

# Final exam solutions

## MAST90058 Elements of Statistics

Semester 2, 2018

1. (a) i.  $x_{(1)} = 2.62$   
ii.  $y_{(4.5)} = 0.5 \times (y_{(4)} + y_{(5)}) = 0.5 \times (4.78 + 5.30) = 5.04$   
iii.  $\bar{x} = 4.193$   
iv.  $\sigma_2$  is a parameter and its value is unknown.  
(b) i. **False**, since  $x_{(1)} = 2.62 < 4.15 = y_{(1)}$ .  
ii. **False**, since  $\bar{x} = 4.193 < 5.212 = \bar{y}$ .  
iii. **Unknown**, since both  $\sigma_1$  and  $\sigma_2$  are parameters and their values are unknown.  
(c) i. Cannot carry out this test, need to know the **sample size** of the observations on  $X$ .  
ii. Under  $H_0$  we have  $T = 9S_2^2/\sigma_2^2 = 9S_2^2/2^2 \sim \chi_9^2$ . Let  $F(x)$  be the cdf of  $\chi_9^2$ . We reject  $H_0$  if  $T < F^{-1}(0.025) = 2.7$  or  $T > F^{-1}(0.975) = 19.0$ . We observe  $t = 9/2^2 \times 0.837^2 = 1.58 < 2.7$ . Therefore, we reject  $H_0$ .
2. (a) Let  $c$  be the 0.975 quantile of  $t_9$ . 95% CI:  $\bar{x} \pm c \frac{s}{\sqrt{n}} = 6.82 \pm 2.262 \times 0.932 / \sqrt{10} = (6.15, 7.49)$ .  
(b) 7 of the 10 observations are greater than 6.5  $\Rightarrow \hat{p} = 7/10 = 0.7$ .  
95% CI:  $\hat{p} \pm 1.96 \sqrt{\hat{p}(1-\hat{p})/n} = (0.42, 0.98)$ .  
(c) Let  $W$  be the number of observations less than  $m$ . We have  $\Pr(X_{(i)} < m < X_{(j)}) = \Pr(i \leq W \leq j-1)$ . By convention, we will choose  $i$  and  $j$  symmetrically from each tail, so possible choices for the CI are  $(X_{(1)}, X_{(10)})$ ,  $(X_{(2)}, X_{(9)})$ ,  $(X_{(3)}, X_{(8)})$ , etc., we just need to find which of these gives a confidence level close to 90%.  
 $\Pr(X_{(1)} < m < X_{(10)}) = \Pr(1 \leq W \leq 9) = 1 - 2 \times \Pr(W \leq 0) = 1 - 2 \times 0.0009766 = 0.998$   
 $\Pr(X_{(2)} < m < X_{(9)}) = \Pr(2 \leq W \leq 8) = 1 - 2 \times \Pr(W \leq 1) = 1 - 2 \times 0.0107422 = 0.9785$   
 $\Pr(X_{(3)} < m < X_{(8)}) = \Pr(3 \leq W \leq 7) = 1 - 2 \times \Pr(W \leq 2) = 1 - 2 \times 0.0546875 = 0.891$   
Therefore, an approximate 90% CI is  $(x_{(3)}, x_{(8)}) = (6.0, 7.3)$ .
3. (a)  $L(\theta) = \prod_{i=1}^n \theta(1-\theta)^{x_i} = \theta^n (1-\theta)^{\sum x_i}$ . Therefore,  $y = \sum x_i$  is sufficient, by the factorisation theorem.  
(b) Need to solve  $\bar{x} = \mathbb{E}(X) = (1-\theta)/\theta$ , which gives the MM estimator  $\bar{\theta} = 1/(1+\bar{X})$ .  
(c) The log-likelihood is  $l(\theta) = n \log \theta + y \log(1-\theta)$ . The score function is,

$$\frac{\partial l}{\partial \theta} = \frac{n}{\theta} - \frac{y}{1-\theta}.$$

Setting this to zero and solving gives the MLE,

$$\hat{\theta} = \frac{1}{1+\bar{X}}.$$

Note that this is the same as the MM estimator.

(d) We first derive the information functions.

$$-\frac{\partial^2 l}{\partial \theta^2} = \frac{n}{\theta^2} + \frac{y}{(1-\theta)^2}.$$

$$I(\theta) = \mathbb{E} \left( -\frac{\partial^2 l}{\partial \theta^2} \right) = \frac{n}{\theta^2} + \frac{n \mathbb{E}(\bar{X})}{(1-\theta)^2} = \frac{n}{\theta^2} + \frac{n}{(1-\theta)^2} \frac{1-\theta}{\theta} = \frac{n}{\theta^2(1-\theta)}.$$

Therefore, the Cramér–Rao lower bound is  $1/I(\theta) = \theta^2(1-\theta)/n$ .

(e) Using the asymptotic variance of the MLE,  $\text{se}(\hat{\theta}) = \sqrt{\hat{\theta}^2(1-\hat{\theta})/n}$ .

(f) From the sample we can calculate  $\bar{x} = 1.1$ . This gives  $\hat{\theta} = 1/(1+\bar{x}) = 0.476$ . We also have  $\text{se}(\hat{\theta}) = 0.0771$  and an approximate 90% CI is given by  $\hat{\theta} \pm 1.645 \text{se}(\hat{\theta}) = (0.35, 0.60)$ .

4. (a) She carried out  $n = 30$  experiments. We can tell this from the output because the degrees of freedom is reported as 28 and we know this is equal to  $n - 2$ .
- (b) i.  $t = -0.7323/1.8852 = -0.388$ ,  $|t| < 2.048$  (0.975 quantile of  $t_{28}$ ). Do not reject  $H_0$ .  
 ii.  $t = 1.1556/0.3818 = 3.03$ ,  $|t| > 2.048$ . Reject  $H_0$ .
- (c) Yes, we rejected  $\beta = 0$ . Also, the R output shows a p-value of 0.005 for the F-test of the same hypotheses (hence, reject).
- (d) We know the degrees of freedom are 1, 28 and 29. We also know that  $MS(E) = \hat{\sigma}^2 = 1.686^2 = 2.84$ , and  $F = 9.159$  is given in the R output. It is straightforward to deduce the rest of the table:

Source	df	SS	MS	F
Model	1	26.02	26.02	9.159
Error	28	79.55	2.84	
Total	29	105.57		

5. (a) The median of an exponential distribution with mean  $\theta$  is  $m = \theta \log 2$ . You can derive this easily from the cdf:  $0.5 = F(m) = 1 - e^{-m/\theta} \Rightarrow m = \theta \log 2$ .
- (b) i. Asymptotically,  $\hat{M} \approx N \left( m, \frac{1}{4nf(m)^2} \right)$ . The pdf is  $f(x) = \frac{1}{\theta} e^{-x/\theta}$ , so we have  $f(m) = f(\theta \log 2) = \frac{1}{\theta} e^{-\log 2} = \frac{1}{2\theta}$ . This gives,  $\hat{M} \approx N \left( \theta \log 2, \frac{\theta^2}{n} \right) = N \left( m, \frac{m^2}{n(\log 2)^2} \right)$ .  
 ii.  $\hat{m} = x_{(5)} = 1.2$   
 iii.  $\text{se}(\hat{m}) = \frac{\hat{m}}{\sqrt{n \log 2}} = 0.577$
- (c) i.  $\mathbb{E}(\bar{X}) = \mathbb{E}(X) = \theta = \frac{m}{\log 2} \neq m$ .  
 ii.  $c = \log 2$  will make  $T$  unbiased.  
 iii. Firstly,  $\text{var}(X) = \theta^2$ . Therefore,  $\text{var}(T) = c^2 \text{var}(\bar{X}) = (\log 2)^2 \theta^2 / n = m^2 / n$ .  
 iv. Both  $\hat{M}$  and  $T$  are asymptotically unbiased but  $T$  has smaller variance:  $\text{var}(\hat{M}) \approx \frac{m^2}{n(\log 2)^2} = \text{var}(T) / (\log 2)^2 \approx 2.08 \text{var}(T) > \text{var}(T)$ . Thus,  $T$  is the better estimator.  
 v.  $t = (\log 2)\bar{x} = 0.693 \times 1.9 = 1.32$   
 vi.  $\text{se}(t) = t/\sqrt{n} = 0.44$
6. (a)  $\theta \sim \text{Gamma}(\alpha, \beta) \implies \theta \mid \text{data} \sim \text{Gamma}(\alpha + n\bar{x}, \beta + n)$   
 (b) We want  $\theta \sim \text{Gamma}(\alpha, \beta)$  that satisfy  $\beta = 10$  and  $\alpha = 10 \times 3.0 = 30$ .  
 (c) The posterior is  $\theta \mid \text{data} \sim \text{Gamma}(\alpha + 45 \times 3.8, \beta + 45) = \text{Gamma}(201, 55)$ .

(d) Posterior mean:  $\mathbb{E}(\theta \mid \text{data}) = 201/55 = 3.65$ .

7. (a) Completing the ANOVA table:

Source	df	SS	MS	F
Treatment (sport)	2	416.4	208.2	5.48
Error	12	456	38	
Total	14	872.4		

The 0.95 quantile of  $F_{2,12}$  is 3.885. We have that  $F = 5.48 > 3.885$ . Therefore, we can reject  $H_0$ , meaning we have evidence that the mean height of students differs across the three sports.

(b) In doing a one-way ANOVA, Ben has assumed that the variance of the heights is the same for each sport.

(c)

$$\frac{12\hat{\sigma}^2}{\sigma^2} \sim \chi_{12}^2$$

(d)

$$\hat{\mu}_1 = \bar{x}_1 \sim N\left(\mu_1, \frac{\sigma^2}{4}\right)$$

(e) Let  $c$  be the 0.975 quantile of  $t_{12}$ . A 95% confidence interval for  $\mu_1$  is,

$$\hat{\mu}_1 \pm c \frac{\hat{\sigma}}{\sqrt{4}} = 177 \pm 2.179 \times \frac{\sqrt{38}}{2} = (170.3, 183.7).$$

8. (a) Using a goodness-of-fit test. The observed and expected frequencies are:

	Genotype		
	AA	Aa	aa
Obs.	153	286	161
Exp.	150	300	150

The test statistic is:

$$\chi^2 = \frac{(153 - 150)^2}{150} + \frac{(286 - 300)^2}{300} + \frac{(161 - 150)^2}{150} = 1.52.$$

The 0.95 quantile of a  $\chi_2^2$  distribution is 5.992. Since  $1.52 < 5.992$ , we cannot reject  $H_0$ .

(b) The corrected counts are: (153, 271, 176). Re-doing the test:

	Genotype		
	AA	Aa	aa
Obs.	153	271	176
Exp.	150	300	150

The test statistic is  $\chi^2 = 7.37$ , therefore this time we reject  $H_0$ .