Goodness of Fit Tests (Year 2) (From OCR 4768)

Q1, (Jun 2006, Q1)

Design engineers are simulating the load on a particular part of a complex structure. They intend that the simulated load, measured in a convenient unit, should be given by the random variable X having probability density function

$$f(x) = 12x^3 - 24x^2 + 12x, \qquad 0 \leqslant x \leqslant 1.$$

(i) Find the mean and the mode of X. [6]

(ii) Find the cumulative distribution function $F(x)$ of X.

Verify that $F\left(\frac{1}{4}\right) = \frac{67}{256}$, $F\left(\frac{1}{2}\right) = \frac{11}{16}$ and $F\left(\frac{3}{4}\right) = \frac{243}{256}$. [3]

The engineers suspect that the process for generating simulated loads might not be working as intended. To investigate this, they generate a random sample of 512 loads. These are recorded in a frequency distribution as follows.

Load x	$0 \leqslant x \leqslant \frac{1}{4}$	$\frac{1}{4} < x \leqslant \frac{1}{2}$	$\frac{1}{2} < x \leqslant \frac{3}{4}$	$\frac{3}{4} < x \leqslant 1$
Frequency	126	209	131	46

(iii) Use a suitable statistical procedure to assess the goodness of fit of X to these data. Discuss your conclusions briefly. [9]

Q2, (Jan 2007, Q4a)

An amateur weather forecaster has been keeping records of air pressure, measured in atmospheres. She takes the measurement at the same time every day using a barometer situated in her garden. A random sample of 100 of her observations is summarised in the table below. The corresponding expected frequencies for a Normal distribution, with its two parameters estimated by sample statistics, are also shown in the table.

Pressure (a atmospheres)	Observed frequency	Frequency as given by Normal model
$a \leqslant 0.98$	4	1.45
$0.98 < a \leqslant 0.99$	6	5.23
$0.99 < a \leqslant 1.00$	9	13.98
$1.00 < a \leqslant 1.01$	15	23.91
$1.01 < a \leqslant 1.02$	37	26.15
$1.02 < a \leqslant 1.03$	21	18.29
$1.03 < a$	8	10.99

Carry out a test at the 5% level of significance of the goodness of fit of the Normal model. State your conclusion carefully and comment on your findings. [9]

Q3, (Jun 2007, Q4i)

A machine produces plastic strip in a continuous process. Occasionally there is a flaw at some point along the strip. The length of strip (in hundreds of metres) between successive flaws is modelled by a continuous random variable X with probability density function $f(x) = \frac{18}{(3+x)^3}$ for $x > 0$. The table below gives the frequencies for 100 randomly chosen observations of X. It also gives the probabilities for the class intervals using the model.

Length x (hundreds of metres)	Observed frequency	Probability
$0 < x \leqslant 0.5$	21	0.2653
$0.5 < x \leqslant 1$	24	0.1722
$1 < x \leqslant 2$	12	0.2025
$2 < x \leqslant 3$	15	0.1100
$3 < x \leqslant 5$	13	0.1094
$5 < x \leqslant 10$	9	0.0874
$x > 10$	6	0.0532

(i) Examine the fit of this model to the data at the 5% level of significance. [9]

Q4, (Jan 2008, Q4a)

In Germany, towards the end of the nineteenth century, a study was undertaken into the distribution of the sexes in families of various sizes. The table shows some data about the numbers of girls in 500 families, each with 5 children. It is thought that the binomial distribution $B(5, p)$ should model these data.

Number of girls	Number of families
0	32
1	110
2	154
3	125
4	63
5	16

(i) Use this information to calculate an estimate for the mean number of girls per family of 5 children. Hence show that 0.45 can be taken as an estimate of p. [3]

(ii) Investigate at a 5% significance level whether the binomial model with p estimated as 0.45 fits the data. Comment on your findings and also on the extent to which the conditions for a binomial model are likely to be met. [12]

Q5, (Jun 2008, Q4a)

A researcher is investigating the feeding habits of bees. She sets up a feeding station some distance from a beehive and, over a long period of time, records the numbers of bees arriving each minute. For a random sample of 100 one-minute intervals she obtains the following results.

Number of bees	0	1	2	3	4	5	6	7	$\geqslant 8$
Number of intervals	6	16	19	18	17	14	6	4	0

(i) Show that the sample mean is 3.1 and find the sample variance. Do these values support the possibility of a Poisson model for the number of bees arriving each minute? Explain your answer. [3]

(ii) Use the mean in part **(i)** to carry out a test of the goodness of fit of a Poisson model to the data. [10]

Q6, (Jan 2010, Q1)

Coastal wildlife wardens are monitoring populations of herring gulls. Herring gulls usually lay 3 eggs per nest and the wardens wish to model the number of eggs per nest that hatch. They assume that the situation can be modelled by the binomial distribution $B(3, p)$ where p is the probability that an egg hatches. A random sample of 80 nests each containing 3 eggs has been observed with the following results.

Number of eggs hatched	0	1	2	3
Number of nests	7	23	29	21

(i) Initially it is assumed that the value of p is $\frac{1}{2}$. Test at the 5% level of significance whether it is reasonable to suppose that the model applies with $p = \frac{1}{2}$. **[10]**

(ii) The model is refined by estimating p from the data. Find the mean of the observed data and hence an estimate of p. **[2]**

(iii) Using the estimated value of p, the value of the test statistic X^2 turns out to be 2.3857. Is it reasonable to suppose, at the 5% level of significance, that this refined model applies? **[3]**

(iv) Discuss the reasons for the different outcomes of the tests in parts **(i)** and **(iii)**. **[2]**

Q7, (Jun 2014, Q3b)

An ornithologist believes that the number of birds landing on the bird feeding station in her garden in a given interval of time during the morning should follow a Poisson distribution. In order to test her belief, she makes the following observations in 60 randomly chosen minutes one morning.

Number of birds	0	1	2	3	4	5	6	$\geqslant 7$
Frequency	2	5	10	17	14	7	4	1

Given that the data in the table have a mean value of 3.3, use a goodness of fit test, with a significance level of 5%, to investigate whether the ornithologist is justified in her belief. **[11]**

Q8, (Jan 2011, Q3)

The masses, in kilograms, of a random sample of 100 chickens on sale in a large supermarket were recorded as follows.

Mass (m kg)	$m < 1.6$	$1.6 \leqslant m < 1.8$	$1.8 \leqslant m < 2.0$	$2.0 \leqslant m < 2.2$	$2.2 \leqslant m < 2.4$	$2.4 \leqslant m < 2.6$	$2.6 \leqslant m$
Frequency	2	8	30	42	11	5	2

(i) Assuming that the first and last classes are the same width as the other classes, calculate an estimate of the sample mean and show that the corresponding estimate of the sample standard deviation is 0.2227 kg. **[3]**

A Normal distribution using the mean and standard deviation found in part **(i)** is to be fitted to these data. The expected frequencies for the classes are as follows.

Mass (m kg)	$m < 1.6$	$1.6 \leqslant m < 1.8$	$1.8 \leqslant m < 2.0$	$2.0 \leqslant m < 2.2$	$2.2 \leqslant m < 2.4$	$2.4 \leqslant m < 2.6$	$2.6 \leqslant m$
Expected frequency	2.17	10.92	f	33.85	19.22	5.13	0.68

(ii) Use the Normal distribution to find f. **[3]**

(iii) Carry out a goodness of fit test of this Normal model using a significance level of 5%. **[9]**

(iv) Discuss the outcome of the test with reference to the contributions to the test statistic and to the possibility of other significance levels. **[3]**

Q9, (Jun 2013, Q3)

The random variable X has the following probability density function, $\mathrm{f}(x)$.

$$\mathrm{f}(x) = \begin{cases} kx(x-5)^2 & 0 \leqslant x < 5 \\ 0 & \text{elsewhere} \end{cases}$$

(i) Sketch $\mathrm{f}(x)$. **[3]**

(ii) Find, in terms of k, the cumulative distribution function, $\mathrm{F}(x)$. **[3]**

(iii) Hence show that $k = \frac{12}{625}$. **[2]**

The random variable X is proposed as a model for the amount of time, in minutes, lost due to stoppages during a football match. The times lost in a random sample of 60 matches are summarised in the table. The table also shows some of the corresponding expected frequencies given by the model.

Time (minutes)	$0 \leqslant x < 1$	$1 \leqslant x < 2$	$2 \leqslant x < 3$	$3 \leqslant x < 4$	$4 \leqslant x < 5$
Observed frequency	5	15	23	11	6
Expected frequency			17.76	9.12	1.632

(iv) Find the remaining expected frequencies. **[3]**

(v) Carry out a goodness of fit test, using a significance level of 2.5%, to see if the model might be suitable in this context. **[8]**
