James K. Pringle
140.673 Theory
Dr. Constantine Frangakis
Assignment 5
April 18, 2014

Assignment 5

Asymptotic normality of the sample median and the delta method

Problem 1

For X_i, $i = 1, ..., n, ...$ i.i.d. Bernoulli trials with probability $\Pr(X_i = 1 \mid \theta) = \theta$, $(\theta \in (0, 1))$, show that the large sample distribution of $\arcsin\sqrt{\hat{\theta}}$, where $\hat{\theta}$ is the MLE of θ, has variance that is free of θ (such a transformation of $\hat{\theta}$ is called variance-stabilizing transformation).

Notes: (i) By large sample distribution of $g(\hat{\theta})$ here we really mean the approximation obtained from finding the asymptotic distribution of $\sqrt{n}(g(\hat{\theta}) - g(\theta))$; (ii) $\alpha = \arcsin\sqrt{\theta}$ means that $\sin(\alpha) = \sqrt{\theta}$; (iii) $\frac{d}{d\theta}\arcsin\sqrt{\theta} = \frac{1}{2\sqrt{\theta(1-\theta)}}$

Proof. The likelihood function is equal to

$$L_n(\underline{X}; \theta) = \prod_{i=1}^{n} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{n\overline{x}}(1-\theta)^{n-n\overline{x}}$$

Where $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Define

$$\ell_n(\underline{X}; \theta) = \log L_n(\underline{X}; \theta) = n\overline{x}\log(\theta) + (n - n\overline{x})\log(1-\theta)$$

Then

$$\frac{\partial \ell_n(\underline{X}; \theta)}{\partial \theta} = \frac{n\overline{x}}{\theta} - \frac{n - n\overline{x}}{1-\theta}$$

Solving for $\hat{\theta}$ that sets the derivative equal to 0,

$$\begin{aligned}
\frac{n\overline{x}}{\hat{\theta}} - \frac{n - n\overline{x}}{1-\hat{\theta}} &= 0 \\
\frac{n\overline{x}}{\hat{\theta}} &= \frac{n - n\overline{x}}{1-\hat{\theta}} \\
\frac{\overline{x}}{\hat{\theta}} &= \frac{1-\overline{x}}{1-\hat{\theta}} \\
\overline{x}(1-\hat{\theta}) &= (1-\overline{x})\hat{\theta} \\
\overline{x} &= \hat{\theta}
\end{aligned}$$

Notice that the second derivative is negative for all values of θ, hence $\hat{\theta}$ is a maximum. We conclude that

$$\hat{\theta} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is the MLE. By the Central Limit Theorem, it follows that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{D} N(0, \theta(1-\theta))$$

Now, define the function and its derivative

$$\begin{aligned}
g(t) &= \arcsin\sqrt{t} \\
\frac{dg(t)}{dt} &= \frac{1}{2\sqrt{t(1-t)}}
\end{aligned}$$

Then by the Delta Method, we know that

$$\begin{aligned}
\sqrt{n}\left(g(\hat{\theta}) - g\left(\theta\right)\right) &\xrightarrow{D} N\left(0, \left(\frac{1}{2\sqrt{\theta(1-\theta)}}\right)^2 \theta(1-\theta)\right) \\
\sqrt{n}\left(g(\hat{\theta}) - g\left(\theta\right)\right) &\xrightarrow{D} N\left(0, \frac{1}{4}\right)
\end{aligned}$$

Therefore the asymptotic variance of is free of θ, as desired. □

Problem 2

Suppose that we are interested in the true median m_0 of the distribution of life times from a population of cells. Assume the distribution is absolutely continuous and that we are about to measure $X_1, \cdots, X_n$ iid measurements from that distribution. Here, we do not want to believe necessarily on a finite parametric likelihood model for X_i, and so we will estimate m_0 with the median, say M_n, of the n observations (you can assume n is odd).

We wish to derive the large sample distribution of $\sqrt{n}(M_n - m_0)$. To do so, you need to assume that a more general version of the central limit theorem (CLT) holds here. Assuming that version of CLT holds, show that the large sample distribution of $\sqrt{n}(M_n - m_0)$ is normal with mean 0 and variance $1/(4f^2(m_0))$, where $f()$ is the true density of the distribution. State what is that version of CLT you need.

Hint: Express the cumulative probability $\text{pr}(M_n < t)$ as a probability of the event that a certain binomial random variable is at a tail. Then, state what approximations to that event are needed to arrive at the result.

Proof. Note that $M_n = X_{(\frac{n+1}{2})}$, the $(n+1)/2$-th order statistic for the sample of size n (assumed odd). To find the distribution of $\sqrt{n}(M_n - m_0)$, we calculate for some $a \in \mathbb{R}$,

$$\text{pr}(\sqrt{n}(M_n - m_0) \leq a) = \text{pr}\left(M_n \leq \frac{a}{\sqrt{n}} + m_0\right)$$

We seek to apply the Central Limit Theorem for Triangular Arrays (see Chung 7.1). Considering row n, define

$$Y_{ni} = I_{\left\{X_i \leq \frac{a}{\sqrt{n}} + m_0\right\}}.$$

Thus, for fixed n, the Y_{ni} are iid with distribution $Bernoulli(p_n)$, where $p_n = F(\frac{a}{\sqrt{n}} + m_0)$ and $F(\cdot)$ is the distribution of the X_i. Since the entries of each row must satisfy certain normalization properties, let entry i of row n, where $i \in \{1, \cdots, n\}$, be

$$Z_{ni} = \frac{Y_{ni} - E[Y_{ni}]}{\sqrt{n \cdot \text{var}(Y_{ni})}} = \frac{Y_{ni} - p_n}{\sqrt{np_n(1-p_n)}}$$

Define $S_n = \sum_{i=1}^n Z_{ni}$. Therefore $E[Z_{ni}] = 0$ and

$$\text{var}(S_n) = \text{var}\left(\sum_{i=1}^n Z_{ni}\right) = \sum_{i=1}^n \text{var}(Z_{ni}) = \sum_{i=1}^n \frac{\text{var}(Y_{ni})}{n \cdot \text{var}(Y_{ni})} = 1$$

by indepdence. Examining the third moments,

$$\begin{aligned}
\gamma_{ni} &= E[|Z_{ni}|^3] \\
&= (n \cdot \text{var}(Y_{ni}))^{-3/2} E[|Y_{ni} - E[Y_{ni}]|^3] \\
&\leq (n \cdot \text{var}(Y_{ni}))^{-3/2}.
\end{aligned}$$

The sum of these absolute third moments is

$$0 \leq \Gamma_n = \sum_{i=1}^n \gamma_{ni} \leq n(n \cdot \text{var}(Y_{ni}))^{-3/2} = \frac{1}{\sqrt{n}} \left(\frac{1}{\text{var}(Y_{ni})}\right)^{3/2}.$$

Calculating,

$$\begin{aligned}
\lim_{n\to\infty} \operatorname{var}(Y_{ni}) &= \lim_{n\to\infty} p_n(1-p_n) \\
&= \lim_{n\to\infty} F\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right) \\
&= F\left(\lim_{n\to\infty} \frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\lim_{n\to\infty} \frac{a}{\sqrt{n}} + m_0\right)\right) \quad (*) \\
&= F(m_0)(1 - F(m_0)) \\
&= \frac{1}{2}\left(1 - \frac{1}{2}\right) \\
&= \frac{1}{4}
\end{aligned}$$

where $(*)$ follows from continuity because $F(1-F)$ is a composition of continuous functions (since F is absolutely continuous by assumption). Again, by composition of continuous functions

$$\lim_{n\to\infty} \left(\frac{1}{\operatorname{var}(Y_{ni})}\right)^{3/2} = 8$$

Therefore,

$$0 \leq \lim_{n\to\infty} \Gamma_n \leq \lim_{n\to\infty} \frac{1}{\sqrt{n}} \left(\frac{1}{\operatorname{var}(Y_{ni})}\right)^{3/2} = 0$$

and we see $\Gamma_n \to 0$. Hence, by Chung Theorem 7.1.2, S_n converges in distribution to Φ, the distribution of a standard normal random variable.

Note that the sample median of n (odd) X_i's is less than a number if and only if $(n+1)/2$ or more of the X_i's are less than that number. Hence

$$\begin{aligned}
\operatorname{pr}\left(M_n \leq \frac{a}{\sqrt{n}} + m_0\right) &= \operatorname{pr}\left(\sum_{i=1}^{n} Y_{ni} \geq \frac{n+1}{2}\right) \\
&= \operatorname{pr}\left(\sum_{i=1}^{n} \frac{Y_{ni} - p_n}{\sqrt{np_n(1-p_n)}} \geq \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) \\
&= \operatorname{pr}\left(S_n \geq \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right)
\end{aligned}$$

Since S_n converges in distribution to Φ, for all $b \in \mathbb{R}$

$$\lim_{n\to\infty} \operatorname{pr}\left(S_n \geq b\right) = 1 - \Phi(b)$$

Furthermore, by Chung 7.4.1, there is a universal constant A_0 such that

$$\sup_x |F_n(x) - \Phi(x)| \leq A_0 \Gamma_n$$

where F_n is the distribution function of S_n. Taking the limit,

$$0 \leq \lim_{n\to\infty} \sup_x |F_n(x) - \Phi(x)| \leq \lim_{n\to\infty} A_0\Gamma_n = 0$$

By squeezing, $\sup_x |F_n(x) - \Phi(x)| \to 0$. This implies that

$$\sup_x |\mu_n([x,\infty)) - (1 - \Phi(x))| \to 0$$

where μ_n is the measure induced by F_n. Therefore,

$$\left|\text{pr}\left(S_n \geq \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) - \left(1 - \Phi\left(\frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right)\right)\right| \leq \sup_x |\mu_n([x,\infty)) - (1-\Phi(x))|. \tag{1}$$

Since the right-hand side tends to 0, the left-hand side also tends to 0.

Next we find the limit of

$$\frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}.$$

We assume that the distribution function is differentiable in a neighborhood of m_0. Then by the Mean Value Theorem,

$$F\left(\frac{a}{\sqrt{n}} + m_0\right) = F(m_0) + f(m^*)\left(\frac{a}{\sqrt{n}} + m_0 - m_0\right) = \frac{1}{2} + f(m^*)\left(\frac{a}{\sqrt{n}}\right)$$

where $m_0 \leq m^* \leq m + \frac{a}{\sqrt{n}}$. Therefore,

$$\begin{aligned}
\frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}} &= \frac{\frac{n+1}{2} - nF\left(\frac{a}{\sqrt{n}} + m_0\right)}{\sqrt{nF\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}} \\
&= \frac{\frac{n}{2} + \frac{1}{2} - n\left(\frac{1}{2} + f(m^*)\left(\frac{a}{\sqrt{n}}\right)\right)}{\sqrt{nF\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}} \\
&= \frac{\frac{1}{2}}{\sqrt{nF\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}} - \frac{f(m^*)a\sqrt{n}}{\sqrt{nF\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}} \\
&= \frac{1}{\sqrt{4nF\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}} - \frac{f(m^*)a}{\sqrt{F\left(\frac{a}{\sqrt{n}} + m_0\right)\left(1 - F\left(\frac{a}{\sqrt{n}} + m_0\right)\right)}}
\end{aligned}$$

By the argument from before about the continuity of $F(1-F)$, we have

$$\lim_{n\to\infty} \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}} = 0 - \frac{f(m_0)a}{\sqrt{\frac{1}{4}}} = -2f(m_0)a$$

since by squeezing, $\lim_{n\to\infty} f(m^*) = f(m_0)$.

It follows by the continuity of Φ that

$$\lim_{n\to\infty} \Phi\left(\frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) = \Phi(-2f(m_0)a) = 1 - \Phi(2f(m_0)a) \tag{2}$$

Now for simplicity, let

$$h_n = \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}$$

Then calculating,

$$\begin{aligned}
|\text{pr}(S_n \geq h_n) - \Phi(2f(m_0)a)| &= |\text{pr}(S_n \geq h_n) - (1 - \Phi(-2f(m_0)a))| \\
&= |\text{pr}(S_n \geq h_n) - (1 - \Phi(h_n)) + (1 - \Phi(h_n)) - (1 - \Phi(-2f(m_0)a))| \\
&\leq |\text{pr}(S_n \geq h_n) - (1 - \Phi(h_n))| + |(1 - \Phi(h_n)) - (1 - \Phi(-2f(m_0)a))| \\
&= |\text{pr}(S_n \geq h_n) - (1 - \Phi(h_n))| + | - \Phi(h_n) + \Phi(-2f(m_0)a))|
\end{aligned}$$

by the triangle inequality. Since both terms in the absolute value signs tend to zero, as shown above at (1) and (2), we have that

$$\lim_{n\to\infty} |\text{pr}(S_n \geq h_n) - \Phi(2f(m_0)a)| = 0$$

Since $\text{pr}(S_n \geq h_n) = \text{pr}(\sqrt{n}(M_n - m_0) \leq a)$ it follows that for all $a \in \mathbb{R}$

$$\text{pr}(\sqrt{n}(M_n - m_0) \leq a) \to \Phi(2f(m_0)a) = \text{pr}\left(\frac{Z}{2f(m_0)} \leq a\right)$$

where Z is a standard normal random variable. This means that

$$\sqrt{n}(M_n - m_0) \sim N\left(0, \frac{1}{4f^2(m_0)}\right)$$

□

Problem 3

Apnea is a condition with which a person cannot breathe normally during sleep. To estimate the prevalence of apnea in children, we conduct the following study. First, we obtain a simple random sample of $n = 5000$ children from all children, and, among the 5000 children, we find out that $N_s = 212$ children snore. Assume a child who does not snore cannot have apnea, so we focus on those who snore. Among those who do snore, we cannot follow all children, so we follow a subset $N_f = 80$ that we get as a simple random sample from the N_s

children. We then study in detail the night sleep pattern of the N_f children and find that $N_a = 60$ of them have apnea. Treat the original sample size n as known (and fixed unless told otherwise), and the other observed data $N = (N_s, N_f, N_a)$ as realizations of random variables. Denote by θ all known or unknown parameters in the problem, and, based on the above design, assume the following:

(1) $\text{pr}(N_s|n, \theta)$ is a Binomial(n, π_s) distribution, where π_s is the probability that a child snores.

(2) $\text{pr}(N_f|N_s, \theta)$ is a Binomial(N_s, π_f) distribution, where π_f is the probability of a child being followed further for detailed study, given that the child snored.

(3) $\text{pr}(N_a|N_f, N_s, \theta) = \text{pr}(N_a|N_f, \theta)$ is a Binomial(N_f, π_a) distribution, where π_a is the probability of a child having apnea given that the child was followed for detailed study.

Address the following

(i) Show that the likelihood of the data, $L(\theta, \underline{N})$ is

$$\pi_s^{N_s}(1-\pi_s)^{(n-N_s)} \times \pi_f^{N_f}(1-\pi_f)^{(N_s-N_f)} \times \pi_a^{N_a}(1-\pi_a)^{(N_f-N_a)}.$$

Proof. Define

$$Y_i^s = \begin{cases} 1 & \text{if person } i \text{ snores} \\ 0 & \text{otherwise,} \end{cases}$$

define

$$Y_i^f = \begin{cases} 1 & \text{if person } i \text{ is selected for the study} \\ 0 & \text{otherwise,} \end{cases}$$

and define

$$Y_i^a = \begin{cases} 1 & \text{if person } i \text{ has apnea} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$\begin{aligned} Y_i^s &\sim Bernoulli(\pi_s) \\ Y_i^f \mid Y_i^s &\sim Bernoulli(\pi_f) \\ Y_i^a \mid Y_i^f, Y_i^s &\sim Bernoulli(\pi_a) \end{aligned}$$

Therefore,

$$\begin{aligned} L(\theta|\underline{N}) &= \Pr(\underline{Y^a}|\theta, \underline{Y^f}, \underline{Y^s}) \Pr(\underline{Y^f}|\theta, \underline{Y^s}) \Pr(\underline{Y^s}|\theta) \\ L(\theta|\underline{N}) &= \prod_{i=1}^{N_f} \pi_a^{Y_i^a}(1-\pi_a)^{1-Y_i^a} \prod_{k=1}^{N_s} \pi_f^{Y_i^f}(1-\pi_f)^{1-Y_i^f} \prod_{j=1}^{n} \pi_s^{Y_j^s}(1-\pi_s)^{1-Y_i^s} \\ &= \left[\pi_a^{N_a}(1-\pi_a)^{N_f-N_a}\right] \pi_f^{N_f}(1-\pi_f)^{N_s-N_f} \left[\pi_s^{N_s}(1-\pi_s)^{n-N_s}\right] \end{aligned}$$

as desired. □

(ii) Find the scores for π_s and π_a, say S_s and S_a, respectively. Hence, find the maximum likelihood estimators for π_s and π_a, assuming the study's researchers tell you π_f. Would the MLEs for π_s and π_a change if you did not know π_f? In the remaining part of the problem, continue to assume you know π_f.

Proof. Define

$$\ell(\theta, \underline{N}) = \log L(\theta, \underline{N}) = C + N_s \log(\pi_s) + (n - N_s)\log(1-\pi_s) + N_a \log(\pi_a) + (N_f - N_a)\log(1-\pi_a)$$

where $C = N_f \log(\pi_f) + (N_s - N_f)\log(1-\pi_f)$, which does not depend on π_s or π_a. Thus,

$$\frac{\partial \ell(\theta, \underline{N})}{\partial \pi_s} = \frac{N_s}{\pi_s} - \frac{n - N_s}{1 - \pi_s}$$
$$\frac{\partial \ell(\theta, \underline{N})}{\partial \pi_a} = \frac{N_a}{\pi_a} - \frac{N_f - N_a}{1 - \pi_a}$$

which implies that the MLEs are the values that set the scores equal to 0, i.e.

$$\hat{\pi_s} = \frac{N_s}{n}$$
$$\hat{\pi_a} = \frac{N_a}{N_f}$$

It is easy to see that second derivatives of the scores are negative, so $\hat{\pi_s}$ and $\hat{\pi_a}$ maximize the likelihood function. Since the scores S_s and S_a do not depend on π_f, the MLEs would not change if we did not know π_f. □

(iii) Find the unconditional expectations $E(N_s|\theta)$, $E(N_f|\theta)$, and $E(N_a|\theta)$ [Hint: for the latter two, use conditional expectations first].

Proof. By assumption (1),

$$E(N_s|\theta) = n\pi_s$$

By assumption (2),

$$E(N_f|N_s, \theta) = N_s\pi_f,$$

so by iterated expectation,

$$E(N_f|\theta) = E(E(N_f|N_s, \theta)|\theta) = E(N_s\pi_f|\theta) = \pi_f E(N_s|\theta) = n\pi_s\pi_f$$

By assumption (3),

$$E(N_a|N_f, N_s, \theta) = N_f\pi_a$$

so by iterated expectation,

$$\begin{aligned} E(N_a|\theta) &= E(E(E(N_a|N_f, N_s, \theta)|N_s, \theta)|\theta) \\ &= E(E(N_f\pi_a|N_s, \theta)|\theta) \\ &= E(N_s\pi_f\pi_a|\theta) \\ &= n\pi_s\pi_f\pi_a \end{aligned}$$

□

(iv) Find the three second-order derivatives of the log-likelihood with respect to (π_s, π_a). Hence, as n increases, find the large sample joint distribution of the vector $[\sqrt{n}(\hat{\pi}_a - \pi_s), \sqrt{n}(\hat{\pi}_a - \pi_a)]$, over hypothetical replications of the study. [Note: the shape of the large sample distribution is, theoretically, in terms of the parameters π_s, π_f, and π_a only, not in terms of the data.]

Proof. Calculating,

$$\begin{aligned} \frac{\partial^2\ell(\theta, \underline{N})}{\partial\pi_s^2} &= -\frac{N_s}{\pi_s^2} - \frac{n - N_s}{(1-\pi_s)^2} \\ \frac{\partial^2\ell(\theta, \underline{N})}{\partial\pi_a^2} &= -\frac{N_a}{\pi_a^2} - \frac{N_f - N_a}{(1-\pi_a)^2} \\ \frac{\partial^2\ell(\theta, \underline{N})}{\partial\pi_a\partial\pi_s} &= \frac{\partial^2\ell(\theta, \underline{N})}{\partial\pi_s\partial\pi_a} = 0 \end{aligned}$$

The information based on n data points is

$$\begin{aligned} I_n(\theta) &= -E\begin{bmatrix} \frac{\partial^2\ell(\theta,\underline{N})}{\partial\pi_s^2} & \frac{\partial^2\ell(\theta,\underline{N})}{\partial\pi_a\partial\pi_s} \\ \frac{\partial^2\ell(\theta,\underline{N})}{\partial\pi_s\partial\pi_a} & \frac{\partial^2\ell(\theta,\underline{N})}{\partial\pi_a^2} \end{bmatrix} \\ &= -E\begin{bmatrix} -\frac{N_s}{\pi_s^2} - \frac{n-N_s}{(1-\pi_s)^2} & 0 \\ 0 & -\frac{N_a}{\pi_a^2} - \frac{N_f-N_a}{(1-\pi_a)^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n\pi_s}{\pi_s^2} + \frac{n-n\pi_s}{(1-\pi_s)^2} & 0 \\ 0 & \frac{n\pi_s\pi_f\pi_a}{\pi_a^2} + \frac{n\pi_s\pi_f - n\pi_s\pi_f\pi_a}{(1-\pi_a)^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n}{\pi_s} + \frac{n}{(1-\pi_s)} & 0 \\ 0 & \frac{n\pi_s\pi_f}{\pi_a} + \frac{n\pi_s\pi_f}{(1-\pi_a)} \end{bmatrix} \\ &= \begin{bmatrix} \frac{n}{\pi_s(1-\pi_s)} & 0 \\ 0 & \frac{n\pi_s\pi_f}{\pi_a(1-\pi_a)} \end{bmatrix} \end{aligned}$$

So the information based on one data point is

$$I_1(\theta) = I_n(\theta)/n = \begin{bmatrix} \frac{1}{\pi_s(1-\pi_s)} & 0 \\ 0 & \frac{\pi_s\pi_f}{\pi_a(1-\pi_a)} \end{bmatrix}$$

Thus,

$$\sqrt{n}\begin{bmatrix}\hat{\pi}_a - \pi_s \\ \hat{\pi}_a - \pi_a\end{bmatrix} \xrightarrow{D} N(0, I_1(\theta)^{-1}) = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\pi_s(1-\pi_s) & 0 \\ 0 & \frac{\pi_a(1-\pi_a)}{\pi_s \pi_f}\end{bmatrix}\right)$$

□

(v) Explain in no more than two sentences why, with the above design, the prevalence of apnea, say p_a, in the population, equals $\pi_s\pi_a$. Hence report the MLE of p_a using your data.

Proof. The prevalence of apnea is the number of people with apnea divided by the total population in consideration. That is equal to the number of snorers divided by the total population (π_s) times the number of people with apnea divided by the number of snorers (π_a).

Alternatively, a description in math terms is

$$\begin{aligned} p_a &= P(Apnea|Snores)P(Snores) + P(Apnea|Snores^c)P(Snores^c) \\ &= \pi_a\pi_s + 0(1-\pi_s) \\ &= \pi_a\pi_s \end{aligned}$$

By the invariance property of the MLE, the MLE for p_a is

$$\hat{p_a} = \hat{\pi}_s\hat{\pi}_a = (N_sN_a)/(nN_f).$$

□

(vi) By approximating, as n increases, the large sample distribution of

$$\sqrt{n}\{\text{logit}(\hat{\pi}_s\hat{\pi}_a) - \text{logit}(\pi_s\pi_a)\},$$

using the delta method (Taylor's expansion), report, using your data, an approximate 95% CI for the prevalence of apnea p_a. [the function logit means: $\text{logit}(x) = \log(x/(1-x))$.]

Proof. Let $g(x,y) = \log(xy/(1-xy))$. Then

$$\begin{aligned} \frac{\partial g}{\partial x} &= \left(y\frac{1}{1-xy} + xy\frac{y}{(1-xy)^2}\right)\frac{1-xy}{xy} \\ &= \left(\frac{y-xy^2}{(1-xy)^2} + \frac{xy^2}{(1-xy)^2}\right)\frac{1-xy}{xy} \\ &= \frac{1}{x(1-xy)} \end{aligned}$$

and by symmetry,

$$\frac{\partial g}{\partial x} = \frac{1}{y(1-xy)}$$

Therefore, by the delta method

$$\sqrt{n}\{\text{logit}(\hat{\pi}_s\hat{\pi}_a) - \text{logit}(\pi_s\pi_a)\} \xrightarrow{D} N(0, \Sigma)$$

where

$$\begin{aligned}
\Sigma &= \nabla g(\pi_s, \pi_a)' I_1(\theta)^{-1} \nabla g(\pi_s, \pi_a) \\
&= \begin{bmatrix} \frac{1}{\pi_s(1-\pi_s\pi_a)} & \frac{1}{\pi_a(1-\pi_s\pi_a)} \end{bmatrix} \begin{bmatrix} \pi_s(1-\pi_s) & 0 \\ 0 & \frac{\pi_a(1-\pi_a)}{\pi_s\pi_f} \end{bmatrix} \begin{bmatrix} \frac{1}{\pi_s(1-\pi_s\pi_a)} \\ \frac{1}{\pi_a(1-\pi_s\pi_a)} \end{bmatrix} \\
&= \frac{1-\pi_s}{\pi_s(1-\pi_s\pi_a)^2} + \frac{1-\pi_a}{\pi_a\pi_s\pi_f(1-\pi_s\pi_a)^2}
\end{aligned}$$

Notice that by composition of continuous functions,

$$\hat{\Sigma} = \frac{1-\hat{\pi}_s}{\hat{\pi}_s(1-\hat{\pi}_s\hat{\pi}_a)^2} + \frac{1-\hat{\pi}_a}{\hat{\pi}_a\hat{\pi}_s\hat{\pi}_f(1-\hat{\pi}_s\hat{\pi}_a)^2} \xrightarrow{P} \frac{1-\pi_s}{\pi_s(1-\pi_s\pi_a)^2} + \frac{1-\pi_a}{\pi_a\pi_s\pi_f(1-\pi_s\pi_a)^2}$$

By Slutsky's theorem,

$$\hat{\Sigma}^{-1/2}\sqrt{n}\{\text{logit}(\hat{\pi}_s\hat{\pi}_a) - \text{logit}(\pi_s\pi_a)\} \xrightarrow{D} N(0,1)$$

Therefore, a 95% CI for $\text{logit}(\pi_s\pi_a)$ is

$$\left[-1.96\sqrt{\hat{\Sigma}/n} + \text{logit}(\hat{\pi}_s\hat{\pi}_a), 1.96\sqrt{\hat{\Sigma}/n} + \text{logit}(\hat{\pi}_s\hat{\pi}_a)\right]$$

Taking the inverse logit (the expit, which is an increasing function), a 95% confidence interval for $\pi_s\pi_a$ is

$$\begin{aligned}
&\text{expit}\left(\left[-1.96\sqrt{\hat{\Sigma}/n} + \text{logit}(\hat{\pi}_s\hat{\pi}_a), 1.96\sqrt{\hat{\Sigma}/n} + \text{logit}(\hat{\pi}_s\hat{\pi}_a)\right]\right) \\
&= \left[\left(\exp\left\{-1.96\sqrt{\hat{\Sigma}/n}\right\}\frac{\hat{\pi}_s\hat{\pi}_a}{1-\hat{\pi}_s\hat{\pi}_a}\right) / \left(1+\exp\left\{-1.96\sqrt{\hat{\Sigma}/n}\right\}\frac{\hat{\pi}_s\hat{\pi}_a}{1-\hat{\pi}_s\hat{\pi}_a}\right),\right. \\
&\qquad \left.\left(\exp\left\{1.96\sqrt{\hat{\Sigma}/n}\right\}\frac{\hat{\pi}_s\hat{\pi}_a}{1-\hat{\pi}_s\hat{\pi}_a}\right) / \left(1+\exp\left\{1.96\sqrt{\hat{\Sigma}/n}\right\}\frac{\hat{\pi}_s\hat{\pi}_a}{1-\hat{\pi}_s\hat{\pi}_a}\right)\right] \\
&= \left[\frac{\hat{\pi}_s\hat{\pi}_a}{\exp\left\{1.96\sqrt{\hat{\Sigma}/n}\right\}(1-\hat{\pi}_s\hat{\pi}_a) + \hat{\pi}_s\hat{\pi}_a}, \frac{\hat{\pi}_s\hat{\pi}_a}{\exp\left\{-1.96\sqrt{\hat{\Sigma}/n}\right\}(1-\hat{\pi}_s\hat{\pi}_a) + \hat{\pi}_s\hat{\pi}_a}\right]
\end{aligned}$$

□