Examiners Report

A1. Most plots were well done. Some candidates muddled hinges and quartiles and gave the wrong one. Generally candidates correctly identified $F_X^{-1}(u)$, but some failed to find $F_X^{-1}(k/(n+1))$ or got the coordinates the wrong way round.

A2. Generally answered correctly.

A3. Parts (i) and (ii) generally answered correctly. Some candidates incorrectly took $P(X = 20) = 0.2$ rather than 0.1, and some computed $E(X^2)$ rather than Var(X). Part (iii) was less well done and some candidates seemed totally confused. Many candidates correctly identified the distribution of $U - T$ to be Normal, but some took it to have variance $\sigma_U^2 - \sigma_T^2$ rather than $\sigma_U^2 + \sigma_T^2$.

A4. One common error was to transpose S and σ, and incorrectly write $\sqrt{n}(\bar{X} - \mu)/S \sim N(0,1)$ or $\sum(X_i - \bar{X})^2/S^2 \sim \chi^2_{n-1}$, in place of the correct statements $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)$ and $\sum(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$. Otherwise both parts were generally answered correctly.

A5. In Part (i) some candidates lost marks by failing to state the distribution of $\sqrt{n}(\bar{X} - \mu)/\sigma_0$. Most candidates correctly computed the end points of the confidence interval, though one or two incorrectly used a t-distribution or even a chi-square distribution. In part (ii), most candidates correctly answered (a) and (b) correctly, but not many got (c) completely correct.

B1. Most candidates found parts (a) and (b) straightforward, though some candidates incorrectly defined the bias as $\mathrm{E}(\theta - \hat{\theta})$ (it should be $\mathrm{E}(\hat{\theta} - \theta)$) and some took mse($\hat{\theta}$) = Var($\hat{\theta}$)+ bias($\hat{\theta}$)^2 as the definition of the mse (it should be mse$(\hat{\theta}) \equiv \mathrm{E}(\hat{\theta} - \theta)^2$). A small number of candidates forgot that θ here is the fixed constant value of the parameter, and tried to compute $\mathrm{E}(\theta)$. Candidates seemed to find part (c) harder; some got as far as $\widehat{\tau} = \exp\{\log(2)/\hat{\theta}\} - 1$, without realising that $\exp\{\log(2)\}$ is just 2. Most candidates answered part (d), but some did not make the step from the median of each distribution (shown in the boxplot) to the mean (required for the bias and variance).

B2. Part (a) was attempted by almost all candidates, and generally answered correctly. Most candidates attempted part (b) – usually correctly identifying the appropriate test (paired t-test) and the correct hypotheses, finding the correct form of test statistic, and using the correct approach to computing the the p-value. However candidates did not always identify the correct model assumptions (i.e. that $d_1, \ldots, d_n$ are the observed values of a random sample from the Normal distribution with unknown mean and variance); a surprising number made numerical slips in computing the variance estimate $s^2 = (\sum d_i^2 - n\bar{d}^2)/(n-1)$; some failed to state the distribution of the test statistic under H_0; and some failed to correctly interpret their p-value or draw appropriate conclusions. Candidates generally found part (c) harder, and many failed to relate the occurrence of, say, a type II error to a range of $\bar{D}$ values for which the probability could then be easily computed under H_1.

B3. This question was significantly less popular than B1 or B2. In part (a), most students correctly computed the estimates, but very few were able to *derive* the equations satisfied by the least squares estimates. Candidates attempting part (b) generally identified the correct formulae for the end points of the confidence intervals, but many made numerical slips in the calculation – often computing an incorrect value for $s_{\hat{\alpha}}$ even having found the correct value for $\hat{\sigma}^2$. Again, candidates found part (c) testing, and only a small number worked their way through to the end.

A1. (i) The ordered observations are:
0.50, 0.66, 0.71, 0.92, 0.94, 1.02, 1.06, 1.12, 1.20, 1.23, 1.47.
The values used to construct the boxplot are: the smallest observation $o_1 = 0.050$; the lower hinge = median of {data values $\leq$ median} $= (0.71+0.92)/2 = 0.815$; the median $o_6 = 1.02$; the upper hinge = median of {data values $\geq$ median} $= (1.12 + 1.20)/2 = 1.16$, and the largest observation $o_{11} = 1.47$.

(ii) $F_X(x) = x/2$ has inverse $F_X^{-1}(u) = 2u$
so the fitted quantiles are $F_X^{-1}(k/(n+1)) = 2k/(n+1)$ for $k = 1, \ldots, n$.
The smallest observation is $o_1 = 0.50$ and the corresponding fitted quantile (as $n = 11$) is $F_X^{-1}(1/12) = 0.167$, so the coordinates of the point are $(0.17, 0.50)$.
The upper tail values are smaller than expected (shorter upper tails) and the lower tail values are larger than expected (shorted lower tails), so overall the data is more concentrated about its centre than expected under the fitted distribution.

A2. (i) For two unknown parameters we use the equations involving the two smallest moments of the distribution. Here $\mathrm{E}(X^2; \mu, \sigma^2) = \mathrm{Var}(X; \mu, \sigma^2) + [\mathrm{E}(X; \mu, \sigma^2)]^2 = \sigma^2 + \mu^2$. Let $m_1 = (x_1 + \cdots + x_n)/n$ and $m_2 = (x_1^2 + \cdots + x_n^2)/n$. Then the method of moments estimators satisfy

$$\mathrm{E}(X; \hat{\mu}, \hat{\sigma^2}) = m_1 \quad \text{and} \quad \mathrm{E}(X^2; \hat{\mu}, \hat{\sigma^2}) = m_2$$
$$\text{i.e.} \quad \hat{\mu} = m_1 \quad \text{and} \quad \hat{\sigma^2} + \hat{\mu}^2 = m_2$$

(ii) Solving the first equation in (a) gives $\hat{\mu} = m_1 = \bar{x}$. Substituting this into the second equation gives $\hat{\sigma}^2 + m_1^2 = m_2$, so $\hat{\sigma^2} = m_2 - m_1^2 = \sum x_i^2/n - \bar{x}^2 = \sum (x_i - \bar{x})^2/n$.

A3. (i) Let X denote the value of a randomly chosen card.
$P(X = 5) = 0.6$; $P(X = 10) = 0.3$, $P(X = 20) = 0.1$
so $\mu_X = 3 + 3 + 2 = 8$, $E(X^2) = 15 + 30 + 40 = 85$, $\sigma_X^2 = E(X^2) - \mu_X^2 = 21$

(ii) The customer collects $n = 50$ cards, so from the central limit theorem $T = \sum_1^n X_i \simeq N(n\mu_X, n\sigma_X^2) = N(400, 1050)$.

(iii) Exactly the same argument gives $U \simeq N(400, 1050)$. Since T and U are independent, $U - T \simeq N(\mu_U - \mu_T, \sigma_U^2 + \sigma_T^2) = N(0, 2100)$ and $(U - T)/\sqrt{2100} \simeq N(0, 1)$. Thus $P(U - T > 20) = P((U - T)/\sqrt{2100} > 20/\sqrt{2100}) = P((U - T)/\sqrt{2100} > 20/45.82576) = P(Z > 0.4364358) = 1 - P(Z < 0.4364358) = 1 - 0.6687 = 0.3313$

A4. (i) Let $U \sim N(0, 1)$ and let $V \sim \chi_r^2$ with U and V independent. Then from notes $W = U/\sqrt{V/r}$ has the t distribution with r degrees of freedom and we write $W \sim t_r$.

(ii) From notes, $\bar{X} \sim N(\mu, \sigma^2/n)$ so that $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0, 1)$. Also from notes, $(n-1)S^2/\sigma^2 = \sum_i (X_i - \bar{X})^2/\sigma^2 \sim \chi_{n-1}^2$.
But $\sqrt{n}(\bar{X} - \mu)/S = [\sqrt{n}(\bar{X} - \mu)/\sigma]/[\sqrt{(n-1)S^2/(n-1)\sigma^2}]$, so $\sqrt{n}(\bar{X} - \mu)/S = U/\sqrt{V/(n-1)}$, where $U \sim N(0, 1)$ and $V \sim \chi_{n-1}^2$. Hence from the result in (i), $\sqrt{n}(\bar{X} - \mu)/S \sim t_{n-1}$.

A5. (i) When σ is known to take value σ_0, from notes $\sqrt{n}(\bar{X} - \mu)/\sigma_0 \sim N(0,1)$.
Thus the end points of a $100(1-\alpha)\%$ confidence interval are $c_L = \bar{x} - z_{\alpha/2}\sigma_0/\sqrt{n}$ and $c_U = \bar{x} + z_{\alpha/2}\sigma_0/\sqrt{n}$, where $z_{\alpha/2}$ is such that $P(Z > z_\alpha) = \alpha$ when $Z \sim N(0,1)$.
Here $n = 10$, $\alpha/2 = 0.05$, $z_{0.05} = 1.645$, $\bar{x} = 5$, $\sigma_0 = 3$ so

$$c_L = 5 - 1.645 \times 3/\sqrt{10} = 5 - 1.56 = 3.44$$
$$c_U = 5 + 1.645 \times 3/\sqrt{10} = 5 + 1.56 = 6.56$$

(ii) a) It would become shorter since $1/\sqrt{n}$ decreases as n increases – intuitively more observations $\Rightarrow$ more information $\Rightarrow$ smaller interval needed for same confidence level;
(b) it would become larger since $z_{\alpha/2}$ increases as α decreases – intuitively more confidence $\Rightarrow$ larger interval needed;
(c) it might increase since $t_{\alpha/2;n-1} > z_{\alpha/2}$ – intuitively more uncertainty over $\sigma \Rightarrow$ less information $\Rightarrow$ larger interval needed. However in this case the end point is also affected by the value of the estimate of σ – if this was smaller than σ_0, it would tend to decrease the interval length.

B1.

(a) For observations from the given Pareto distribution,

$$\begin{aligned} f(x;\theta) &= \theta/(1+x)^{\theta+1} \\ \log f(x;\theta) &= \log(\theta) - (\theta+1)\log(1+x) \\ \frac{\partial}{\partial\theta}\log f(x;\theta) &= \frac{1}{\theta} - \log(1+x) \\ \sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(x_i;\theta) &= \left(\frac{1}{\theta} - \log(1+x_1)\right) + \cdots + \left(\frac{1}{\theta} - \log(1+x_n)\right) \\ &= \frac{n}{\theta} - \sum_{1}^{n} log(1+x_i) \end{aligned}$$

so the likelihood equation satisfied by the maximum likelihood estimate $\hat{\theta}$ is

$$\frac{n}{\hat{\theta}} - \sum_{1}^{n}\log(1+x_i) = 0$$

giving $$\hat{\theta} = n/\sum_1^n \log(1+x_i)$$

(b) bias$(\hat{\theta}) = \mathrm{E}(\hat{\theta}-\theta)$, mse$(\hat{\theta}) = \mathrm{E}(\hat{\theta}-\theta)^2$, and (from notes) mse($\hat{\theta}$) = Var($\hat{\theta}$)+ bias$(\hat{\theta})^2$.
Here $\mathrm{E}(\hat{\theta}) = n\theta/(n-1)$, so bias$(\hat{\theta}) = \mathrm{E}(\hat{\theta}-\theta) = \mathrm{E}(\hat{\theta}) - \theta = n\theta/(n-1) - \theta = \theta/(n-1)$.

(c) You are given that $(1+\tau)^\theta = 2$, so $\tau = \tau(\theta) = 2^{1/\theta} - 1$.

The maximum likelihood estimate has the property that $\widehat{\tau(\theta)} = \tau(\hat{\theta})$

Thus $\widehat{\tau} = 2^{1/\hat{\theta}} - 1 = 2^{\sum \log(1+x_i)/n} - 1$

(d) Despite the outliers in the boxplot for the sample median, both boxplots appear roughly symmetric about the median, possibly with some positive skew. This would imply that, in the distribution of each estimator, the median is approximately equal to, or slightly less than, the mean. The plot indicates that for each distribution, the median is close to 1, so for each distribution the mean is also (approximately) close to 1, so both estimators are approximately unbiased for $\tau = 1$ (though the plot seems to indicate that the median/mean of the distribution of τ_{mle} is slightly less than 1, indicating a slight negative bias).

However, the plot indicates that estimates based on the sample median have substantially greater variability about their median than the maximum likelihood estimates, with the hinges and whisker ends in the sample median boxplot noticeably further away from the median than in the mle plot. Since the median and the mean are not that dissimilar here, this indicates that the sample median has a greater variability about its mean. Since both estimators are unbiased, this indicates the sample median has greater variability about τ and hence greater mean square error than the mle, as an estimator of τ.

B2.

(a) (i) Type I error occurs if we reject H_0 when it is in fact true;
(ii) Type II error occurs if we accept H_0 when it is in fact false;
(iii) The significance level of the test is P(Type I error) = P(Reject $H_0|H_0$ true);
(iv) The power is $1-$ P(Type II error) = P(Reject $H_1|H_1$ true) = $1-$ P(Accept $H_0|H_1$ true).

(b) A student's body fat percentage both before and after the course is likely to depend on the student as well as the effect of the course – some students may have naturally high levels of body fat, others some may have naturally lower levels. However, the difference in the percentage of body fat for each student may well be independent of the difference for other subjects. Thus it is sensible to use a paired t-test based on the differences, say $d_i = x_i - y_i$.

Model assumptions: Let X_i denote the percentage of body fat before the course for the ith student, let Y_i denote the percentage of body fat after the course for the same student, and let $D_i = X_i - Y_i$. Assume $D_1, \ldots, D_{10}$ are a simple random sample from the $N(\mu, \sigma^2)$ distribution, where μ and σ^2 are unknown.

Hypotheses: H_0: $\mu = 0$ versus H_A: $\mu > 0$. H_0 corresponds to a zero mean difference between the before and after body fat percentages; H_1 corresponds to the percentages after the course being systematically smaller than before the course, so the mean difference is positive.

Test Statistic: Let $T = \sqrt{n}\bar{D}/\hat{\sigma}_D$, where here $n = 10$, $\hat{\sigma}_D^2 = S_D^2 = \sum_{i=1}^{10}(D_i - \bar{D})^2/(10-1)$, so T has the t_9 distribution when H_0 is true. For the given data, the d_i values have mean $\bar{d} = 1.71$ and variance $s_D^2 = 3.8299$. so the observed test statistic is $t_{obs} = 2.7631$.

***p*-value**: For H_A: $\mu > 0$, the values of T which are at least as extreme as t_{obs} are the set $\{T \geq t_{obs}\}$. Thus the p-value = $P(T \geq t_{obs}|H_0$ true) = $P(t_9 \geq 2.7631) = 1 - P(t_9 \leq 2.7631)$. From tables $P(t_9 \leq 2.7) = 0.9878$ and $P(t_9 \leq 2.8) = 0.9896$, so linear interpolation gives $P(t_9 \leq 2.76) = 0.9878 + 0.6 \times (0.0018) = 0.9888$, giving a p-value of 0.0112.

[*Not required but may replace the computation of the p-value.*]
Critical region: The critical region of values where an α-level test would reject H_0 has form $C = \{T \geq c^*\}$, where c^* is defined by: $\alpha = P(\text{Reject } H_0|H_0 \text{ true}) = P(T \geq c^* \,|\, H_0 \text{ true}) = P(t_9 \geq c^*)$. Thus, for $\alpha = 0.05$, $c^* = t_{9;0.05} = 1.833$ giving $C = \{T \geq 1.833\}$.]

Conclusions: The p-value is very small, so there is very strong evidence that H_0 is not true. [*or* The observed test statistic value $t_{obs} = 2.7631$ falls well within the critical region of the 0.05-level test.] Thus we would reject H_0 in favour of H_A, and conclude that the mean body fat percentages after taking the course are indeed smaller those before taking the course.

(c) Under H_1, the power of the test is 1 - P(Type II error) = 1 - P(Accept $H_0|H_1$ true) = P(Reject $H_0|H_1$ true) = P($T > 1.645|$ H_1 true) = 1 - P($T < 1.645 \,|$ H_1 true).

But when H_1: $\mu = 1$ is true, $D \sim N(1, 4)$, so $\bar{D} \sim N(1, 4/10)$, so $\sqrt{10}(\bar{D} - 1)/2 \sim N(0, 1)$ and $T < 1.645 \Leftrightarrow \sqrt{10}\bar{D}/2 < 1.645 \Leftrightarrow \bar{D} < 1.0404 \Leftrightarrow \sqrt{10}(\bar{D} - 1)/2 < 0.0638$.

So Power $= 1 - P(\sqrt{10}(\bar{D} - 1)/2 < 0.0638|\, H_1 \text{ true}) = 1 - P(Z < 0.0638)$ [where $Z \sim N(0, 1)$] $= 1 - \Phi(0.0638) = 1 - 0.5254 = 0.4746$.

B3.

(a) The least squares estimates minimise the sum of squares $\sum(y_i - \alpha - \beta x_i)^2$, so they satisfy the equations obtained by setting $\partial/\partial\alpha = 0$ and $\partial/\partial\beta = 0$.

The first equation gives $-2(\sum y_i - n\hat{\alpha} - \sum \hat{\beta} x_i) = 0$, i.e. $\bar{y} - \hat{\alpha} - \hat{\beta}\bar{x} = 0$, so $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$.

The second eqn gives $-2(\sum y_i x_i - \sum \hat{\alpha} x_i - \sum \hat{\beta} x_i^2) = 0$, i.e. $\sum y_i x_i - \hat{\alpha} n\bar{x} - \hat{\beta}\sum x_i^2 = 0$.
Substituting in for $\hat{\alpha}$ gives $\sum y_i x_i - (\bar{y} - \hat{\beta}\bar{x})n\bar{x} - \hat{\beta}\sum x_i^2 = 0$,
and rearranging gives $\hat{\beta}(\sum x_i^2 - n\bar{x}^2) = \sum y_i x_i - n\bar{y}\bar{x}$, so $\hat{\beta} = (\sum y_i x_i - n\bar{y}\bar{x})/(\sum x_i^2 - n\bar{x}^2)$
or equivalently $\hat{\beta} = (\sum y_i x_i - \sum y_i \sum x_i/n)/(\sum x_i^2 - (\sum x_i)^2/n)$.

Here $\sum x_i = 30$; $\sum y_i = 73.62$; $\sum x_i^2 = 110$; $\sum y_i^2 = 599.8738$; $\sum y_i x_i = 254.3$ giving:
$\hat{\beta} = (254.3 - 30 \times 73.62/10)/(110 - 30^2/10) = 1.672$,
$\hat{\alpha} = \sum y_i/n - \hat{\beta}\sum x_i/n = 73.62/10 - 1.672 \times 30/10 = 2.346$.

The definitions of x_i and y_i were inadvertently swapped at some point when the question was being revised, so 'times' became 'distances' and vice versa. The intended answer was that the estimated time y to reach an address a distance x from the base is $y = \hat{\alpha} + \hat{\beta}x$, so the estimated time taken to reach an address 2.5 miles from the base is $\hat{\alpha} + 2.5\hat{\beta} = 6.526$. However, as printed, x denoted the 'time' and $y = 2.5$ the 'distance', so the estimated time became $x = (y - \hat{\alpha})/\hat{\beta} = 0.092$. Both answers were awarded full marks.

(b) From the **R** output we can read off that the estimate of σ is 0.4964 and the corresponding value of $s_{\hat{\alpha}}$ is 0.3682.

Alternatively, $\hat{\sigma}^2 = [\sum y_i^2 - n\bar{y}^2 - (\sum x_i y_i - n\bar{x}\bar{y})^2/(\sum x_i^2 - n\bar{x}^2)]/(n-2] = 0.24646$ giving
$s_{\hat{\alpha}}^2 = \hat{\sigma}^2(1/n + \bar{x}^2/\sum(x_i - \bar{x})^2) = 0.1356$ and $s_{\hat{\alpha}} = 0.3682$.

You are given that $(\hat{\alpha} - \alpha)/s_{\hat{\alpha}}$ has the t-distribution with $n-2$ degrees of freedom, so the end points (c_L, c_U) of a $100(1-\gamma)\%$ confidence interval for α are given by

$$c_L = \hat{\alpha} - t_{n-2;\gamma/2} \times s_{\hat{\alpha}} \qquad \text{and} \qquad c_U = \hat{\alpha} + t_{n-2;\gamma/2} \times s_{\hat{\alpha}}.$$

For a 95% confidence interval, $\gamma/2 = 0.025$ and from tables $t_{8;0.025} = 2.30$. Thus the 95% confidence interval for α has end points

$$c_L = 2.346 - 2.3 \times 0.368 = 1.4996 \qquad c_U = 2.346 + 2.3 \times 0.368 = 3.1924.$$

(c) For fixed x, you are given that $\hat{\alpha} + \hat{\beta}x$ can be expressed in the form

$$\hat{\alpha} + \hat{\beta}x = \sum_{i=1}^{n} Y_i(1/n + b_i(x - \bar{x})) \quad \text{where} \quad b_i = (x_i - \bar{x})/s_{xx} \quad \text{and} \quad s_{xx} = \sum(x_i - \bar{x})^2.$$

Now $\sum_1^n b_i = \sum(x_i - \bar{x})/s_{xx} = 0/s_{xx} = 0$
and $\sum_1^n b_i^2 = \sum_1^n(x_i - \bar{x})^2/(s_{xx})^2 = s_{xx}/(s_{xx})^2 = 1/s_{xx}$.

Since the Y_i are independent and each has the same variance σ^2, we have
$\text{Var}(\hat{\alpha} + \hat{\beta}x) = \text{Var}(\sum Y_i(1/n + b_i(x - \bar{x})) = \sum \text{Var}(Y_i(1/n + b_i(x - \bar{x}))$
$= \sum \text{Var}(Y_i)(1/n + b_i(x - \bar{x}))^2 = \sigma^2 \sum(1/n + b_i(x - \bar{x}))^2$
$= \sigma^2 \sum(1/n^2 + 2b_i(x - \bar{x})/n + b_i^2(x - \bar{x})^2)$
$= \sigma^2[\sum 1/n^2 + 2(x_i - \bar{x})(\sum b_i)/n + (x - \bar{x})^2 \sum b_i^2]$
$= \sigma^2[1/n + (x - \bar{x})^2/s_{xx}]$ since $\sum b_i = 0$ and $\sum b_i^2 = 1/s_{xx}$.