MATH10282 Introduction to Statistics

Mark scheme for main exam

2016/17

A1.

(i) The density histogram is defined by

$$\text{Hist}(x) = \frac{\nu_k}{nh}\,, \quad \text{for } x \in B_k\,,$$

where ν_k is the number of observations in B_k.

(2 marks)

(ii)

$$\begin{aligned}
\hat{\mu}_{\text{Hist}} &= \int_{a_1}^{a_{K+1}} x\,\text{Hist}(x)\,dx \\
&= \sum_{k=1}^{K} \int_{a_k}^{a_{k+1}} x\,\frac{\nu_k}{nh}\,dx = \sum_{k=1}^{K} \frac{\nu_k}{nh}\left[\frac{x^2}{2}\right]_{a_k}^{a_{k+1}} \\
&= \sum_{k=1}^{K} \frac{\nu_k}{2nh}(a_{k+1}^2 - a_k^2) \\
&= \sum_{k=1}^{K} \frac{\nu_k}{2nh}(a_{k+1} + a_k)(a_{k+1} - a_k) \\
&= \sum_{k=1}^{K} \frac{\nu_k}{n}\frac{(a_{k+1} + a_k)}{2}\,.
\end{aligned}$$

(5 marks)

(iii) Note that the total number of observations is $n = 7 + 19 + 38 + 23 + 12 + 1 = 100$. The interval mid points are 15, 25, 35, 45, 55, 65, therefore $\hat{\mu}_{\text{Hist}} = \frac{1}{100}(7 \times 15 + 19 \times 25 + 38 \times 35 + 23 \times 45 + 12 \times 55 + 1 \times 65) = 36.7$.

(3 marks)

BOOKWORK. Similar to Example Sheet 4, Question 4 TOTAL FOR A1, 10 MARKS

A2. (i)

$$\begin{aligned}
S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i^2-2\bar{X}X_i+\bar{X}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2\bar{X}\sum_{i=1}^{n}X_i+\sum_{i=1}^{n}\bar{X}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-2n\bar{X}^2+n\bar{X}^2\right)\\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2-n\bar{X}^2\right)
\end{aligned}$$

(5 marks)

BOOKWORK. Chapter 2.

(ii) $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$. A $100(1-\alpha)\%$ CI can be obtained as follows:

$$\begin{aligned}
1-\alpha &= \mathrm{P}\left(\chi^2_{1-\alpha/2} < (n-1)S^2/\sigma^2 < \chi^2_{\alpha/2}\right)\\
&= \mathrm{P}\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right)
\end{aligned}$$

with $\chi^2_{\alpha/2}$, $\chi^2_{1-\alpha/2}$ denoting the *upper* $\alpha/2$ and $1-\alpha/2$ points of the χ^2 distribution with $n-1=9$ d.f.. Hence a $100(1-\alpha)\%$ CI is given by

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

Note:

- Statement of interval without full derivation is fine.
- NB if quantiles/percentage points are used rather than the upper points of the distribution, the formula will be $\left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2}}\right]$ instead. This is fine if the student clearly states what they mean by $\chi^2_{\alpha/2}$. Otherwise give only partial credit.
- If wrong number of d.f. stated/d.f. not given, deduct 1 mark.

(3 marks)

(iii) As $n=10$, we use the critical values from a χ^2_9 distribution. These are $\chi^2_{0.025}=19.023$ and $\chi^2_{0.975}=2.7$. Here $S^2=\frac{1}{9}(74227-10\times 85.9^2)=48.76667$ and so we obtain the following 95% CI for σ^2:

$$\left[\frac{9\times 48.76667}{19.023}, \frac{9\times 48.76667}{2.7}\right] = [23.07, 162.56]$$

(2 marks)

- If they use d.f.=10, they will obtain [23.81, 150.19]. In this case deduct 1 mark.

BOOKWORK. Chapter 7. TOTAL FOR A2, 10 MARKS

A3. (i) If $Y \sim U[a,b]$ then $\mathrm{E}\,Y = (a+b)/2$ and $\mathrm{Var}\,Y = (b-a)^2/12$. Hence $\mathrm{E}\,X = \theta$ and $\mathrm{Var}\,X = 1/12$. Thus $\mathrm{E}\,\bar{X} = \mathrm{E}(\sum_{i=1}^{n} X_i/n) = (1/n)\sum_i \mathrm{E}\,X_i = n\theta/n = \theta$ and $\bar{X}$ is an unbiased estimator of θ. Moreover, $\mathrm{Var}(\bar{X}) = \mathrm{Var}\left(\frac{1}{n}\sum_i X_i\right) = \frac{1}{n^2}\mathrm{Var}\sum_i X_i = \frac{1}{n^2}\sum_i \mathrm{Var}\,X_i = \frac{1}{n}\mathrm{Var}\,X = 1/(12n)$, where we have used independence of the X_i.

(4 marks)

BOOKWORK. Chapter 5. Similar to Example 5.5 in written notes.

(ii) For large n, $\bar{X} \sim N(\theta, 1/(12n))$ approximately by the central limit theorem.

(2 marks)

BOOKWORK. Chapter 4.

(iii) From the past sample of size 75, $\hat{\theta}_{75} = \bar{x} = 1147/75 = 15.293$. For a future sample of size $m = 80$,

$$\mathrm{P}(\bar{X}_m \geq 15.25) = \mathrm{P}\left(\frac{\bar{X}-\theta}{\sqrt{1/(12m)}} \geq \frac{15.25-\theta}{\sqrt{1/(12m)}}\right) \approx 1 - \Phi\left(\frac{15.25-\theta}{\sqrt{1/(12n)}}\right).$$

We estimate the above by plugging in $\hat{\theta}_{75}$ in place of θ, to obtain an estimated probability of

$$1 - \Phi(-1.3426) = \Phi(1.3426) = 0.9103 \quad \text{or} \quad \Phi(1.34) = 0.9099$$

If they round $\hat{\theta}_{75}$ to 15.293 and use $m = 80$ they will obtain $1 - \Phi(-1.3323) \approx 1 - \Phi(-1.33) = 0.9082$.

If they use the correct value of $\hat{\theta}$ but set $m = 75$ they will obtain $1 - \Phi(-1.3) = 0.9032$; in this case award 3 out of 4.

(4 marks)

BOOKWORK. Chapter 4.

TOTAL FOR A3, 10 MARKS

A4. (i) An unbiased estimator of σ^2 is given by

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

and we have that $(n+m-2)\hat{\sigma}^2/\sigma^2 \sim \chi^2_{n+m-2}$.

(3 marks)

(ii) A suitable test statistic is

$$T = \frac{\bar{X} - \bar{Y} - \theta_0}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

and we have that $T \sim t_{n+m-2}$ exactly when H_0 is true. We reject H_0 when $|T| > t_{\alpha/2}$ where $t_{\alpha/2}$ is the upper $\alpha/2$ point of a t_{n+m-2} distribution.

(3 marks)

(iii) Here

$$\hat{\sigma}^2 = \frac{9 \times 1.25 + 11 \times 1.16}{10 + 12 - 2} = 1.2005$$

and

$$t = \frac{23.7 - 21.6 - 1}{\sqrt{1.2005 \times \left(\frac{1}{10} + \frac{1}{12}\right)}} = 2.345$$

If $\alpha = 0.05$, the critical value is $t_{0.025} = 2.086$ on 20 d.f. Thus we reject H_0 as $|t| > t_{0.025}$.

If $\alpha = 0.01$, the critical value is $t_{0.005} = 2.845$ on 20 d.f. Thus we do not reject H_0 as $|t| < t_{0.005}$.

(4 marks)

BOOKWORK. Chapter 7, Part II.

TOTAL FOR A4, 10 MARKS

B5. (i) A Type I error occurs if H_0 is rejected when it is actually true.

A Type II error occurs if we fail to reject H_0 when it is in fact false.

The significance level (or size) of the test is the probability of rejecting H_0 when it is actually true, in other words the probability of a Type I error.

BOOKWORK [3 marks]

(ii) An appropriate test statistic is

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}},$$

where $\bar{X} = \sum_{i=1}^{n} X_i/n$. We reject H_0 if $Z > z_\alpha$, where z_α is the upper α point of a $N(0,1)$ distribution, i.e. $\Phi(z_\alpha) = 1 - \alpha$, with Φ the standard normal cdf.

Note that under H_0, $Z \sim N(0,1)$, and so the probability of rejecting H_0 if it is true is $P(Z > z_\alpha) = 1 - \Phi(z_\alpha) = \alpha$. Thus, the test has significance level α as claimed.

BOOKWORK [4 marks]

(iii) In this case $\bar{x} = \sum_i x_i/n = 10.6$. Hence $Z = (10.6 - 10)/\sqrt{1/10} = 0.6\sqrt{10} = 1.897$. Also note that $z_{0.05} = 1.645$ and $z_{0.01} = 2.326$. As $Z > z_{0.05}$ but $Z < z_{0.01}$, H_0 is rejected at the 5% significance level but not at the 1% significance level.

BOOKWORK [3 marks]

(iv) In this case $\bar{x} = \mu_0 + 0.2\sigma$, and so $Z = 0.2\sqrt{n}$. At the level α, H_0 is rejected when $Z > z_\alpha$ i.e. when

$$0.2\sqrt{n} > z_\alpha \implies n > 25 z_\alpha^2$$

For $\alpha = 0.05$, H_0 is rejected when $n > 67$. For $\alpha = 0.01$, H_0 is rejected when $n > 135$.

UNSEEN [3 marks]

(v) We use the fact that $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$. Hence

$$\begin{aligned}
\mathrm{P}\left(\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}} > z_\alpha\right) &= \mathrm{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + \frac{\mu-\mu_0}{\sigma/\sqrt{n}} > z_\alpha\right) = \mathrm{P}\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} > z_\alpha + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right) \\
&= 1 - \Phi\left(z_\alpha + \frac{\mu_0-\mu}{\sigma/\sqrt{n}}\right).
\end{aligned}$$

When $\mu_0 = 10$, $\mu = 10.5$, $\sigma = 1$, $\alpha = 0.05$ and $n = 10$, we have that the probability of correctly rejecting H_0 is

$$1 - \Phi\left(1.645 + \frac{-0.5}{1/\sqrt{10}}\right) = 1 - \Phi(0.0639) = 1 - 0.5255 = 0.4745$$

Students will most likely round 0.0639 to 0.06 when using the normal table and so obtain an overall probability of 0.4761, or a rounded version thereof e.g. 0.48.

BOOKWORK, all Chapter 9. [7 marks]

TOTAL FOR B5, 20 MARKS

B6.

(i) The likelihood is the joint probability of the data, considered as a function of p. By independence this is

$$L(p) = \prod_{i=1}^{n} p_X(x_i) = \prod_{i=1}^{n} \left[\binom{x_i + r - 1}{x_i} (1-p)^r p^{x_i} \right] = \left[\prod_{i=1}^{n} \binom{x_i + r - 1}{x_i} \right] (1-p)^{nr} p^{\sum_{i=1}^{n} x_i}$$

as claimed.

(2 marks) BOOKWORK

(ii) The log-likelihood is

$$\ell(p) = \log L(p) = \sum_{i=1}^{n} (\log \binom{x_i + r - 1}{x_i}) + nr \log(1-p) + \sum_{i=1}^{n} x_i \log p$$

This function has a turning point at $\hat{p}$ if

$$\begin{aligned} 0 = \left. \frac{d\ell}{dp} \right|_{\hat{p}} &= -\frac{nr}{(1-\hat{p})} + \frac{\sum_i x_i}{\hat{p}} \\ \implies -nr\hat{p} &+ (1-\hat{p}) \sum_i x_i = 0 \\ \implies \sum_i x_i &= \hat{p}(nr + \sum_i x_i) \\ \implies \hat{p} &= \frac{\sum_i x_i}{nr + \sum_i x_i} = \frac{\bar{x}}{r + \bar{x}} . \end{aligned}$$

To check it is indeed a maximum, consider the 2nd derivative

$$\left. \frac{d^2 \ell}{dp^2} \right|_{\hat{p}} = -\frac{nr}{(1-\hat{p})^2} - \frac{\sum_i x_i}{\hat{p}^2}$$

As $x_i \geq 0$, the above is < 0 and so the t.p. is indeed a maximum.

(7 marks)

BOOKWORK. Given the expression for the likelihood in part (i), this is a simple application of technique from Chapter 6.

(iii)

$$\begin{aligned} \mathrm{P}(a < \hat{p} < b) &= \mathrm{P}\left(a < \frac{\bar{X}}{r + \bar{X}} < b\right) = \mathrm{P}\left(a(r + \bar{X}) < \bar{X} < b(r + \bar{X})\right) \\ &= \mathrm{P}\left(ar < \bar{X}(1-a) \text{ and } \bar{X}(1-b) < br\right) \\ &= \mathrm{P}\left(\frac{ar}{1-a} < \bar{X} \text{ and } \bar{X} < \frac{br}{1-b}\right) \\ &= \mathrm{P}\left(\frac{ar}{1-a} < \bar{X} < \frac{br}{1-b}\right) \end{aligned}$$

(6 marks)

UNSEEN algebraic manipulation.

(iv) First, using part (iii), with $a = 0.45$, $b = 0.55$ and $r = 3$,

$$\mathrm{P}(0.45 < \hat{p} < 0.55) = \mathrm{P}(2.4545 < \bar{X} < 3.6666...) .$$

Note that $\mathrm{E}(\bar{X}) = \mathrm{E}(X) = 3$ and $\mathrm{Var}(\bar{X}) = \mathrm{Var}(X)/n = 6/100$. As n is large, using the central limit theorem $\frac{\bar{X}-3}{\sqrt{6/100}} \sim N(0,1)$ approximately and so

$$\begin{aligned}
\mathrm{P}(0.45 < \hat{p} < 0.55) = \mathrm{P}(2.4545 < \bar{X} < 3.6666...) &= \mathrm{P}\left(\frac{2.454545-3}{\sqrt{6/100}} < \frac{\bar{X}-3}{\sqrt{6/100}} < \frac{3.6666..-3}{\sqrt{6/100}}\right) \\
&\approx \Phi(2.7217) - \Phi(-2.2268) = \Phi(2.7217) - (1 - \Phi(2.2268)) \\
&= 0.9968 - (1 - 0.9870) = 0.9838
\end{aligned}$$

If rounding to 2 d.p. in normal tables

$$\Phi(2.72) - (1 - \Phi(2.23)) = 0.9967 - (1 - 0.9871) = 0.9838$$

Students who round early might get a slightly different answer - allow anything reasonable.

(5 marks - 2 marks for correctly using part (iii), 3 marks for normal calculations)

(SOMEWHAT) UNSEEN. Once it has been realised how to use part (iii), as suggested in the question, this is a standard application of technique from Chapter 4.

TOTAL FOR B6, 20 MARKS

B7. (a) (i) A suitable estimator is $\hat{\theta} = \hat{p}_1 - \hat{p}_2$, where $\hat{p}_1 = X_1/n$ is the proportion of defectives in the sample of batteries of Type 1, and $\hat{p}_2 = X_2/n$ is the proportion of defectives in the sample of batteries of Type 2.

(2 marks) BOOKWORK, cf. Chapter 7 Part II.

(ii) Note that assuming independence of the different batteries, the number of defectives in the first sample satisfies $X_1 \sim \mathrm{Bi}(n, p_1)$, which has expectation $E(X) = np_1$. Moreover $\hat{p}_1 = X_1/n$ hence $E(\hat{p}_1) = E(X_1/n) = E(X_1)/n = p_1$. Similarly, $E(\hat{p}_2) = p_2$. Thus

$$E\hat{\theta} = E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2 = \theta\,,$$

and so $\hat{\theta}$ is unbiased. The variance satisfies

$$\begin{aligned}\mathrm{Var}(\hat{\theta}) &= \mathrm{Var}(\hat{p}_1 - \hat{p}_2) = \mathrm{Var}\,\hat{p}_1 + \mathrm{Var}\,\hat{p}_2 \text{ by independence}\\ &= \mathrm{Var}(X_1/n) + \mathrm{Var}(X_2/m)\\ &= np_1(1-p_1)/n^2 + mp_2(1-p_2)/m^2\\ &= \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\end{aligned}$$

i.e.

$$v(n,m) = \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{m}\,.$$

(6 marks) BOOKWORK. Chapter 5.

(b) (i) The variance under Design 1 is $v(150, 150)$ and that under Design 2 is $v(100, 200)$. Hence the variance is smaller under Design 1 rather than Design 2 if

$$v(150,150) = \frac{p_1(1-p_1)}{150} + \frac{p_2(1-p_2)}{150} \le \frac{p_1(1-p_1)}{100} + \frac{p_2(1-p_2)}{200} = v(100,200)$$

The above is true if and only if

$$\begin{aligned}&p_2(1-p_2)\left(\frac{1}{150} - \frac{1}{200}\right) \le p_1(1-p_1)\left(\frac{1}{100} - \frac{1}{150}\right)\\ \iff\ &p_2(1-p_2)\left(\frac{4}{600} - \frac{3}{600}\right) \le p_1(1-p_1)\left(\frac{6}{600} - \frac{4}{600}\right)\\ \iff\ &p_2(1-p_2)\frac{1}{600} \le p_1(1-p_1)\frac{2}{600}\\ \iff\ &\frac{1}{2} \le \frac{p_1(1-p_1)}{p_2(1-p_2)}\end{aligned}$$

(4 marks) UNSEEN

(ii) When $p_1 = 0.1$, it is better to use Design 1 rather than Design 2 if $\mathrm{Var}\,\hat{\theta}$ is smallest under Design 1, i.e. if

$$\frac{1}{2} \le \frac{p_1(1-p_1)}{p_2(1-p_2)} \iff -p_2^2 + p_2 - 0.18 \le 0$$

The roots of this quadratic occur when $(p_2 - 1/2)^2 = 0.25 - 0.18 = 0.07$, i.e. when $p_2 = 0.5 \pm \sqrt{0.07} = 0.235, 0.765$. The quadratic displayed above is a 'sad face' shape and so the inequality above is satisfied if $p_2 \le 0.235$ or $p_2 \ge 0.765$, as required.

(4 marks) UNSEEN

(c) (i) A 95% CI is given by the end-points

$$\hat{\theta} \pm z_{0.025}\widehat{\mathrm{Var}}(\hat{\theta}) = \hat{p}_1 - \hat{p}_2 \pm z_{0.025}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}$$

Here $\hat{p}_1 = 13/150 = 0.08666$, $\hat{p}_2 = 17/150 = 0.11333$ and $z_{0.025} = 1.960$. Hence the CI is $(-0.0945, 0.0412)$.

(4 marks) BOOKWORK, Chapter 7 part II.

TOTAL FOR B7, 20 MARKS