Exam 16 November 2017, answers

Statistics (University of Melbourne)

Final exam solutions

MAST20005 Statistics

Semester 2, 2017

1. (a) $\mathbb{E}(X) = 3\theta$ and $\text{var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = 5\theta - 9\theta^2 = \theta(5 - 9\theta)$

   (b) i. Let the sample frequencies of the three values be n_0, n_1, n_2. The likelihood is $L(\theta) = (1-2\theta)^{n_0}\theta^{n_1}\theta^{n_2} = (1-2\theta)^{n_0}\theta^{n_1+n_2}$, and therefore the log-likelihood is:

   $$l(\theta) = n_0 \log(1-2\theta) + (n_1 + n_2)\log\theta.$$

   Note that $n_0 + n_1 + n_2 = n$, so an alternative expression for the log-likelihood is:

   $$l(\theta) = n_0 \log(1-2\theta) + (n - n_0)\log\theta.$$

   ii. $\frac{\partial l}{\partial \theta} = -\frac{2n_0}{1-2\theta} + \frac{n-n_0}{\theta} = 0 \quad \Rightarrow \quad \hat{\theta} = \frac{n-N_0}{2n} = \frac{N_1+N_2}{2n}$.

   iii. A sufficient statistic is n_0, the number of 0's. An equivalent alternative is $n_1 + n_2 = n - n_0$, which is the sum of the number of 1's and 2's.

   iv. Note that $N_0 \sim \text{Bi}(n, 1-2\theta)$, which means that $\mathbb{E}(\hat{\theta}) = \frac{n-n(1-2\theta)}{2n} = \theta$, and therefore the MLE is unbiased.

   v. $-\frac{\partial^2 l}{\partial\theta^2} = \frac{4n_0}{(1-2\theta)^2} + \frac{n-n_0}{\theta^2} \Rightarrow I(\theta) = \mathbb{E}\left(-\frac{\partial^2 l}{\partial\theta^2}\right) = \frac{4n(1-2\theta)}{(1-2\theta)^2} + \frac{n-n(1-2\theta)}{\theta^2} = \frac{4n}{1-2\theta} + \frac{2n}{\theta} = \frac{4n\theta+2n(1-2\theta)}{\theta(1-2\theta)} = \frac{2n}{\theta(1-2\theta)}$. Therefore, the lower bound is $\frac{1}{I(\theta)} = \frac{\theta(1-2\theta)}{2n}$.

   vi. $\text{var}(\hat{\theta}) = \frac{1}{4n^2}\text{var}(N_0) = \frac{1}{4n^2}n(1-2\theta)(2\theta) = \frac{\theta(1-2\theta)}{2n}$. Therefore, the MLE achieves the lower bound.

   (c) i. We have $n_0 = 8$, $n_1 = 4$, $n_2 = 8$. This gives $\hat{\theta} = \frac{12}{2\times 20} = 0.30$. To get a standard error we substitute this for θ into the expression for the variance of $\hat{\theta}$, which gives, $\text{se}(\hat{\theta}) = \sqrt{0.3 \times 0.4/40} = 0.055$.

   ii. Based on the above results, we know that $\hat{\theta} \approx \text{N}(\theta, 0.055^2)$. Therefore, an approximate 95% confidence interval is given by $\hat{\theta} \pm 1.96 \times 0.055 = (0.19, 0.41)$.
   Note: Using either the observed information function or the Fisher information function will lead to the same result. In particular, $J(\hat{\theta}) = I(\hat{\theta}) = \frac{4n^3}{n_0(n-n_0)} = 1000/3$. An exact alternative is an interval based on F_0, which follows a binomial distribution. This gives a very similar result:

   ```
   > round((1 - rev(binom.test(8, 20)$conf.int)) / 2, 2)
   [1] 0.18 0.40
   ```

   iii. The expected counts (based on $\hat{\theta}$) are: 8, 6, 6. The test statistic is:

   $$\chi^2 = \frac{(8-8)^2}{8} + \frac{(4-6)^2}{6} + \frac{(8-6)^2}{6} = \frac{4}{3} = 1.33.$$

   We have estimated one parameter so have a test with 1 degree of freedom. The 0.95 quantile of a χ^2_1 distribution is 3.84. Since $1.33 < 3.84$, we cannot reject H_0.

iv. It will not be possible, there will be no degrees of freedom remaining to carry out the test. The resulting model will be simply a binomial distribution. (Once we estimate the probability parameter for the binomial, we get a 'perfect' (saturated) fit to the data and cannot test for deviations.)

2. (a) i. Let the sample frequencies of the three values be N_0, N_1, N_2. The MM estimate is obtained by solving $3\theta = \bar{X}$, which gives $\tilde{\theta} = \frac{1}{3}\bar{X} = \frac{N_1+2N_2}{3n}$.

ii. $\mathbb{E}(\tilde{\theta}) = \frac{1}{3}\mathbb{E}(\bar{X}) = \frac{1}{3}\mathbb{E}(X) = \theta$, and therefore the MM estimator is unbiased.

iii. $\text{var}(\tilde{\theta}) = \frac{1}{9}\text{var}(\bar{X}) = \frac{1}{9}\frac{\text{var}(X)}{n} = \frac{\theta(5-9\theta)}{9n}$. Note that,

$$\text{var}(\tilde{\theta}) = \frac{\theta(\frac{5}{9}-\theta)}{n} > \frac{\theta(\frac{1}{2}-\theta)}{n} = \text{var}(\hat{\theta}) = \frac{1}{I(\theta)},$$

So we can see that the MM estimator does **not** achieve the lower bound.

iv. From above, we see that both estimators are unbiased but the MLE has lower variance, so the MLE is a better estimator in this scenario.

(b) i. We have $n_0 = 8$, $n_1 = 4$, $n_2 = 8$. This gives $\tilde{\theta} = \frac{20}{3\times 20} = \frac{1}{3} = 0.333$. To get a standard error we substitute this for θ into the expression for the variance of $\tilde{\theta}$, which gives,

$$\text{se}(\tilde{\theta}) = \sqrt{\frac{1}{3}\times\frac{2}{180}} = \frac{1}{\sqrt{270}} = 0.061.$$

ii. The MM estimator is based on the sample mean and so due to the Central Limit Theorem will be approximately normally distributed. Based on the above results, we know that $\tilde{\theta} \approx \text{N}(\theta, 0.061^2)$. Therefore, an approximate 95% confidence interval is given by $\tilde{\theta} \pm 1.96 \times 0.061 = (0.21, 0.45)$.

3. (a) These are paired samples so we first calculate the differences and treat those as coming from a single normal distribution. We get the following summary statistics for the differences: $\bar{d} = 3.4$ and $s_d = 3.75$. The sample size is $n = 10$ and we need the 0.975 quantile of t_9 which is 2.26. A 95% confidence interval is given by: $3.4 \pm 2.26 \times 3.75/\sqrt{10} = (0.72, 6.1)$.

(b) i. H_0: $\mu_1 = \mu_2$ and H_1: $\mu_1 \neq \mu_2$.

ii. Use the t-test statistic on the differences, $T = \frac{\bar{D}}{S_D/\sqrt{10}}$.

iii. $T \sim t_9$ under the null and its 0.975 quantile is 2.26. Therefore, reject H_0 if $|t| > 2.26$.

iv. $t = 3.4/(3.75/\sqrt{10}) = 2.87 > 2.26$ so therefore reject H_0.

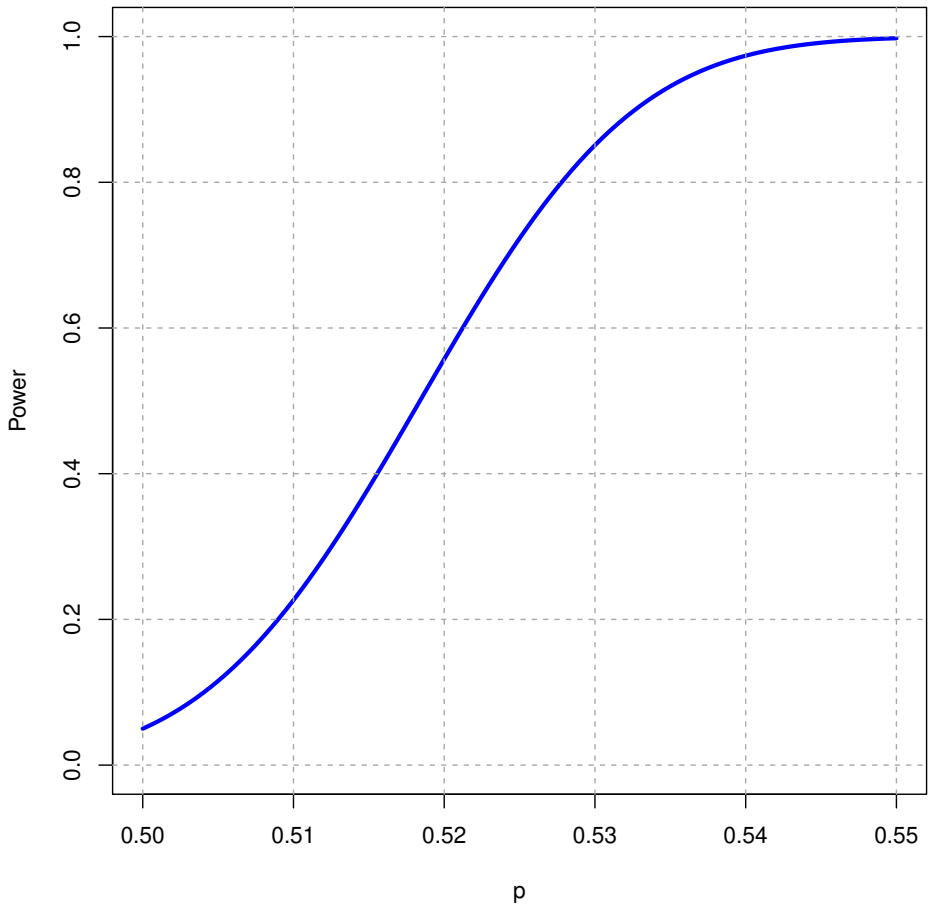
4. (a) Using the standard approximation for a single proportion, $\hat{p}_1 = 260/500 = 0.520$, $\text{se}(\hat{p}_1) = \sqrt{\frac{0.52\times 0.48}{500}} = 0.0223$, so a 95% confidence interval is $\hat{p}_1 \pm 1.96\,\text{se}(\hat{p}_1) = (0.48, 0.56)$.

(b) We need to account for the sampling variation in both polls in any comparison. Using the standard comparison of two proportions, $\hat{p}_2 = 255/520 = 0.490$, $\hat{p}_2 - \hat{p}_1 = -0.0296$, $\text{se}(\hat{p}_2 - \hat{p}_1) = \sqrt{\frac{0.52\times 0.48}{500} + \frac{0.49\times 0.51}{520}} = 0.0313$, so a 95% confidence interval is $\hat{p}_2 - \hat{p}_1 \pm 1.96\,\text{se}(\hat{p}_2 - \hat{p}_1) = (-0.091, 0.032)$. Differences in both a positive and negative direction are plausible, which means we do not have strong evidence for a change between the two polls. We certainly cannot draw as strong a conclusion as what the newspaper reported.

(c) It needs to be $n \geqslant 1.96^2/\left(4 \times (0.01)^2\right) = 9604$.

(d) Let p be the proportion of the population who support the Purple Party. The test is $H_0 : p = 0.5$ versus $H_1 : p > 0.5$. Use the standard test for a single proportion: reject H_0 if $\frac{\hat{p}-p_0}{\sqrt{p_0(1-p_0)/n}} > c$. We have $p_0 = 0.5$, $n = 2000$ and $c = \Phi^{-1}(0.95) = 1.645$, so the rejection rule simplifies to: $2\sqrt{n}(\hat{p} - 0.5) > 1.645$. Under H_1, we have $\hat{p} \approx \mathrm{N}(p, p(1-p)/n)$. With a bit of algebraic manipulation, we can show the power function is:

$$K(p) = \Pr(2\sqrt{n}(\hat{p} - 0.5) > 1.645 \mid p) = \Phi\left(\frac{2\sqrt{n}(p - 0.5) - 1.645}{2\sqrt{p(1-p)}}\right)$$



(e) Use $n = 2000$ and $p = 0.53$ to give $\beta = 1 - K(0.53) = 1 - \Phi(1.04) = 1 - 0.851 = 0.149$

5. (a) This is a sample from a binomial distribution so the MLE is $\hat{p} = \frac{1}{40} = 0.025$.

(b) The conjugate prior for a binomial likelihood is a **beta distribution**.

(c) We want $p \sim \mathrm{Beta}(\alpha, \beta)$ where the α and β can be interpreted as pseudocounts. We require that $\alpha + \beta = 5$ and also that $\mathbb{E}(p) = \frac{\alpha}{\alpha+\beta} = 0.01$. Solving these gives $\alpha = 0.05$ and $\beta = 4.95$.

(d) The posterior is $p \mid \text{data} \sim \mathrm{Beta}(1 + \alpha, 39 + \beta) = \mathrm{Beta}(1.05, 43.95)$.

(e) Posterior mean: $\mathbb{E}(p \mid \text{data}) = \frac{1.05}{45} = 0.023$. Central 95% credible interval: $(0.0007, 0.0826)$.

6. This is a standard contingency table scenario. The observed and expected frequencies are:

		Symptoms		
		Worse	Same	Better
Placebo	O	15	10	18
	E	9.46	6.45	27.09
Drug	O	7	5	45
	E	12.54	8.55	35.91

The test statistic is:

$$\chi^2 = \frac{(15-9.46)^2}{9.46} + \cdots + \frac{(45-35.91)^2}{35.91} = 14.47.$$

The 0.99 quantile of a χ^2_2 distribution is 9.21. Since $14.47 > 9.21$, we reject H_0.

7. (a) i. According to the exponential model, $\mathbb{E}(X) = \text{sd}(X) = \theta$. The Central Limit Theorem then implies $\hat{\theta} = \bar{X} \approx \text{N}(\theta, \theta^2/n)$.
   ii. $\hat{\theta} = \bar{x} = 3.9$.
   iii. $\text{se}(\hat{\theta}) = \hat{\theta}/\sqrt{n} = 3.9/\sqrt{9} = 1.3$. An alternative solution is to use the sample standard deviation directly, $\text{se}(\hat{\theta}) = s/\sqrt{n} = 0.82$. This avoids making the exponential distribution assumption so is more robust, but is less accurate if the exponential distribution is indeed true. Both are valid solutions to the question.

   (b) i. The median of an exponential distribution with mean θ is $\theta \log 2$. You can derive this easily from the cdf: $0.5 = F(m) = 1 - e^{-m/\theta} \Rightarrow m = \theta \log 2$. The sample median is asymptotically unbiased, which means $\mathbb{E}(\hat{M}) \approx \theta \log 2$. Therefore, it is biased for θ.
   ii. According to the above, letting $c = 1/\log 2$ makes T asymptotically unbiased.
   iii. Asymptotically, $\hat{M} \approx \text{N}(m, \frac{1}{4nf(m)^2})$. 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, $\hat{M} \approx \text{N}(\theta \log 2, \frac{\theta^2}{n})$. Therefore, $T \approx \text{N}(\theta, \frac{\theta^2}{n(\log 2)^2})$.
   iv. Both $\bar{X}$ and T are asymptotically unbiased but $\bar{X}$ has smaller variance: $\text{var}(T) \approx \frac{\theta^2}{n(\log 2)^2} = 2.08\frac{\theta^2}{n} \approx 2.08\,\text{var}(\bar{X}) > \text{var}(\bar{X})$. Therefore, $\bar{X}$ is the better estimator.
   v. $t = \hat{m}/\log 2 = 1.443 x_{(5)} = 1.443 \times 3.5 = 5.05$
   vi. $\text{se}(t) = \frac{t}{\sqrt{n}\log 2} = 2.43$

8. (a) $L(\theta) = \prod_{i=1}^n \frac{2x_i}{\theta^2} \propto \theta^{-2n}$ with $\theta \geqslant x_{(n)}$. This is maximised on the boundary, $\hat{\theta} = X_{(n)}$.

   (b) $F(x) = \int_0^x f(y)\,dy = \left(\frac{x}{\theta}\right)^2$. The pdf of the MLE is $G(x) = \Pr(X_{(n)} \leqslant x) = \Pr(X \leqslant x)^n = F(x)^n = \left(\frac{x}{\theta}\right)^{2n}$. The p-quantile of this distribution satisfies $p = G(\pi_p) = \left(\frac{\pi_p}{\theta}\right)^{2n}$. Rearranging gives $\pi_p = \theta p^{1/(2n)}$. For a central 95% confidence interval, we need bounds based on $p = 0.025$ and $p = 0.975$,

$$\begin{aligned} 0.95 &= \Pr(\theta \times 0.025^{1/(2n)} \leqslant X_{(n)} \leqslant \theta \times 0.975^{1/(2n)}) \\ &= \Pr(X_{(n)} \times 0.975^{-1/(2n)} \leqslant \theta \leqslant X_{(n)} \times 0.025^{-1/(2n)}). \end{aligned}$$

   Therefore, a 95% confidence interval is given by $(x_{(n)} \times 0.975^{-1/(2n)}, x_{(n)} \times 0.025^{-1/(2n)})$.

   (c) $n = 6$, $x_{(6)} = 4.8$, $0.975^{-1/12} = 1.002112$, $0.025^{-1/12} = 1.359894$, and a central 95% confidence interval for θ is $(4.81, 6.53)$.