Statistical Disclosure Control Using the ϵ-uncertainty Intervals and the Grouped Likelihood Method

Jinfang Wang

Department of Mathematics and Informatics, Chiba University, Japan

Contents

1. ϵ-Random Uncertainty Intervals

   (Wang (2002))

2. Disclosure of Real Data

   (Box-Cox Transformation + Wang (2002))

3. Analysing Published Interval Data

   (Grouped likelihood)

Contents

1. ϵ-Random Uncertainty Intervals

   (Wang (2002))

2. Disclosure of Real Data

   (Box-Cox Transformation + Wang (2002))

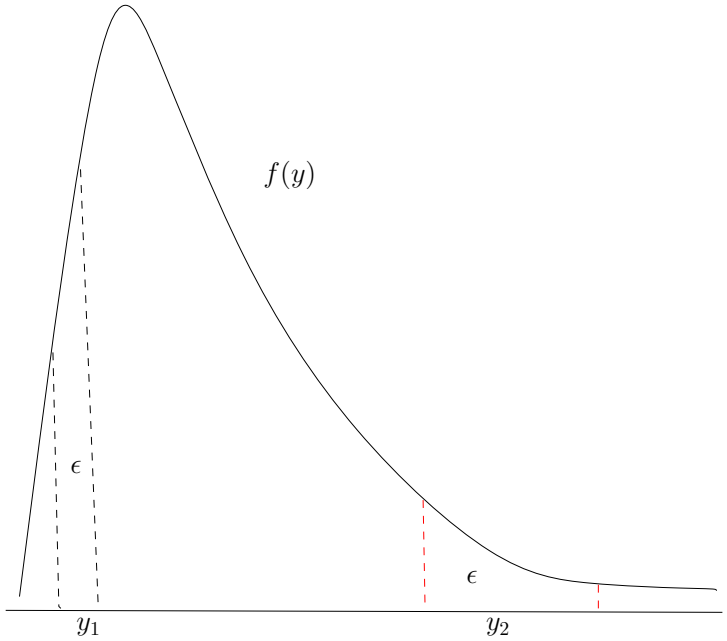
3. Analysing Published Interval Data

   (Grouped likelihood)

ϵ-Random Uncertainty Intervals

DEFINITION 1 (Wang, 2002). *Let $f(y)$ be the p.d.f. of y. Let $0 < \epsilon < 1/2$. An ϵ-uncertainty interval of y, $i_\epsilon(y)$, is any interval containg y and satisfying*

$$\int_{y \in i_\epsilon(y)} f(y)\, dy = \epsilon$$

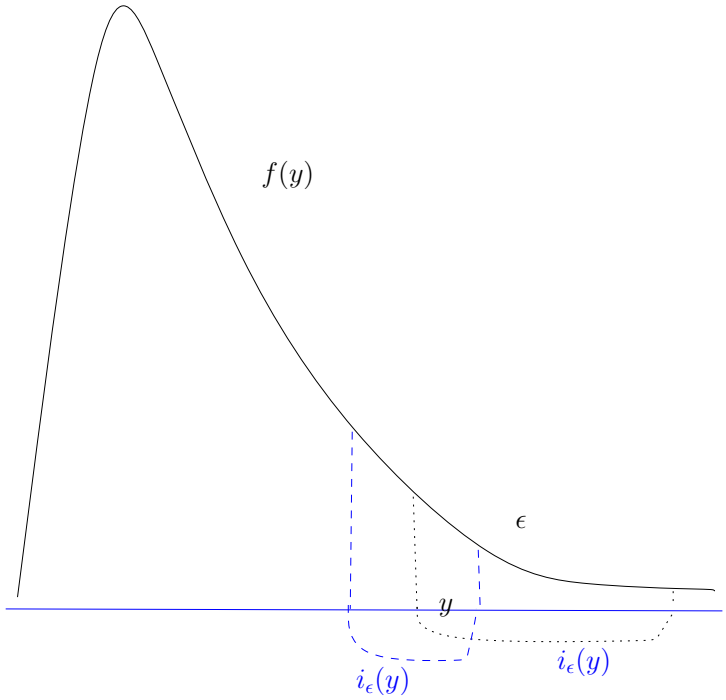


ϵ-uncertainty Interval Space

DEFINITION 2 (Wang, 2002). *The set of all ϵ-uncertainty intervals of y, $i_\epsilon(y)$*

$$I_\epsilon(y) = \{i_\epsilon(y) | i_\epsilon(y) : \epsilon\text{-uncertainty interval of } y\}$$

is called the ϵ-uncertainty interval space of y.



Representation of an ϵ-uncertainty IS

THEOREM 1. *Let $F(y)$ be the cumulative distribution function of Y. Let $0 < \epsilon < 1/2$. Suppose that y is in the interior of the support of $F(y)$. Then there exists a 1-1 correspondence between the uncertainty interval space $I_\epsilon(y)$ and the interval $[y_\epsilon^-, y_\epsilon^+]$, which is defined as follows.*

(i) If $F(y) \leq \epsilon$, then define

$$[y_\epsilon^-, y_\epsilon^+] = [F^{-1}(\epsilon), F^{-1}(F(y) + \epsilon)]\,;$$

(ii) If $F(y) > \epsilon$, then define

$$[y_\epsilon^-, y_\epsilon^+] = [F^{-1}(F(y) - \epsilon), \min\{y, F^{-1}(1 - \epsilon)\}]$$

Representation of ϵ-uncertainty IS, ctd.

DEFINITION 3 (representative interval). *The interval* $[y_\epsilon^-, y_\epsilon^+]$ *is called the representative interval of the* ϵ*-uncertainty interval space of* y.

REMARK 1. *Each point in the representative interval* $[y_\epsilon^-, y_\epsilon^+]$ *corresponds to the lower end of the interval space* $I_\epsilon(y)$.

Disclosure Policy

- Disclosure policy:

  DEFINITION 4 (Disclosure Policy). *For a given ϵ, a disclosure policy is a predetermined probability distribution over $[y_\epsilon^-, y_\epsilon^+]$ for each datum y.*

- Remark: In the rest of the paper, we shall use the uniform distribution on each $[y_\epsilon^-, y_\epsilon^+]$.

Contents

1. ϵ-Random Uncertainty Intervals

   (Wang (2002))

2. Disclosure of Real Data

   (Box-Cox Transformation + Wang (2002))

3. Analysing Published Interval Data

   (Grouped likelihood)

Form of the Data

- We consider data having the following form

$$\mathcal{D} = (\boldsymbol{y}, \boldsymbol{x}) = \begin{pmatrix} y_1 & x_{11} & \cdots & x_{1p} \\ y_2 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ y_n & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- y_j are scalar variables, which are sensitive for publication.
- $\boldsymbol{x}$ are covariates, which are not sensitive for publication.

A Fundamental Assumption

- We assume that there exists a λ so that

$$z(y_j, \lambda) \sim N(\boldsymbol{x}_j'\beta, \sigma^2)$$

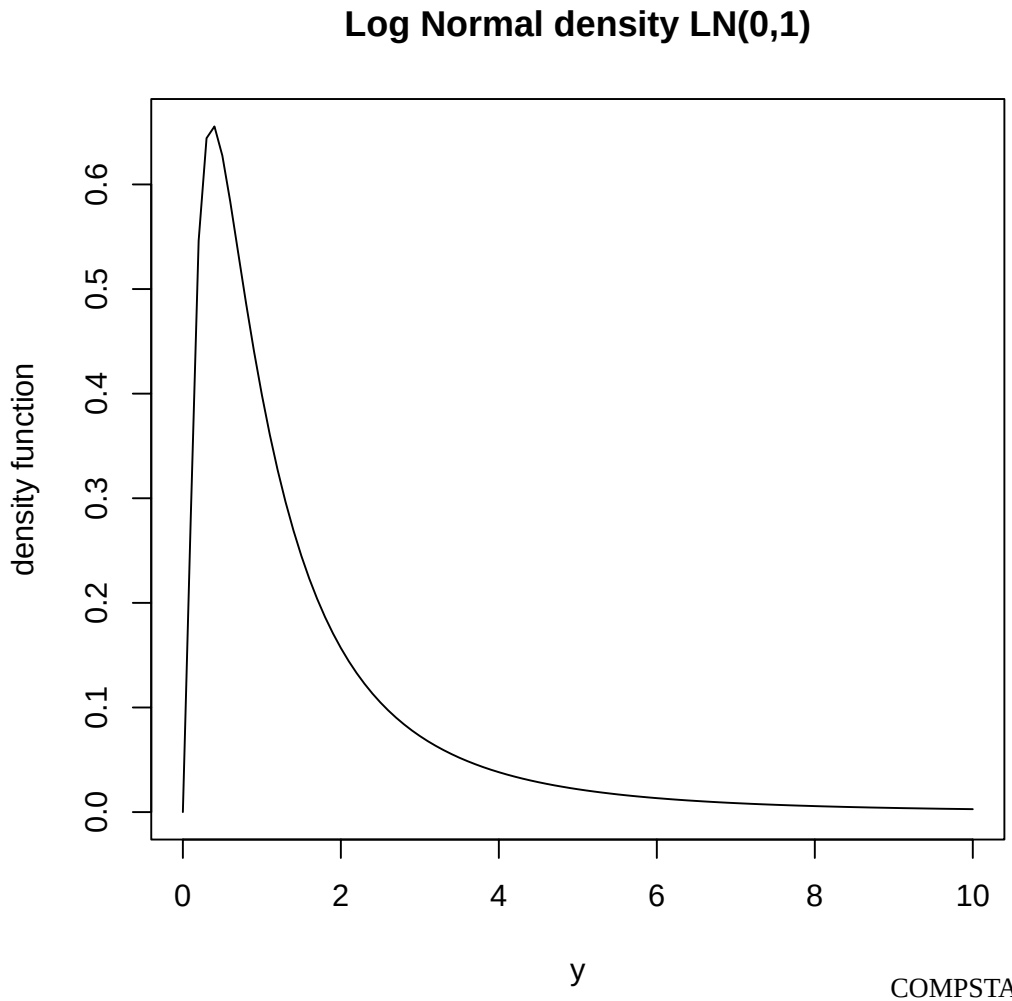
where $z(y_j, \lambda)$ is the Box-Cox transformation

$$z(y_j, \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0 \\ \log y, & \text{if } \lambda = 0 \end{cases}$$

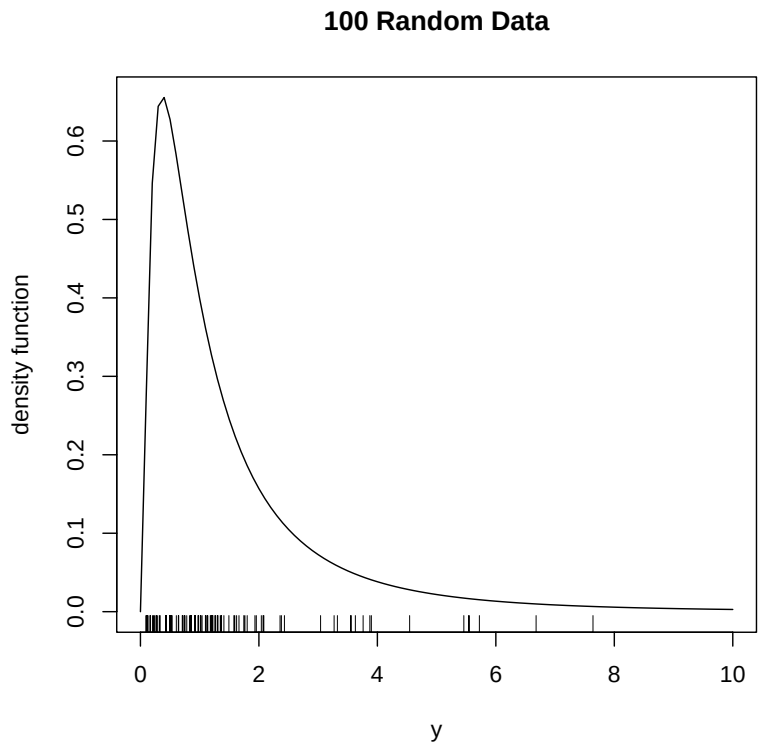
- For simplicity, I assume that $\mu = \boldsymbol{x}_j'\beta$
- Generalization to $\mu_j = \boldsymbol{x}_j'\beta$ is immediate.

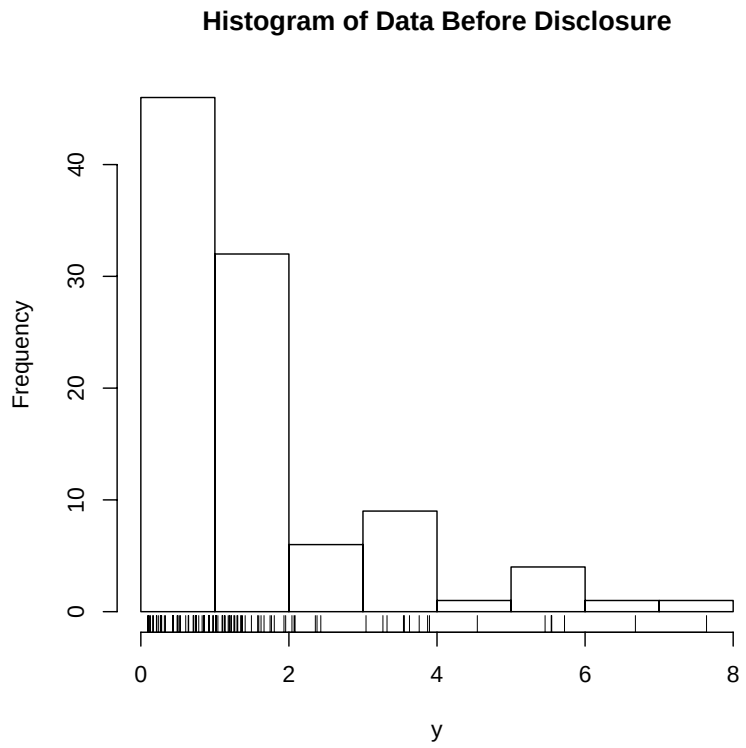
The Log Normal Population

- Assume that $Z = \log(Y) \sim N(\mu, \sigma^2)$
- Density function with $(\mu, \sigma^2) = (0, 1)$:



Artificial Data Before Publication





ML Estimators of β, σ^2 and λ

- The log-likelihood for β and σ^2 conditional on λ:

$$\begin{aligned} \ell &= \sum_{j=1}^{n} \log \phi(z(y_j, \lambda); \boldsymbol{x}_j'\beta, \sigma^2) \\ &= -\frac{1}{2\sigma^2} \sum_{j=1}^{n} \left(z(y_j, \lambda) - \boldsymbol{x}_j'\beta\right)^2 - n \log(\sqrt{2\pi}\sigma) \end{aligned}$$

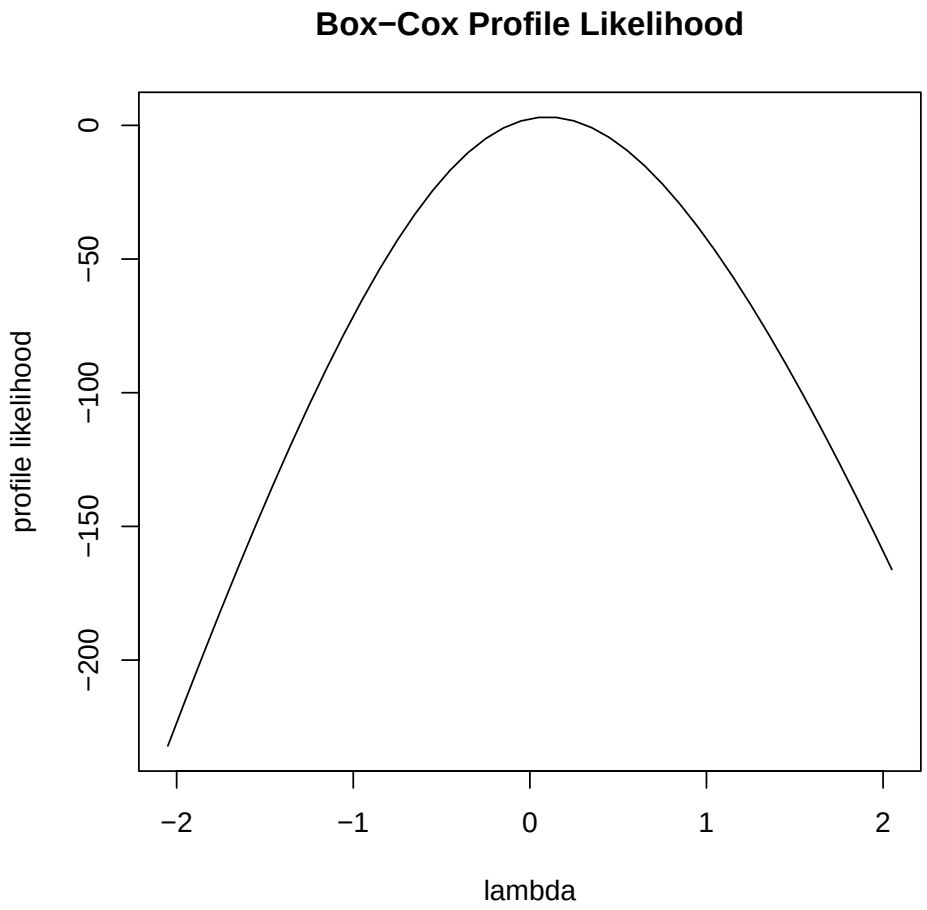
- The profile log-likelihood for λ:

$$\ell_p(\lambda) = -\frac{n}{2} \log(\hat{\sigma}^2(\lambda)) + (\lambda - 1) \sum_{j=1}^{n} \log z(y_j, \lambda) + C$$

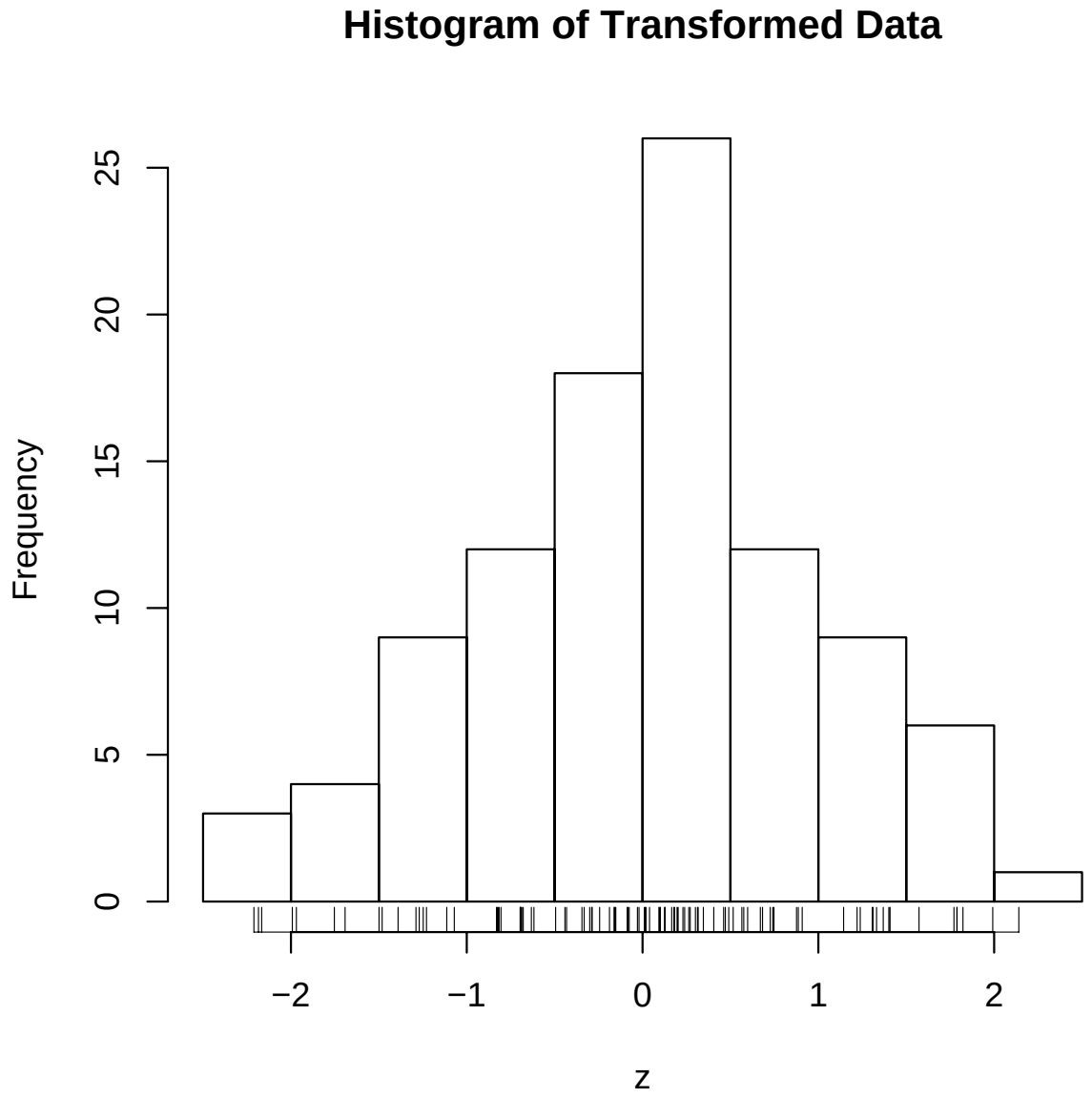
Box-Cox Profile Likelihood

The profile log-likelihood for λ:

$$\ell_p(\lambda) = -\frac{n}{2}\log(\hat{\sigma}^2(\lambda)) + (\lambda - 1)\sum \log(z_j) + C$$



Box-Cox Transformed Data



Computing Percentiles in the y-scale

- Estimated density function of y

$$\hat{f}(y) = \phi\left(z(y,\hat{\lambda});\, \boldsymbol{x}_j'\hat{\beta}, \hat{\sigma}\right) |\partial\, z(y,\hat{\lambda})/\partial y|$$

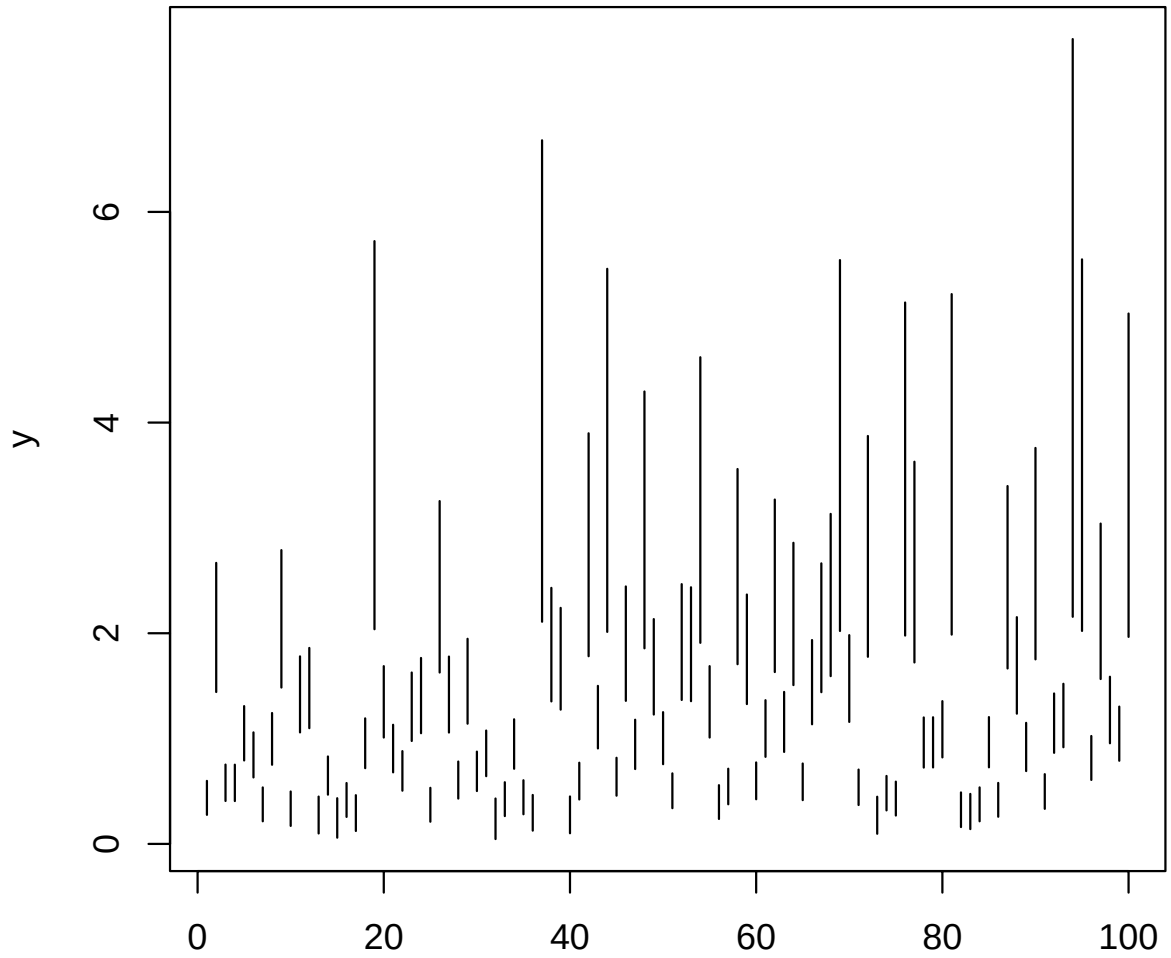
- For a given α we can find the 100αth percentile of $F(y)$ through the relation

$$z(y,\lambda) = \hat{\sigma}\Phi^{-1}(\alpha) + \boldsymbol{x}_j'\hat{\beta}$$

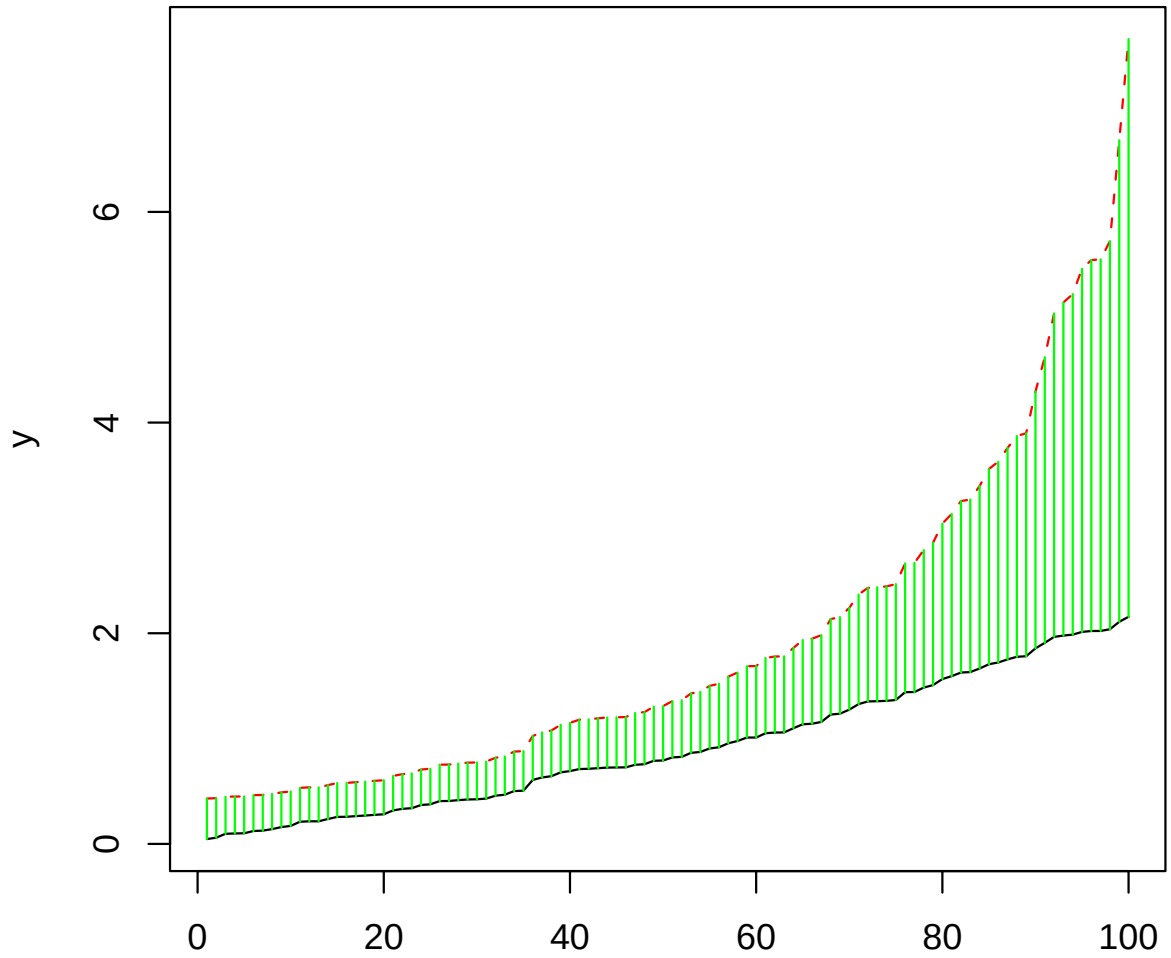
and the inverse of the Box-Cox transformation $z = z(y,\hat{\lambda})$ is given by

$$y = (\hat{\lambda}z + 1)^{1/\hat{\lambda}}$$

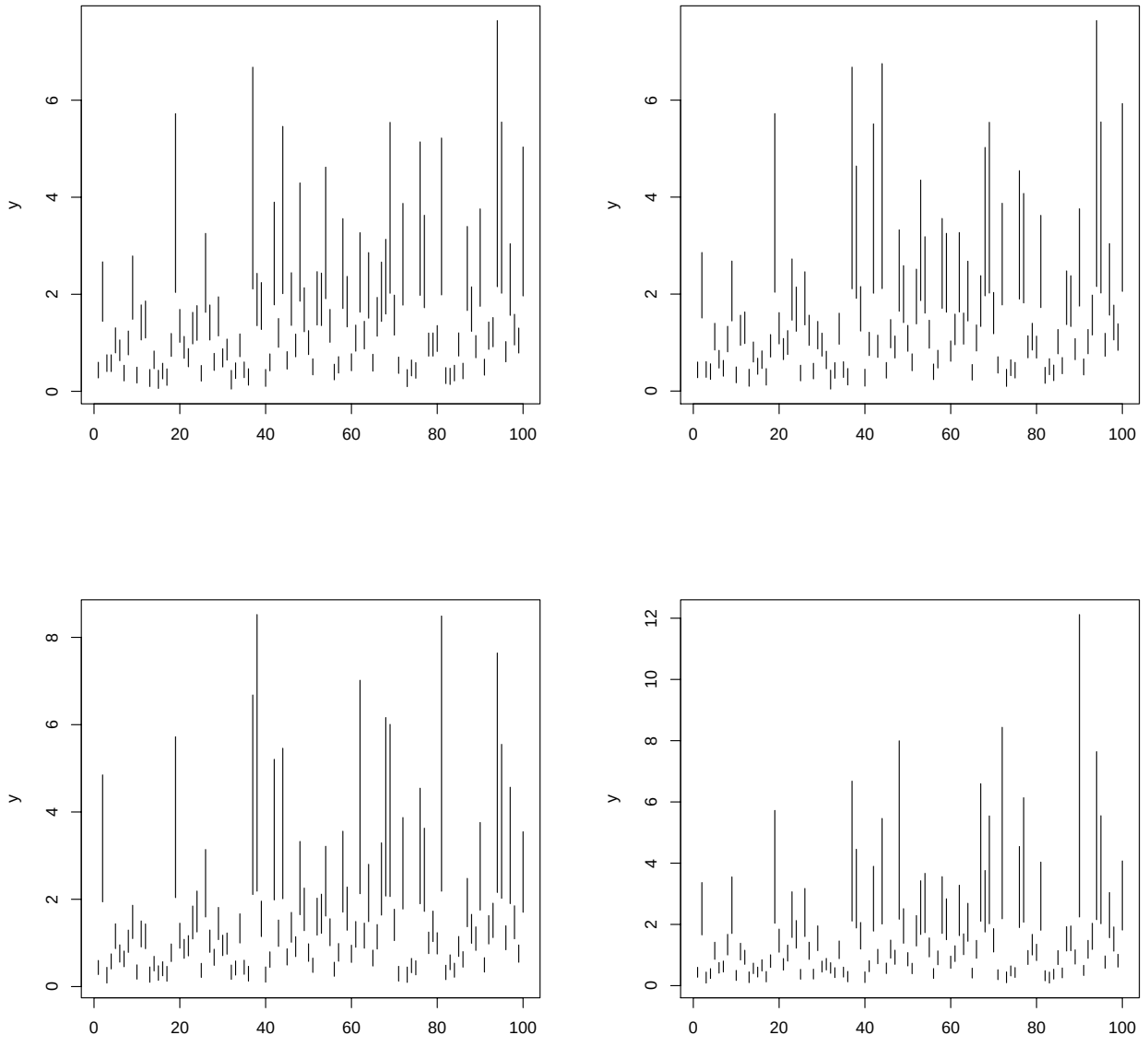
20%-uncertainty Intervals:1/2



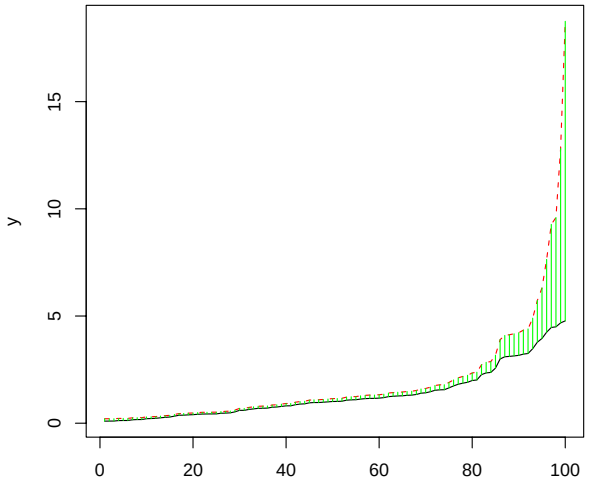
20%-uncertainty Intervals:2/2

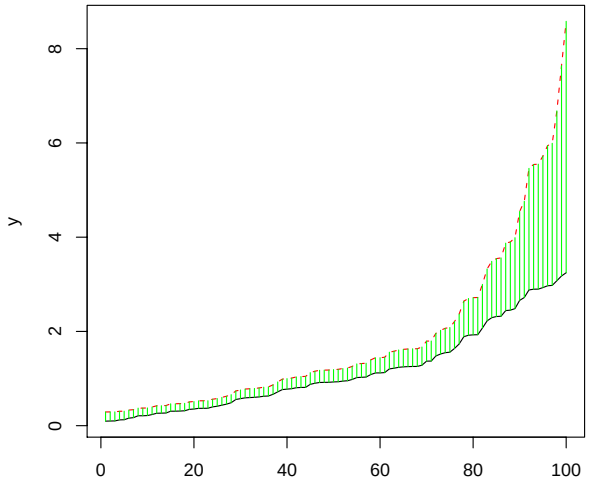


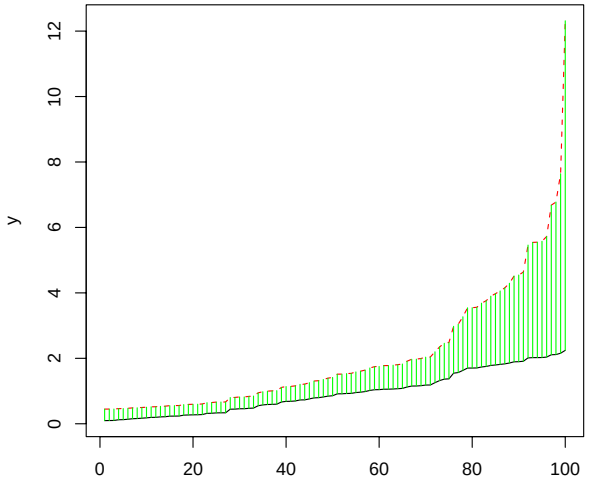
20%-uncertainty In.: disclosed 4 times

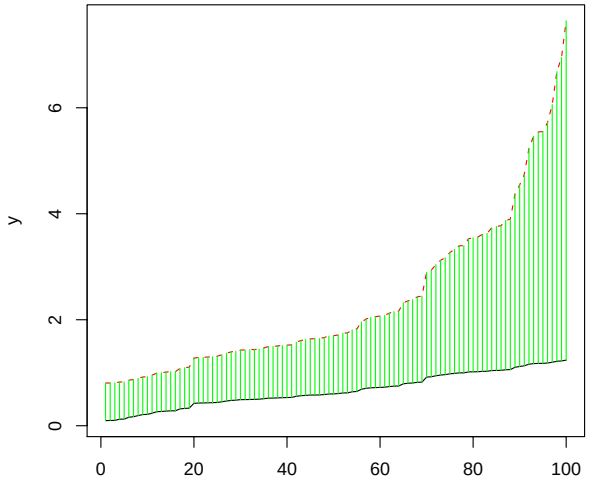


Uncertainty Intervals at Various Levels









Contents

1. ϵ-Random Uncertainty Intervals

   (Wang (2002))

2. Disclosure of Real Data

   (Box-Cox Transformation + Wang (2002))

3. Analysing Published Interval Data

   (Grouped likelihood)

The Published Data

- Data published in the following form

$$\mathcal{D}_p = \begin{pmatrix} (y_1^-, y_1^+) & x_{11} & \cdots & x_{1p} \\ (y_2^-, y_2^+) & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ (y_n^-, y_n^+) & x_{n1} & \cdots & x_{np} \end{pmatrix}$$

- Inferential goal: estimates of β, σ and λ

The Grouped Likelihood Method

- Literature: Barnard (1965), Kempthorne (1966) and Giesbrecht and Kempthorne (1976), Atkinson et al. (1991) and Nadal and Pericchi (1998)

  proposed mainly for irregular parametric problems

The Grouped Likelihood Method, ctd.

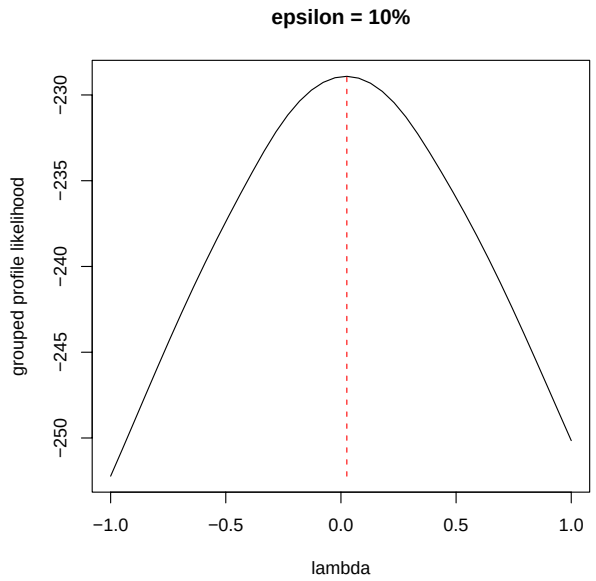
- Let $\Phi_0(c) = \begin{cases} 1 - \Phi(c) & \lambda > 0 \\ 1 & \lambda = 0 \\ \Phi(c) & \lambda < 0 \end{cases}$
- Contribution of (y_j^-, y_j^+) to the likelihood:

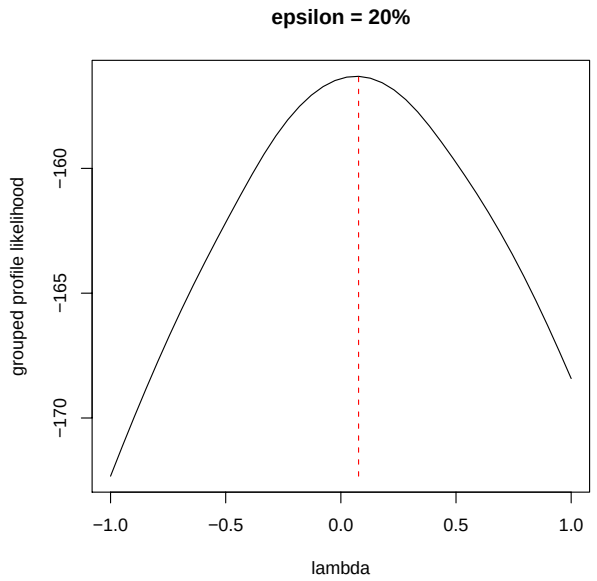
$$p_j = \frac{\Phi\left(\frac{z(y_j^+,\lambda)-\boldsymbol{x}_j'\beta}{\sigma}\right) - \Phi\left(\frac{z(y_j^-,\lambda)-\boldsymbol{x}_j'\beta}{\sigma}\right)}{\Phi_0(-(\lambda\boldsymbol{x}_j'\beta+1)/\lambda\sigma)}$$

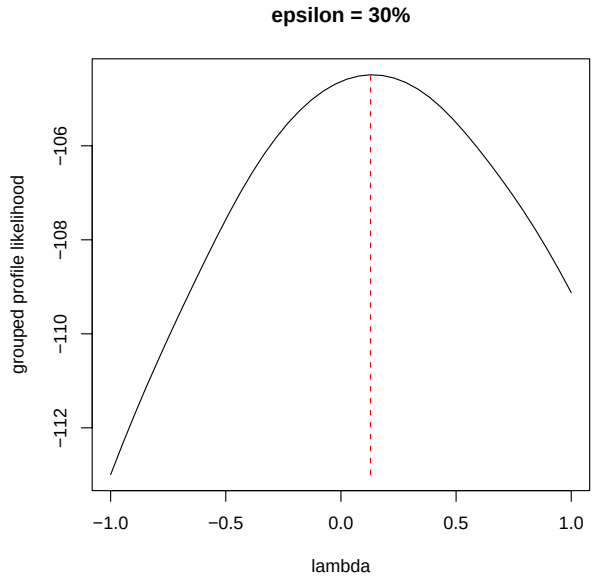
- The grouped log-likelihood:

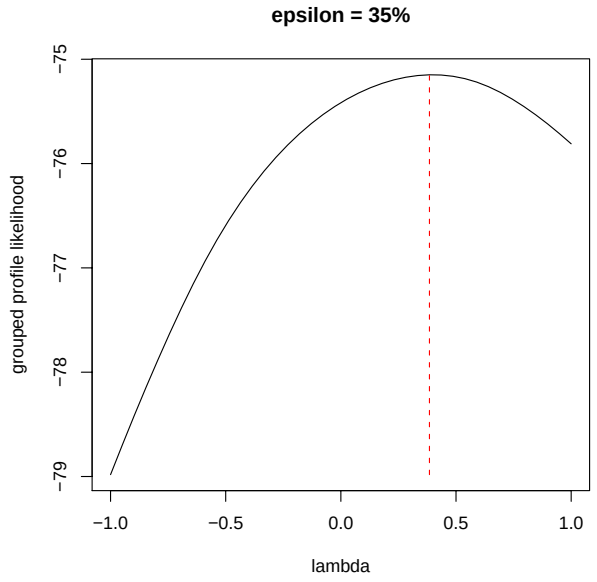
$$\ell = \sum_{j=1}^{n} \log\left\{p_j/(y_j^+ - y_j^-)\right\} = \sum_{j=1}^{n} \log\{p_j\} - \sum_{j=1}^{n} \log\left\{y_j^+ - y_j^-\right\}$$

Grouped Profile Likelihoods

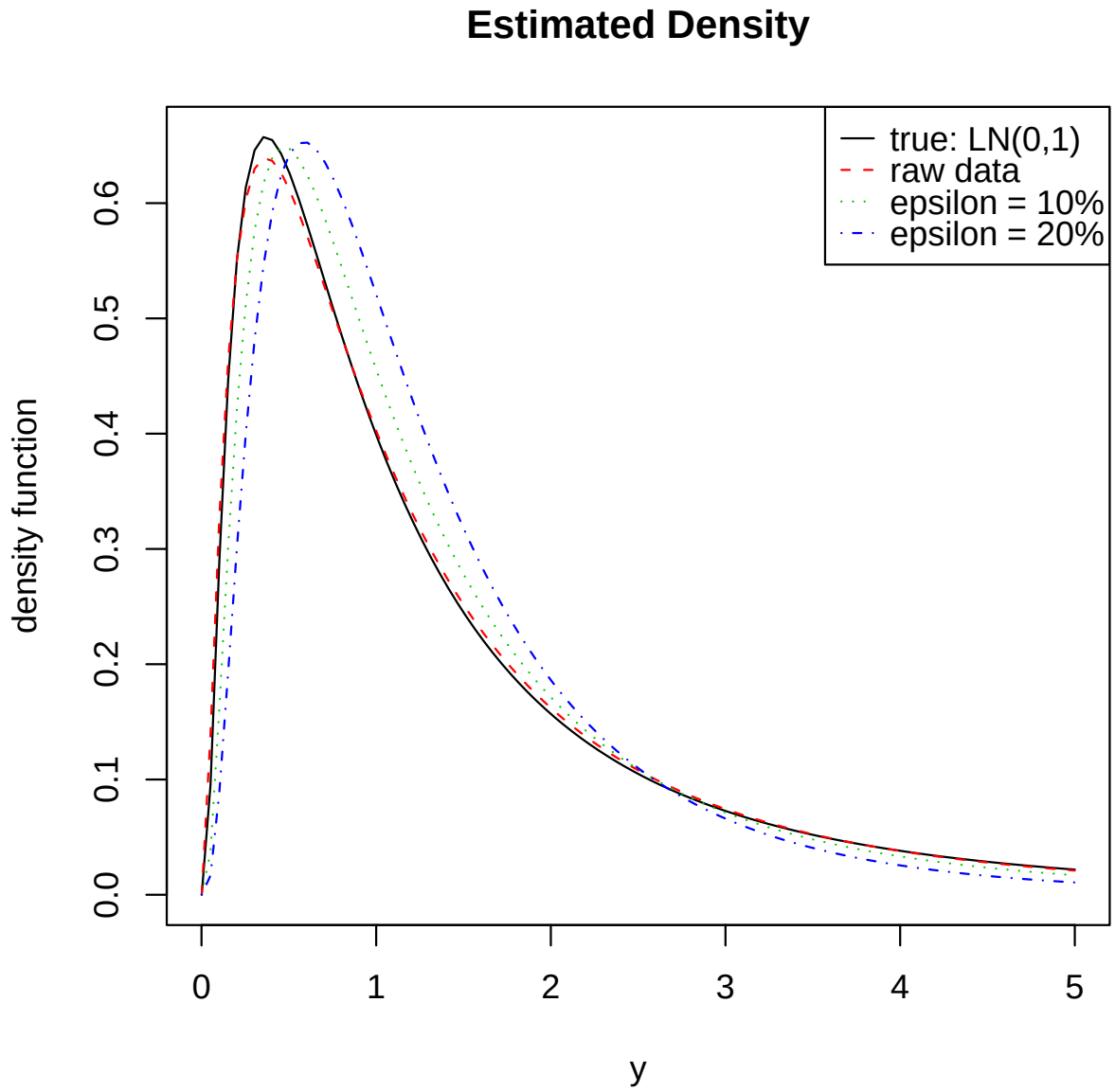








Density Estimation



ϵ-asymptotics of Density Estimation