Probability and Statistics I 2004–05

Progress Test

Wednesday 12 January 2005, 6.00–7.30 pm

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

1. A group of ten wine-tasters is asked to give an overall rating out of 50 for a bottle of red wine. The values recorded are

23 36 34 39 28 33 35 40 37 40.

A second group, of nine tasters, is asked to record ratings for a bottle of white wine. Their results are

28 32 34 31 26 29 35 42 32.

Use a stem-and-leaf diagram and a box-and-whisker plot to compare the two data sets. Comment on location, spread and skewness of the data. **[8 marks]**

2. A group of Engineering students is given a Maths test on arrival at university and another test, of equivalent level of difficulty, six weeks later. Their perfomances on the second test, compared with the first, are tabulated as follows:

Range	**Frequency**
15% worse – 6% worse	3
0 – 5% worse	12
1 – 5% better	15
6 – 10% better	10
11 – 20% better	10
21 – 30% better	8
31 – 50% better	2

Construct a histogram to display these results and calculate the average improvement in the students' scores over the six weeks. **[7 marks]**

3. A pack of 40 cards consists of the numbers 1 to 10 in each of 4 colours. A hand of three cards is dealt from the pack. Define the following events:

$A =$"the numbers on the three cards form a sequence"
$B =$"all three cards are the same colour"
$C =$"the numbers on the three cards are all different"
$D =$"the numbers on the three cards are all the same"
$E =$"two of the cards show the same number but the third is different"

(i) Define a *partition* and explain why $\{C, D, E\}$ forms a partition of the sample space. **[2 marks]**

(ii) Draw a Venn diagram illustrating the relationships between the five events listed above. (You are **not** expected to include the probabilities on the diagram.) **[3 marks]**

(iii) Calculate the probabilities of the events C, D and E. **[3 marks]**

(iv) Determine whether A and B are independent. **[3 marks]**

4. A motor insurance company classifies a quarter of all policyholders as Low Risk, a third as High Risk and the rest as Medium Risk. The probability of making one or more claims in a year is 10% for Low Risk policyholders, 20% for Medium Risk and 30% for High Risk.

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

(ii) What proportion of policyholders make no claims during a year? **[1 marks]**

(iii) If a policyholder makes one or more claims, what is the probability that he was originally classified as High Risk? **[2 marks]**

(iv) A manager inspects the records of five policyholders randomly selected from all those who have not made a claim during the year. Find the probability that at least three of them were classified as Medium Risk. **[3 marks]**

5. A block of flats contains 50 flats of identical size. According to the records, the number of occupants in each flat is as follows:

Number of occupants, x	0	1	2	3	4	>4
Frequency, f_x	1	5	10	21	13	0

(i) Calculate the sample mean and variance of the number of occupants in a flat. **[3 marks]**

(ii) If the distribution of the number of occupants were Poisson, with mean equal to the sample mean you have calculated, find the expected proportion of flats with x occupants, for $x = 0, 1, 2, 3, 4$. **[2 marks]**

(iii) Use a suitable graph to compare the expected proportions under the Poisson model with the proportions actually observed. Comment on any differences you notice. **[2 marks]**

(iv) It is suggested that Binomial$(4, \theta)$ might be a better model than Poisson. Calculate a suitable value of θ and, by comparing the variance of the Binomial model with that of the Poisson model and the observed sample variance, state whether you think this suggestion is reasonable. **[2 marks]**

6. A density function, f, is given by

$$f(x) = \begin{cases} \frac{5}{6} + x - x^2 & \text{if } 0 < x < K \\ 0 & \text{otherwise} \end{cases}.$$

(i) Sketch the function f. **[1 marks]**

(ii) Derive a cubic equation satisfied by K. Show that $K = 2$ is one solution and determine the others. Giving your reason, state which of these values is the required solution. **[3 marks]**

(iii) If X has density f, calculate $\mathbb{E}(X)$ and $\mathrm{Var}\,(X)$. **[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	Bin(k, θ)	$k\theta$	$k\theta(1-\theta)$	$\binom{k}{x}\theta^x (1-\theta)^{k-x}$	$x = 0, 1, \ldots, k$
Poisson	Pois(λ)	λ	λ	$e^{-\lambda}\frac{\lambda^x}{x!}$	$x = 0, 1, \ldots$
Geometric	Geom(θ)	θ^{-1}	$\theta^{-2}(1-\theta)$	$\theta\,(1-\theta)^{x-1}$	$x = 1, 2, \ldots$

Probability and Statistics I 2004–05

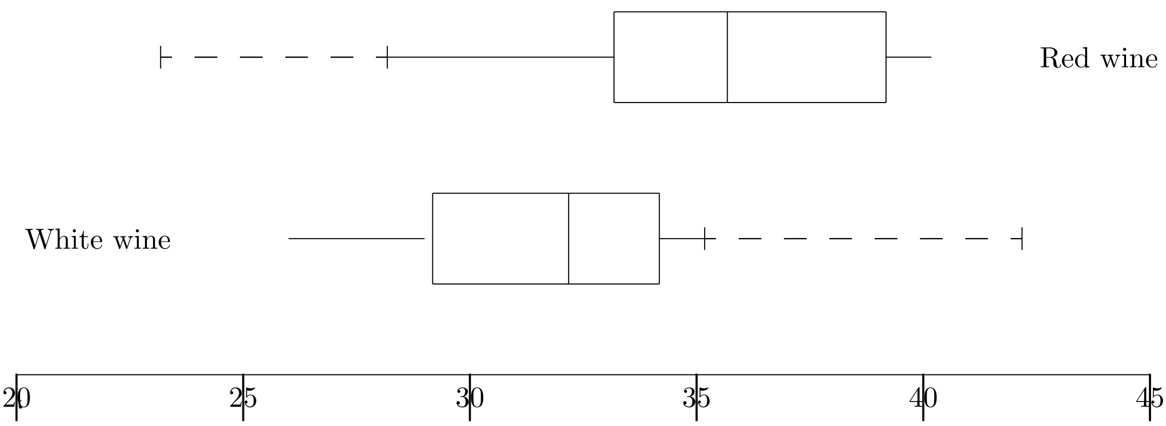
Solutions to Progress Test

1.

```
White         Red
    2 | 40 | 00
    5 | 30 | 5679
 4221 | 30 | 34
  986 | 20 | 8
      | 20 | 3
```

$\boxed{1}$

Red wine: $n = 10$, so quartiles are 3rd and 8th values, i.e. 33 and 39. $\boxed{0.5}$
Median: 5th and 6th equally valid so choose mid-point, which is 35.5. $\boxed{0.5}$
IQR $= 6$, so possible outliers are anything below 24 or above 48. The only possible outlier is 23. $\boxed{0.5}$
White wine: $n = 9$, so quartiles are 3rd and 7th values, i.e. 29 and 34. $\boxed{0.5}$
Median: 5th value, which is 32. $\boxed{0.5}$
IQR $= 5$; possible outliers are > 41.5 or < 21.5. Therefore 42 is the only possible outlier. $\boxed{0.5}$
Box-and-whisker plot. $\boxed{2}$

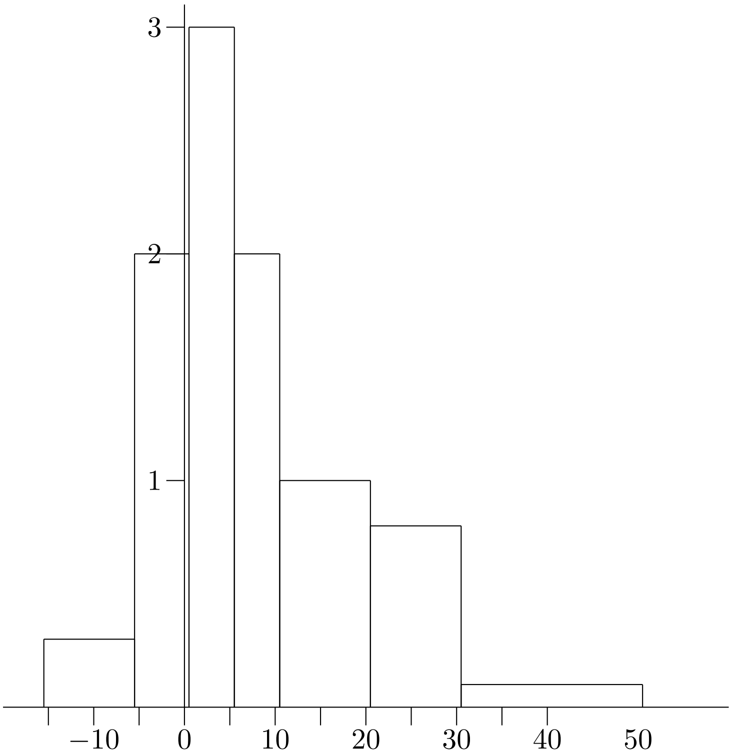


Comment: in general red wine has higher ratings than white. $\boxed{0.5}$
The spread of the two data sets, either in terms of range or as measured by IQR, is about the same. $\boxed{0.5}$
Judging from the stem-and-leaf diagram the white wine data appears slightly right-skewed, whereas the red wine data is more left-skewed. Looking at the quartiles would suggest the exact opposite. For the red wine the sample mean is slightly less than the median; for white wine they are virtually equal. It is reasonable to say that the data sets are roughly symmetric. $\boxed{1}$

2. Width and frequency density columns on table. $\boxed{2}$

Range	**Frequency,** f	**Width**	**Freq density**	**Midpoint,** x	xf
15% worse – 6% worse	3	10	0.3	−10.5	−31.5
0 – 5% worse	12	6	2.0	−2.5	−30.0
1 – 5% better	15	5	3.0	3.0	45.0
6 – 10% better	10	5	2.0	8.0	80.0
11 – 20% better	10	10	1.0	15.5	155.0
21 – 30% better	8	10	0.8	25.5	204.0
31 – 50% better	2	20	0.1	40.5	81.0
Totals	**60**				**503.5**

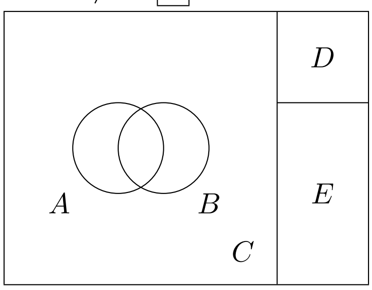
Execution of histogram. $\boxed{3}$



The sample mean is $503.5/60 = 8.39$. $\boxed{2}$

3. (i) A partition is a collection of events which are mutually exclusive and between them cover the whole sample space. $\boxed{1}$
In the given instance, either all three numbers are the same, or all are different, or two are the same with the third being different: exactly one of these three events must occur. $\boxed{1}$

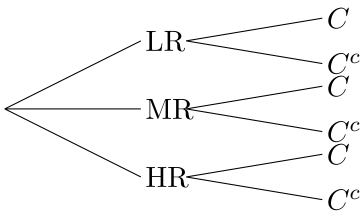
(ii) Venn diagram: C, D and E are disjoint; $\boxed{1}$
A and B are subsets of C; $\boxed{1}$
$A \cap B \neq \emptyset$. $\boxed{1}$



(iii) $\mathbb{P}(C) = 1 \times \frac{36}{39} \times \frac{32}{38} = \frac{192}{247}$. $\boxed{1}$
$\mathbb{P}(D) = 1 \times \frac{3}{39} \times \frac{2}{38} = \frac{1}{247}$. $\boxed{1}$
$\mathbb{P}(E) = 1 - \mathbb{P}(C) - \mathbb{P}(D) = \frac{54}{247}$. $\boxed{1}$

(iv) A and B are independent if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$. $\boxed{1}$
$\mathbb{P}(B) = 1 \times \frac{9}{39} \times \frac{8}{38} = \frac{12}{247}$. $\boxed{1}$
Whatever the probability of A, we know that $\mathbb{P}(A \cap B) = \frac{1}{16}\mathbb{P}(A)$, since any sequence of cards in A will belong to $A \cap B$ if and only if the colours of the second and third match the colour of the first. So they are not independent. $\boxed{1}$

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



(ii) $\mathbb{P}(C) = \mathbb{P}(L)\mathbb{P}(C \mid L) + \mathbb{P}(M)\mathbb{P}(C \mid M) + \mathbb{P}(H)\mathbb{P}(C \mid H) = \frac{1}{4} \times 0.1 + \frac{5}{12} \times 0.2 + \frac{1}{3} \times 0.3 = \frac{2.5}{12} = 0.208$. Therefore $\mathbb{P}(C^c) = 0.792$. $\boxed{1}$

(iii) $\mathbb{P}(H \mid C) = \mathbb{P}(C \mid H)\mathbb{P}(H)/\mathbb{P}(C) = 0.1/0.208 = \frac{12}{25} = 0.48$. $\boxed{2}$

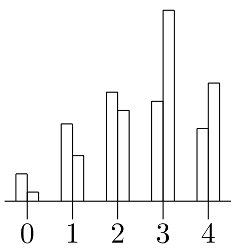
(iv) Consider one policyholder picked at random. $\mathbb{P}(M \mid C^c) = \mathbb{P}(C^c \mid M)\mathbb{P}(M)/\mathbb{P}(C^c) = 0.8 \times \frac{5}{12}/0.792 = \frac{8}{19}$.
If five such policyholders are chosen, the number, X, who are classified as Medium risk has $\text{Bin}(5, \frac{8}{19})$ distribution, so $\mathbb{P}(X \geq 3) = 0.2502 + 0.0910 + 0.0132 = 0.3544$. $\boxed{3}$

5. (i) $\overline{x} = \sum x f_x/50 = (5 + 20 + 63 + 52)/50 = 2.8$. $\boxed{1}$
$s^2 = \frac{1}{49}\left(\sum x^2 f_x - 50 \times 2.8^2\right) = \frac{1}{49}(5 + 40 + 189 + 208 - 392) = \frac{50}{49} = 1.02$. $\boxed{1}$

(ii) We need the probability function of Poisson with mean 2.8: $\boxed{2}$

Number of occupants, x	0	1	2	3	4
Probability, $p(x)$	.0608	.1703	.2384	.2225	.1557
Obs rel frequency, f_x	.02	.10	.20	.42	.26

(iii) Line graph or bar chart is best here. There are major differences between the two, with the observed frequencies being much more concentrated. $\boxed{2}$



(iv) If $4\theta = 2.8$ then $\theta = 0.7$. The variance of $\text{Bin}(4, 0.7)$ is $4 \times 0.7 \times 0.3 = 0.84$. This is certainly much closer to the sample variance, so it may well be a better model. $\boxed{2}$

6. (i) Sketch (f is just a quadratic with maximum at $x = 0.5$). $\boxed{1}$

(ii) $1 = \frac{5}{6}K + \frac{1}{2}K^2 - \frac{1}{3}K^3$, or $2K^3 - 3K^2 - 5K + 6 = 0$. $\boxed{1}$
If we put $K = 2$ we find that this is equal to zero. So we can factorise $2K^3 - 3K^2 - 5K + 6 = (K - 2)(2K^2 + K - 3)$. Observe that $K = 1$ is a factor of the quadratic element, so we have $(K - 2)(K - 1)(2K + 3) = 0$, with solutions $K = 2$, $K = 1$ and $K = -1.5$. $\boxed{1}$
K must be positive, so -1.5 is excluded. $f(2)$ is negative, so $K = 2$ is no good. Therefore $K = 1$ is the one required. $\boxed{1}$

(iii) $\mathbb{E}[X] = \int_0^1 \frac{5}{6}x + x^2 - x^3 = \frac{5}{12} + \frac{1}{3} - \frac{1}{4} = \frac{1}{2}$ (or by symmetry). $\boxed{1}$
$\mathbb{E}[X^2] = \int_0^1 \int_0^1 \frac{5}{6}x^2 + x^3 - x^4 = \frac{5}{18} + \frac{1}{4} - \frac{1}{5} = \frac{59}{180}$. Therefore $\text{Var}[X] = \frac{59}{180} - \frac{1}{4} = \frac{7}{90}$. $\boxed{2}$
(Note: if the student mistakenly chooses $K = 2$, they can still get marks for doing the integration: they should get an expectation of $\frac{1}{3}$ and a variance of $-\frac{8}{45} - \frac{1}{9} = -\frac{13}{45}$. But they can only get full marks if they point out that a negative variance is an impossibility, so there must be a mistake somewhere.)