Solutions for Statistics 1
Mock Examination 2011
Dr James Abdey

SECTION A

1. (a) i.
   - Non-random way of sampling.
   - No list/we use quota controls to ensure representativeness of sample.

   ii. Used in random sampling when we wish to reduce the sampling error for a given number of observations.

   iii.
   - In hypothesis testing we test the null against the alternative hypothesis.
   - We can only accept (fail to reject) or reject the null and assume that this can mean we reject/accept the alternative, **or** they may be one-sided or represent an inequality.

(Total for 1.(a) 6 marks)

(b) The median — it is a measure of location rather than dispersion. (Could accept 'variance' — it is a square.)

(Total for 1.(b) 2 marks)

(c) i. $x = 4$.

ii. $1.5 = \frac{33+3y}{26+y}$, so $y = 4$.

(Total for 1.(c) 4 marks)

(d) i.
- Number of equally likely results with a total of 7 is $N = 6$
- Number of these in which no six appears is $n = 4$
- Probability is $\frac{2}{3}$

ii.
- Let R = number of empty bowls
- Probability all in one bowl is $5 \times \left(\frac{1}{5}\right)^3 = \frac{1}{25} = 0.04$
- Probability all in three different bowls is $1 \times \frac{4}{5} \times \frac{3}{5} = \frac{12}{25} = 0.48$
- Hence $P(R = 3) = 1 - \frac{1}{25} - \frac{12}{25} = \frac{12}{25} = 0.48$
- Probability distribution of R is

r	2	3	4
$P(R = r)$	0.48	0.48	0.04

(Total for 1.(d) 8 marks)

(e) i. No, a non-linear relationship: y increases with x and then decreases with x

ii. Approximately 0

iii.
- 0.7 to 0.9 (sign and magnitude)
- -0.9 to -0.7 (sign and magnitude)

(Total for 1.(e) = 7 marks)

(f)
- $H_0 : \pi = 0.18$ v. $H_1 : \pi > 0.18$
- Standard error: $\sqrt{\frac{0.18 \times 0.82}{220}} = 0.0259$
- Test statistic: $\frac{p-\pi}{\text{S.E.}_{(p)}} = \frac{0.25-0.18}{0.0259} = 2.702$
- Critical value at 5% level is 1.6449
- Reject H_0 at the 5% level
- Proportion buying cereal appears greater after campaign
- It appears advertising campaign has been effective
- **Bonus mark:** Still significant at the 1% level or provide p-value of 0.003

(Total for 1.(f) = 9 marks)

(g) i. $\sum_{i=1}^{2} x_i^3 = 1^3 + 3^3 = 28.$

ii. $\sum_{i=2}^{4} (x_i - 4)^2 = (3-4)^2 + (4-4)^2 + (2-4)^2 = 5.$

iii. $\sum_{i=3}^{4} 2x_i = (2 \times 4) + (2 \times 2) = 12.$

Minor error = 1 mark — e.g., correct use of $\sum$ but wrong i values used.

Note: no marks awarded if i values confused with x_i values.

(Total for 1.(g) = 6 marks)

(h) Two marks each but **only** if a reasonable explanation given.

i. Not possible. As either John or Peter must get a prize — the total probability must be 1.

ii. Not possible. As the numerical value of a correlation coefficient cannot be less than -1.

iii. Not possible. Variance always has a positive value.

iv. True (possible). By definition, the larger the sample, the smaller the variability in the sample mean. Or draw a suitable diagram to demonstrate.

(Total for 1.(h) = 8 marks)

SECTION B

2. (a) i.
- Informative title.
- Definition of stem.
- Definition of leaf.
- Diagram (accuracy & vertical alignment)

Stem-and-leaf plot of lengths of fish caught in one day in a river

Stem = inches | Leaf = 0.1 inches

10	2335666699
11	01667
12	115689
13	5778889999
14	0156789999
15	00112236
16	
17	
18	4

ii.
- Median = 13.8
- Lower quartile = 11.6
- Upper quartile = 14.9

iii. Any sensible three comments, such as:
- Most fish are under 16 inches
- Concentration of very small fish
- Modal lengths are 10.6, 13.9 and 14.9 inches
- Outlier of 18.4

(Total for 2. (a) = 12 marks)

(b) i. $P(Y > 290) = P\left(Z > \frac{290-150}{10}\right) = P(Z > 14) \approx 0$

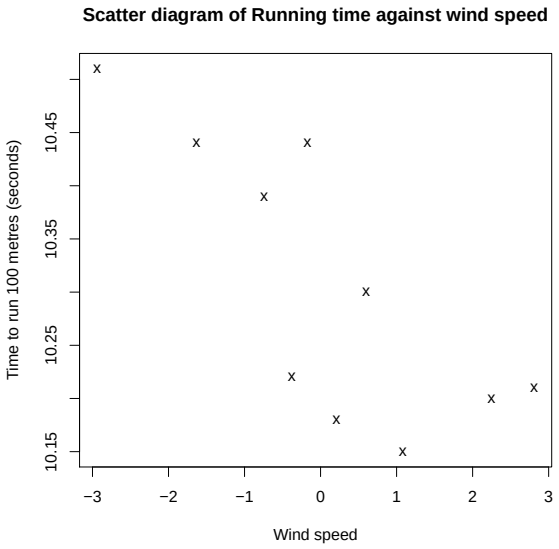
ii.
- $X - 2Y \sim N(260 - 300, 17^2 + 4 \times 10^2) = N(-40, 689)$
- $P(X - 2Y > 0) = P\left(Z > \frac{0--40}{\sqrt{689}}\right) = P(Z > 1.52) = 0.064$

iii.
- $X + Y \sim N(410, 389)$
- $P(X + Y < 400) = P\left(Z < \frac{400-410}{\sqrt{389}}\right) = P(Z < -0.51) = 0.305$

(Total for 2. (b) = 13 marks)

3. (a) i. **No marks if not on graph paper**

- Informative title
- Axis labels
- Accuracy (penalise heavily if x and y transposed)



ii. 'Strong' negative r / clear downward-sloping relationship (or similar)

iii. **Note a maximum of two marks if correct answer with no formula**
- Formula: $r = \frac{\sum xi - n\bar{x}\bar{y}}{\sqrt{(\sum x^2 - n\bar{x}^2)(\sum y^2 - n\bar{y}^2)}}$
- Correct calculation: $\frac{-1.57506}{\sqrt{26.43168 \times 0.15264}} = -0.784$

iv. It looks as if a following wind is beneficial to the runner while having the wind against you is disadvantageous (or other sensible comment).

v. There is such a close relationship between x and y that a regression line should be useful here for predicting the time a runner takes from the wind speed and direction (or similar).

(Total for 3. (a) = 15 marks)

(b)
- Must be a **random** sample (because of frame/accuracy)
- Available **sampling frame** (email addresses provided)
- Discuss use of stratification
- Give stratification factors (e.g. by country, age)
- Justify use of each stratification factor
- Other comments as appropriate (e.g. discussion of multi-stage sampling, cluster sampling etc. with valid justification)

(Total for 3. (b) = 10 marks)

4. (a) i.
- H_0 : No association v. H_1 : Association.
- Correct expected values.

- Test statistic formula: $\sum \frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}$.
- Correct test statistic value: 73.92.
- Degrees of freedom: 4.
- $\alpha = 0.05 \Rightarrow$ critical value of 9.488, hence reject H_0.
- Second (smaller) α, say 1%, hence critical value of 13.28, hence still reject H_0.
- Result is highly significant/strong evidence of an association.

ii. Any sensible comments accepted, such as:

- Middle class more likely to be for or against (others more likely to be indifferent)
- Upper class group least likely to approve
- These 'random' residents are very upper class

(Total for 4. (a) = 12 marks)

(b)
- Sample proportion $\frac{90}{350} = 0.2571$
- Correct z value: 1.96
- CI formula (or implied): $p \pm 1.96\sqrt{\frac{p\times(1-p)}{n}}$
- Correct CI: $(0.2113, 0.3029)$

(Total for 4. (b) = 4 marks)

(c) i.
- $H_0 : \mu_1 = \mu_2$ v. $H_1 \mu_1 < \mu_2$
- Test statistic formula: (accept either)

$$\frac{\bar{x}_2 - \bar{y}_1}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \text{or} \quad \frac{\bar{x}_2 - \bar{y}_1}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

- Test statistic value: 6.434 (or 6.466 if pooled variance used).
- For $\alpha = 0.05$, critical value is 1.645 (1.98 if t_{120} used).
- Decision: reject H_0.
- Choose smaller α, say 1%, hence 2.326, hence again reject H_0.
- Strong evidence that extended contact caused the mother to have a greater attachment to her baby.

ii.
- Although these figures are very striking, they do not go beyond the first week of the babies' lives and so we could not say that the figures show anything about the long term relationship between mothers and their children.

(Total for 4. (c) = 9 marks)