Fall 2009

1. Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter $p \in (0,1)$, that is,

$$P(X_i = 1) = p, \qquad \text{and} \qquad P(X_i = 0) = 1 - p.$$

a) Determine the UMVUE of

$$q(p) = p(1-p).$$

Solution. The joint density is

$$f(\boldsymbol{x}; p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum x_i}(1-p)^{n-\sum x_i},$$

and so, by the Factorization theorem, $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is sufficient. We can also write the joint density as

$$f(\boldsymbol{x}; p) = (1-p)^n \left(\frac{p}{1-p}\right)^{\sum x_i} = (1-p)^n \exp\left(\sum_{i=1}^{n} x_i \log\left(\frac{p}{1-p}\right)\right),$$

and since $\{\log p - \log(1-p) : p \in (0,1)\}$ contains an open set in $\mathbb{R}$, we have that $T(\boldsymbol{X}) = \sum_{i=1}^{n} X_i$ is a complete and sufficient statistic. Note that $\mathbf{1}(X_1 = 1, X_2 = 0)$ is an unbiased estimator of $p(1-p)$. Also,

$$\begin{aligned} P\left(X_1 = 1, X_2 = 0 \,\middle|\, \sum_{i=1}^{n} X_i = k\right) &= \frac{P(X_1 = 1, X_2 = 0, \sum_{i=3}^{n} X_i = k-1)}{P(\sum_{i=1}^{n} X_i = k)} \\ &= \frac{p(1-p)\binom{n-2}{k-1}p^{k-1}(1-p)^{n-k-1}}{\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{\binom{n-2}{k-1}p^k(1-p)^{n-k}}{\binom{n}{k}p^k(1-p)^{n-k}} \\ &= \frac{\binom{n-2}{k-1}}{\binom{n}{k}} \qquad \text{(yay sufficiency!)} \\ &= \frac{k(n-k)}{n(n-1)}, \end{aligned}$$

and so,

$$\mathbb{E}\left(\mathbf{1}(X_1 = 1, X_2 = 0) \,\middle|\, \sum_{i=1}^{n} X_i\right) = P\left(X_1 = 1, X_2 = 0 \,\middle|\, \sum_{i=1}^{n} X_i\right) = \frac{\sum X_i(n - \sum X_i)}{n(n-1)}.$$

Hence,

$$\boxed{\frac{n}{n-1}\frac{\sum X_i}{n}\left(1 - \frac{\sum X_i}{n}\right)}$$

is the UMVUE for $p(1-p)$.

b) Prove that the odds ratio

$$q(p) = \frac{p}{1-p}$$

is not unbiasedly estimatable.

Solution. Suppose that $S(\boldsymbol{X})$ is an unbiased estimator of $q(p)$. Then,

$$S^*(\boldsymbol{X}) := \mathbb{E}\left[S(\boldsymbol{X}) \mid T(\boldsymbol{X})\right]$$

is a UMVUE of $q(p)$, and so, $S^*(\boldsymbol{X})$ can be written as a function of $T(\boldsymbol{X})$, say $S^*(\boldsymbol{X}) = g(T(\boldsymbol{X}))$. Then,

$$\frac{p}{1-p} = \mathbb{E}S^*(\boldsymbol{X}) = \sum_{k=0}^{n} g(k)\binom{n}{k}p^k(1-p)^{n-k}.$$

Note that the RHS is a polynomial in p of degreen at most n, while $p/(1-p)$ cannot be written as a polynomial, which is a contradiction. Hence, no unbiased estimator of $p/(1-p)$ exists. □

c) Determine necessary and sufficient conditions on $q(p)$ such that the UMVUE of $q(p)$ exists.

Solution. If $q(p)$ is a polynomial in p of degree at most n, then UMVUE of $q(p)$ exists. Is this condition sufficient?...

2. Let $Y_1, \dots, Y_n$ be independent with distribution $Y_i \sim \mathcal{N}(\theta_0 + \theta_1 x_i, 1)$, where $x_1, \dots, x_n$ are known real numbers.

   a) Determine the maximum likelihood estimator $\left(\hat{\theta}_0, \hat{\theta}_1\right)$ of (θ_0, θ_1).

      Solution. The likelihood function is

$$\mathcal{L}(\theta_0, \theta_1; \boldsymbol{y}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y_i - \theta_0 - \theta_1 x_i)^2}{2}\right) = (2\pi)^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2\right),$$

      and so, the log-likelihood function is

$$\log \mathcal{L}(\theta_0, \theta_1; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2.$$

      Now, the first partial derivatives are

$$\frac{\partial}{\partial \theta_0}\log \mathcal{L} = \sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i) = \sum_{i=1}^{n} y_i - n\theta_0 - \theta_1 \sum_{i=1}^{n} x_i, \text{ and}$$
$$\frac{\partial}{\partial \theta_1}\log \mathcal{L} = \sum_{i=1}^{n} x_i(y_i - \theta_0 - \theta_1 x_i) = \sum_{i=1}^{n} x_i y_i - \theta_0 \sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2.$$

      Setting them equal to zero gives us the linear system

$$n\theta_0 + \theta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$
$$\theta_0 \sum_{i=1}^{n} x_i + \theta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

      The second partial derivatives are

$$\frac{\partial^2}{\partial \theta_0^2}\log \mathcal{L} = -n, \quad \frac{\partial^2}{\partial \theta_0 \partial \theta_1}\log \mathcal{L} = -\sum_{i=1}^{n} x_i, \text{ and } \frac{\partial^2}{\partial \theta_1^2}\log \mathcal{L} = -\sum_{i=1}^{n} x_i^2.$$

      The second derivative test gives us that

$$\left(\hat{\theta}_0, \hat{\theta}_1\right) = \boxed{\left(\frac{\sum x_i^2 \sum Y_i - \sum x_i \sum (x_i Y_i)}{n\sum x_i^2 - (\sum x_i)^2}, \frac{n\sum(x_i Y_i) - \sum x_i \sum Y_i}{n\sum x_i^2 - (\sum x_i)^2}\right)}$$

      is the MLE of (θ_0, θ_1).

   b) Calculate the Fisher information matrix for (θ_0, θ_1).

      Solution. Using the second derivatives we found in part (a), we see that the Fisher information matrix is

$$\mathcal{I}(\theta_0, \theta_1) = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

c) Compare the Cramer-Rao lower bound for the estimation of θ_1 when θ_0 is unknown to the case where θ_0 is known, and show this second lower bound is the smaller.

Solution. The inverse of the Fisher information matrix from part (b) is

$$\mathcal{I}^{-1}(\theta_0, \theta_1) = \frac{1}{n\sum x_i^2 - (\sum x_i)^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}.$$

Hence, the Cramer-Rao lower bound for the estimation of θ_1 when θ_0 is unknown is

$$\frac{n}{n\sum x_i^2 - (\sum x_i)^2}.$$

The log-likelihood function, when θ_0 is known, is

$$\log \mathcal{L}(\theta_1; \boldsymbol{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - \theta_0 - \theta_1 x_i)^2,$$

and so,

$$\frac{d}{d\theta_1}\log \mathcal{L} = \sum_{i=1}^{n} x_i y_i - \theta_0 \sum_{i=1}^{n} x_i - \theta_1 \sum_{i=1}^{n} x_i^2.$$

Hence,

$$\frac{d^2}{d\theta_1^2}\log \mathcal{L} = -\sum i = 1^n x_i^2,$$

and so, the Fisher information is

$$\mathcal{I}(\theta_1) = \sum_{i=1}^{n} x_i^2,$$

and thus, the Cramer-Rao lower bound for the estimation of θ_1 when θ_0 is known is

$$\frac{1}{\sum x_i^2}.$$

We see that the bound when θ_0 is known is indeed smaller:

$$\frac{1}{\sum x_i^2} = \frac{n}{n\sum x_i^2} < \frac{n}{n\sum x_i^2 - (\sum x_i)^2}.$$

□

d) With both parameters unknown, find a simple necessary condition on a sequence of real numbers $x_1, x_2, \ldots$ such that $\left(\hat{\theta}_0, \hat{\theta}_1\right)$ is consistent for (θ_0, θ_1).