COMPARISON OF METHODS FOR ANALYZING LEFT-CENSORED OCCUPATIONAL EXPOSURE DATA

Supplemental Materials

Tran Huynh, Gurumurthy Ramachandran, Sudipto Banerjee, Joao Monteiro, Mark Stenzel, Dale P. Sandler, Lawrence S. Engel, Richard K. Kwok, Aaron Blair and Patricia A. Stewart

1 Details for β-substitution

The details for the following derivation of β_{GM} were provided by Gary Ganser. The derivation for β_{AM} proceeds in a similar fashion.

1.1 Distributional properties

We let x_i be a lognormally distributed random variable with LOD L for $i = 1, \ldots, n$, where n denotes the sample size. This implies that $y_i = \log x_i$ is Normally distributed with mean μ and variance σ^2. Assume k of our observations are below the limit of detection such that we only observe y_i for $i = 1, \ldots, (n-k)$. That is, $y_i \in (\log L, \infty)$ follows a truncated Normal distribution with pdf

$$f(y_i \,|\, \mu, \sigma^2) = \frac{\exp\left[-\left(y_i - \mu\right)^2 / 2\sigma^2\right]}{\int_{\log L}^{\infty} \exp\left[-\left(u - \mu\right)^2 / 2\sigma^2\right] du}, \tag{1}$$

which has $E[y_i \,|\, \mu, \sigma^2] = \mu + f(z)\sigma$, where $z = (\log L - \mu)/\sigma$; for the sake of notation, we shall suppress the conditioning on μ and σ^2 from our remaining equations. This implies that if we let $\bar{y} = \sum_i^{n-k} y_i/(n-k)$, then $E[\bar{y}] = \mu + f(z)\sigma$, as well. Another property of the truncated Normal distribution is that the moment generating function is

$$E\left[e^{ry}\right] = \exp\left[r\mu + \frac{\sigma^2 r^2}{2}\right] \times \left[\frac{\int_{\frac{\log L - (r\mu + \sigma^2 r^2)}{\sigma}}^{\infty} \exp\left[-v^2/2\right] dv}{\int_{\frac{\log L - \mu}{\sigma}}^{\infty} \exp\left[-v^2/2\right] dv}\right] = \exp\left[r\mu + \frac{\sigma^2 r^2}{2}\right] X(r). \tag{2}$$

1.2 Derivation for β_{GM}

Given these properties, we will derive an estimate for β_{GM}. First note that

$$\exp\left[\widehat{\mu}(\beta_{GM})\right] = (x_1 x_2 \cdots x_{n-k})^{1/n} (\beta_{GM} L)^{k/n} = \exp\left[\sum_{i=1}^{n-k} y_i/(n-k)\right] (\beta_{GM} L)^{k/n}. \quad (3)$$

Using (2), we find

$$\begin{aligned}
E\left(\exp\left[\widehat{\mu}(\beta_{GM})\right]\right) &= E\left(\exp\left[\sum_{i=1}^{n-k} y_i/(n-k)\right] (\beta_{GM} L)^{k/n}\right) = \prod_{i=1}^{n-k} E\left[e^{(1/n)y_i}\right] (\beta_{GM} L)^{k/n} \\
&= \left(E\left[e^{(1/n)y_1}\right]\right)^{n-k} (\beta_{GM} L)^{k/n} \\
&= \left(\exp\left[(1/n)\mu + \sigma^2 (1/n)^2/2\right] X(1/n)\right)^{n-k} (\beta_{GM} L)^{k/n} \\
&= \exp\left[\frac{n-k}{n}\mu + \frac{n-k}{2n^2}\sigma^2 + (n-k)\log X\left(\frac{1}{n}\right) + \frac{k}{n}\log\beta_{GM} + \frac{k}{n}\log L\right] \\
&= e^{\mu}\exp\left[\frac{-k}{n}\mu + \frac{n-k}{2n^2}\sigma^2 + (n-k)\log X\left(\frac{1}{n}\right) + \frac{k}{n}\log\beta_{GM} + \frac{k}{n}\log L\right] \\
&= e^{\mu}\exp\left[g\left(\beta_{GM}\right)\right]. \qquad (4)
\end{aligned}$$

If we can solve $g\left(\beta_{GM}\right) = 0$ for β_{GM} in (4, then $E\left(\exp\left[\widehat{\mu}(\beta_{GM})\right]\right) = e^{\mu}$ as we desire. That is, we need

$$\begin{aligned}
\log\beta_{GM} &= -\frac{n}{k}\left(\frac{-k}{n}\mu + \frac{n-k}{2n^2}\sigma^2 + (n-k)\log X(1/n) + (k/n)\log L\right) \\
&= \mu - \log L - \frac{n-k}{2kn}\sigma^2 - \frac{n(n-k)}{k}\log X(1/n) \\
&= -z\sigma - \frac{n-k}{2kn}\sigma^2 - \frac{n(n-k)}{k}\log X(1/n). \qquad (5)
\end{aligned}$$

To do this, we first note that, for large n,

$$
\begin{aligned}
k/n \approx P(X < L) &= P\left(\frac{\log X - \mu}{\sigma} < \frac{\log L - \mu}{\sigma} = z\right) = \frac{1}{\sqrt{2\pi}} \int_{\infty}^{z} e^{v^2/2} dv \\
\implies z \approx \widehat{z} &= \Phi^{-1}(k/n), \qquad (6)
\end{aligned}
$$

where $\Phi(\cdot)$ denotes the cdf of a Normal distribution and $\Phi^{-1}(\cdot)$ denotes the inverse cdf. Then, using the expression for $E[\bar{y}]$ and the definition of z, we find

$$
\begin{aligned}
\sigma &= \frac{\log L - \mu}{z} = \frac{\log L - [E(\bar{y}) - \sigma f(z)]}{z} = \frac{\log L - E(\bar{y})}{z} + \sigma f(z)/z \\
\implies \sigma &= \frac{\log L - E(\bar{y})}{z} \frac{z}{z - f(z)} = \frac{\log L - E(\bar{y})}{z - f(z)} \\
&\approx \widehat{s}_y = \frac{\log L - \bar{y}}{\widehat{z} - f(\widehat{z})}. \qquad (7)
\end{aligned}
$$

Lastly, we can approximate

$$
\begin{aligned}
X(1/n) &= \left[\frac{\int_{\left(\log L - (\frac{1}{n}\mu + \sigma^2 \frac{1}{n^2})\right)/\sigma}^{\infty} \exp\left[-v^2/2\right] dv}{\int_{(\log L - \mu)\sigma}^{\infty} \exp\left[-v^2/2\right] dv}\right] \\
&\approx f(\widehat{s}_y, \widehat{z}) = \frac{1 - \Phi(\widehat{z} - \widehat{s}_y/n)}{1 - \Phi(\widehat{z})} \qquad (8)
\end{aligned}
$$

Finally, plugging the expressions from (6), (7), and 8 into (5), we obtain

$$
\begin{aligned}
\beta_{GM} &= \exp\left[-z\sigma - \frac{n-k}{2kn}\sigma^2 - \frac{n(n-k)}{k} \log X(1/n)\right] \\
&\approx \exp\left[-\widehat{z}\widehat{s}_y - \frac{n-k}{2kn}\widehat{s}_y^2 - \frac{n(n-k)}{k} \log f(\widehat{s}_y, \widehat{z})\right].
\end{aligned}
$$

2 Details for Kaplan-Meier Method

The equations for the K-M method can be found in the EPA ProUCL Technical Guide (EPA, 2007). Beal (2009) also described the method and provided SAS macro for it.

Suppose out of a sample of n exposure concentrations, k have values below their respective limits of detection. We can define indicator variables, $d_1, \ldots, d_n$, such that $d_i = 1$ if the ith observation is above its limit of detection and $d_i = 0$ if the ith observation is censored. We then let $x_1, \ldots, x_n$ be a vector of concentrations such that x_i equals the ith concentration when $d_i = 1$ and equals LOD_i when $d_i = 0$. Without loss of generality, we assume the x_i are sorted in ascending order such that $x_1 \leq x_2 \leq \ldots \leq x_n$.

To compute the K-M estimate of the arithmetic mean, AM, we let $x_1^*, \ldots, x_p^*$ denote p distinct values (sorted in ascending order) at which concentrations above the limit of detection are observed such that $p \leq (n-k)$. For $j = 1, \ldots, p$, let

$$m_j = \sum_i d_i \times I(x_i = x_j^*)$$
$$n_j = \sum_i I(x_i \leq x_j^*),$$

where $I(\cdot)$ is a 0/1 indicator function; i.e., m_j is the number of detected concentrations equal to x_j^* and n_j counts the number of concentrations (both detected and nondetected) less than or equal to x_j^*. We then define an empirical cdf, $F(\cdot)$ as

$$F(x) = \begin{cases} 1 & x \geq x_p^* \\ \prod_{j:x_j^*>x}^{p} \frac{n_j - m_j}{n_j} & x_1^* \leq x \leq x_{p-1}^* \\ \prod_{j=1}^{p} \frac{n_j - m_j}{n_j} & x_1 \leq x \leq x_1^* \\ 0 & 0 \leq x \leq x_1 \end{cases}.$$

From this, we can estimate AM as

$$\widehat{AM} = \sum_{j=1}^{p} x_j^* \left[F(x_j^*) - F(x_{j-1}^*) \right], \tag{9}$$

where $F(x_0^*) = 0$.