MATH10282 Introduction to Statistics

Mark scheme for main exam

2015/16

A1. (a) (i) First identify any outliers. x_i is classified as an outlier if

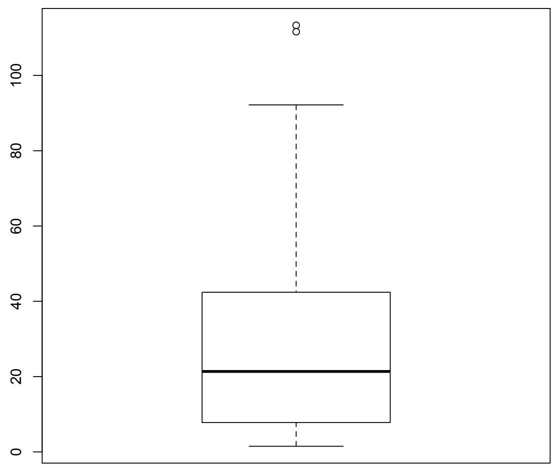
$$x_i \geq \hat{Q}(0.75) + 1.5 \times \text{IQR} \quad \text{or}$$
$$x_i \leq \hat{Q}(0.25) - 1.5 \times \text{IQR}$$

The IQR is $42.4075 - 7.81 = 34.5975$. The thresholds are

$$42.4075 + 1.5 \times 34.5975 = 94.30375$$
$$7.81 - 1.5 \times 34.5975 = -44.08625\,,$$

Thus there are 2 outliers: 113.32, 111.62. Thus the upper adjacent value is therefore 92.17, and the lower adjacent value is 1.49. (3 marks)

The box plot is as follows: (3 marks)



(ii) The distribution is skewed to the right, indicated by the fact that the upper whisker is longer than the lower whisker. (1 mark)

(iii) A normal distribution is unlikely to be a good fit. Applying a log transformation may enable a normal model to be fitted. (1 mark)

(b) The bin containing $x = 9$ is $(0, 20)$. The height of the histogram is given by

$$\text{Hist}(x) = \nu_k/(nh)\,,$$

where ν_k is the number of data points in the corresponding bin. Here $\nu_k = 10$ and so $\text{Hist}(9) = 10/(20 \times 20) = 0.025$.

(2 marks)

ALL BOOKWORK. Boxplots/histograms covered in Chapter 2. Goodness of fit/transformations in Chapter 3. Fairly similar to Example Sheet 4, Qs 2,3,5.

TOTAL FOR A1, 10 MARKS

A2. (i) The likelihood is the joint probability of the data, considered as a function of the parameter λ. By independence,

$$L(\lambda) = \mathrm{P}(X_1, \ldots, X_n | \lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda} \lambda^{X_i}}{X_i!} = \frac{e^{-n\lambda} \lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!} \,.$$

(1 mark) BOOKWORK. This example is given in lectures, Chapter 6.

(ii) The log-likelihood is

$$l(\lambda) = -n\lambda + \left(\sum_{i=1}^{n} X_i \right) \log \lambda - \log \left(\prod_{i=1}^{n} X_i! \right) .$$

Solving $\frac{dl(\lambda)}{d\lambda} = 0$, we obtain

$$\left. \frac{dl}{d\lambda} \right|_{\lambda=\hat{\lambda}} = -n + \frac{\sum_{i=1}^{n} X_i}{\hat{\lambda}} = 0 \,, \qquad \text{which implies that } \hat{\lambda} = \bar{X} \,.$$

Checking the second derivatives, we see that

$$\left. \frac{d^2 l}{d\lambda^2} \right|_{\lambda=\hat{\lambda}} = \frac{-\sum_{i=1}^{n} X_i}{\hat{\lambda}^2} = \frac{-n}{\bar{X}} < 0 \,.$$

Therefore, $\hat{\lambda} = \bar{X}$ is indeed the maximum likelihood estimator of λ.

(4 marks) BOOKWORK. This example is given in the lectures, Chapter 6.

(iii) Note that $\mathrm{E}(\hat{\lambda}) = \mathrm{E}(\bar{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(X_i) = \mathrm{E}(X_1) = \lambda$. Hence $\mathrm{bias}(\hat{\lambda}) = \mathrm{E}(\hat{\lambda}) - \lambda = 0$. Also, $\mathrm{Var}(\bar{X}) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(X_i)$, by independence. However, $\mathrm{Var}(X_i) = \lambda$ and so $\mathrm{Var}\, \hat{\lambda} = \lambda / n$.

(2 marks) UNSEEN example of variance and bias, but similar to those in Chapter 5.

(iv) For large n, $\bar{X}$ has approximately a $N(\lambda,\, \lambda/n)$ distribution, by the Central Limit Theorem.

$$\begin{aligned} \mathrm{P}(9.9 < \bar{X} < 10.1) &= \mathrm{P}\left(\frac{9.9 - 10}{\sqrt{10/100}} < \frac{\bar{X} - \lambda}{\sqrt{\lambda/n}} < \frac{10.1 - 10}{\sqrt{10/100}} \right) \\ &= \mathrm{P}(-0.316 < Z < 0.316) \\ &\approx \Phi(0.32) - \Phi(-0.32) \\ &= 0.6255 - (1 - 0.6255) = 0.25 \,. \end{aligned}$$

(3 marks) UNSEEN example of normal approximation, though similar uses of the CLT are numerous in Chapters 3,7,9.

TOTAL FOR A2, 10 MARKS

A3. (i) An appropriate unbiased estimator is

$$\hat{\sigma}^2 = \frac{(n-1)S_1^2 + (m-1)S_2^2}{n+m-2}$$

A suitably scaled version of $\hat{\sigma}^2$ has a χ^2 distribution, namely:

$$\frac{(n+m-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n+m-2)\,.$$

(1 mark for correct estimator; 1 mark for correct distributional statement including correct scaling and d.f.)

(ii) Under H_0,

$$\bar{X}_1 - \bar{X}_2 \sim N\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right),$$

independently of $(n+m-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2(n+m-2)$. Thus, the test statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2)\,.$$

(1 mark for correct test statistic, 1 mark for correct distribution including d.f.)

(iii) For a two-tailed test with significance level $100\alpha\%$, we reject if

$$T \geq t_{\alpha/2} \text{ or } T \leq -t_{\alpha/2}\,,$$

where $t_{\alpha/2}$ is the upper $\alpha/2$ point of a $t(n+m-2)$ distribution, i.e. $\mathrm{P}(T > t_{\alpha/2}) = \alpha/2$.

(1 mark for a symmetric test; 1 mark for correct critical values, must state number of d.f.)

(iv) In this case, $\hat{\sigma}^2 = 2.04^2/2 + 1.92^2/2 = 3.924$. Thus,

$$t = \frac{46.0 - 48.1}{\sqrt{3.924}\sqrt{\frac{1}{10} + \frac{1}{10}}} = -2.37\,.$$

For $\alpha = 0.05$, we have $t_{\alpha/2} = 2.101$ on 18 d.f., and for $\alpha = 0.01$ we have $t_{\alpha/2} = 2.878$ on 18 d.f.. Therefore we reject at the 5% significance level but not at the 1% significance level.

(1 mark for value of $\hat{\sigma}^2$; 1 mark for correct value of test statistic; 1 mark for correct critical values; 1 mark for correct interpretation)

ALL BOOKWORK. This test is illustrated in the course notes, Chapter 10 on two sample hypothesis tests, where a numerical example is given.

TOTAL FOR A3, 10 MARKS

A4. (i) Let $X_1, \ldots, X_n \sim \mathrm{Bi}(n, p)$ and let $\hat{p}$ be the sample proportion of successes, i.e. $\hat{p} = (1/n)\sum_{i=1}^{n} X_i$. Asymptotic results show that

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \sim N(0,1) \quad \text{approximately for large } n\,.$$

This is the standardized version of $\hat{p}$ with the standard deviation in the denominator replaced by a sample estimate.

Thus

$$\left[\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\,,\, \hat{p} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right]$$

is an approximate $100(1-\alpha)\%$ confidence interval for p (for large n), where $z_{\alpha/2}$ is the upper $\alpha/2$ point of a $N(0,1)$ distribution, i.e. $\mathrm{P}(Z > z_{\alpha/2}) = \alpha/2$.

- 1 mark for correct approximate pivot
- 1 mark for a confidence interval of the correct confidence level - still award if $\alpha/2$ not specified explicitly, but not if incorrect
- 1 mark if $\hat{p}$ and $z_{\alpha/2}$ defined

(ii) In this case $\hat{p} = 0.3$ and so the 95% confidence interval is

$$(0.3 - 1.96\sqrt{0.3 \times 0.7/1000},\, 0.3 + 1.96 \times \sqrt{0.3 \times 0.7/1000}) = (0.272, 0.328)\,.$$

(1 mark for correct z-value; 1 mark for correct interval)

PARTS (i) AND (ii) - BOOKWORK. This asymptotic method of constructing confidence intervals for a binary proportion is covered in lectures (Chapter 7 Part I). Theory and a numerical example were given.

(iii) Let $X \sim \mathrm{Bi}(n, p)$ be the number of individuals in the sample supporting Labour. By the normal approximation to the binomial,

$$X \sim N[np, np(1-p)] \quad \text{approximately}\,,$$

Here, $np = 500 \times 0.28 = 140$ and $np(1-p) = 140 \times (1 - 0.28) = 100.8$.

This approximation is valid provided $n \geq 9\max\{p/(1-p), (1-p)/p\} = 23.1$. Here $n = 500$ and so the normal approximation is valid.

Thus,

$$\begin{aligned}
\mathrm{P}(X \geq 150) &= \mathrm{P}\left(\frac{X - 140}{\sqrt{100.8}} \geq \frac{150 - 140}{\sqrt{100.8}}\right) \\
&\approx \mathrm{P}\left(Z \geq \frac{149.5 - 140}{\sqrt{100.8}}\right) \quad \text{using continuity correction} \\
&= 1 - \Phi(0.9462) = 0.1720 \\
&[\text{alternatively } 1 - \Phi(0.95) = 0.17\,,\ \text{if rounding}]
\end{aligned}$$

- 2 marks for approximating with the correct normal distribution
- 1 mark for correct check of validity
- 1 mark for correct normal probability calculations
- 1 mark for correct use of continuity correction

BOOKWORK. The normal approximation to the binomial, including continuity correction was covered in Chapter 4 of the lecture notes. A similar example for an opinion poll was given (with different numbers).

TOTAL MARKS FOR A4, 10 marks

B5. (a) (i)

$$\begin{aligned}
\mathrm{E}(S^2) &= \frac{1}{(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] \\
&= \frac{1}{(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{n}[(X_i-\mu)-(\bar{X}-\mu)]^2\right] \\
&= \frac{1}{(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{n}[(X_i-\mu)^2-2(X_i-\mu)(\bar{X}-\mu)+(\bar{X}-\mu)^2]\right] \\
&= \frac{1}{(n-1)}\,\mathrm{E}\left[\sum_{i=1}^{n}(X_i-\mu)^2-2n(\bar{X}-\mu)(\bar{X}-\mu)+n(\bar{X}-\mu)^2\right] \\
&= \frac{1}{(n-1)}\left[\sum_{i=1}^{n}\mathrm{E}\left[(X_i-\mu)^2\right]-2n\,\mathrm{E}[(\bar{X}-\mu)^2]+n\,\mathrm{E}[(\bar{X}-\mu)^2]\right] \\
&= \frac{1}{(n-1)}\left[n\sigma^2-2n\frac{\sigma^2}{n}+n\frac{\sigma^2}{n}\right] \\
&\quad \text{since } \mathrm{E}[(\bar{X}-\mu)^2]=\mathrm{Var}(\bar{X})=\frac{\sigma^2}{n} \\
&= \frac{1}{(n-1)}\left[(n-1)\sigma^2\right]=\sigma^2\,.
\end{aligned}$$

(7 marks) BOOKWORK - this derivation appears in Chapter 4.

(ii) $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. (1 mark) BOOKWORK - Chapter 4

(b) (i) Firstly, the point estimate is as follows:

$$\begin{aligned}
s^2 &= \frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right) \\
&= \frac{1}{9}\left(1474.5-10\times 11.32^2\right) \\
&= 21.453
\end{aligned}$$

(2 marks) BOOKWORK - formula given in Chapter 2

In general, a $100(1-\alpha)\%$ confidence interval for σ^2 is given by

$$\left[\frac{(n-1)S^2}{\chi^2_{\alpha/2}},\ \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right],$$

where χ^2_α is the upper α point of a $\chi^2(n-1)$ distribution. In this case, $\alpha=0.01$, $\chi^2_{0.005}=23.59$, $\chi^2_{0.995}=1.735$, and so the confidence interval is

$$\left[\frac{9\times 21.453}{23.59},\ \frac{9\times 21.453}{1.735}\right]=[8.184, 111.287]\,.$$

(3 marks) BOOKWORK - similar to example in Chapter 7, Part I

(ii) In general, with σ^2 unknown, a $100(1-\alpha)\%$ confidence interval is given by

$$\left[\bar{X}-\frac{t_{\alpha/2}s}{\sqrt{n}},\ \bar{X}+\frac{t_{\alpha/2}s}{\sqrt{n}}\right],$$

where t_α is the upper α point of a $t(n-1)$ distribution. In this case, $\alpha = 0.01$ and $t_{0.005} = 3.250$. Thus the 99% confidence interval is

$$(11.32 - 3.250 \times \sqrt{21.453/10}\,,\ 11.32 + 3.250 \times \sqrt{21.453/10}) = (6.56, 16.08)\,.$$

(4 marks) BOOKWORK - similar to example in Chapter 7, Part I

(iii) From the data the fitted model is $X_i \sim N(11.32, 21.453)$, from which $\bar{X} \sim N(11.32, 2.1453)$. Using this fitted model, we estimate the probability

$$\begin{aligned} \mathrm{P}(\bar{X} > 11.0) &= \mathrm{P}\left(\frac{\bar{X} - 11.32}{\sqrt{2.1453}} > \frac{11 - 11.32}{\sqrt{2.1453}}\right) \\ &= \mathrm{P}(Z > -0.218) \\ &\approx \Phi(0.22) = 0.59 \end{aligned}$$

(3 marks) BOOKWORK. Similar example given in Chapter 4.

TOTAL FOR B5, 20 MARKS

B6. (i)

$$p_0 = \mathrm{P}(N(6, 2^2) \leq 7) = \Phi\left(\frac{7-6}{2}\right) = \Phi(1/2) = 0.6915$$

(2 marks for showing $p_0 = 1/2$; 1 mark for numerical value)

UNSEEN - the calculation is bookwork similar to examples in Chapter 3, but the application in this context is unfamiliar.

(ii) Let $\hat{p}$ denote the sample proportion of patients who recover within 7 days. We use a suitably scaled version of $\hat{p}$, namely

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

(2 marks) BOOKWORK. This test has been seen in Chapter 9 of the lectures with an example.

(iii) Under the null hypothesis, the distribution of this test statistic is approximately $N(0, 1)$ for large n. This approximation is reasonably accurate since $60 = n \geq 9 \max\{p_0/(1-p_0), (1-p_0)/p_0\} = 20.17$.

(1 mark for correct null distribution; 1 mark for checking accuracy of normal approximation)

BOOKWORK. Chapter 9.

(iv) The significance level is α if $\mathrm{P}(\text{reject } H_0 \mid H_0) = \alpha$. This is achieved by the one-tailed rejection region

$$Z \geq z_\alpha$$

where z_α is the upper α point of the $N(0, 1)$ distribution, i.e. $1 - \Phi(z_\alpha) = \alpha$. For $\alpha = 0.05$, we have $z_\alpha = 1.645$. Thus we reject H_0 if

$$\hat{p} \geq p_0 + 1.645\sqrt{p_0(1-p_0)/n} = 0.7896\,.$$

If 52 out of 60 patients were to recover, then $\hat{p}$ would be equal to $52/60 = 0.8667$ and so H_0 would be rejected.

(4 marks).

BOOKWORK. Chapter 9.

(v) Note that here, since $X_i \sim N(5, 2^2)$ the probability that a treated patient recovers within 7 days is

$$p = \mathrm{P}(X_i \leq 7) = \Phi\left(\frac{7-5}{2}\right) = 0.8413.$$

(3 marks)

The probability of rejecting the null hypothesis under this test is

$$\begin{aligned}\mathrm{P}\left(\hat{p} \geq 0.7896\right) &= \mathrm{P}\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \geq \frac{0.7896 - 0.8413}{\sqrt{0.8413 \times 0.1587/60}}\right) \\ &\approx 1 - \Phi\left(-1.10\right) = 0.86 \quad \text{(to 2 d.p.)}\,,\end{aligned}$$

since under the alternative hypothesis $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \sim N(0, 1)$ approximately for large n.

(6 marks)

UNSEEN, but fairly similar to Example Sheet 10, Q7.

B7. (a) (i) $\mathrm{E}(\hat{\mu}) = \mu$ and so bias$(\hat{\mu}) = 0$. Var$(\hat{\mu}) = \sigma^2/n$. (1 mark for bias, 1 mark for variance)
BOOKWORK - example given in Chapter 5.

(ii) For the bias, note that

$$\begin{aligned}\mathrm{E}(\tilde{\mu}) &= \mathrm{E}\left(\frac{1}{2n}(X_1 + \ldots + X_n)\right) \\ &= \frac{1}{2n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{n\mu}{2n} = \mu/2.\end{aligned}$$

and so bias$(\tilde{\mu}) = \mu/2 - \mu = -\mu/2$. For the variance

$$\begin{aligned}\mathrm{Var}(\tilde{\mu}) &= \mathrm{Var}\left(\frac{1}{2n}\sum_{i=1}^{n}X_i\right) = \frac{1}{4n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i)\,, \quad \text{by independence.} \\ &= \frac{\sigma^2}{4n}\,.\end{aligned}$$

(2 marks for bias, 2 marks for variance)

UNSEEN example - technique is in Chapter 5.

(b) (i)

$$\begin{aligned}\mathrm{P}(-\epsilon < \hat{\mu} - \mu < \epsilon) &= \mathrm{P}\left(-\frac{\epsilon}{\sigma/\sqrt{n}} < \frac{\hat{\mu}-\mu}{\sigma/\sqrt{n}} < \frac{\epsilon}{\sigma/\sqrt{n}}\right) \\ &= \mathrm{P}\left(-0.1\sqrt{n} < Z < 0.1\sqrt{n}\right) \\ &= \Phi\left(0.1\sqrt{n}\right) - \Phi\left(-0.1\sqrt{n}\right) \\ &= 2\Phi(0.1\sqrt{n}) - 1\,, \text{ by symmetry of } N(0,1)\end{aligned}$$

(4 marks)

(ii)

$$\begin{aligned}\mathrm{P}\left(\mu - \epsilon < \tilde{\mu} < \mu + \epsilon\right) &= \mathrm{P}\left(\frac{\mu/2-\epsilon}{\sigma/\sqrt{4n}} < \frac{\tilde{\mu}-\mu/2}{\sigma/\sqrt{4n}} < \frac{\mu/2+\epsilon}{\sigma/\sqrt{4n}}\right) \\ &= \mathrm{P}\left(0 < Z < \frac{\sqrt{4n}(0.1\sigma + 0.1\sigma)}{\sigma}\right) \\ &= \Phi(0.4\sqrt{n}) - \Phi(0) = \Phi(0.4\sqrt{n}) - 0.5.\end{aligned}$$

(5 marks) UNSEEN example. Techniques for calculations are in Chapter 4.

(iii) $p_1(10) = 2 \times \Phi(0.316) - 1 = 0.25$ to 2 d.p., and $p_2(10) = \Phi(1.265) - 0.5 = 0.40$ to 2 d.p.. Thus, $\tilde{\mu}$ has the greatest probability of being within ϵ of μ when $n = 10$. (3 marks)

(iv) For small n (< 45), $p_2(n) \geq p_1(n)$ and so the experiment has a higher probability of success if $\tilde{\mu}$ is used. Thus, for small n, $\tilde{\mu}$ is preferable. For large n (≥ 45), $p_1(n) \geq p_2(n)$ and so the experiment has a higher probability of success if $\hat{\mu}$ is used. Thus for large n, $\hat{\mu}$ is preferable.

However, the experimenter does not know which estimator is best, as they do not know the values of μ and σ. (1 mark for each point up to a max of 2 marks)

UNSEEN. The idea that there are certain circumstances under which biased estimators may be preferred was alluded to in lectures (Chapter 5), but not discussed in detail.

TOTAL FOR B7, 20 MARKS