e}_i x_i = 0.$$

[1 mark for $\sum e_i = 0$, 3 marks for $\sum x_i \widehat{e}_i = 0$, 2 marks for $r = 0$] For the first part

$$\begin{aligned} \sum e_i &= \sum (y_i - a - bx_i) \\ &= n(\overline{y} - a - b\overline{x}) \\ &= 0 \end{aligned}$$

since $a = \overline{y} - b\overline{x}$.

Then

$$
\begin{aligned}
\sum x_i \widehat{e}_i &= \sum (y_i - abx_i) x_i \\
&= \sum (y_i - \overline{y} - b(x_i - \overline{x})) x_i \\
&= \sum (y_i - \overline{y} - b(x_i - \overline{x}))(x_i - \overline{x}) + \overline{x} \sum (y_i - \overline{y} - b(x_i - \overline{x})) \\
&= S_{xy} - bS_{xx} + 0 \quad (\text{since } \textstyle\sum (y_i - \overline{y}) = \sum (x_i - \overline{x}) = 0) \\
&= S_{xy} - \frac{S_{xy}}{S_{xx}} S_{xx} \\
&= 0
\end{aligned}
$$

Subpart ii

Hence, show that $\{x_i\}$ and $\{\widehat{e}_i\}$ are uncorrelated.

To calculate the (sample) correlation of x_i and $\widehat{e}_i$ we have

$$r = \frac{S_{xe}}{\sqrt{S_{ee}S_{xx}}}$$

where

$$S_{xe} = (\sum x_i\widehat{e}_i) - (\sum x_i)(\sum \widehat{e}_i)/n = 0 - (\sum x_i) \times 0/n = 0$$

since $\sum e_i = 0$ and $\sum x_i\widehat{e}_i = 0$. Hence $r = 0$.

Question 2 (18 marks in all)

Part a (9 Marks)

The clinically accepted value for mean blood pressure in healthy males aged 18 to 22 years is 120 mm Hg. It is widely claimed that examination stress causes blood pressure to rise above 120 mm Hg. To test this theory, 10 healthy male students have their blood pressure taken just prior to a Statistics quiz. The sample mean and sample standard deviation of these measurements are 135.1 and 17.42 respectively. Assume that the measurements for each student are independent and normally distributed.

Note: We have from the description $n = 10$, $\overline{x} = 135.1$ and $s = 17.42$.

Subpart i

What are appropriate null and alternative hypotheses to test this claim?

[1 mark for H_0, 1 mark for justifying H_1] Since "it is widely believed that exams increase blood pressure" the appropriate null and alternative hypotheses are:

$$H_0\colon \mu = 120 \qquad \text{vs} \qquad H_1\colon \mu > 120.$$

Subpart ii

State an appropriate test statistic to test this hypothesis and the null distribution of this test statistic.

[1 mark for T, 1 mark for justifying t_9] The appropriate test statistic is

$$T = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

where $\mu_0 = 120$. Assuming independence and normality under the null hypothesis $T \sim t_{n-1} = t_9$.

Subpart iii

Calculate the test statistic chosen in (ii), the corresponding P-value and form an appropriate conclusion.

[2 mark for T, 1 mark for $P = P(t_9 > 2.74112472)$, 1 mark for bounding P-value, 1 mark for rejecting null, 1 mark for relating to relating to question] The value of the test statistic is

$$T = \frac{135.1 - 120}{17.42/\sqrt{10}} = 2.74112472$$

Note

$$P(t_9 > 2.262) = 0.025 \quad \text{and} \quad P(t_9 > 2.821) = 0.01$$

Hence the P-value is between 1% and 2.5% which is below 5%. Hence, we reject the null hypothesis. There is sufficient evidence to reject the claim that exam stress does not cause blood pressure to rise.

Part b (9 Marks) – Subpart i

Consider the geometric distribution $P(X = x) = pq^x$, $x = 0, 1, 2, \ldots$, $0 \leq p \leq 1$ *and* $q = 1 - p$. *Show that the probability generating function is given by* $\pi(s) = p/(1 - qs)$ *for* $|s| < 1/q$.

[1/2 mark for defining $\pi(s)$, 1/2 for $p\sum_{i=0}^{\infty}(sq)^i$, 1/2 mark for $\frac{p}{1-qs}$, 1/2 mark for $|s| < 1/q$] We have

$$P(X = x) = pq^x, \qquad x = 0, 1, 2, \ldots$$

The probability generating function is

$$\pi(s) = \sum_{i=0}^{\infty} s^i P(X - i) = \sum_{i=0}^{\infty} s^i pq^i = p\sum_{i=0}^{\infty}(sq)^i = \frac{p}{1 - qs}$$

provided $|qs| < 1$ or equivalently $|s| < 1/q$ (since $q > 0$).

Subpart ii

Use Part (i) to show that $E(X) = q/p$ and Var$(X) = q/(p^2)$.

[1 mark for $\pi'(s)$, 1 mark for $\pi''(s)$, 1 mark for $E(X)$, 1 mark for Var(X)] Note that

$$\pi'(s) = \frac{pq}{(1-qs)^2} \qquad \text{and} \qquad \pi''(s) = \frac{2pq^2}{(1-qs)^3}.$$

Now

$$E(X) = \pi'(1) = \frac{pq}{(1-q)^2} = \frac{pq}{p^2} = \frac{q}{p}$$

and

$$\text{Var}(X) = \pi''(1) + \pi'(1) - \pi'(1)^2 = \frac{2pq^2}{(1-q)^3} + \frac{p}{q} - \frac{q^2}{p^2} = \frac{q(p+q)}{p^2} = \frac{q}{p^2}.$$

Subpart iii

Use Part (ii) and Chebyshev's inequality to bound $P(|X - q/p| > 1)$.

[1/2 mark for Chebyshev's inequality, 1/2 mark for identifying σ, 1 for identifying c, 1 mark for bound] Chebyshev's inequality is

$$P(|X - \mu| > c\sigma) < 1/c^2.$$

We have $\mu = q/p$ and $\sigma = \sqrt{q}/p$.

$$P(|X - p/q| > 1) = P(|X - q/p| > p/\sqrt{q} \times \sqrt{q}/p) < 1/(p/\sqrt{q})^2 = q/p^2$$

where $c = p/\sqrt{q}$.

Question 3 (15 marks in all)

Part a (10 marks)

It has been claimed that at least 60% of all purchasers of a certain computer program will call the manufacturer's hotline within one month of purchase. A random sample of 12 purchasers of this software is drawn and 3 of those in the sample had contacted the hotline within one month of purchase. Does this provide evidence that the claim of a 60% contact rate is an overestimate? Let p be the true proportion of all purchasers who contact the hotline.

From the question description we have $p_0 = 0.6$, $n = 12$ and $X = 3$.

Subpart i

Calculate an approximate 95% confidence interval for p.

[1 mark for formula, 1 mark for $\widehat{p}$, 1 mark for calculation] The 95% confidence interval for p is given by

$$\widehat{p} \pm 1.96\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = \frac{1}{4} \pm 1.96\sqrt{\frac{3/16}{12}} = 0.25 \pm 0.245 = (0.005, 0.495).$$

where $\widehat{p} = 3/12 = 1/4$.

Subpart ii

Form an appropriate hypotheses and perform a statistical test for the above situation stating any assumptions you may require.

[1 mark for assuming $X \sim B(n, p)$, 1 mark for hypotheses, 1 mark for null distribution, 1 mark for P-value, 1 mark for conclusion] Assume constant probability over each trial, each trial is independent so that $X \sim B(n, p)$. Appropriate null and alternative hypotheses are:

$$H_0\colon p = 0.6 \qquad \text{and} \qquad H_1\colon p < 0.6.$$

The test statistic is $X = 3$ with small values of X supporting H_1. Under H_0 we have $X \sim B(12, 0.6)$. Then

$$\begin{aligned} P(X \leq 3) &= P(X = 3) + P(X = 2) + P(X = 1) + P(X = 0) \\ &\approx 0.01246 + 0.0025 + 0.0003 + 10^{-5} \\ &\approx 0.0153 \end{aligned}$$

Also, using the normal approximation $X \approx Y \sim N(np, np(1-p)) = N(7.2, 2.88)$ and

$$\begin{aligned} P(X \leq 3) &\approx P(Y \leq 3.5) \\ &= P(Z \leq (3.5 - 7.2)/\sqrt{(}2.88)) \\ &= P(Z \leq -2.180246) \\ &= 1 - P(Z \leq 2.18) \\ &= 1 - 0.9854 \\ &= 0.0146 \end{aligned}$$

(from tables) where $Z \sim N(0,1)$. Either way the P-value is less than 5%. Hence we reject H_0 and conclude that there is sufficient to suggest that 60% is too high.

Subpart iii

Do the results from part (i) and part (ii) agree? Justify your answer.

[1 mark for stating CI does not contain 0.6, 1 mark for comment] Since that 95% CI does not contain the value $p_0 = 0.6$ and we reject $H_0\colon p = 0.6$ the conclusions are consistent.

Part b (5 marks)

Suppose that the probability that a randomly chosen individual having a particular disease is 0.02.

From the problem description we have $n = 100$ and $p = 0.02$ with $X \sim B(100, 0.02)$.

Subpart i

Write an expression for the exact probability that at most 2 cases in a random sample of 100 has the disease.

[1/2 mark for n, 1/2 mark for p, 1 mark for $P(X \leq 2)$] The expresssion for $P(X \leq 2)$ is

$$P(X \leq 2) = \binom{100}{2}(0.02)^2(0.98)^{98} + 100(0.02)(0.98)^{99} + (0.98)^{100}$$

Subpart ii

Use the Poisson approximation to the Binomial distribution to approximate this probability. Why is this approximation appropriate?

[1 mark for $Y \sim P(2)$, 1 mark for $P(Y \leq 2) = 5e^{-2}$, 1 mark for comment] The Poisson approximation is $X \approx Y \sim P(np) = P(2)$. Then

$$P(Y \leq 2) = \frac{2^2 e^{-2}}{2} + \frac{2e^{-2}}{1} + \frac{2^0 e^{-2}}{1} = 5e^{-2} = 0.6766764.$$

This approximation is appropriate since $n = 100$ is large and $p = 0.02$ is small.

Question 4 (18 marks in all) – Part a (5 marks)

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 101 times, with the following observed counts:

Sixes Rolled	*0*	*1*	*2*	*3*
Number of Rolls	*48*	*35*	*15*	*3*

You may use the R output:

```
> dbinom(0:3,3,1/6)
[1] 0.57870370 0.34722222 0.06944444 0.00462963
```

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. Use a statistical test to determine where whether the dice are fair or not.

[1/2 mark for g, 1/2 mark for E_i, 1/2 mark for combining, 1 mark for X^2, 1/2 mark for $X^2 \sim \chi^2_2$, 1/2 mark for P-value, 1/2 mark for conclusion] There are $g = 4$ groups and

O_i	48	35	15	3
E_i	58.449	35.069	7.014	0.468

Since the last column contains both expected and observed frequencies less than 5 we combine the last two columns to obtain the table:

O_i	48	35	18
E_i	58.449	35.069	7.481

The observed value of the Pearson's chi-square statistic is

$$X^2 = \sum_{i=1}^{g} \frac{O_i^2}{E_i} - n = 38.81 + 35.28 + 43.74 - 101 = 16.66 \quad \text{(to 2 d.p.)}$$

Here $g = 3$. Under the null hypothesis $X^2 \sim \chi^2_{g-1} = \chi^2_2$ and the P-value is given by $P(\chi^2_2 > 16.66) < 0.01$ from tables. This we reject the claim that the dice are fair.

Part b (9 marks)

Two pathology labs, lab A and lab B, are compared to see which of the labs report their results for a specific test faster. Samples from 20 patients are collected and then 10 of these samples are sent to lab A and 10 of these samples are sent to lab B. Summary values for the times, in days, for each lab to report its results are summarised in the table below:

	Size	Mean	Median	Variance
Lab A	10	20.23	19.45	2.74
Lab B	10	18.68	17.98	1.64

You may assume that the measurements for lab A and lab B are normally distributed. Suppose we wish to test whether there is a difference in the times each lab reports its results.

Subpart i

State an appropriate null and alternative hypothesis, defining any parameters used.

[$1/2$ mark for H_0, $1/2$ mark for H_1, 1 mark for definitions]

$$H_0\colon \mu_x = \mu_y \qquad \text{vs} \qquad H_1\colon \mu_x \neq \mu_y$$

where $\mu_x = \{$mean time form lab A$\}$ and $\mu_y = \{$mean time form lab B$\}$.

Subpart ii

State an appropriate test statistic to test this hypothesis and the null distribution of this test statistic, stating any additional assumptions required.

[1 mark for justifying 2 sample t-test, 1/2 mark for test statistic, 1/2 mark for null distribution, 1 mark for equal variance assumption] The two sets of samples are independent so that a 2 sample t-test is appropriate. The appropriate test statistic is:

$$T = \frac{\overline{X} - \overline{Y}}{S_p\sqrt{1/n_x + 1/n_y}}$$

Assuming equal variance the null distribution of T is $t_{n_x+n_y-2} = t_{18}$.

Subpart ii

Calculate the test statistic chosen in (ii), the corresponding P-value and form an appropriate conclusion.

[1 mark for s_p^2, 1 mark for T, 1 mark for P-value, 1 mark for conclusion] The pooled variance is

$$s_p^2 = \frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2} = \frac{1}{2}(2.74 + 1.64) = 2.19$$

The observed value of the test statistic is

$$T = \frac{20.23 - 18.68}{\sqrt{2.19}\sqrt{2/10}} = 2.342.$$

The P-value is given by

$$P = P(|t_{18}| > 2.2342) = 2P(t_{18} > 2.2342).$$

From tables $P(t_{10} > 2.228) = 0.025$ and $P(t_{20} > 2.528) = 0.01$. Hence $0.02 < P < 0.05$. We reject H_0 and conclude that the two labs have different mean times.

Part b (4 marks)

Let Ω be a sample space, $A \subset \Omega$, $B \subset \Omega$ and $P(\cdot)$ be a probability function satisfying the axioms of probability

- *For any event $A \subset \Omega$, $P(A) \geq 0$,*
- $P(\Omega) = 1$
- *If A and B are mutually exclusive events ($A \cap B = \emptyset$), then*

$$\mathrm{P}(A \cup B) = \mathrm{P}(A) + \mathrm{P}(B).$$

$1 = P(\Omega)$	by axiom 2 (1/2 mark)
$P(\Omega) = P(A \cup A^c)$	by set properties (1 mark)
$P(A \cup A^c) = P(A) + P(A^c)$	by axiom 3 (1 mark)
$P(A) = 1 - P(A^c)$	(1/2 mark)
$P(A) = 1 - P(A^c) \leq 1$	since $P(A^c) \geq 0$ by axiom 1 (1 mark)