Probability and Statistics 1 2005–06

Progress Test

Wednesday 4 January 2005, 10.00–11.30 am

Answer all 6 questions.
The total number of marks available is 50.

1. A sample of 18 companies is taken from the Eurotop 100. In each case the Price/Earnings ratio is noted down, with the following results (sorted into ascending order):
9.4 9.7 12.0 12.8 12.9 14.5 14.7 15.3 16.1 16.3 16.4 17.2 17.4 17.7 23.1 23.3 27.3 33.6

(i) Use a stem-and-leaf diagram to illustrate the data. **[2 marks]**

(ii) Calculate the median and quartiles. Are outliers present? **[4 marks]**

(iii) Draw a box-and-whisker plot to illustrate the data set. **[2 marks]**

(iv) The sample mean of the data is 17.2. Use **three** different methods to assess the symmetry or skewness of the sample. **[3 marks]**

2. A schoolchild is undertaking a statistical project.
He has noted down the numbers of buses passing along a busy road for half an hour around 11am and half an hour around 11pm. He has drawn a stem-and-leaf diagram (right) to illustrate the data.

```
    Night |     | Day
    73 91 |   0 | 73 73 28 14 73 28 73 14 73 28 73 91
 01 13 26 | 100 | 01 26 01 13 26 13 01 26
    19 19 | 200 | 19 37 19 19 19
          | 600 | 76 76 76 76
  N676 N3 |   ? | B1 W4 W4 W4 B1
```

The child tells you that he wants to draw a histogram but is having problems with it.

(i) Explain why a histogram is inappropriate and comment on the child's stem-and-leaf diagram. **[2 marks]**

(ii) What would be the best form of diagram to summarise the data on daytime buses? Give a reason for your answer. **[1 mark]**

(iii) Which of the sample mean, the median and the mode would be a suitable summary measure, and why? **[1 mark]**

(iv) Calculate the modes of the daytime data and of the nighttime data. What can you learn by comparing the two values? **[2 marks]**

3. A gambler rolls two ordinary dice. If the two scores are equal (event A) he wins £2; if the two scores sum to 3 or 11 (event B), he wins £3; and if one or both of the dice shows a six (event C), he wins £1. If the outcome of the dice roll implies that he wins more than one prize, the prizes are added together.

(i) Calculate $\mathbb{P}(A)$, $\mathbb{P}(B)$, $\mathbb{P}(C)$, $\mathbb{P}(A \cap B)$, $\mathbb{P}(B \cap C)$ and $\mathbb{P}(A \cap C)$. State whether any two of A, B and C are independent or mutually exclusive. **[4 marks]**

(ii) Draw a Venn diagram to show the relationships between A, B and C. (You are **not** expected to include the probabilities on the diagram.) **[1 mark]**

(iii) Let X be the random variable which denotes the amount the gambler wins on one roll of the dice. Write down the probability function $p_X(x)$. **[2 marks]**

(iv) If it costs £1 to play, does the gambler win or lose money on average? **[1 mark]**

4. Marvin has coursework to finish, but it is dinner time. There is a 40% chance that he has the necessary ingredients to cook dinner for himself, in which case he will certainly be able to finish the coursework. Otherwise he eats in the refectory (probability $\frac{2}{3}$) or goes to the supermarket and then cooks for himself (probability $\frac{1}{3}$). Eating in the refectory will leave Marvin time to finish the coursework unless he meets friends and they persuade him to go to the pub (probability 30%; no chance of finishing the coursework). A visit to the supermarket will take time and leave him with a one in four chance of being unable to finish the coursework.

(i) Draw a tree diagram to illustrate the situation. **[2 marks]**

(ii) What is the probability that Marvin finishes the coursework? **[2 marks]**

(iii) Given that the coursework does not get finished, how likely is it that Marvin visited the pub? **[2 marks]**

5. A company has a new automated system for handling customer feedback on a scale from 1 (very good) to 5 (very poor). In order to test the system the company simulates feedback by generating observations from a discrete distribution with probability function

$$p(x) = \begin{cases} \frac{c}{x} & \text{for } x = 1, 2, 3, 4, 5 \\ 0 & \text{all other } x \end{cases}$$

where c is a constant.

(i) Calculate the value of the constant c. **[1 mark]**

(ii) Calculate the expectation and variance of a random variable X whose probability function is $p(x)$. **[3 marks]**

(iii) 10 independent feedback responses are generated, each with probability function $p(x)$. Let Y be the number which are "poor" (4) or "very poor" (5). Find $\mathbb{P}(Y \leq 2)$. **[2 marks]**

(iv) Use a Poisson approximation to calculate $\mathbb{P}(Y \leq 2)$ and comment on the result. **[3 marks]**

6. The density function, f_X, of a continuous random variable X is given by

$$f_X(x) = \begin{cases} e^{-2x} & \text{if } x > 0 \\ e^{2x} & \text{if } x < 0 \end{cases}.$$

(i) Sketch the function f_X and find the expectation of X. **[2 marks]**

(ii) Calculate the cumulative distribution function, F_X, of X. (You will need to obtain one formula for $x < 0$ and a different one for $x > 0$.) **[3 marks]**

(iii) Describe the principal difference between the distribution function of a continuous random variable and that of a discrete random variable. **[1 mark]**

(iv) Find $\mathbb{P}(-1 < X < +2)$. **[1 mark]**

(v) Define $Y = |X|$, so that $\{Y \leq y\} \iff \{-y \leq X \leq +y\}$. For $y > 0$ calculate $\mathbb{P}(Y \leq y)$. Use your answer to identify the distribution of Y. **[3 marks]**

Formulae

$$\mathbb{P}(A \cup B \cup C) = \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \mathbb{P}(A \cap B \cap C).$$

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}$$

Distribution	Notation	Mean	Variance	$p(x)$	Domain
Binomial	$\text{Bin}(k, \theta)$	$k\theta$	$k\theta(1-\theta)$	$\binom{k}{x}\theta^x (1-\theta)^{k-x}$	$x = 0, 1, \ldots, k$
Poisson	$\text{Pois}(\lambda)$	λ	λ	$e^{-\lambda}\frac{\lambda^x}{x!}$	$x = 0, 1, \ldots$
Geometric	$\text{Geom}(\theta)$	θ^{-1}	$\theta^{-2}(1-\theta)$	$\theta(1-\theta)^{x-1}$	$x = 1, 2, \ldots$

Probability and Statistics I 2005–06

Solutions to Progress Test

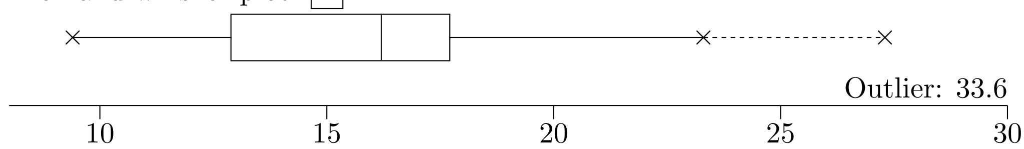
1. (i) A stem interval of 5 looks best.

```
30 | 36                Key 30 | 36 = 33.6
20 | 73
20 | 31 33
10 | 53 61 63 64 72 74 77
10 | 20 28 29 45 47
 0 | 94 97
```

$\boxed{2}$

(ii) Since there are 18 values, the quartiles are the 5th from the bottom and 5th from the top (12.9 and 17.7). The median is the average of the 9th and 10th (16.2). $\boxed{2}$
With an IQR of 4.8, anything outside $(-1.5, 32.5)$ is a definite outlier: this includes 33.6. Any other values outside $(5.7, 24.9)$ are possible outliers: this includes 27.3. $\boxed{2}$

(iii) Box-and-whisker plot: $\boxed{2}$



(iv) The sample mean is larger than the median; this indicates right (positive) skew.
UQ−Median = 1.5, whereas Median−LQ = 3.3. This tends to indicate left (negative) skew.
Judging the stem-and-leaf diagram by eye seems to suggest a slight right skew.
Finally, it is conceivable that students could calculate the coefficient of skewness. I get 1.16, indicating a definite positive skew. $\boxed{\text{1 each, maximum 3}}$

2. (i) The data are qualitative, not quantitative. Thus histogram is not appropriate. The grouping of buses by the numbers on the front is illusory: similar numbers do not indicate "closeness" in any reasonable sense, and so the idea of using a stem-and-leaf diagram is also flawed. $\boxed{2}$

(ii) A bar chart or a pie chart would be best to illustrate the relative frequencies of the various buses. These are the only suitable methods for nominal (unordered qualitative) data. $\boxed{1}$

(iii) For qualitative data we can only use the mode. $\boxed{1}$

(iv) Daytime 73, nighttime 219. We can't really conclude anything except that nighttime bus schedules are different from daytime schedules. $\boxed{2}$

3. The following diagram of the sample space allows us to see which outcomes belong to A, B and C. Also included on the diagram is an indication of the value of X for that outcome.

		Second die					
		1	2	3	4	5	6
	1	A 2	B 3	0	0	0	C 1
	2	B 3	A 2	0	0	0	C 1
First	3	0	0	A 2	0	0	C 1
die	4	0	0	0	A 2	0	C 1
	5	0	0	0	0	A 2	BC 4
	6	C 1	C 1	C 1	C 1	BC 4	AC 3

(i) The probabilities can be read off the diagram: $\boxed{3}$

$\mathbb{P}(A) = \dfrac{6}{36}, \mathbb{P}(B) = \dfrac{4}{36}, \mathbb{P}(C) = \dfrac{11}{36}, \mathbb{P}(A \cap B) = 0, \mathbb{P}(B \cap C) = \dfrac{2}{36}$ and $\mathbb{P}(A \cap C) = \dfrac{1}{36}$.

A and B are clearly mutually exclusive. No two of the events are independent. $\boxed{1}$

(ii) Venn diagram. $\boxed{1}$

(iii) Again from the diagram, or derived from the list of probabilities: $\boxed{2}$

x	$p_X(x)$
0	$\frac{18}{36}$
1	$\frac{8}{36}$
2	$\frac{5}{36}$
3	$\frac{3}{36}$
4	$\frac{2}{36}$

(iv) Calculate the average winnings using $\mathbb{E}(X) = \sum_x x p_X(x) = \frac{35}{36}$. If it costs 1 to play, the gambler loses money on average (though not much). $\boxed{1}$

4. (i) Tree diagram. $\boxed{2}$

(ii) Let I be the event that he has ingredients already, R the event that he eats in the refectory, C the event that he finishes the coursework on time. Then $\mathbb{P}(I) = 0.4$, $\mathbb{P}(C \mid I) = 1$, $\mathbb{P}(R \mid I^c) = \frac{2}{3}$, $\mathbb{P}(C \mid I^c \cap R) = 0.7$ and $\mathbb{P}(C \mid I^c \cap R^c) = 0.75$.
Therefore

$$\begin{aligned}\mathbb{P}(C) &= \mathbb{P}(I)\mathbb{P}(C \mid I) + \mathbb{P}(I^c)\{\mathbb{P}(R \mid I^c)\mathbb{P}(C \mid R \cap I^c) + \mathbb{P}(R^c \mid I^c)\mathbb{P}(C \mid R^c \cap I^c)\} \\ &= 0.4 + 0.6\left[\frac{2}{3} \times 0.7 + \frac{1}{3} \times 0.75\right] = 0.83. \quad \boxed{2}\end{aligned}$$

(iii) We want $\mathbb{P}(R \mid C^c) = \mathbb{P}(R \cap C^c)/\mathbb{P}(C^c) = 0.12/(1 - 0.83) = \frac{12}{17}$. $\boxed{2}$

5. (i) $1 = c\left(1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5}\right) = \frac{137}{60}c$, so that $c = \frac{60}{137}$. $\boxed{1}$

(ii) $\mathbb{E}(X) = \sum_1^5 x p(x) = 5c = \frac{300}{137} = 2.190$. $\boxed{1}$
$\mathbb{E}(X^2) = \sum_1^5 x^2 p(x) = 15c = \frac{900}{137} = 6.569$. $\boxed{1}$
$\mathrm{Var}\,(X) = 6.569 - 2.190^2 = 1.774$. $\boxed{1}$

(iii) $Y \sim \mathrm{Bin}\left(10, \frac{27}{137}\right)$. $\boxed{1}$
We require $\mathbb{P}(Y \le 2) = p_Y(0) + p_Y(1) + p_Y(2) = 0.1114 + 0.2733 + 0.3019 = 0.6866$. $\boxed{1}$

(iv) According to the Poisson approximation, $Y \approx \mathrm{Pois}\left(\frac{270}{137}\right) = \mathrm{Pois}(1.971)$. Thus $\mathbb{P}(Y \le 2) \approx e^{-1.971}(1 + 1.971 + \frac{1}{2} \times 1.971^2 = 0.1393 + 0.2746 + 0.2706 = 0.6845$. $\boxed{2}$
This is pretty close to the true probability, but we notice that the Poisson approximations to $p(0)$ and $p(2)$ are not very close to the true figures. $\boxed{1}$

6. (i) Sketch rises to a point (not rounded) at $x = 0$ and is symmetric about 0. $\boxed{1}$
By symmetry, $\mathbb{E}(X) = 0$. $\boxed{1}$

(ii) For $x < 0$, $F(x) = \int_{-\infty}^{x} e^{2y}\,dy = \frac{1}{2}e^{2x}$. $\boxed{1}$
For $x > 0$, $F(x) = \frac{1}{2} + \int_0^x e^{-2y}\,dy = 1 - \frac{1}{2}e^{-2x}$. $\boxed{2}$

(iii) The cdf of a continuous r.v. is a continuous function, whereas the cdf of a discrete r.v. is flat almost everywhere and changes only by means of jumps. $\boxed{1}$

(iv) $F(2) - F(-1) = 1 - \frac{1}{2}e^{-4} - \frac{1}{2}e^{-2} = 1 - 0.0092 - 0.0677 = 0.9215$. $\boxed{1}$

(v) $\mathbb{P}(Y \le y) = \mathbb{P}(-y \le X \le y) = F(y) - F(-y) = 1 - e^{-2y}$. $\boxed{2}$
Differentiating, we obtain $f_Y(y) = 2e^{-2y}$, the density of an exponential random variable with rate 2. $\boxed{1}$