Minnesota Radon Levels

Statistics Project, Spring 2008
Mei Yin
Department of Mathematics, University of Arizona

1 Project Description

"Radon is the second-leading cause of lung-cancer, after smoking, according to the U.S. surgeon general. According to the EPA, an estimated 14,000 people nationwide die each year from radon-causing cancer..."(San Francisco Chronicle, July 5, 1995).

In this project, to reduce computing, consider only three Minnesota counties: Hennepin, Ramsey, and St. Louis, as constituting a mini-Minnesota state. Here are the data collected on these counties. Note that pCi/l denotes picoCuries per litre.

Hennepin	119	3925	4.64	3.4
Ramsey	42	1809	4.54	4.9
St. Louis	122	81	3.06	3.6
Total	$283 = n$	5815 (in 100s)$= N$		

In the above table, the 1st column denotes the Number of Houses Sampled (n_i), the 2nd column denotes the Total Number (N_i) of Houses (in 100s), the 3rd column denotes the Sample Mean $\hat{m}_i$ (pCi/l) and the 4th column denotes the Sample Standard Deviation $s_i = \sqrt{\hat{v}_i}$ (pCi/l).

2 Task

2.1 Nonparametric Model

(i) Estimate the average (or mean) radon concentration m per household in the mini-state.
Solution: $\bar{Y} = \sum_{1\le i\le 3} \frac{N_i}{N}\hat{m}_i$ is an unbiased estimator of the average radon concentration m per household in the mini-state. We have,

$$\begin{aligned}\bar{Y} &= \sum_{1\le i\le 3} \frac{N_i}{N}\hat{m}_i \\ &= \frac{3925}{5815}\cdot 4.64 + \frac{1809}{5815}\cdot 4.54 + \frac{81}{5815}\cdot 3.06 \\ &= 4.59\end{aligned}$$

(ii) Estimate the standard error σ_E, (i.e., the square root of the variance) of the estimate Y-bar, say.
Solution: The variance of the estimate $\bar{Y}$ is given by $E(\bar{Y}-m)^2 = \sum_{1\le i\le 3}(\frac{N_i}{N})^2 \frac{\hat{v_i}}{n_i}$. We have,

$$\begin{aligned} E(\bar{Y}-m)^2 &= \sum_{1\le i\le 3}(\frac{N_i}{N})^2 \frac{\hat{v_i}}{n_i} \\ &= (\frac{3925}{5815})^2 \cdot \frac{3.4^2}{119} + (\frac{1809}{5815})^2 \cdot \frac{4.9^2}{42} + (\frac{81}{5815})^2 \cdot \frac{3.6^2}{122} \\ &= 0.10 \end{aligned}$$

Therefore, the standard error σ_E, (i.e., the square root of the variance) is easily obtained, $\sqrt{0.10} = 0.32$.

(iii) Compute a 95% confidence interval for the mean concentration (based on Normal approximation) of the estimate.
Solution: Based on Normal approximation, the 95% confidence interval for the mean concentration is given by $[\bar{Y} - 1.96\sigma_E, \bar{Y} + 1.96\sigma_E]$. We have,

$$\begin{aligned} [\bar{Y} - 1.96\sigma_E, \bar{Y} + 1.96\sigma_E] &= [4.59 - 1.96 \cdot 0.32, 4.59 + 1.96 \cdot 0.32] \\ &= [3.96, 5.22] \end{aligned}$$

2.2 Parametric Model

Assume that in each county the distribution of radon concentration is a two-parameter gamma, possibly different in different counties (This is a six-parameter model). In addition to the quantities given in the table, we are given the values 1.386, 1.335, 0.916, of $\sum_{1\le j\le n_i} \log x_{j,i}/n_i$, where $x_{j,i}$ is the radon concentration in the j-th house sampled in the i-th county ($i = 1, 2, 3$) (Here and elsewhere, "log" denotes natural logarithm).

(i) Find the UMVU estimates of the mean radon concentrations m_i in the three counties, as well as that in the mini-state.
Solution: In each county the distribution of radon concentration is a two-parameter gamma,

$$\begin{aligned} p_i(x|\theta) &= \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} e^{-\frac{x}{\alpha_i}} x^{\beta_i - 1} 1_{(0,\infty)}(x) \\ &= \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} \frac{1}{x} 1_{(0,\infty)}(x) e^{-\frac{1}{\alpha_i}x + \beta_i \log x} \end{aligned}$$

where $i = 1, 2, 3$ and $\theta = (\alpha, \beta) \in (0, \infty) \times (0, \infty) = \Theta$.

We see explicitly that the natural parameters are $\pi_1 = -\frac{1}{\alpha}$ and $\pi_2 = \beta$. The natural parameter space $\Pi = (-\infty, 0) \times (0, \infty)$.

The (joint) distribution P_θ of $X_i = (x_{j,i})$, $1 \le j \le n_i$ has pdf (wrt. Lebesgue measure on $\mathbf{R}^{n_i}$),

$$f_i(X|\theta) = (\frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)})^{n_i}\frac{1}{\prod_j x_{j,i}}e^{-\frac{1}{\alpha}\sum_j x_{j,i}+\beta\sum_j \log x_{j,i}}\prod_j 1_{(0,\infty)}(x_{j,i})$$

A sufficient statistic for $\{P_\theta : \theta \in \Theta\}$ is $T = (\sum_j X_{j,i}, \sum_j \log X_{j,i})$.
The natural parameter space Π has a nonempty interior, thus T is actually complete sufficient.

By Lehmann-Scheffé Theorem, $\frac{\sum_{1\le j\le n_i} X_{j,i}}{n_i}$ is the UMVU estimate of the mean radon concentration m_i in each of the three counties.

The UMVU estimate of the mean radon concentration m in the mini-state is a weighted sum, i.e., $\sum_{1\le i\le 3}\frac{N_i}{N}\frac{\sum_{1\le j\le n_i} X_{j,i}}{n_i}$.

(ii) Find (estimates of) the standard errors of these estimates.
Solution: This is analogous to the calculations done in Nonparametric Model case.

The average (or mean) radon concentration m_i per household in the three counties, Hennepin, Ramsey and St. Louis are estimated to be 4.64, 4.54 and 3.06 respectively. And the average (or mean) radon concentration m per household in the mini-state is estimated to be 4.59.

The standard error of these estimates are $3.4/\sqrt{119} = 0.31$, $4.9/\sqrt{42} = 0.76$ and $3.6/\sqrt{122} = 0.33$ respectively in the three counties and 0.32 in the mini-state.

(iii) Find a 95% confidence interval for the mean concentration m in the mini-state (based on Normal approximation for this UMVU estimate of m).
Solution: Based on Normal approximation, the 95% confidence interval for the mean concentration is given by

$$[4.59 - 1.96\cdot 0.32, 4.59 + 1.96\cdot 0.32] = [3.96, 5.22]$$

(iv) Compute the MLEs α_i, β_i of the gamma parameters in the three counties. You may perhaps use the Newton-Raphson or the gradient method, beginning with the trial solution provided by the method of moments; or use some other algorithm such as EM.
Solution: The trial solution is provided by the method of moments.

$E(X_i) = \int_0^\infty \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} x^{\beta_i} e^{-\frac{x}{\alpha_i}} = \alpha_i\beta_i$

$Var(X_i) = \int_0^\infty \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} x^{\beta_i+1} e^{-\frac{x}{\alpha_i}} = \alpha_i^2\beta_i$

Thus the estimates are,

$$\tilde{\alpha_i} = \frac{v_i}{m_i} \qquad\qquad \tilde{\beta_i} = \frac{m_i^2}{v_i}$$

We perform the calculations for each of the three counties.

$$\tilde{\alpha_1} = \frac{s_1^2}{m_1} = \frac{3.4^2}{4.64} = 2.49$$

$$\tilde{\beta_1} = \frac{m_1^2}{s_1^2} = 1.86$$

$$\tilde{\alpha_2} = \frac{s_2^2}{m_2} = \frac{4.9^2}{4.54} = 5.29$$

$$\tilde{\beta_2} = \frac{m_2^2}{s_2^2} = 0.86$$

$$\tilde{\alpha_3} = \frac{s_3^2}{m_3} = \frac{3.6^2}{3.06} = 4.24$$

$$\tilde{\beta_3} = \frac{m_3^2}{s_3^2} = 0.72$$

Recall that the (joint) distribution P_θ of $X_i = (x_{j,i})$, $1 \leq j \leq n_i$ has pdf (wrt. Lebesgue measure on $\mathbf{R}^{n_i}$),

$$f_i(X|\theta) = (\frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)})^{n_i} e^{-\frac{1}{\alpha}\sum_j x_{j,i}} (\prod_j x_{j,i})^{\beta-1} \prod_j 1_{(0,\infty)}(x_{j,i})$$

i.e., the likelihood function restricted to $(0,\infty)^{n_i}$ is given by,

$$l_i(\alpha_i, \beta_i) = (\frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)})^{n_i} e^{-\frac{1}{\alpha}\sum_j x_{j,i}} (\prod_j x_{j,i})^{\beta-1}$$

It is easier, equivalently, to consider maximizing the log-likelihood function,

$$\log l_i(\alpha_i, \beta_i) = -n_i\beta_i \log(\alpha_i) - n_i \log(\Gamma(\beta_i)) + (\beta - 1)\sum_j \log x_{j,i} - \frac{1}{\alpha_i}\sum_j x_{j,i}$$

Setting the partial derivatives wrt. to α_i and β_i to zero yields the following:

$$-n_i\beta_i\frac{1}{\alpha_i} + \frac{1}{\alpha_i^2}\sum_j x_{j,i} = 0$$

$$-n_i \log(\alpha_i) - n_i\psi(\beta_i) + \sum_j \log x_{j,i} = 0$$

where $\psi(\beta_i) = \frac{d\log(\Gamma(\beta_i))}{d\beta_i}$ by definition.

We employ the Newton-Raphson method to find the zeros, beginning with the trial solution provided by the method of moments, as calculated above.

And we arrive at the following (see appendix for MATLAB code):

$\hat{\alpha_1} = 1.71$, $\hat{\beta}_1 = 2.71$

$\hat{\alpha_2} = 3.19$, $\hat{\beta}_2 = 1.42$

$\hat{\alpha_3} = 2.55$, $\hat{\beta}_3 = 1.20$

(v) Find the MLEs $\hat{m}_i$ of the mean radon concentrations m_i in the three counties($i = 1, 2, 3$), and the corresponding estimate $\hat{m}$ of the mean radon concentration m in the mini-state. Note that these are not in general the same as the UMVU estimates in (i).

Solution: For gamma distribution, $E(X_i) = \int_0^\infty \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} x^{\beta_i} e^{-\frac{x}{\alpha_i}} = \alpha_i\beta_i$

Thus the mean m_i is estimated by $\alpha_i\beta_i$.

From (iv),

$$\hat{m_1} = \alpha_1\beta_1 = 1.71 \cdot 2.71 = 4.63$$

$$\hat{m_2} = \alpha_2\beta_2 = 3.19 \cdot 1.42 = 4.53$$

$$\hat{m_3} = \alpha_3\beta_3 = 2.55 \cdot 1.20 = 3.06$$

The estimate $\hat{m}$ of the mean radon concentration m in the mini-state is a weighted sum,

$$\begin{aligned} \hat{m} &= \sum_{1\le i\le 3} \frac{N_i}{N}\hat{m}_i \\ &= \frac{3925}{5815}\cdot 4.63 + \frac{1809}{5815}\cdot 4.53 + \frac{81}{5815}\cdot 3.06 \\ &= 4.58 \end{aligned}$$

(vi) Use these MLEs to estimate the proportions of households in each of the three counties with radon concentration more than 4pCi/l.

Solution: What we are looking for is

$$\int_4^\infty \frac{1}{\alpha_i^{\beta_i}\Gamma(\beta_i)} e^{-\frac{x}{\alpha_i}} x^{\beta_i - 1}$$

Using the gamcdf function in MATLAB, this is easily achieved. (The subtlety here is that in MATLAB, the roles of the parameters α_i and β_i are reversed)

The proportion of households in Hennepin with radon concentration more than 4pCi/l is 49%.

The proportion of households in Ramsey with radon concentration more than 4pCi/l is 56%.

The proportion of households in St. Louis with radon concentration more than 4pCi/l is 73%.

(vii) Use (vi) to obtain an estimate of the proportion of households in the mini-state with radon concentration more than 4pCi/l.

Solution: The proportion of households in the mini-state with radon concentration more than 4pCi/l is a weighted sum,

$$\frac{3925}{5815}\cdot 0.49 + \frac{1809}{5815}\cdot 0.56 + \frac{81}{5815}\cdot 0.73 = 52\%$$