Data Uncertainty, MCML and Sampling Density

Graham Byrnes

International Agency for Research on Cancer

27 October 2015

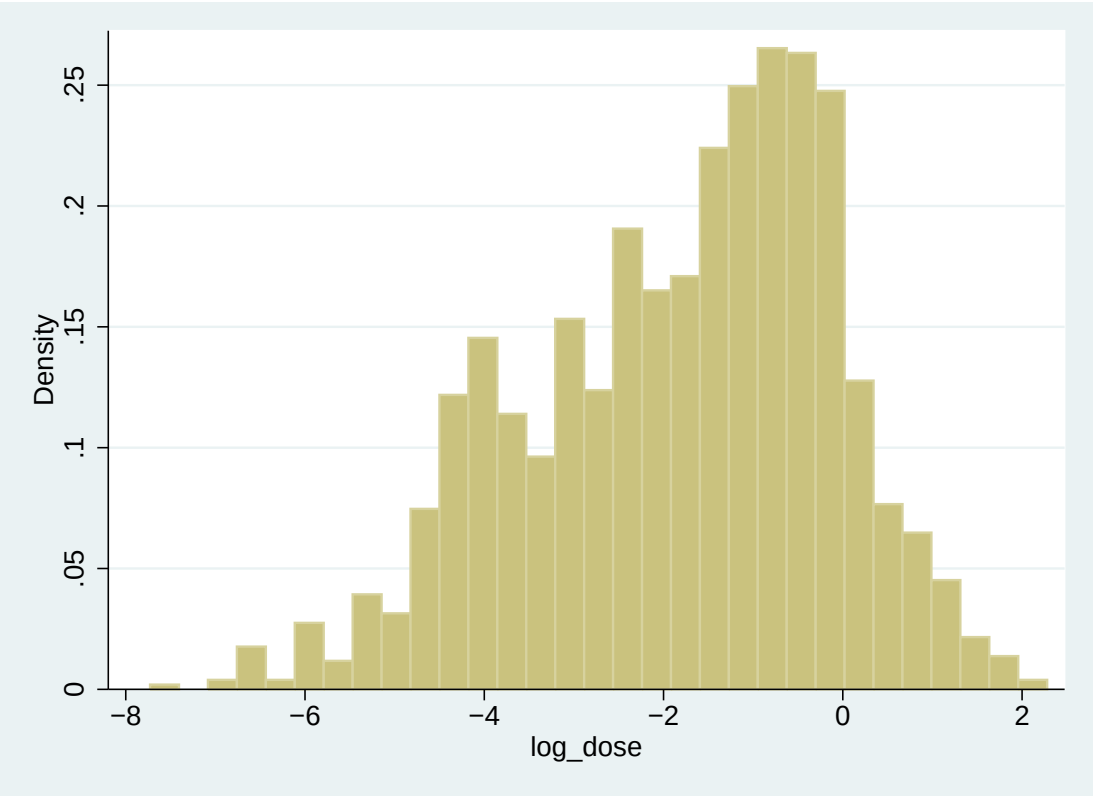
Outline...

- Correlated Measurement Error
- Maximal Marginal Likelihood
- Monte Carlo Maximum Likelihood
- Sampling Density

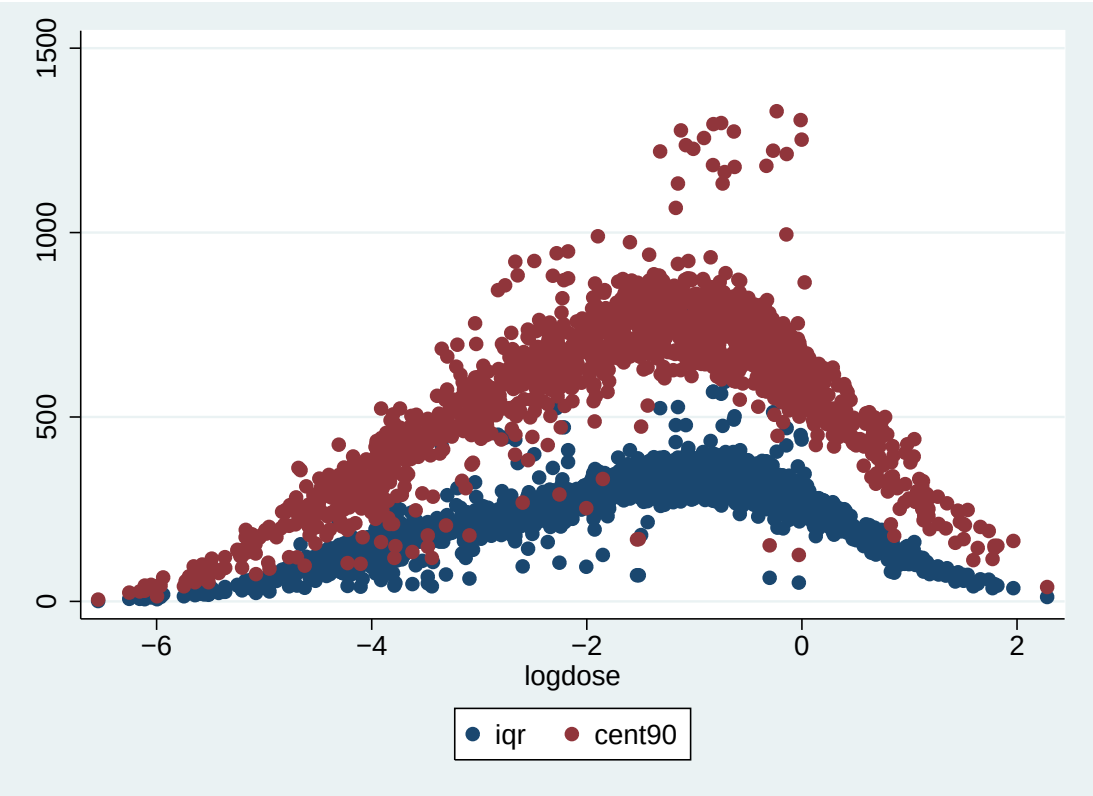
Example Data: Chernobyl Thyroid doses

- Range of doses is large;
- Dose rank varies dramatically by draw;
- Also rank within matched (risk) set.
- EPI-CT estimates still not ready

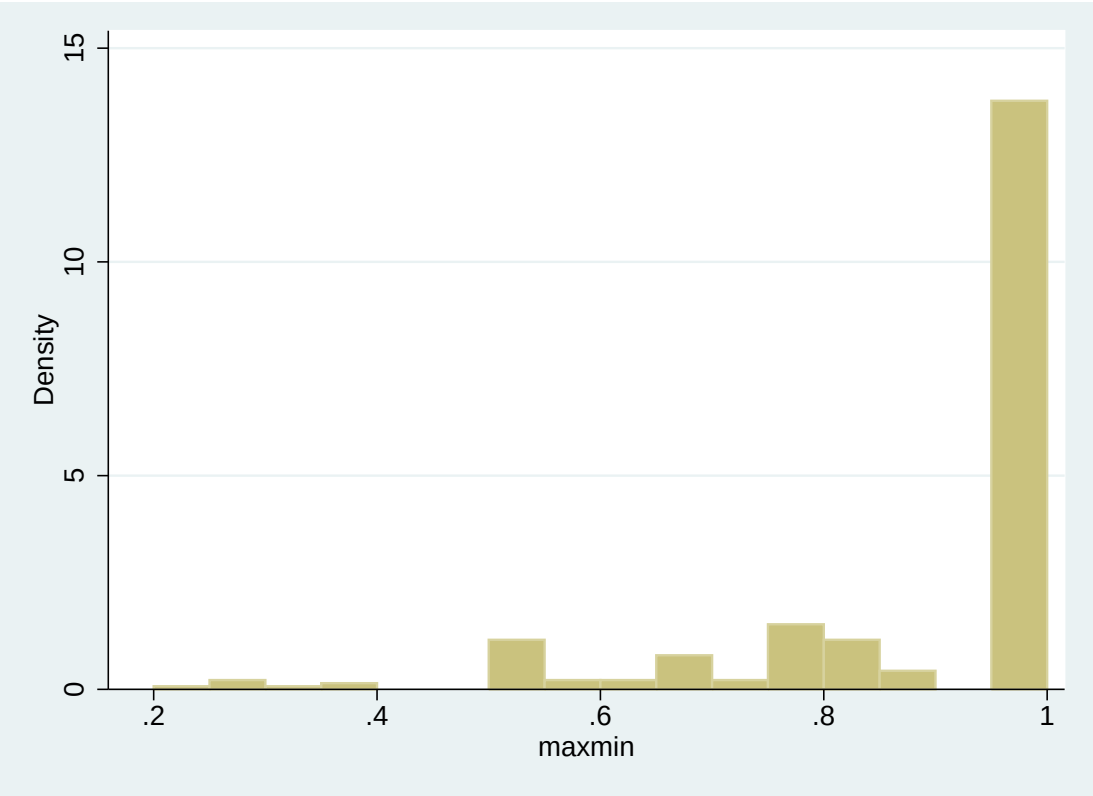
Log dose (central estimate)



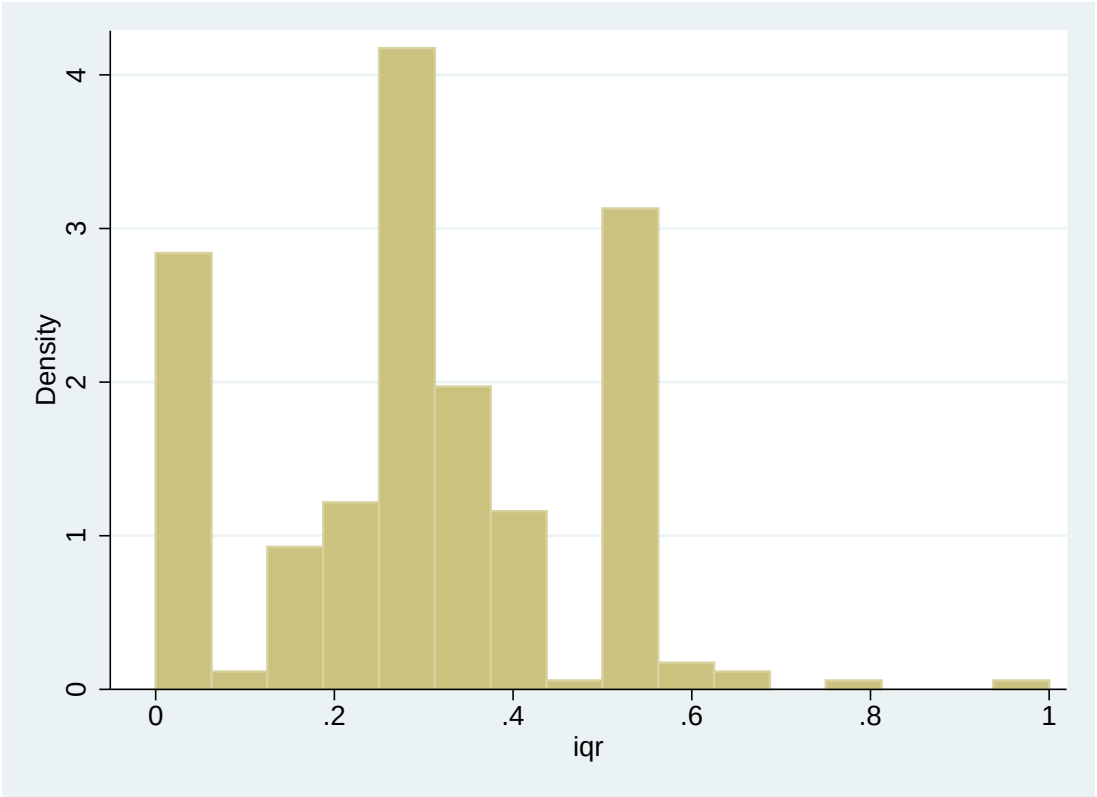
Rank variation



Within-set rank of case: max-min



Within-set rank of case: IQR



Is it a problem?

- Clearly the errors are present, and large
- If X is exact value, and Z is measurement, $E(Z|X) = X$ is classical, if eg we add some zero-mean error to the exact value
- $E(X|Z) = Z$ is Berkson error, if for *e.g.* we attribute local average to each member of a community
- "Classical" error causes point estimates to be biased, usually towards the null, Berkson error does not
- Either type biases variance estimates (we have less information than we think)

Classical Error

Estimator of β in a linear regression, with classical, independent error

$$\hat{\beta} = \frac{\langle Y, Z \rangle}{\langle Z, Z \rangle} = \frac{\langle Y, X \rangle + \langle Y, \epsilon \rangle}{\langle X, X \rangle + \langle \epsilon, \epsilon \rangle}$$
$$= \frac{\langle Y, X \rangle}{\langle X, X \rangle + \tau^2}$$

So the point estimate is biased towards the null, and the information matrix $\langle Z, Z \rangle = \langle X, X \rangle + \tau^2$ is biased upwards.

Berkson Error

Here $X = Z - \epsilon$ and $\langle Z, \epsilon \rangle = 0$, so $\langle X, \epsilon \rangle = -\tau^2$.
Also $\langle Y, \epsilon \rangle = \langle Y, X \rangle \langle X, \epsilon \rangle / \langle X, X \rangle$, so we have

$$\begin{aligned} \hat{\beta} &= \frac{\langle Y, Z \rangle}{\langle Z, Z \rangle} = \frac{\langle Y, X \rangle \left(1 - \tau^2 / \langle X, X \rangle\right)}{\langle X, X \rangle - \tau^2} \\ &= \frac{\langle Y, X \rangle}{\langle X, X \rangle} \end{aligned}$$

The point estimate is unbiased but the information matrix $\langle Z, Z \rangle = \langle X, X \rangle - \tau^2$ is biased downwards.

Regression Calibration

- In the case of classical error, the bias can be corrected given a method of estimating τ^2
- This is usually done via a "gold-standard" sub-sample
  - 24HDR in EPIC
  - Film badges for external radiation in an industrial setting
- If we believe we have a precise model of the uncertainty, that can be used
- Confidence intervals corrected using sandwich estimator

General case

- Presence of both Berkson and classical error
- Error is not independent due to shared missing data
- Require detailed model of exposure process and error sources
  - May be simple: one of two machines used at random
  - Or more complex requiring assumption of continuous prior distribution
- We can write out full likelihood

Marginal Likelihood

- Suppose by the magic of modelling, we know the distribution of the X conditional on $Z = z$ defined by the measure $dP(X|z) = f(X|z)dX$ with corresponding density
- Suppose we know the correct likelihood conditional on X:

$$\mathcal{L}(Y, x; \beta) = \sum_i L(Y_i, x_i; \beta)$$

- Suppose that we have the usual GLM form $(Y_i, x_i; \beta) = g(Y_i, \beta x_i)$ for some smooth g.

Marginal Likelihood

- Marginal Likelihood is then

$$\mathcal{L}_M(Y, z; \beta) = \log \int \exp\left(\sum_i L(Y_i, x_i; \beta)\right) \, dP(x|z).$$

- Includes special cases:
  - If Z is a proxy measure of X such that $E(X|z) = z$ this is Berksonian error;
  - If $dP(X|z) = \prod_i dP_i(X_i|z)$, error uncorrelated error, could consider regression calibration.

Marginal Likelihood

- Expected value of score is still zero
- Covariance matrix is still estimated by $-H^{-1}$
- Normality of estimators **does not** hold.

Approximation

- Define $\tilde{x}_z = E(X|z)$, or usually $\tilde{x}$.
- Form an *imputed likelihood*

$$\mathcal{L}_I(y, z; \beta) = \sum_i L(y_i, \tilde{x}_i; \beta)$$

- Let this be maximized at $\hat{\beta}_\sim$, so that $0 = \sum_i L_{,\beta}(y_i, \tilde{x}_i, \hat{\beta}_\sim)$.
- Expand around $\mathcal{L}_M(Y, z; \beta)$ around $(\tilde{x}, \hat{\beta}_\sim)$.

Approximation

$$
\begin{aligned}
L(y_i, x_i, \beta) \sim\ & L(y_i, \tilde{x}_i, \hat{\beta}_\sim) + (x_i - \tilde{x}_i) L_{i,x}(y_i, \tilde{x}_i, \hat{\beta}_\sim) \\
& + (\beta - \hat{\beta}_\sim) L_{,\beta}(y_i, \tilde{x}_i, \hat{\beta}_\sim) \\
& + \frac{1}{2}(x_i - \tilde{x}_i)^2 L_{,xx}(y_i, \tilde{x}_i, \hat{\beta}_\sim) \\
& + \frac{1}{2}(\beta - \hat{\beta}_\sim)^2 L_{,\beta\beta}(y_i, \tilde{x}_i, \hat{\beta}_\sim) \\
& + (x_i - \tilde{x}_i)(\beta - \hat{\beta}_\sim) L_{,x\beta}(y_i, \tilde{x}_i, \hat{\beta}_\sim) + R_2(y_i, x_i, \beta)
\end{aligned}
$$

Approximation

After summing over observations and integrating, the first β derivative vanishes

$$\mathcal{L} = \log \int \exp\left(\sum_i (A_i + (x_i - \tilde{x}_i)B_i + (x_i - \tilde{x}_i)^2 C_i\right) \, dP(x|z)$$

$$= \sum_i A_i + \frac{1}{2} B^T \operatorname{Var}_P(X) B + \operatorname{diag}(\operatorname{Var}_P(X)) \cdot C$$

$$+ \log \int O((x - \tilde{x})^3) \, dP(x|z).$$

Approximation

where

$$\sum_i A_i = \mathcal{L}(Y, \tilde{x}, \hat{\beta}_\sim) + \frac{1}{2}(\beta - \hat{\beta}_\sim)^2 \mathcal{L}_{,\beta\beta}(Y, \tilde{x}, \hat{\beta}_\sim);$$

$$\sum_i B_i = \mathcal{L}_{,x}(Y, \tilde{x}, \hat{\beta}_\sim) + (\beta - \hat{\beta}_\sim)\mathcal{L}_{,x\beta}(Y, \tilde{x}, \hat{\beta}_\sim);$$

$$\sum_i C_i = \mathcal{L}_{,xx}(Y, \tilde{x}, \hat{\beta}_\sim).$$

Approximation

- A is the second order expansion of the imputed likelihood, the limiting where $P(X|z)$ is concentrated on a single exposure vector.
- If both $\mathcal{L}_{,x}(Y, \tilde{x}, \hat{\beta}_{\sim})$ and $\mathcal{L}_{,x\beta}(Y, \tilde{x}, \hat{\beta}_{\sim})$ are non-zero, then $\frac{\partial}{\partial\beta}\mathcal{L}_M(Y, z)|_{\beta=\hat{\beta}_{\sim}} \neq 0$.
- Consequently, MCML does not give the same point estimate as regression calibration even if the errors are uncorrelated: the estimate "pulls" x toward a better fitting value.

Monte-Carlo Marginal Likelihood

Implausible that we can evaluate the marginal likelihood analytically. Instead, make uniform draws from $P(X)$ and use a Monte-Carlo approximation:

$$\begin{aligned}\mathcal{L}_{MC}(Y, z; \beta) &= \frac{1}{s} \sum_{k}^{s} \mathcal{L}(Y, x^k; \beta) \\ &= \frac{1}{s} \sum_{k}^{s} e^{\sum_i L(Y_i, x_i^k; \beta)}.\end{aligned}$$

Monte Carlo Likelihood: history

- Proposed by Duncan Thomas, Stram, Dwyer (Annu Rev Pub Health 1993). More details by Stram, Kopecky (Radiation Research 2003).
- Applied to Chernobyl thyroid data by Cardis *et. al.* (JNCI 2005), fitting a 1-parameter logistic model.
- Cardis *et. al.* (Radiat Res. 2007; 168: 757-763) estimated the CI's by the Likelihood Ratio Test and shot-gun searching.
- Fearn, Hill and Darby (*op. cit.*) investigating household radon exposure split the data into independent risk-sets.

Shotgun?

- Cardis *et. al.* evaluated the likelihood on a grid of values of β, then did a second stage with finer spacing.
- Coded in Stata, extremely slow. Recoded in Fortran, still slow.
- The likelihood was skewed around the maximum. Almost symmetric using log-dose.
- Fearn *et. al.* were obliged to also use a grid, since maximum of sum is not the sum of the maxima...

Monte Carlo Likelihood: issues

- $E(D \ln \mathcal{L}_{MC}) = 0$, if it exists. Newton-Raphson code starting from the central estimate converges reliably on simulated data after removing extreme dose estimates ($>$ 10Gy to 98Gy);
- Log-Likelihood is not the sum of IID contributions, so CLT does not apply, so quadratic approximation is not guaranteed;
- LRT is "generally considered" more reliable than Wald test in such circumstances;
- But what is the null distribution? Not addressed in any of the above papers.

What should we expect?

- Within each draw: CLT applies, approximately χ^2
- Exponentiate and sum. Does the CLT apply?
- Experimentally, if the draws are independent, no convergence.
- Density of exponentiated χ^2_1 distribution is

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{x^{3/2}\sqrt{\ln(x)}}.$$

- Need some dependence

Go Bayesian?

- Still need to evaluate the integral over a range of values to convolve with prior on β.
- May be faster than grid-search of β values
- Unlikely to be significantly faster than properly coded search algorithm
- In any case, issues of convergence and precision of the numerical integral still apply.

Sampling Density: test case

Try estimating the area under the standard normal density

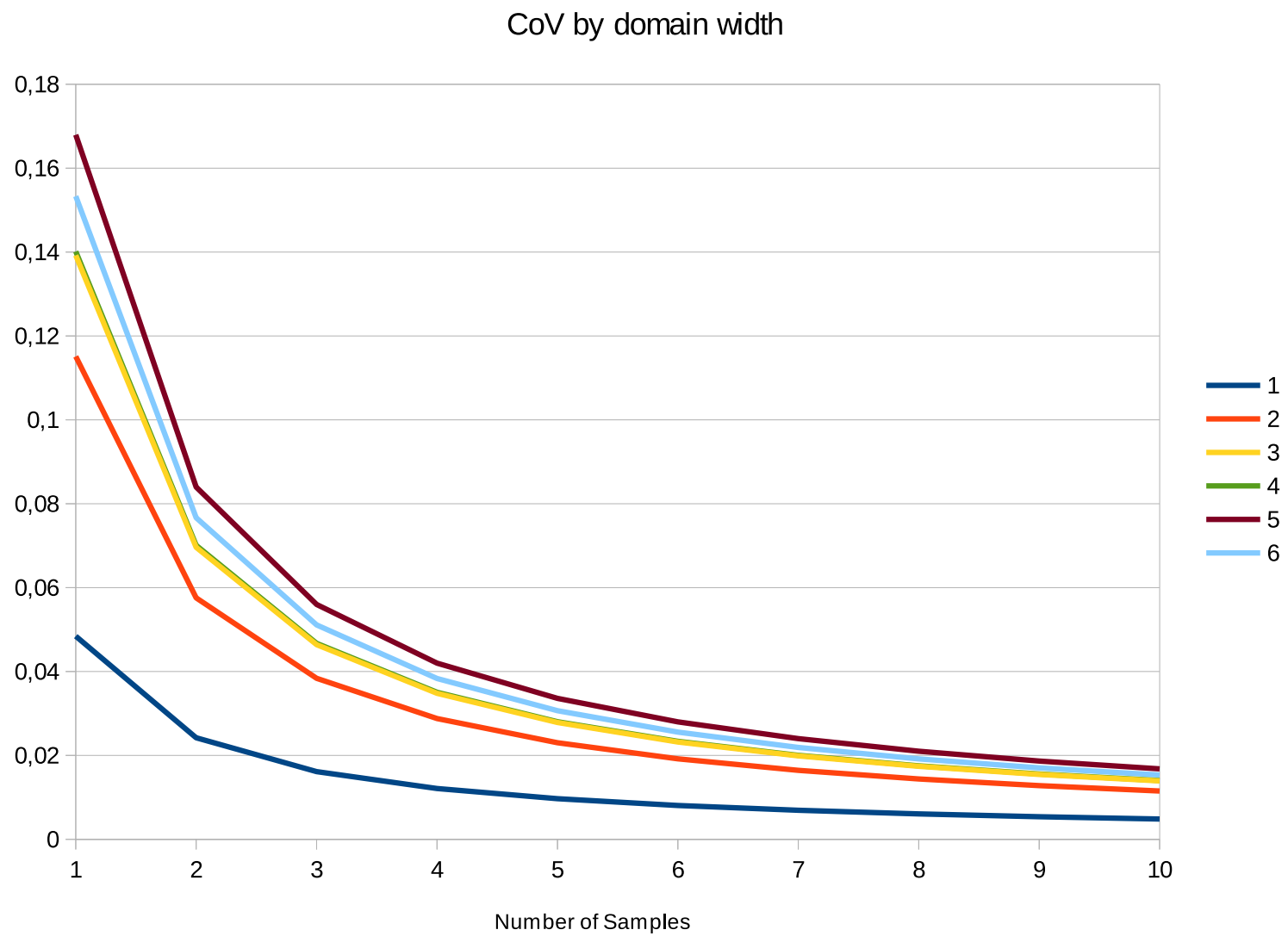
$$I_s = \frac{2a}{s} \sum_{j=1}^{s} \phi(x_j), \; x_j \sim U(-a, a)$$

Unbiased, but variance depends on a and s

$$\begin{aligned} \mathrm{Var}(I_s) &= \frac{2a}{s} \left(\int_{-a}^{a} \phi^2(x)\, dx - (EI)^2 \right) \\ &= \frac{a}{s} \left(\frac{2\Phi(a\sqrt{2}) - 1}{\sqrt{\pi}} - (2\Phi(a) - 1)^2 \right). \end{aligned}$$

Coeff of Variation

Determined by number of samples under the “fat part” of the density



Integral Transform View

- Can think of the integral over $P(x)$ as an iterated integral transform (approximately Laplace)
- Maps from density function on $x \in \mathbb{R}^n$ to posterior on $(\beta_1, \ldots, \beta_n)$
- Then we restrict to the sub-domain $\beta_1 = \cdots = \beta_n = \beta$.

Sequential Transforms

Consider the partial transform

$$F^{(k)}(\beta_1, \ldots, \beta_k) = \int_{\mathbb{R}^k} e^{\sum_i^k L(Y_i, x_i; \beta_i)} \, dP(x),$$

so that

$$F^{(k+1)}(\beta_1, \ldots, \beta_{k+1}) = \int_{\mathbb{R}} e^{L(Y_{k+1}, x_{k+1}; \beta_{k+1})} F^{(k)} \, dP(x_{k+1} | x_1, \ldots, x_k).$$

The concern is that we will "lose" some proportion of samples at each iterate.

Sample scaling

- In EPI-CT where $n \approx 10^6$ this could easily result in retaining finally only a single draw from the dose-set distribution.
- If this proportional loss remains constant, we would require the number of initial draws to increase exponentially with n.

Curse of dimensionality...

- Draw randomly from a multi-variate standard normal distribution $N(0, I_n)$
- Squared distance from the origin is distributed as χ^2_n, with expectation n and variance $2n$.
- As n increases, sampled points will cluster tightly about a shell of radius $\sqrt{n}$.
- Very difficult to draw a sample near the maximum!

Close our eyes?

- It may be tempting to say that worrying about measurement error is too difficult for large n.
- However for large n, $\mathrm{Var}(\hat{\beta}) \longrightarrow 0$ if we ignore measurement error. With error, it does not.
- Measurement error is the dominant source of uncertainty for sufficiently large n.

Thanks

- Ausra Kesminiene
- Deukwoo Kwon
- Elizabeth Cardis