

**Question 1 (a)**  $\hat{\theta}$  is an unbiased estimator of  $\theta$  if  $E(\hat{\theta}) = \theta$ .

We have  $E(Y_i) = \mu$  for all  $i = 1, \dots, n$ , hence

$$\begin{aligned} E(\hat{\mu}_1) &= \frac{1}{2}(\mu + \mu) = \mu \\ E(\hat{\mu}_2) &= \frac{1}{4}\mu + \frac{1}{2(n-2)} \times (n-2)\mu + \frac{1}{4}\mu = \mu \\ E(\hat{\mu}_3) &= \frac{1}{n} \times n\mu = \mu \end{aligned}$$

Hence, all three estimators are unbiased.

**(b)** The random variables  $Y_i$  are independent and  $\text{var}(Y_i) = \sigma^2$  for all  $i = 1, \dots, n$ , hence we have:

$$\begin{aligned} \text{var}(\hat{\mu}_1) &= \frac{1}{4}(\sigma^2 + \sigma^2) = \frac{1}{2}\sigma^2 \\ \text{var}(\hat{\mu}_2) &= \frac{1}{16}\sigma^2 + \frac{1}{4(n-2)^2} \times (n-2)\sigma^2 + \frac{1}{16}\sigma^2 \\ &= \frac{1}{8}\sigma^2 + \frac{1}{4(n-2)}\sigma^2 \\ &= \frac{n}{8(n-2)}\sigma^2 \\ \text{var}(\hat{\mu}_3) &= \frac{1}{n^2} \times n\sigma^2 = \frac{\sigma^2}{n} \end{aligned}$$

The estimators are unbiased, hence the condition for consistency  $\text{MSE}(\hat{\mu}) \rightarrow 0$  as  $n \rightarrow \infty$  simplifies to  $\text{var}(\hat{\mu}) \rightarrow 0$  as  $n \rightarrow \infty$ .

We can immediately see that  $\hat{\mu}_1$  is not consistent.

The variance of  $\hat{\mu}_2$  can be written as

$$\text{var}(\hat{\mu}_2) = \frac{1}{8}\sigma^2 + \frac{1}{4(n-2)}\sigma^2 \xrightarrow{n \rightarrow \infty} \frac{\sigma^2}{8}.$$

Hence,  $\hat{\mu}_2$  is not consistent.

The variance of  $\hat{\mu}_3$  tends to zero with  $n \rightarrow \infty$ , hence  $\hat{\mu}_3$  is consistent.

**(c)** The efficiency of  $\hat{\theta}$  is defined as

$$\text{eff}(\hat{\theta}) = \lim_{n \rightarrow \infty} \frac{\text{CRLB}(\theta)}{\text{var}(\hat{\theta})},$$

where

$$\text{CRLB}(\theta) = \frac{1}{\mathbb{E} \left\{ -\frac{d^2 \log f(\underline{Y}; \theta)}{d\theta^2} \right\}},$$

and  $f(\underline{Y}; \theta)$  is the joint pdf. Here we have

$$\begin{aligned} f(\underline{y}; \mu, \sigma^2) &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - \mu)^2}{2\sigma^2}} \\ &= \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{n}{2}} e^{-\sum \frac{(y_i - \mu)^2}{2\sigma^2}} \end{aligned}$$

Then

$$\begin{aligned} \log f &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum (y_i - \mu)^2 \\ \frac{d \log f}{d\mu} &= \frac{1}{\sigma^2} \sum (y_i - \mu) \\ \frac{d^2 \log f}{d\mu^2} &= -\frac{n}{\sigma^2} \end{aligned}$$

This gives

$$\text{CRLB}(\mu) = \frac{\sigma^2}{n}.$$

Hence we obtain:

$$\begin{aligned} \text{eff}(\hat{\mu}_1) &= \lim_{n \rightarrow \infty} \left\{ \frac{\sigma^2}{n} \frac{2}{\sigma^2} \right\} = 0 \\ \text{eff}(\hat{\mu}_2) &= \lim_{n \rightarrow \infty} \left\{ \frac{\sigma^2}{n} \frac{8(n-2)}{n\sigma^2} \right\} = \lim_{n \rightarrow \infty} \frac{8(n-2)}{n^2} = 0 \\ \text{eff}(\hat{\mu}_3) &= \lim_{n \rightarrow \infty} \left\{ \frac{\sigma^2}{n} \frac{n}{\sigma^2} \right\} = 1 \end{aligned}$$

An (asymptotically) unbiased estimator  $\hat{\theta}$  is efficient if  $\text{eff}(\hat{\theta}) = 1$ . Hence, out of the three, the only efficient estimator of  $\mu$  is  $\hat{\mu}_3 = \bar{Y}$ .

**Question 2 (a)** The likelihood is

$$\begin{aligned} L(\phi; \underline{y}) &= \prod_{i=1}^n \frac{2y_i}{\phi} e^{-\frac{y_i^2}{\phi}} \\ &= \frac{2^n}{\phi^n} e^{-\frac{1}{\phi} \sum_{i=1}^n y_i^2} \times \prod_{i=1}^n y_i. \end{aligned}$$

Hence, by Neyman's factorisation theorem,  $\sum_{i=1}^n Y_i^2$  is a sufficient statistic for  $\phi$ .

**(b)** The log-likelihood is

$$\ell(\phi; \underline{y}) = n \log 2 - n \log \phi + \sum_{i=1}^n \log y_i - \frac{1}{\phi} \sum_{i=1}^n y_i^2,$$

and so

$$\frac{d\ell}{d\phi} = -\frac{n}{\phi} + \frac{1}{\phi^2} \sum_{i=1}^n y_i^2 = 0 \Rightarrow n\hat{\phi} = \sum_{i=1}^n y_i^2 \Rightarrow \hat{\phi} = \frac{1}{n} \sum_{i=1}^n y_i^2.$$

Thus, the maximum likelihood estimator of  $\phi$  is  $\hat{\phi} = \sum_{i=1}^n Y_i^2/n$ , which is clearly a function of the sufficient statistic.

**Question 3 (a)** For the one-parameter exponential family, it should be possible to write the pdf as

$$f(y|\theta) = h(y) \exp\{a(\theta)b(y) + c(\theta)\}.$$

Here we have

$$\begin{aligned} f(y|\theta) &= \frac{\frac{1}{\theta}}{y^{\frac{1}{\theta}+1}} \\ &= \frac{1}{y} \exp\left\{\log \frac{1/\theta}{y^{1/\theta}}\right\} \\ &= \frac{1}{y} \exp\left\{-\frac{1}{\theta} \log y - \log \theta\right\} \end{aligned}$$

This in the form as above with

$$h(y) = \frac{1}{y}, \quad a(\theta) = -\frac{1}{\theta}, \quad b(y) = \log y, \quad c(\theta) = -\log \theta.$$

Hence, the distribution belongs to the exponential family.

- (b) By Lehmann's Theorem the complete sufficient statistic for  $\theta$  is  $S(\underline{Y}) = \sum_{i=1}^n b(Y_i) = \sum_{i=1}^n \log Y_i$ .
- (c) We need to calculate expectation of  $\log Y_i$ . It is the same for all  $i$ , so I drop index  $i$  in the derivation below.

$$E(\log Y) = \int_1^{\infty} \log y \frac{\frac{1}{\theta}}{y^{\frac{1}{\theta}+1}} dy.$$

The method by parts:

$$\begin{aligned} u &= \log y & du &= \frac{1}{y} dy \\ dv &= \frac{\frac{1}{\theta}}{y^{\frac{1}{\theta}+1}} dy & v &= -y^{-\frac{1}{\theta}} \end{aligned}$$

This gives

$$E(\log Y) = \left[ -\log y \frac{1}{y^{\frac{1}{\theta}}} \right]_1^{\infty} + \int_1^{\infty} \frac{1}{y} \frac{1}{y^{\frac{1}{\theta}}} dy = 0 + \left[ -\theta \frac{1}{y^{\frac{1}{\theta}}} \right]_1^{\infty} = \theta.$$

Hence,

$$E\left(\sum_{i=1}^n \log Y_i\right) = \sum_{i=1}^n E(\log Y_i) = n\theta.$$

This gives the unbiased estimator of  $\theta$  to be:

$$\frac{1}{n} \sum_{i=1}^n \log Y_i$$

which is a function of the complete sufficient statistic  $S(\underline{Y}) = \sum_{i=1}^n \log Y_i$ .

- (d) An unbiased estimator of  $\theta$  which is a function of a complete sufficient statistic is a MVUE( $\theta$ ).

**Question 4 (a)** We have

$$\hat{p}_i \sim \mathcal{AN}\left(p_i, \frac{p_i(1-p_i)}{n_i}\right), \quad i = 1, 2.$$

The difference  $\hat{p}_1 - \hat{p}_2$  is also asymptotically normally distributed, with mean and variance, respectively, equal to:

$$\begin{aligned} E(\hat{p}_1 - \hat{p}_2) &= p_1 - p_2, \\ \text{var}(\hat{p}_1 - \hat{p}_2) &= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}. \end{aligned}$$

Hence, we can write

$$\hat{p}_1 - \hat{p}_2 \sim \mathcal{AN}\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right).$$

After standardization we get

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{AN}(0, 1).$$

Hence, for large samples, the approximate pivot for  $p_1 - p_2$  is

$$Q(\underline{Y}, p_1 - p_2) = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

(b) For  $z_{\frac{\alpha}{2}}$  such that  $P(|Z| < z_{\frac{\alpha}{2}}) = 1 - \alpha$ ,  $Z \sim \mathcal{N}(0, 1)$ , we can write

$$P(-z_{\frac{\alpha}{2}} < Q(\underline{Y}, p_1 - p_2) < z_{\frac{\alpha}{2}}) \cong 1 - \alpha$$

that is

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} < z_{\frac{\alpha}{2}}\right) \cong 1 - \alpha.$$

Rearranging the expression within the brackets, we obtain

$$P\left(\hat{p}_1 - \hat{p}_2 - z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} < p_1 - p_2 < \hat{p}_1 - \hat{p}_2 + z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right) \cong 1 - \alpha$$

This gives the lower and upper limits of the approximate confidence interval for  $p_1 - p_2$  as:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

(c) Denote the random variables as follows:

$Y_i$  - output of the poll from voter  $i$  in campaign for the candidate in Ohio State. Then  $Y_i \sim \text{Bernoulli}(p_1)$ .

$X_i$  - output of the poll from voter  $i$  in campaign for the candidate in Texas. Then  $X_i \sim \text{Bernoulli}(p_2)$ .

The estimates of  $p_1$  and  $p_2$  are

$$\hat{p}_1 = \bar{y} = \frac{330}{560} = 0.5893$$
$$\hat{p}_2 = \bar{x} = \frac{290}{510} = 0.5686$$

This gives the estimate of the 95% confidence interval:

$$0.5893 - 0.5686 \pm 1.96\sqrt{0.000432 + 0.000481}$$
$$0.0207 \pm 1.96 \times 0.0302184$$
$$0.0207 \pm 0.059228$$

That is,  $[-0.03857, 0.07989]$ .

The interval includes zero, hence, at the significance level  $\alpha = 0.05$ , there is no evidence to reject the hypothesis  $H_0 : p_1 - p_2 = 0$  against  $H_1 : p_1 - p_2 \neq 0$ , that is there is no evidence to say that the candidate is more popular in one of the two states.

**Question 5 (a)** The likelihood is

$$\begin{aligned} L(p; \underline{y}) &= \prod_{i=1}^n \binom{20}{y_i} p^{y_i} (1-p)^{20-y_i} \\ &= \prod_{i=1}^n \binom{20}{y_i} p^{\sum_{i=1}^n y_i} (1-p)^{20n - \sum_{i=1}^n y_i}, \end{aligned}$$

and so the likelihood ratio is

$$\begin{aligned} \lambda(\underline{y}) = \frac{L(p_0; \underline{y})}{L(p_1; \underline{y})} &= \frac{\prod_{i=1}^n \binom{20}{y_i} p_0^{\sum_{i=1}^n y_i} (1-p_0)^{20n - \sum_{i=1}^n y_i}}{\prod_{i=1}^n \binom{20}{y_i} p_1^{\sum_{i=1}^n y_i} (1-p_1)^{20n - \sum_{i=1}^n y_i}} \\ &= \left\{ \frac{p_0(1-p_1)}{p_1(1-p_0)} \right\}^{\sum_{i=1}^n y_i} \left( \frac{1-p_0}{1-p_1} \right)^{20n}. \end{aligned}$$

(b) We have  $p_0 < p_1$  and  $1-p_1 < 1-p_0$ . Hence, we have  $p_0(1-p_1)/\{p_1(1-p_0)\} < 1$ , so that

$$\log \left\{ \frac{p_0(1-p_1)}{p_1(1-p_0)} \right\} < 0.$$

(c) Since the critical region is  $R = \{\underline{y} : \lambda(\underline{y}) \leq a\}$ , we reject  $H_0$  if

$$\left\{ \frac{p_0(1-p_1)}{p_1(1-p_0)} \right\}^{\sum_{i=1}^n y_i} \leq b,$$

where  $a$  and  $b$  are constants chosen to give significance level  $\alpha$ . Thus, we reject  $H_0$  if

$$\sum_{i=1}^n y_i \log \left\{ \frac{p_0(1-p_1)}{p_1(1-p_0)} \right\} \leq \log b \Rightarrow \sum_{i=1}^n y_i \geq c,$$

where  $c$  is a constant chosen to give significance level  $\alpha$ . It follows that the critical region is of the form  $R = \{\underline{y} : \bar{y} \geq d\}$ , where  $d$  is a constant chosen to give significance level  $\alpha$ . Now, by the central limit theorem, we have  $\bar{Y} \sim \mathcal{N}\{20p, 20p(1-p)/n\}$  approximately for large  $n$ . Thus, if  $H_0$  is true,

$$\frac{\bar{Y} - 20p_0}{\sqrt{\frac{20p_0(1-p_0)}{n}}} \sim \mathcal{N}(0, 1)$$

approximately for large  $n$ . So we reject  $H_0$  at the 1% level of significance if

$$\bar{y} \geq 20p_0 + z_{0.01} \sqrt{\frac{20p_0(1-p_0)}{n}} = 20p_0 + 2.3263 \sqrt{\frac{20p_0(1-p_0)}{n}}.$$

(d) Consider the alternative hypothesis  $H_1^* : p \in (p_0, 1]$ . Then, since the test of  $H_0$  against  $H_1 : p = p_1$  has the same critical region for all  $p_1 \in (p_0, 1]$ , a uniformly most powerful test exists and its critical region is  $R = \{\underline{y} : \bar{y} \geq 20p_0 + z_\alpha \sqrt{20p_0(1-p_0)/n}\}$ .