Solution 1

(a) We have

$$m(y) = m(\eta + \tau\varepsilon) = \eta + \tau m(\varepsilon), \quad s(y) = s(\eta + \tau\varepsilon) = \tau s(\varepsilon),$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, so

$$q_1(y, \eta) = \frac{m(y) - \eta}{s(y)} = \frac{\eta + \tau m(\varepsilon) - \eta}{\tau s(\varepsilon)} = \frac{m(\varepsilon)}{s(\varepsilon)}, \quad q_2(y, \tau) = s(y)/\tau = \tau s(\varepsilon)/\tau = s(\varepsilon),$$

so both $q_1(y, \eta)$ and $q_2(y, \tau)$ are functions of the data y and parameters that have known distributions, as those of $m(\varepsilon)/s(\varepsilon)$ and $s(\varepsilon)$ are both known (at least in principle). If $Q_1 = q_1(\varepsilon, 0)$ and $Q_2 = q_2(\varepsilon)$ have respective α quantiles $q_1'(\alpha, n)$ and $q_2'(\alpha, n)$, i.e., $\mathrm{P}\{Q_1 \le q_1'(\alpha, n)\} = \alpha$ and $\mathrm{P}\{Q_2 \le q_2'(\alpha, n)\} = \alpha$ for $\alpha \in (0, 1)$, then we can write

$$1 - 2\alpha = \mathrm{P}\{q_1'(\alpha, n) < Q_1 \le q_1'(1 - \alpha, n)\} = \mathrm{P}\left\{q_1'(\alpha, n) < \frac{m(Y) - \eta}{s(Y)} \le q_1'(1 - \alpha, n)\right\},$$

and rearrangement of the inequality in the right-hand probability shows that

$$L = m(Y) - s(Y)q_1'(1 - \alpha, n), \quad U = m(Y) - s(Y)q_1'(\alpha, n),$$

are the limits of a $(1 - 2\alpha)$ confidence interval for η. Likewise

$$1 - 2\alpha = \mathrm{P}\{q_2'(\alpha, n) < Q_2 \le q_2'(1 - \alpha, n)\} = \mathrm{P}\left\{q_2'(\alpha, n) < s(Y)/\tau \le q_2'(1 - \alpha, n)\right\},$$

and rearrangement of the inequality in the right-hand probability shows that

$$L = (s(Y)/q_2'(1 - \alpha, n), \quad U = s(Y)/q_2'(\alpha, n)$$

are the limits of a $(1 - 2\alpha)$ confidence interval for τ.

(b) Clearly $m_1(y) = m_1(\eta + \tau\varepsilon) = n^{-1}\sum_j(\eta + \tau\varepsilon_j) = \eta + \tau\overline{\varepsilon}$ and a similar calculation shows that $s_1(y) = \tau s_1(\varepsilon)$, and likewise for $m_2(y)$ and $s_2(y)$, leading to pivots.

Situation (i) corresponds to the t and χ^2 statistics used for inference on η and τ when $y_1, \ldots, y_n \overset{\text{iid}}{\sim} \mathcal{N}(\eta, \tau^2)$. Situation (ii) should give intervals that are highly robust to outliers.

(c) As $\{Y_+ - M(Y)\}/s(Y)$ is easily shown to be independent of the parameters, with a known distribution, it can be used to make prediction intervals for Y_+.

Solution 2

(a) According to Bayes' theorem

$$f(\theta \mid y_1) = \frac{f(y_1 \mid \theta)f(\theta)}{f(y_1)} \propto \theta^{y_1 + a - 1}(1 - \theta)^{n_1 - y_1 + b - 1}, \quad 0 < \theta < 1,$$

where the constant of proportionality ensures that the right-hand side has unit integral. Since the beta density has unit integral for any $a, b > 0$, and since $y_1 + a, n_1 - y_1 + b > 0$, the constant of proportionality must be obtained by replacing a and b by $y_1 + a$ and $n_1 - y_1 + b$, and thus must equal $\Gamma(n_1 + a + b)/\{\Gamma(y_1 + a)\Gamma(n_1 - y_1 + b)\}$. This gives the stated posterior density.

Note that this argument avoids any need for integration, and that the constants in the densities cancel from the numerator and denominator of the posterior.

(b) Here

$$f(n_2 \mid \theta)f(\theta) \propto \theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}, \quad 0<\theta<1,$$

and we see at once using the argument from (a) that

$$f(\theta \mid n_2) = \frac{\Gamma(n_2+a+b)}{\Gamma(y_2+a)\Gamma(n_2-y_2+b)}\theta^{y_2+a-1}(1-\theta)^{n_2-y_2+b-1}, \quad 0<\theta<1.$$

(c) The two posterior densities will be the same, so any Bayesian inferences based on the two experiments will be identical.

(d) In (a) the number of successes Y_1 will tend to be small if $\theta < \frac{1}{2}$, so the observed significance level is the binomial probability

$$\mathrm{P}_0(Y_1 \le y_1) = \sum_{y=0}^{y_1} \mathrm{P}_0(Y=y) = \sum_{y=0}^{y_1} \binom{n_1}{y} 2^{-n_1},$$

where P_0 denotes probability computed under the null hypothesis $\theta = \frac{1}{2}$. Similarly if $\theta < \frac{1}{2}$ then it will take longer to attain y_2 successes than if $\theta = \frac{1}{2}$, so we compute the negative binomial probability

$$\mathrm{P}_0(N_2 \ge n_2) = \sum_{n=n_2}^{\infty} \mathrm{P}_0(N_2 = n) = \sum_{n=n_2}^{\infty} \binom{n-1}{y_2-1} 2^{-n}.$$

The following R code shows the computation:

```
pbinom(x=3,size=12,prob=1/2)
[1] 0.07299805
nbinom(x=11,size=3,prob=1/2,lower.tail=F)
[1] 0.006469727
```

Unlike with (c), these suggest quite different evidence against the null hypothesis, because they sum probabilities over two different reference sets.

Solution 3

(a) We note that $\mathrm{E}(I_j)=0$ and $\mathrm{var}(I_j)=1$, so $\mathrm{E}(\overline{D})=\theta$, and as the I_j are independent,

$$\mathrm{var}(\overline{D}) = m^{-2}\sum_{j=1}^{m}\mathrm{var}(\theta + I_j c_j) = m^{-2}\sum_{j=1}^{m} c_j^2 \mathrm{var}(I_j) = m^{-2}\sum_{j=1}^{m} c_j^2 = \sigma^2.$$

The c_j are unknown and therefore so is σ^2, which must be estimated from the data $D_1, \dots, D_m$.

(b) To estimate σ^2, we use the problems for week 3 to write

$$S^2 = \frac{1}{m(m-1)}\sum_{j=1}^{m}(D_j - \overline{D})^2 = \frac{1}{2m^2(m-1)}\sum_{j,k=1}^{m}(D_j-D_k)^2 = \frac{1}{2m^2(m-1)}\sum_{j\neq k}^{m}(D_j-D_k)^2,$$

and note that as $D_j - D_k = I_j c_j - I_k c_k$, $\mathrm{E}(I_j)=0$ and $\mathrm{var}(I_j)=1$, and the I_j are independent, the right-most expression has expectation

$$\frac{2(m-1)}{2m^2(m-1)}\sum_{j=1}^{m} c_j^2 \mathrm{E}(I_j^2) - \frac{2}{2m^2(m-1)}\sum_{j\neq k} c_j c_k \mathrm{E}(I_j I_k) = \frac{1}{m^2}\sum_{j=1}^{m} c_j^2 = \sigma^2,$$

(c) To ease the notation, let $m = 2n$. Under this randomization scheme the number of possible allocations is $\binom{m}{n}$, which equals 252 when $m = 10$; this is appreciably lower than the number 1024 obtained before.

The expectations and variances of the I_j are the same as in (a), but if $j \neq k$ then by symmetry

$$\mathrm{cov}(I_j, I_k) = \mathrm{E}(I_j I_k) = 2\mathrm{P}(I_j = I_k) - 2\mathrm{P}(I_j \neq I_k) = 2\frac{n(n-1)}{2n(2n-1)} - 2\frac{n^2}{2n(2n-1)} = -\frac{1}{m-1}.$$

Under this randomisation scheme, $\sum_{j=1}^m I_j = 0$, so $\overline{D} = \theta + m^{-1}\sum_{j=1}^m I_j(c_j - \overline{c})$. Obviously $\mathrm{E}(\overline{D}) = \theta$ and

$$\begin{aligned}
m^2 \mathrm{var}(\overline{D}) &= \sum_{j=1}^m (c_j - \overline{c})^2 \mathrm{var}(I_j) + \sum_{i \neq j} (c_i - \overline{c})(c_j - \overline{c}) \mathrm{cov}(I_i, I_j) \\
&= \sum_{j=1}^m (c_j - \overline{c})^2 - \frac{1}{m-1} \sum_{j=1}^m (c_j - \overline{c}) \sum_{i \neq j} (c_i - \overline{c}) \\
&= \sum_{j=1}^m (c_j - \overline{c})^2 + \frac{1}{m-1} \sum_{j=1}^m (c_j - \overline{c})^2 \\
&= \frac{m}{m-1} \sum_{j=1}^m (c_j - \overline{c})^2,
\end{aligned}$$

where the step from the second to third lines used the fact that $\sum_{i=1}^m (c_i - \overline{c}) = 0$ implies that $\sum_{i \neq j} (c_i - \overline{c}) = -(c_j - \overline{c})$. Hence

$$\mathrm{var}(\overline{D}) = \frac{1}{m(m-1)} \sum_{j=1}^m (c_j - \overline{c})^2 = \tau^2,$$

say; note that subtracting $\overline{c}$ from the c_j will mean that it is very likely that $\tau^2 < \sigma^2$.

To find an estimator of the unknown τ^2 we write $\varepsilon_j = I_j c_j$ and note that

$$\sum_{j=1}^m (D_j - \overline{D})^2 = \sum_{j=1}^m (\varepsilon_j - \overline{\varepsilon})^2 = \sum_{j=1}^m \varepsilon_j^2 - m\overline{\varepsilon}^2 = \sum_{j=1}^m I_j^2 c_j^2 - \frac{1}{m} \sum_{i,j=1}^m I_j I_i c_j c_i$$

has expected value

$$\sum_{j=1}^m c_j^2 - \frac{1}{m}\left\{\sum_{j=1}^m c_j^2 + \sum_{i \neq j}^m c_i c_j \mathrm{cov}(I_i, I_j)\right\} = \sum_{j=1}^m c_j^2 - \frac{1}{m}\left\{\sum_{j=1}^m c_j^2 - \frac{1}{m-1}\sum_{j=1}^m c_j(m\overline{c} - c_j)\right\}$$

which equals

$$\sum_{j=1}^m (c_j - \overline{c})^2 - \frac{1}{m}\sum_{j=1}^m (c_j - \overline{c})^2 - \frac{1}{m(m-1)}\sum_{j=1}^m (c_j - \overline{c})^2 = \frac{m-2}{m-1}\sum_{j=1}^m (c_j - \overline{c})^2.$$

Hence τ^2 is estimated by

$$\frac{1}{m(m-2)} \sum_{j=1}^m (D_j - \overline{D})^2,$$

which can be computed from the observed differences $D_1, \ldots, D_m$.