Mark Scheme and Solutions for Statistics 1
Mock Examination 2013

Section A

1 (a) (Total 8 marks)

(i) False. For sufficiently large n, the CLT justifies assuming the sample mean to be normally distributed.

(ii) False. Least squares estimation minimises the sum of *squared* deviations/residuals.

(iii) False. The power of a test is $1 - P(\text{Type II error})$.

(iv) True. Correlation measures the strength of a linear relationship.

1 (b) (Total 6 marks)

(i) $\sum_{i=2}^{i=4} x_i^2 = 3^2 + 1^2 + 7^2 = 59$

(ii) $\sum_{i=1}^{i=3} (x_i - 2) = (2-2) + (3-2) + (1-2) = 0$

(iii) $\sum_{i=2}^{i=5} 2x_i = (2 \times 3) + (2 \times 1) + (2 \times 7) + (2 \times 9) = 40$

1 (c) (Total 2 marks)

$$\frac{\sum x_i}{19} = 7, \quad \frac{\sum y_i}{25} = 5.1, \quad \implies \quad \frac{\sum x_i + \sum y_i}{19 + 25} = \frac{(19 \times 7) + (25 \times 5.1)}{44} = 5.92$$

1 (d) (Total 8 marks)

(i) $P(\text{at least one odd \#}) = 1 - P(\text{no odd \#s}) = 1 - P((2,2) \cup (2,4) \cup (2,6) \cup (4,2) \cup (4,4) \cup (4,6) \cup (6,2) \cup (6,4) \cup (6,6)) = 1 - \frac{9}{36} = 1 - \frac{3^2}{6^2} = \frac{27}{36} = 0.75$

Extending the pattern, $P(\text{at least one odd \#}) = 1 - P(\text{no odd \#s}) = 1 - \frac{3^4}{6^4} = 0.9375$

(ii) Let D = defective. Hence $P(\text{D}) = P(\text{D}|\text{A}) \cdot P(\text{A}) + P(\text{D}|\text{B}) \cdot P(\text{B})$

So $P(\text{D}) = (0.1 \times 0.1) + (0.05 \times 0.9) = 0.055$

Using Bayes' Theorem to obtain correct answer:

$$\begin{aligned} P(\mathrm{B}|\mathrm{D}) &= \frac{P(\mathrm{D}|\mathrm{B}) \cdot P(\mathrm{B})}{P(\mathrm{D}|\mathrm{B}) \cdot P(\mathrm{B}) + P(\mathrm{D}|\mathrm{A}) \cdot P(\mathrm{A})} \\ &= \frac{0.05 \times 0.9}{0.055} \\ &= 0.8182 \end{aligned}$$

1 (e) (Total 8 marks)

(i) A test which is significant at the 1% level can be thought of as 'highly significant' and the data provide strong evidence to support rejection of H_0.

(ii) When significant at the 10% level but not the 5% level there is, at best, weak evidence in support of rejection of H_0. Re-sampling/increasing n would be recommended.

(iii) H_0: $\pi = 0.01$, H_1: $\pi > 0.01$ — need H_0 $\underline{\text{and}}$ H_1

Test statistic: $Z = \frac{p-0.01}{\sqrt{\frac{\pi\cdot(1-\pi)}{n}}} \sim N(0,1)$ since n is large.

$p = \frac{3}{144} = 0.0208$

$z = 1.306$

Only 1.306 is $\underline{\text{weakly}}$ significant at the 10% level, so insufficient evidence to reject H_0, hence no evidence to suggest that more than 1% of transcations contain errors.

1 (f) (Total 4 marks)

The 90% confidence interval for μ is $\bar{x} \pm 1.645\sigma/\sqrt{n}$ (correct z-value)

We require $1.645\sigma/\sqrt{n} = 1.645 \cdot 8.5/\sqrt{n} \leq 1.5$

Solving for n yields $n = \left(\frac{1.645\times 8.5}{1.5}\right)^2 = 86.89$

n must be an integer, so $n = 87$

1 (g) (Total 6 marks)

(i) $P(X < 12) = P(Z < \frac{12-22}{6}) = P(Z < -1.67) = 0.0478$

(ii) $P(X > 37) = P(Z > \frac{37-22}{6}) = P(Z > 2.5) = 0.0062$. Hence $\underline{\text{percentage}}$ of students $= 0.62\%$

(iii) $\bar{X} \sim N\left(22, \frac{36}{225}\right)$. $P(21 \leq \bar{X} \leq 23) = P(-2.5 \leq Z \leq 2.5) = \Phi(2.5) - \Phi(-2.5) = 0.9938 - 0.0062 = 0.9876$

1 (h) (Total 4 marks)

- Quota sampling is a form of *non-probablitiy/non-random* sampling, i.e. the probability of an individual's selection is not known.

- Interviewers are given *quota controls* to interview by sex, social class, age etc.

- Basic rules of inference do not apply

- Use when in a hurry to collect sample data

- Use when no available list/sampling frame

- Use to reduce survey cost

- Use when detailed accuracy of results not important

1 (i) (Total 4 marks)

- Use stratified random sampling to reduce standard errors

- Stratify study population to ensure a representative sample

- Administrative reason — i.e. list is divided into groups already

- Unlike quota sampling, there is a known non-zero (not necessarily equal) chance of being selected

- Usual inference can be applied

- Survey cost higher than with quota sampling

Section B

2 (a) (Total 13 marks)

i. H_0: No association between gender and party affiliation
H_1: Association between gender and party affiliation

Compute expected values for each cell (row total × column total / N):

O_{ij}/E_{ij}	Party Identification			
Gender	Democrat	Independent	Republican	Total
Male	279 / 261.4163	73 / 70.6531	225 / 244.9306	577
Female	165 / 182.5837	47 / 49.3469	191 / 171.0694	403
Total	444	120	416	980

Test statistic, $T = \sum_{i,j} \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ and $t = 7.01$

Degrees of freedom $= (c-1)(r-1) = 2 \times 1 = 2$

At $\alpha = 0.05$, critical value is 5.991, hence reject H_0 and conclude there is an association between gender and party affilitation

Then test at $\alpha = 0.01$, critical value is 9.210, hence we do not reject H_0 and conclude there is no association at the 1% level

Comment: Rejection of H_0 depends on the significance level used as the result is significant at 5% but not 1% thus indicating some support for H_1, but not overwhelmingly so

ii. Obtain sample proportions for all cases: $p_{DEM} = 0.37$, $p_{IND} = 0.39$, $p_{REP} = 0.46$ and $p_{ALL} = 0.41$

Since Republican proportion is highest, we should test $\text{H}_0 : \pi_{REP} = 0.41$ vs. $\text{H}_1 : \pi_{REP} > 0.41$

$z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} = \frac{0.46-0.41}{\sqrt{\frac{0.41*0.59}{416}}} = 2.07$

At the 5% level, the critical value is 1.645, hence reject H_0.

Therefore women are more likely to identify with Republicans than other parties.

2 (b) (Total 12 marks)

Guideline suggestions:

'Random sample design' specified in question, so the student should mention **random** sampling techniques.

Should specify that each unit in a random sample has a known (though not necessarily equal) probability of being selected.

Mention of potential sampling frame / list, e.g. electoral register, national list of addresses etc.

An *interviewer* survey is required, hence recommend using a pilot survey — e.g. to assess clarity of survey questions/interviewer.

Asked to provide **at least three** stratification and clustering factors, hence candidates should define stratification and cluster sampling and give examples of suitable strata, e.g. gender, social status and employment status etc. and cluster factors

Individual respondents selected in a random manner. As this is an interviewer survey, cluster sampling helps to reduce costs as the interviewer is limited to a small geographical area. May have to sacrifice accuracy due to intra-class correlation. Clusters themselves might represent constituencies

Discussion of contact method, i.e. face-to-face interview or telephone interview, with discussion of advantages and disadvantages.

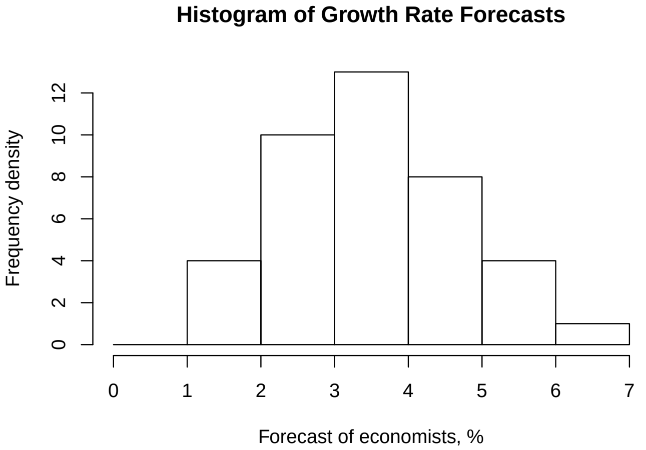
Potential questions to include in the survey relevant to researching 'political affiliation'.

Minimise non-response by offering incentives, e.g. enter into prize draw

3 (a) (Total 12 marks)

Award marks for histogram as follows:

- Informative title
- 'Frequency density' axis label
- x-axis label
- Sensible number of classes
- Plotting of frequency densities
- Accuracy



Median = 3.35, lower quartile = 2.55, upper quartile = 4.25

Population mean forecast > median forecast if the sample distribution is positively/right-skewed, as it is here, mean affected by extreme values, unlike median (sample mean = 3.46)

3 (b) (Total 13 marks)

(i) Correct calculation of difference between the means: $7.0 - 5.1 = 1.9$. Do not accept $5.1 - 7.0 = -1.9$ as question specifies $\mu_x - \mu_y$

Correct standard error computation: 0.4501

Use t distribution with $n_x + n_y - 2 = 42$ degrees of freedom

Confidence interval formula for observed sample means $\bar{x}$ and $\bar{y}$:

$$\bar{x} - \bar{y} \pm t_{\alpha/2, n_x+n_y-2} \cdot s\sqrt{\frac{1}{n_x} + \frac{1}{n_y}}$$

Correct t value: t_{42} value = 2.018, rounded to t_{40} = 2.021 from tables

Correct C.I. : (0.9903, 2.8097) if t_{40} used

Since the confidence interval does not include ‘0’, this does **not** support the view that there is no true difference between the population means.

ii. Test $\mathrm{H}_0 : \mu_x = \mu_y$ vs. $\mathrm{H}_1 : \mu_x \neq \mu_y$

$t = \frac{\bar{x}-\bar{y}}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{7.0-5.1}{\sqrt{2.187\left(\frac{1}{19}+\frac{1}{25}\right)}} = 4.22$

p-value $= 2 \times P(Z > 4.22) \approx 0$

Hence reject H_0 at any sensible significance level α

This agrees with our original deduction that there was no evidence of no difference between the means.

4 (a) (Total 10 marks)

i. $\mathrm{H}_0 : \pi_A = \pi_Z$ vs. $\mathrm{H}_1 : \pi_A \neq \pi_Z$

Compute sample proportions: $p_A = 0.6,\ p_Z = 0.55$

Test statistic value: $z = 0.4529$

For $\alpha = 0.05$, two-tailed critical z-value $= \pm 1.96$

Hence result is **not** significant and we do not reject H_0. There is no evidence to suggest there exists a difference in the population proportions

ii. If $n_A = n_Z = 100$, only the standard error is affected

Standard error now 0.0698 resulting in a test statistic value of $0.05/0.0698 = 0.7161$, again not statistically significant

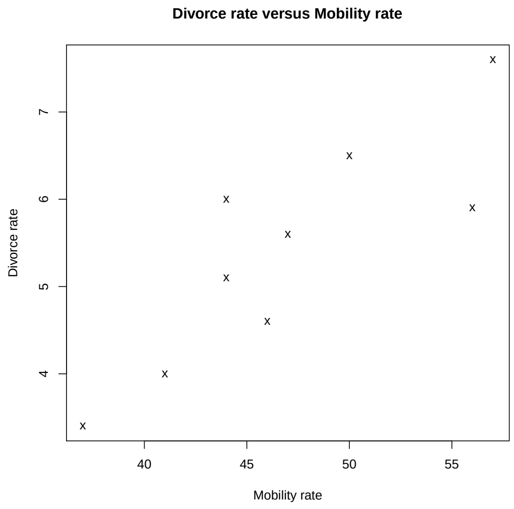
iii. Standard error is 0.0349, hence $z = 0.05/0.0349 = 1.432$, still not significant

4 (b) (Total 15 marks)

i. Award scatter diagram marks for:

- Informative title
- Axis labels
- Accuracy

Comment: Plot shows a fairly strong *positive*, *linear* relationship



ii. Calculation of least squares regression line: $y = \alpha + \beta x + \epsilon$

$$\hat{\beta} = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n\bar{x}^2} = 0.1685$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} = -2.4893$$

Correct parameter estimates

Hence the estimated regression line is $\hat{y} = -2.4893 + 0.1685x$

For $x = 40$, the expected divorce rate is $-2.4893 + 0.1685(40) = 4.25$

iii. Use of divorce rate as the response variable is reasonable due to the likely disruptive effect moving home may have on relationships. (Or similar)

iv. Calculation of sample correlation coefficient:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{(n\sum x^2 - (\sum x)^2)(n\sum y^2 - (\sum y)^2)}}$$

Compute $r = 0.8552$

$r = 0.8552$ suggests a strong positive correlation/linear relationship between divorce rate and mobility rate

v. Using mobility as the dependent variable would imply that the divorce rate is the driver of mobility.

Some sensible comment about the feasibility of this argument.