Ph.D. Written Prelim
Spring 2000

Stat 542/3 - I

A problem of interest in medicine is determining if some families are at higher risk for certain diseases (e.g., cancer). Suppose we pick a single disease and examine data in a disease registry for that disease. Data is usually accumulated into a disease registry by having a patient diagnosed with a disease give information about the members of the family who also have the disease. To investigate this problem further, define θ as the probability that an individual has the disease (we assume θ is quite small, say less than .01) and let N be the size of the family. Assume that θ is the same for all individuals and that disease occurrence is independent for the members of a family and across families.

(i) Let X denote the number of diseased individuals in a family. Identify a plausible probability distribution for the conditional distribution of X given that $N = n$?

(ii) Derive the mean and variance of X given that $N = n$ for the probability distribution you specify in (i).

(iii) Suppose that we observe m families at random from the population with sizes $n_1, n_2, \ldots, n_m$. Conditional on these family sizes, we observe numbers of diseased individuals $x_1, x_2, \ldots, x_m$.

  (a) Give the likelihood function for θ.

  (b) Find the maximum likelihood estimator for θ.

  (c) Give the variance (exact or asymptotic) for the estimator you found in (b).

(iv) Suppose that the distribution of family size in the population is given by the probability mass function $p(n)$ for $n \geq 1$, i.e., $\Pr(N = n) = p(n)$.

  (a) Derive the distribution of family size for families in the disease registry by finding $\Pr(N = n | X > 0)$.

  (b) Find an approximation to the result in part (a) that is linear in θ. (You can ignore terms of order θ^2 in the final answer.)

  (c) Are families in the registry generally larger or smaller than families in the population? Justify your answer.

(v) Recall that the primary question of interest is whether some families are at higher risk for certain diseases.

  (a) Calculate $\Pr(X > 1 | X > 0)$ under the assumptions of this problem.

  (b) Explain how you might use the calculation from (a) to tell if some families are at higher risk for the disease in question.

To investigate this problem, define θ as the probability that an individual has the disease (we assume θ is quite small, say less than .01) and let N be the size of the family. Assume that θ is the same for all families and that disease occurrence is independent for the members of a family and across families.

(i) Given $N = n$ and the assumptions, X is a binomial random variable with sample size n and probability of success (disease) θ.

(ii) Given the observation in (i) we have $E(X|N = n) = n\theta$ and $\mathrm{Var}(X|N = n) = n\theta(1-\theta)$. This can be derived using MGFs or directly from the definitions.

(iii.a) $L(\theta) = \prod_{i=1}^{m} \left(\theta^{x_i}(1-\theta)^{n_i - x_i}\right)$

(iii.b) Maximizing $L(\theta)$ gives $\hat{\theta}_{mle} = \sum x_i / \sum n_i$

(iii.c) In this case the exact variance can be found as $\theta(1-\theta)/\sum n_i$. Alternatively one can derive the same answer as the Cramer-Rao Lower Bound using the Fisher Information.

(iv.a) It is natural to apply Bayes' theorem here. Then

$$\begin{aligned}\Pr(N = n|X > 0) &= \Pr(X > 0|N = n)\Pr(N = n)/\Pr(X > 0)\\ &= (1-(1-\theta)^n)p(n)/\left(\sum_{n=1}^{\infty}(1-(1-\theta)^n)p(n)\right)\end{aligned}$$

(iv.b) Note that $(1-\theta)^n \approx 1 - n\theta + n(n-1)\theta^2/2$ plus a term that is roughly the size of θ^3. Remember that although the target is an expression linear in θ we should keep terms of order θ^2 around during intermediate calculations. Plugging the approximation in and simplifying yields

$$\begin{aligned}\Pr(N = n|X > 0) &\approx \frac{(n\theta - n(n-1)\theta^2/2)p(n)}{\sum_{n=1}^{\infty}(n\theta - n(n-1)\theta^2/2)p(n)}\\ &= \frac{(n\theta - n(n-1)\theta^2/2)p(n)}{\theta E(N) - \theta^2 E(N(N-1))/2)}\\ &\approx \frac{np(n)}{E(N)}\left(1 - \frac{(n-1)\theta}{2} + \frac{\theta E(N(N-1))}{2E(N)}\right)\end{aligned}$$

(iv.c) Ignoring for the moment the θ terms in (b) we find that $\Pr(N = n|X > 0) = np(n)/E(N)$. Then $E(N|X > 0) = \sum_n n^2 p(n)/E(N) = (Var(N) + E(N)^2)/E(N) > E(N)$. It is natural to expect that families selected in this way should be a bit larger than average because larger families are more likely to have someone with the disease.

(v.a) Key point here is that one needs to introduce family size.

$$\begin{aligned}
\Pr(X > 1|X > 0) &= \frac{\Pr(X > 1)}{\Pr(X > 0)} = \frac{\sum_{n=1}^{\infty} \Pr(X > 1|N = n)p(n)}{\sum_{n=1}^{\infty} \Pr(X > 0|N = n)p(n)} \\
&= \frac{\sum_{n=1}^{\infty}(1 - (1 - \theta)^n - n\theta(1 - \theta)^{n-1})p(n)}{\sum_{n=1}^{\infty}(1 - (1 - \theta)^n)p(n)} \\
&\approx \frac{\sum_{n=1}^{\infty} n(n - 1)\theta^2/2)p(n)}{\sum_{n=1}^{\infty}(n\theta - n(n - 1)\theta^2/2)p(n)} \approx \theta E(N(N - 1))/(2EN) + O(\theta^2)
\end{aligned}$$

(v.b) The answer in (v.a) is the conditional probability of more than one case given at least one case under the model where disease rate is constant across families. To address the question we could compute the empirical probability and compare to the above (assuming an estimate of θ is available).

Ph.D. Prelim 542-543II Spring 2000

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with common pdf

$$f(x;\theta) = \begin{cases} cx^2 & \text{if } \ 0 \le x \le \theta \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$.

1. Find the value of c.
2. Find the cumulative distribution function $F(x;\theta)$ of X_1.
3. Let
$$F^{-1}(u;\theta) \equiv \min\{x \in I\!R : F(x;\theta) \ge u\}, \quad u \in (0,1),$$
denote the quantile transform of $F(x;\theta)$. Find $F^{-1}(u;\theta)$.
4. Find the maximum likelihood estimator (mle) $\hat{\theta}_n$ of θ based on $X_1, \ldots, X_n$.
5. Find the method of moments estimator (mme) $\tilde{\theta}_n$ of θ based on $X_1, \ldots, X_n$.
6. Let q_θ denote the median of $F(x;\theta)$, i.e., $q_\theta = F^{-1}(\frac{1}{2};\theta)$.
   (a) Find $P_\theta(\hat{\theta}_n \le q_\theta)$ exactly.
   (b) Using the Central Limit Theorem, find an approximate value of $P_\theta(\bar{\theta}_n \le E_\theta X_1)$.

SOLUTION / Ph.D. Prelim, 2010 / STAT 542-543-II.

1. $$1 = \int_0^\theta c x^2 \, dx = c \cdot \frac{\theta^3}{3}$$

$$\Rightarrow \quad c = 3/\theta^3$$

2. For $0 < x < \theta$,

$$F(x;\theta) = \int_0^x c t^2 \, dt = \frac{3}{\theta^3} \cdot \frac{x^3}{3} = \left(\frac{x}{\theta}\right)^3$$

Hence,

$$F(x,\theta) = \begin{cases} 0 & \text{if } x < 0 \\ (x/\theta)^3 & \text{if } 0 \leq x \leq \theta \\ 1 & \text{if } x > \theta . \end{cases}$$

3. Note that the function $F(x;\theta)$ is strictly increasing over the support $(0,\theta)$ of X_1. Hence,

$$F^{-1}(u;\theta) = \min\{x \in \mathbb{R} : F(x;\theta) \geq u\}$$

is the unique solution to the equation

$$F(x;\theta) = u$$

$$\Leftrightarrow \quad (x/\theta)^3 = u \quad \Longleftrightarrow \quad x = \theta \cdot u^{1/3}, \quad 0 < u < 1$$

Hence, $F^{-1}(u;\theta) = u^{1/3}.\theta, \qquad 0<u<1.$

4. The likelihood function is given by

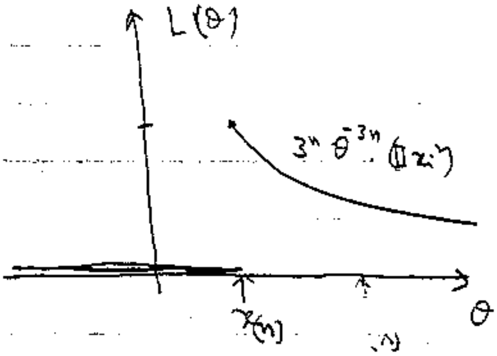
$$L(\theta) = \Pi_{i=1}^{n} f(x_i;\theta)$$

$$= \left(\frac{3}{\theta^3}\right)^n \Pi_{i=1}^{n} \left\{x_i^2 \, \mathbb{1}(0<x_i<\theta)\right.$$

$$= 3^n \theta^{-3n} \cdot \left(\Pi_{i=1}^{n} x_i^2\right) \cdot \mathbb{1}(x_{(1)}>0) \cdot \mathbb{1}(x_{(n)}<\theta)$$

where $\mathbb{1}(\cdot)$ denotes the indicator function

and $x_{(1)} = \min_{1\le i\le n} x_i$ and $x_{(n)} = \max_{1\le i\le n} x_i$



From the graph, it is clear that $L(\theta)$ is maximized at $\theta = x_{(n)}$.

Hence,

$$\hat{\theta}_n = \text{mle of } \theta = X_{(n)}.$$

$$E_\theta X_1 = \int_0^\theta x \cdot cx^2\,dx = c \cdot \int_0^\theta x^3\,dx$$

$$= c \cdot \frac{\theta^4}{4} = \frac{3\theta}{4}.$$

Hence, the MME of θ is a solution of

$$\frac{3\theta}{4} = \bar{X}_n \iff \theta = \frac{4}{3}\bar{X}_n.$$

Thus, $\tilde{\theta}_n = \frac{4}{3} \cdot \bar{X}_n$.

6. (a).

$$P_\theta\left(\hat{\theta}_n \le F^{-1}(\tfrac{1}{2};\theta) \right)$$

$$= P_\theta\left(X_{(n)} \le (\tfrac{1}{2})^{1/3} \cdot \theta \right)$$

$$= \left[P_\theta\left(X_1 \le (\tfrac{1}{2})^{1/3}\,\theta\right) \right]^n, \text{ since } X_i\text{'s are iid}$$

$$= \left[F\left((\tfrac{1}{2})^{1/3} \cdot \theta ; \theta \right) \right]^n$$

$$= \left[\left\{ \frac{(\frac{1}{2})^{1/3} \cdot \theta}{\theta} \right\}^3 \right]^n \quad \text{from part 2.}$$

$$= [\tfrac{1}{2}]^n = 2^{-n}$$

6(b). Here, $E_\theta X_1^2 = \int_0^\theta x^2 \cdot c x^2 \, dx = c \cdot \frac{\theta^5}{5}$

$$= \frac{3}{5} \cdot \theta^2$$

Hence, $Var_\theta(X_1) = \frac{3}{5}\theta^2 - \left(\frac{3}{4}\theta\right)^2$

$$= \frac{3}{80} \cdot \theta^2$$

By the CLT,

$$\sqrt{n}\,(\bar{X}_n - E_\theta X_1) \xrightarrow{d} N(0, Var_\theta(X_1))$$

i.e. $\sqrt{n}\,(\bar{X}_n - \frac{3}{4}\theta) \xrightarrow{d} N(0, \frac{3}{80}\theta^2)$

Hence, $P_\theta(\tilde{\theta}_n \le E_\theta X_1)$

$$= P_\theta\left(\frac{4}{3}\bar{X}_n \le \frac{3}{4}\theta\right)$$

$$= P_\theta\left(\bar{X}_n \le \frac{9}{16}\theta\right)$$

$$= P_\theta\left(\sqrt{n}\left(\bar{X}_n - \frac{3}{4}\theta\right) \le -\frac{3\theta}{16}\sqrt{n}\right)$$

$\simeq P\left(Z \le -\frac{3\theta\sqrt{n}}{16} \Big/ \sqrt{\frac{3}{80}\theta^2}\right)$, where $Z \sim N(0,1)$

$$= P\left(Z \le -\sqrt{n} \cdot \sqrt{\frac{15}{16}}\right)$$

Ph.D. Prelim STAT 542-543 III Spring 2000

Let $X_1, \ldots, X_n$ be a collection of independent and identically distributed (iid) Exponential (θ) random variables with common pdf

$$f(x;\theta) = \begin{cases} \theta^{-1}\exp(-x/\theta) & \text{if} \quad x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$.

1. Consider the random equation $t^2 - 2X_1 t + 4X_2 = 0$ in the variable t and let A be the event that this equation has real roots. Find $P_\theta(A)$.

2. Next define the random variables $Y_1, \ldots, Y_n$ by $Y_i = X_i^2 + 1$, $1 \le i \le n$.

   (a) Find the pdf $g(y;\theta)$ of Y_1.

   (b) Show that the family of pdfs $\{\prod_{i=1}^n g(y_i;\theta) : \theta > 0\}$ has the monotone likelihood ratio property in $T_n = \sum_{i=1}^n (Y_i - 1)^{1/2}$

   (c) Show that $n^{-1/2}(T_n - n\theta) \to^d N(0, \theta^2)$ as $n \to \infty$, where $\to^d$ denotes convergence in distribution.

   (d) Write down a uniformly most powerful (UMP) test of size $\alpha \in (0,1)$ for testing the hypotheses $H_0 : \theta \le 2$ against $H_1 : \theta > 2$ based on $Y_1, \ldots, Y_n$, and find the constant(s) of your UMP test *approximately* using part (c).

SOLUTION / PhD. Prelim, 2000 / STAT 542-543-III.

1. The solutions to the equation

$$t^2 - 2X_1 t + 4X_2^2 = 0$$

are $\quad t = X_1 \pm \sqrt{X_1^2 - 4X_2^2}\,,$

which are real $\iff$ $X_1^2 - 4X_2^2 \geq 0$. Hence,

$$P_\theta(A) = P_\theta\left(X_1^2 \geq 4X_2^2\right) = P_\theta(X_1 > 2X_2)$$

$$= \int_0^\infty \int_0^{x_1/2} \theta^{-2} \exp(-(x_1+x_2)/\theta)\, dx_2\, dx_1$$

$$= \theta^{-2} \int_0^\infty e^{-x_1/\theta} \left[\int_0^{x_1/2} e^{-x_2/\theta}\, dx_2\right] dx_1$$

$$= \theta^{-1} \int_0^\infty e^{-x_1/\theta} \left[1 - e^{-x_1/2\theta}\right] dx_1$$

$$= \theta^{-1} \int_0^\infty e^{-x_1/\theta}\, dx_1 - \theta^{-1} \int_0^\infty e^{-3x_1/\theta}\, dx_1$$

$$= 1 - \theta^{-1} \left[-\frac{e^{-3x_1/\theta}}{(3/\theta)} \Bigg|_0^\infty \right]$$

$$= 1 - \tfrac{1}{3}[1 - 0] = 2/3.$$

2. (a)

Let $h(x) = x^2+1$. Then, h is a 1-1 function from $(0,\infty)$, onto $(1,\infty)$, with inverse

$$h^{-1}(y) = \sqrt{y-1} \quad , \quad y \in (1,\infty).$$

Hence, the pdf of Y_1 is given by

$$g(y;\theta) = f(h^{-1}(y)) \cdot \left| \frac{d}{dy} h^{-1}(y) \right| , \quad y \in (1,$$

$$= \begin{cases} \left(2\theta \sqrt{y-1}\right)^{-1} \exp\left(-\sqrt{y-1}/\theta\right), & y > \\ 0 & \text{otherwis} \end{cases}$$

(b) Fix $\theta_1 < \theta_2$. Then, $B \equiv \{ (y_1,\ldots,y_n)' : \prod_{i=1}^n g(y_i;\theta_1)$ $\prod_{i=1}^n g(y_i;\theta_2) > 0 \} = \{ (y_1,\ldots,y_n)' : y_{(1)} > 1 \}$, where $y_{(1)} = \min_{1\le i\le n} y_i$. For any $(y_1,\ldots,y_n)' \in B$,

$$\frac{\prod_{i=1}^n g(y_i;\theta_2)}{\prod_{i=1}^n g(y_i;\theta_1)} = \frac{\prod_{i=1}^n \left\{ \left(2\theta_2\sqrt{y_i-1}\right)^{-1} \exp\left(-\sqrt{y_i-1}/\theta_2\right) \right\}}{\prod_{i=1}^n \left\{ \left(2\theta_1\sqrt{y_i-1}\right)^{-1} \exp\left(-\sqrt{y_i-1}/\theta_1\right) \right\}}$$

$$= \left(\frac{\theta_1}{\theta_2}\right)^n \cdot \exp\left(\left[\frac{1}{\theta_1} - \frac{1}{\theta_2}\right] \cdot \left\{\Sigma_{i=1}^n \sqrt{y_i - 1}\right\}\right),$$

which is an increasing function of $\Sigma_{i=1}^n \sqrt{y_i - 1}$, as $\theta_1^{-1} - \theta_2^{-1} > 0$.

Hence, $\{ \Pi_{i=1}^n g(y_i; \theta) : \theta > 0 \}$ has MLR in T_n.

(c) Note that $X_i = \sqrt{Y_i - 1}$, $i \geq 1$ are iid with $E_\theta X_1 = \theta$ and $Var_\theta(X_1) = \theta^2$. Hence, by the central limit theorem,

$$n^{-\frac{1}{2}}(T_n - n\theta) = n^{-\frac{1}{2}}\left(\Sigma_{i=1}^n X_i - n\theta\right)$$

$$= \sqrt{n}\,(\bar{X}_n - E_\theta X_1) \xrightarrow{d} N(0, \underbrace{Var_\theta(X_1)}_{\theta^2})$$

(d) A size α UMP test for testing $H_0: \theta \leq 2$ vs. $H_1: \theta > 2$ is given by

$$\textcircled{1} \leftarrow \quad \phi(\underline{y}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} (y_i-1)^{1/2} > k \\ \gamma & \qquad\qquad = \\ 0 & \qquad\qquad < \end{cases}$$

where $\gamma \in [0,1]$ and $k \in (0,\infty)$ are such that

$$E_{\theta=2}\, \phi(\underset{\sim}{Y}) = \alpha . \quad \longrightarrow \textcircled{2}$$

Note that $T_n = \sum_{i=1}^{n} (Y_i - 1)^{1/2} = \sum_{i=1}^{n} X_i$ has a continuous distribution under $\theta = 1$. Hence, we may take $\gamma = 0$. To choose k, from $\textcircled{2}$ we have

$$\alpha = P_{\theta=2}(T_n > k) + \gamma .\ P_{\theta=2}(T_n = k)$$

$$= P_{\theta=2}\left(\frac{(T_n - 2n)}{2\sqrt{n}} > \frac{k-2n}{2\sqrt{n}} \right) + 0$$

$$\approx P\left(Z > \frac{k-2n}{2\sqrt{n}}\right), \text{ where } Z \sim N(0,1)$$

Thus, $\frac{k-2n}{2\sqrt{n}} \approx z_{1-\alpha}$, the $(1-\alpha)$ quantile of $N(0,1)$

$$\Rightarrow \quad k \approx 2n + 2\sqrt{n} . z_{1-\alpha} .$$