Serik Sagitov, Chalmers and GU, February 6, 2018

Solutions chapter 8

Problem 8.3

Number X of yeast cells on a square. Test the Poisson model $X \sim \text{Pois}(\lambda)$.

Concentration 1.

$$\bar{X} = 0.6825, \quad \overline{X^2} = 1.2775, \quad s^2 = 0.8137, \quad s = 0.9021, \quad s_{\bar{X}} = 0.0451.$$

Approximate 95% CI for μ: 0.6825 ± 0.0884.
Pearson's chi-square test based on $\hat{\lambda} = 0.6825$:

x	0	1	2	3	4+	Total
Observed	213	128	37	18	4	400
Expected	202.14	137.96	47.08	10.71	2.12	400

Observed test statistic $X^2 = 10.12$, df $= 5 - 1 - 1 = 3$, $P < 0.025$. Reject the model.

Concentration 2.

$$\bar{X} = 1.3225, \quad \overline{X^2} = 3.0325, \quad s = 1.1345, \quad s_{\bar{X}} = 0.0567.$$

Approximate 95% CI for μ: 1.3225 ± 0.1112.
Pearson's chi-square test: observed test statistic $X^2 = 3.16$, df $= 4$, $P > 0.10$. Accept the model.

Concentration 3.

$$\bar{X} = 1.8000, \quad s = 1.1408, \quad s_{\bar{X}} = 0.0701.$$

Approximate 95% CI for μ: 1.8000 ± 0.1374.
Pearson's chi-square test: observed test statistic $X^2 = 7.79$, df $= 5$, $P > 0.10$. Accept the model.

Concentration 4.

$$n = 410, \quad \bar{X} = 4.5659, \quad s^2 = 4.8820, \quad s_{\bar{X}} = 0.1091.$$

Approximate 95% CI for μ: 4.566 ± 0.214.
Pearson's chi-square test: observed test statistic $X^2 = 13.17$, df $= 10$, $P > 0.10$. Accept the model.

Problem 8.4

Population distribution: X takes values $0, 1, 2, 3$ with probabilities

$$p_0 = \frac{2}{3} \cdot \theta, \quad p_1 = \frac{1}{3} \cdot \theta, \quad p_2 = \frac{2}{3} \cdot (1 - \theta), \quad p_3 = \frac{1}{3} \cdot (1 - \theta), \quad \theta \in [0, 1].$$

Two independent coin model: 1/3-coin and θ-coin. IID sample with $n = 10$

$$3, 0, 2, 1, 3, 2, 1, 0, 2, 1, \quad \bar{X} = 1.5, \quad s = 1.08.$$

Observed counts $(O_0, O_1, O_2, O_3) \sim \text{Mn}(n, p_0, p_1, p_2, p_3)$:

x	0	1	2	3	Total
O_x	2	3	3	2	10

Observe that $T = O_0 + O_1$ has $\text{Bin}(n, \theta)$ distribution.

(a) Method of moments. Using

$$\mu = \frac{1}{3} \cdot \theta + 2 \cdot \frac{2}{3} \cdot (1 - \theta) + 3 \cdot \frac{1}{3} \cdot (1 - \theta) = \frac{7}{3} - 2\theta,$$

derive an equation

$$\bar{X} = \frac{7}{3} - 2\tilde{\theta}.$$

It gives an unbiased estimate

$$\tilde{\theta} = \frac{7}{6} - \frac{\bar{X}}{2} = \frac{7}{6} - \frac{3}{4} = 0.417.$$

(b) To find $s_{\tilde{\theta}}$, observe that

$$\text{Var}(\tilde{\theta}) = \frac{1}{4}\text{Var}(\bar{X}) = \frac{\sigma^2}{40}.$$

Thus we need to find $s_{\tilde{\theta}}$, which estimates $\sigma_{\tilde{\theta}} = \frac{\sigma}{6.325}$. Next we estimate σ using two methods.

Method 1. From

$$\sigma^2 = \text{E}(X^2) - \mu^2 = \frac{1}{3} \cdot \theta + 4 \cdot \frac{2}{3} \cdot (1 - \theta) + 9 \cdot \frac{1}{3} \cdot (1 - \theta) = \frac{7}{3} - 2\theta - \left(\frac{7}{3} - 2\theta\right)^2 = \frac{2}{9} + 4\theta - 4\theta^2,$$

we estimate σ as

$$\sqrt{\frac{2}{9} + 4\tilde{\theta} - 4\tilde{\theta}^2} = 1.093.$$

This gives

$$s_{\tilde{\theta}} = \frac{1.093}{6.325} = 0.173.$$

Method 2:

$$s_{\tilde{\theta}} = \frac{s}{6.325} = \frac{1.08}{6.325} = 0.171.$$

(c) Likelihood function

$$L(\theta) = \left(\frac{2}{3} \cdot \theta\right)^{O_0} \left(\frac{1}{3} \cdot \theta\right)^{O_1} \left(\frac{2}{3} \cdot (1 - \theta)\right)^{O_2} \left(\frac{1}{3} \cdot (1 - \theta)\right)^{O_3} = \text{const } \theta^T (1 - \theta)^{n - T},$$

where $T = O_0 + O_1$ is a sufficient statistic. Log-likelihood and its derivative

$$\ln L(\theta) = \text{const } + T \ln \theta + (n - T) \ln(1 - \theta),$$
$$(\ln L(\theta))' = \frac{T}{\theta} - \frac{n - T}{1 - \theta}.$$

Setting the latter to zero, we find

$$\frac{T}{\hat{\theta}} = \frac{n - T}{1 - \hat{\theta}}, \quad \hat{\theta} = \frac{T}{n} = \frac{2 + 3}{10} = \frac{1}{2}.$$

The MLE is the sample proportion, an unbiased estimate of the population proportion θ.

(d) We find $s_{\hat\theta}$ using the formula for the standard error of sample proportion

$$s_{\hat\theta} = \sqrt{\frac{\hat\theta(1-\hat\theta)}{n-1}} = 0.167.$$

A similar answer is obtained using the formula

$$s_{\hat\theta} = \sqrt{\frac{1}{nI(\hat\theta)}}, \quad I(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(Y|\theta)\right),$$

where $Y \sim \mathrm{Ber}(\theta)$ and $f(1|\theta) = \theta$, $f(0|\theta) = 1-\theta$. Since

$$\frac{\partial^2}{\partial\theta^2}\ln f(1|\theta) = \frac{\partial^2}{\partial\theta^2}\ln\theta = -\frac{1}{\theta^2}, \quad \frac{\partial^2}{\partial\theta^2}\ln f(0|\theta) = \frac{\partial^2}{\partial\theta^2}\ln(1-\theta) = -\frac{1}{(1-\theta)^2},$$

we get

$$I(\theta) = -\mathrm{E}\left(\frac{\partial^2}{\partial\theta^2}\ln f(Y|\theta)\right) = \frac{1}{\theta^2}\cdot\theta + \frac{1}{(1-\theta)^2}\cdot(1-\theta) = \frac{1}{\theta(1-\theta)}.$$

(e) Assume uniform prior $\theta \sim \mathrm{U}(0,1)$ and find the posterior density. Since

$$f(x|\theta) \propto \theta^5(1-\theta)^5,$$

and the prior is flat, we get

$$h(\theta|x) \propto f(x|\theta) \propto \theta^5(1-\theta)^5.$$

We conclude that the posterior distribution is Beta $(6,6)$. This yields

$$\hat\theta_{\mathrm{MAP}} = \hat\theta_{\mathrm{PME}} = \frac{1}{2}.$$

Problem 8.6

Likelihood function of $X \sim \mathrm{Bin}(n,p)$ for a given n and $X = x$ is

$$L(p) = \binom{n}{x}p^x(1-p)^{n-x} \propto p^x(1-p)^{n-x}.$$

(a) To maximise $L(p)$ we minimise

$$\ln p^x(1-p)^{n-x}) = x\ln p + (n-x)\ln(1-p).$$

Since

$$\frac{\partial}{\partial p}(x\ln p + (n-x)\ln(1-p)) = \frac{x}{p} - \frac{n-x}{1-p},$$

we have to solve $\frac{x}{p} = \frac{n-x}{1-p}$, which brings the MLE formula $\hat p = \frac{x}{n}$.

(b) We have $X = Y_1 + \ldots + Y_n$, where $(Y_1, \ldots, Y_n)$ is an IID sample from a Bernoulli distribution

$$f(y|p) = p^y (1-p)^{1-y}, \quad y = 0, 1.$$

By Cramer-Rao, if $\tilde{p}$ is an unbiased estimate of p, then

$$\mathrm{Var}(\tilde{p}) \geq \frac{1}{nI(p)},$$

where

$$I(p) = -\mathrm{E}\left(\frac{d^2}{dp^2} \ln f(Y|p)\right).$$

Using

$$\begin{aligned} \ln f(y|p) &= y \ln p + (1-y)\ln(1-p), \\ \frac{d}{dp} \ln f(y|p) &= \frac{y}{p} - \frac{1-y}{1-p}, \\ \frac{d^2}{dp^2} \ln f(y|p) &= -\frac{y}{p^2} - \frac{1-y}{(1-p)^2}, \end{aligned}$$

we find

$$I(p) = \mathrm{E}\left(\frac{Y}{p^2} + \frac{1-Y}{(1-p)^2}\right) = \frac{1}{p(1-p)},$$

and conclude that the sample proportion $\hat{p}$ has the smallest variance

$$\mathrm{Var}(\tilde{p}) \geq \frac{1}{nI(p)} = \frac{p(1-p)}{n} = \mathrm{Var}(\hat{p}).$$

(c) Plot $L(p) = 252p^5(1-p)^5$.

Problem 8.8

Number of bird hops $X \sim \mathrm{Geom}(p)$

$$f(x|p) = (1-p)^{x-1}p, \quad x = 1, 2, \ldots.$$

Data

$$\boldsymbol{x} = (x_1, \ldots, x_{130}).$$

(d) Using a uniform prior $p \sim \mathrm{U}(0,1)$, we find the posterior to be

$$h(p|\boldsymbol{x}) \propto f(x_1|p) \cdots f(x_n|p) = (1-p)^{n\bar{X}-n}p^n, \quad n = 130, \quad n\bar{X} = 363.$$

It is a beta distribution

$$\mathrm{Beta}(n+1, n\bar{X} - n + 1) = \mathrm{Beta}(131, 234).$$

Posterior mean

$$\mu = \frac{a}{a+b} = \frac{131}{131+234} = 0.36, \quad \mu = \frac{1+\frac{1}{n}}{\bar{X}+\frac{2}{n}},$$

and standard deviation

$$\sigma = \sqrt{\frac{\mu(1-\mu)}{a+b+1}} = \sqrt{\frac{0.36 \cdot 0.64}{366}} = 0.025.$$

Problem 8.26

Capture-recapture method: N fish in the lake. Estimate N by first capturing and tagging $n = 100$ fish, then releasing them in the lake and capturing $k = 50$ fish. Suppose among $k = 50$ fish $X = 20$ fish were tagged.

Statistical model: sampling without replacement of $k = 50$ balls from an urn with N balls of which n balls are black. Hypergeometric distribution

$$\mathrm{P}(X = 20) = \frac{\binom{n}{20}\binom{N-n}{30}}{\binom{N}{50}}.$$

The likelihood function

$$L(N) = \frac{\binom{100}{20}\binom{N-100}{30}}{\binom{N}{50}} = \text{const} \cdot \frac{(N-100)(N-101)\cdots(N-129)}{N(N-1)\cdots(N-49)}.$$

To find the maximum consider the ratio

$$\frac{L(N)}{L(N-1)} = \frac{(N-100)(N-50)}{N(N-130)}.$$

Solving the equation

$$(\hat{N}-100)(\hat{N}-50) = \hat{N}(\hat{N}-130),$$

we arrive at the MLE estimate $\hat{N} = \frac{5000}{20} = 250$.

Intuitively,

$$100 : N \approx 20 : 50.$$

Problem 8.32

An IID sample of size $n = 16$ from a normal distribution.

(a) $\bar{X} = 3.6109$, $s^2 = 3.4181$, $s_{\bar{X}} = 0.4622$.

(b), (c) Three exact CIs

	90%	95%	99%
μ	3.61 ± 0.81	3.61 ± 0.98	3.61 ± 1.36
σ^2	(2.05; 7.06)	(1.87; 8.19)	(1.56; 11.15)
σ	(1.43; 2.66)	(1.37; 2.86)	(1.25; 3.34)

(d) Find sample size x to halve the CI length:

$$t_{15}(\alpha/2) \cdot \frac{s}{\sqrt{16}} = 2 \cdot t_{x-1}(\alpha/2) \cdot \frac{s'}{\sqrt{x}},$$

implies $x \approx (2 \cdot 4)^2 = 64$. Further adjustment for 95% CI:

$$t_{15}(\alpha/2) = 2.13, \quad t_{x-1}(\alpha/2) \approx 2,$$

therefore $x \approx (2 \cdot 4 \cdot \frac{2}{2.13})^2 = 56.4$.

Problem 8.53

An IID sample $(X_1, \dots, X_n)$ from the uniform distribution $\mathrm{U}(0, \theta)$ with density

$$f(x) = \frac{1}{\theta} 1_{\{0 \le x \le \theta\}}.$$

(a) Method of moments estimate $\tilde{\theta}$ is unbiased

$$\mu = \theta/2, \quad \tilde{\theta} = 2\bar{X}, \quad \mathrm{E}(\tilde{\theta}) = \theta, \quad \mathrm{Var}(\tilde{\theta}) = \frac{4\sigma^2}{n} = \frac{\theta^2}{3n}.$$

(b) Denote $X_{(n)} = \max(X_1, \dots, X_n)$. Likelihood function

$$L(\theta) = \frac{1}{\theta^n} \text{ for } \theta \ge X_{(n)},$$

and $L(\theta) = 0$ otherwise. This yields MLE $\hat{\theta} = X_{(n)}$.

(c) Sampling distribution of the MLE $\hat{\theta} = X_{(n)}$:

$$\mathrm{P}(X_{(n)} \le x) = (\frac{x}{\theta})^n$$

with pdf

$$f_{\hat{\theta}}(x) = \frac{n}{\theta^n} \cdot x^{n-1}, \quad 0 \le x \le \theta.$$

The MLE is biased

$$\mathrm{E}(\hat{\theta}) = \frac{n}{n+1}\theta, \quad \mathrm{E}(\hat{\theta}^2) = \frac{n}{n+2}\theta^2, \quad \mathrm{Var}(\hat{\theta}) = \frac{\theta^2}{(n+1)^2(n+2)}.$$

Compare two mean square errors:

$$\mathrm{MSE}(\hat{\theta}) = \left(-\frac{\theta}{n+1}\right)^2 + \frac{\theta^2}{(n+1)^2(n+2)} = \frac{n+3}{n+2} \cdot \frac{\theta^2}{(n+1)^2},$$
$$\mathrm{MSE}(\tilde{\theta}) = \frac{\theta^2}{3n}.$$

(d) Corrected MLE $\hat{\theta}_c = \frac{n+1}{n} \cdot X_{(n)}$ becomes unbiased $\mathrm{E}(\hat{\theta}_c) = \theta$ with $\mathrm{Var}(\hat{\theta}_c) = \frac{\theta^2}{n^2(n+2)}$.

Problem 8.55

Genetic model: $p_1 = \frac{2+\theta}{4}$, $p_2 = \frac{1-\theta}{4}$, $p_3 = \frac{1-\theta}{4}$, $p_4 = \frac{\theta}{4}$, where $0 < \theta < 1$. In particular, if $\theta = 0.25$, then the genes are unlinked and the genotype frequencies are

	Green	White	Total
Starchy	$\frac{9}{16}$	$\frac{3}{16}$	$\frac{3}{4}$
Sugary	$\frac{3}{16}$	$\frac{1}{16}$	$\frac{1}{4}$
Total	$\frac{3}{4}$	$\frac{1}{4}$	1

(a) Sample counts $(X_1, X_2, X_3, X_4) \sim \text{Mn}(n, p_1, p_2, p_3, p_4)$ with $n = 3839$. Likelihood

$$L(\theta) = \binom{n}{x_1, x_2, x_3, x_4}(2+\theta)^{x_1}(1-\theta)^{x_2+x_3}\theta^{x_4}4^{-n}.$$

Putting

$$\frac{d}{d\theta}\ln L(\theta) = \frac{x_1}{2+\theta} - \frac{x_2+x_3}{1-\theta} + \frac{x_4}{\theta}$$

equal to zero, we solve the equation

$$\frac{x_1}{2+\theta} + \frac{x_4}{\theta} = \frac{x_2+x_3}{1-\theta}$$

or equivalently

$$\theta^2 n + \theta u - 2x_4 = 0,$$

where $u = 2x_2 + 2x_3 + x_4 - x_1$. We find the MLE to be

$$\hat{\theta} = \frac{-u + \sqrt{u^2 + 8nx_4}}{2n} = 0.0357.$$

Asymptotic variance

$$\text{Var}(\hat{\theta}) \approx \frac{1}{I(\theta)}, \quad I(\theta) = -\text{E}(\frac{d^2}{d\theta^2}\ln f(X_1, X_2, X_3, X_4|\theta)).$$

Since

$$\frac{d^2}{d\theta^2}\ln L(\theta) = -\frac{x_1}{(2+\theta)^2} - \frac{x_2+x_3}{(1-\theta)^2} - \frac{x_4}{\theta^2},$$

$$I(\theta) = \frac{n}{4(2+\theta)} + \frac{2n}{4(1-\theta)} + \frac{n}{4\theta} = \frac{n(1+2\theta)}{2\theta(2+\theta)(1-\theta)},$$

we get $I(\hat{\theta}) = 29345.8$, so that $s_{\hat{\theta}} = 0.0058$.

(b) $0.0357 \pm 1.96 \cdot 0.0058 = 0.0357 \pm 0.0114$

(c) Parametric bootstrap using Matlab:

```
p1=0.5089, p2=0.2411, p3=0.2411, p4=0.0089,
n=3839; B=1000; b=ones(B,1);
x1=binornd(n,p1,B,1);
x2=binornd(n*b-x1,p2/(1-p1));
x3=binornd(n*b-x1-x2,p3/(1-p1-p2));
x4=n*b-x1-x2-x3;
u=2*x2+2*x3+x4-x1;
t=(-u+sqrt(u.^2+8*n*x4))/(2*n);
std(t)
histfit(t)
```

gives std(t)=0.0058.

(d) Two ends of interval covering 95% of the components of the vector t produced by bootstrapping:

```
c1=prctile(t,2.5)
c2=prctile(t,97.5)
```

are c1=0.0250 and c2=0.0473, yielding a 95% CI for θ:

$$(2\hat{\theta} - c_2, 2\hat{\theta} - c_1) = (0.0241, 0.0464).$$

Problem 8.61

Laplace's rule of succession.

Binomial model $X \sim \text{Bin}(n, p)$. Conjugate prior $p \sim \text{Beta}(1, 1)$. Given $X = n$, the posterior becomes $p \sim \text{Beta}(n + 1, 1)$. Since the posterior mean is $\frac{n+1}{n+2}$, we get

$$\hat{p}_{\text{PME}} = \frac{n + 1}{n + 2}.$$