Statistics 1:

Solutions to 2012 mock examination, prepared by Dr James Abdey

Section A

1. (a) **Total = 6 marks**

   Ensure:

   - Title (variable not mentioned, so cannot be informative)
   - Stem/leaf labels
   - Vertical alignment of leaves
   - Ordered leaves
   - Accuracy
   - Calculation of median: $(303 + 305)/2 = 304$

   Stem-and-leaf plot of dataset

   Stem = 10s | Leaf = 1s

27	1279
28	147
29	13578
30	2357
31	12258
32	46
33	37
34	04
35	
36	
37	
38	
39	4

   (b) **Total = 4 marks**

   - Mean: 36.75
   - Median: 30
   - Range: 74
   - IQR (slight quartile variations accepted): $51.5 - 18.5 = 33$

   (c) **Total = 2 marks**

   i. $\hat{y} = -5$

   ii. $\hat{y} = 3$

(d) **Total = 6 marks**

Suggested possible comments:

- Ball 3 seems to be driven the longest; little obvious differences between the other balls.
- The spread of distances seems to be greatest for ball 2.
- Suggestion of skewness for ball 1; no obvious asymmetry for the other balls.
- Two outliers for ball 4 – dud drives?

(e) **Total = 5 marks**

- It means we are 95% confident that the population mean mark lies between 25% and 80%. In other words, the probability that the randomly obtained confidence interval contains the true mean is 95%
- A 99% confidence interval has a higher coverage probability (will include more results/exclude fewer)
- However, it is also wider and so less accurate
- It depends how you want to use these figures/the application to say which is to be preferred

(f) **Total = 5 marks**

- z-score for type A: $z_A = \frac{70-61}{5} = 1.8$
- z-score for type B: $z_B = \frac{70-64}{4} = 1.5$
- Could say that since $z_A > z_B$, then type B schools have higher proportion with marks above 70
  - **Alternatively**, could calculate actual proportions: $P(Z > 1.8) = 0.0359$ and $P(Z > 1.5) = 0.0668$, hence type B schools have higher proportion

(g) **Total = 6 marks**

i. $\sum_{i=1}^{i=3}(x_i - 1)^2 = (1-1)^2 + (4-1)^2 + (6-1)^2 = 34$

ii. $\sum_{i=1}^{i=4}(x_i + 1) = (1+1) + (4+1) + (6+1) + (2+1) = 17$

iii. $\sum_{i=2}^{i=5} 3x_i = (3 \times 4) + (3 \times 6) + (3 \times 2) + (3 \times 3) = 45$

(h) **Total = 4 marks**

i. Any 2×2 contingency table with $\chi^2 \approx 0$

ii. Any 2×2 table with $\chi^2 > 6.635$ (1% critical value)

(i) **Total = 8 marks**

i. Possible – interviewee not necessarily drawn from target population, also non-respondents hence sample may not be representative

ii. Not possible (not *always*) – can have non-sampling errors

iii. Not possible – since $P(\text{take umbrella}) \leq 1$

iv. Not possible – r and b always have the same sign

(j) **Total = 4 marks**

i. $P(PT) = P(PT|F) \cdot P(F) + P(PT|M) \cdot P(M) = (0.15 \times 0.6) + (0.2 \times 0.4) = 0.17$

ii. $P(M|PT) = P(PT|M) \cdot P(M)/P(PT) = (0.2 \times 0.4)/0.17 = 0.4706$

Section B

2. (a) **Total = 13 marks**

i.
- $\mathrm{H}_0 : \pi = 0.6$ and $\mathrm{H}_1 : \pi > 0.6$

ii.
- Sample proportion $p = \frac{320}{400} = 0.8$
- Test statistic and distribution: $\frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \sim N(0,1)$
- Correct test statistic value: $\frac{0.8-0.6}{\sqrt{\frac{0.6\times0.4}{400}}} = 8.165$
- At $\alpha = 0.05$, $z_c = 1.645$
- Hence reject H_0
- Second level, $\alpha = 0.01$, $z_c = 2.33 < 8.165$, hence reject H_0
- Result is highly significant
- Strong indication that the new treatment is more effective than the conventional treatment

iii.
- Type I error: thinking new treatment is more effective when it is not; implication that new drug bought when unnecessary
- Type II error: thinking new treatment is not more effective when it is; implication that new drug not bought when it should be to help more patients

(b) **Total = 12 marks**

i. $P(X > 40) = P(Z > -2) = 0.9772$

ii. $P(56 < X < 60) = P(2 < Z < 3) = 0.9987 - 0.9772 = 0.0215$

iii. $0.05 = P(Z > 1.645) \Rightarrow \frac{x-48}{4} = 1.645$, hence $x = 54.58$

iv.
- $X_1 - X_2 \sim N(0, 32)$
- $P(X_1 - X_2 \geq 3) = P(Z \geq 0.53) = 0.2981$

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

   i.
   - H_0 : No association between districts having a high crime rate and having the death penalty
   - H_1 : There is an association between districts having a high crime rate and having the death penalty
   - Method for calculating expected values (can be implied)
   - Correct expected values: (No, Low) = (Yes, Low) = 45, (No, High) = (Yes, High) = 55
   - χ^2 test statistic formula, $\sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$
   - Test statistic value: 18.18
   - Degrees of freedom: $(2-1)(2-1)=1$
   - At $\alpha = 0.05$, critical value is 3.841
   - Since $3.841 < 18.18$, we reject H_0
   - At $\alpha = 0.01$, critical value is 6.635, again reject H_0
   - Result is highly significant/strong evidence of association between crime rate and death penalty

   ii.
   - Since both chi-squared statistics are less than 3.841, this time there is no association at the 5% significance level
   - So for both rich and poor districts, there is no significant evidence that district crime rate is associated with whether the death penalty is used or not
   - Overall we can say that, when we take into account the wealth of a district, the association between crime rate and whether the death penalty is used disappears (or may say that this now contradicts part 'i'.)

   (b) **Total = 10 marks**

   i. ∗ Strengths
   - Can explain what is meant to interviewee
   - Can make analysis easier or cheaper if we use laptop/coding sheets in the field
   - Interviewee can't see the questionnaire, so can use order-sensitive form of questioning
   - Essential in quota surveys

   ∗ Weaknesses
   - Interviewer may cause bias by the way s/he asks questions
   - Dependent on contact of those targeted (particularly in random samples), so bias through not-at-home
   - Interviewees may need time to reflect (if survey on very detailed or technical questions)

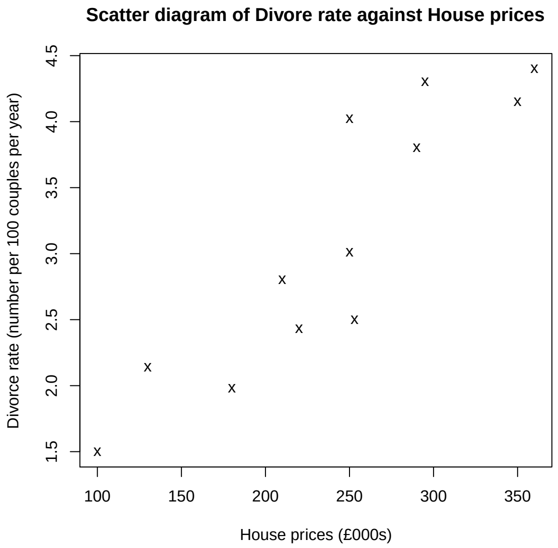
ii.

- Adhering to the criteria given in the question
- This age group may find it difficult to complete forms (bad eyesight), hence use interviews (told have a realistic budget)
- Propose a *random* survey given budget and accuracy wanted
- State and justify realistic stratification or clustering factors

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

i. Ensure:

- Informative title
- Axis labels
- Scale/units
- Accuracy of points



ii. $r = 0.901$, which is a very high value

iii. There is a strong, positive linear relationship between 'House price' and 'Divorce rate'

iv. $\hat{y} = 0.3219 + 0.0115x$ and draw on graph

v. For $x = 150$, $\hat{y} = 2.0469$ divorces per 100 married couples

(b) **Total = 13 marks**

i. - Confidence interval formula:

$$(p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

- Sample proportions: $p_1 = \frac{150}{300} = 0.5$ and $p_2 = \frac{80}{100} = 0.8$

- Difference in sample proportions: $p_2 - p_1 = 0.8 - 0.5 = 0.3$
- Correct z-value: 1.96
- Correct standard error (implied): 0.0493
- Correct confidence interval: $(0.2033, 0.3967)$
- As confidence interval does not include 0, it is unlikely that there is no difference between the two
- Suggests postgraduates are more likely to own a laptop

ii. Not a good estimate because:

- Not told what proportions of the student population are undergraduate and postgraduate
- Not told how the students were selected
- Although told selection was random, the exact sampling scheme is unknown
- Does the ratio 2:1 represent the population ratio
- The figures could only be true for this university, not all