UNSW SYDNEY

Estimating Equations for data summaries

Tom Whitaker, Boris Beranger, Scott A. Sisson,

UNSW & ACEMS

New Perspectives in Data Science
IHP, 6th March, 2020

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

1987: **Edwin**'s first SDA paper and . . .**I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

1987: **Edwin**'s first SDA paper and . . . **I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

1987: **Edwin**'s first SDA paper and . . .**I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

1987: **Edwin**'s first SDA paper and . . .**I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

1987: **Edwin**'s first SDA paper and . . .**I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Edwin Diday

SDA Seminal contributor
Citations: 9282
H-index: 41

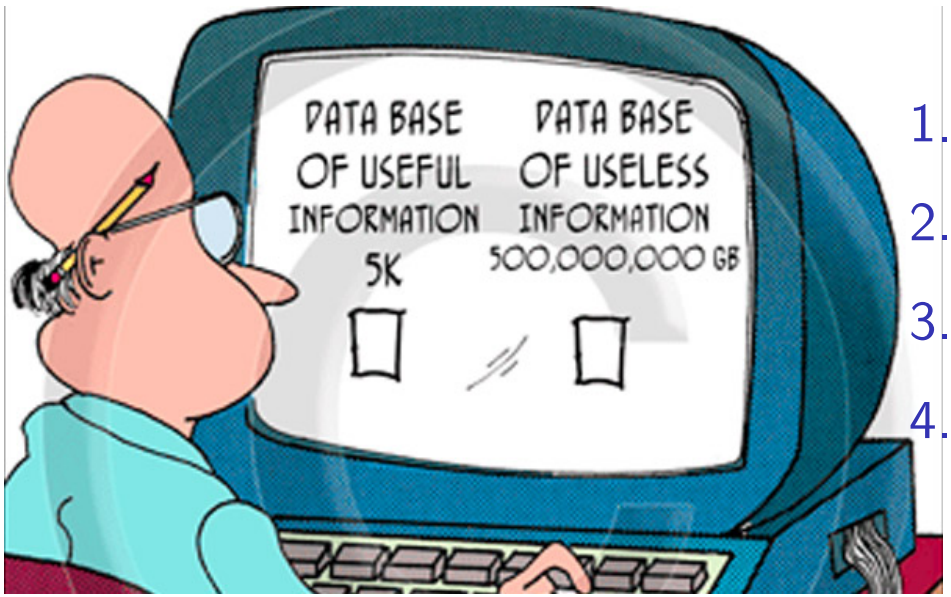
1987: **Edwin**'s first SDA paper and . . .**I** was born!

2016: I started research in SDA

2017: **Edwin** and **I** met at the SDA workshop in Ljubljana

2020: **HAPPY BIRTHDAY EDWIN!**

Talk Outline



1. Background Information
2. Estimating Equations for symbolic data
3. Examples
4. Discussion

Background Information (1)

- Let $X = (X_{[1]}, \ldots, X_{[D]}) \in \mathcal{D}_X \subset \mathbb{R}^D$ with d.f. F_X (unknown).
- Let $\boldsymbol{X} = (X_1, \ldots, X_N)$ be the collection of N i.i.d. replicates of X with realisation given by $\boldsymbol{x} = (x_1, \ldots, x_N)$

Without any parametric assumption about F_X:

Make statistical inference on $\theta \in \mathcal{D}_\theta \subset \mathbb{R}^M$ using $R \geq M$ functionally independent **estimating equations (EE)**:

$$g(X, \theta) = (g_1(X, \theta), \ldots, g_R(X, \theta))^\top,$$

with the condition

$$\mathbb{E}_{F_X}[g_r(X, \hat{\theta})] = 0, \text{ for all } r = 1, \ldots, R.$$

Background Information (2)

- No assumption on $F_X \Rightarrow$ Empirical alternative

The **empirical likelihood (EL)** associated with the observed sample $\boldsymbol{x}$ is

$$L(F_X; \boldsymbol{x}) = \prod_{n=1}^{N} \mathrm{d}F_X(x_n) = \prod_{n=1}^{N} P(X = x_n).$$

F_X can be seen as a discrete distribution on $\{x_1, \ldots, x_N\}$ with probability vector $p = (p_1, \ldots, p_N)$ defined such that:

(C1): $p_n = P(X = x_n) > 0,\ n = 1, \ldots, N.$
(C2): $\sum_{n=1}^{N} p_n = 1$

If there are no other conditions on $\boldsymbol{x}$ other than **(C1)** and **(C2)** then

$$\hat{p}_n = \underset{p_n | C1, C2}{\arg\max} \prod_{n=1}^{N} p_n = \frac{1}{N},$$

and the estimating equation becomes:

$$\frac{1}{N} \sum_{n=1}^{N} g_r(x_n, \hat{\theta}) = 0, \text{ for all } r = 1, \ldots, R.$$

Background Information (3)

Example:

Assume we are interested in estimating the **mean** and **variance**, i.e. for $M = 2$:

$$\theta = (\mu, \sigma^2) = (\mathbb{E}(X), \mathbb{V}(X)).$$

A natural choice of estimating function with $R = 2$ is then

$$g(X, \theta) = \left(X - \mu, (X - \mu)^2 - \sigma^2\right),$$

for which the estimating equations give

$$\begin{cases} \frac{1}{N} \sum_{n=1}^{N} (x_n - \hat{\mu}) = 0 \\ \frac{1}{N} \sum_{n=1}^{N} \left((x_n - \hat{\mu})^2 - \hat{\sigma}^2\right) = 0 \end{cases}$$

$$\Leftrightarrow \begin{cases} \hat{\mu} = \frac{1}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{x_n} \\ \hat{\sigma}^2 = \frac{1}{\mathbf{N}} \sum_{\mathbf{n=1}}^{\mathbf{N}} (\mathbf{x_n} - \hat{\mu})^2 \end{cases}$$

Symbolic Data

For a class $c = 1, \ldots, C$, let $\boldsymbol{X}^{(c)} = \{X_1^{(c)}, \ldots, X_{n_c}^{(c)}\}$ denote the c-th subset of $\boldsymbol{X}$ of size n_c s.t.

$$\begin{cases} \bigcup_{c=1}^{C} \boldsymbol{X}^{(c)} = \boldsymbol{X} \\ \sum_{c=1}^{C} n_c = N \end{cases}$$

The *Symbolic object* $S_c \in \mathcal{D}_{S_c}$ is a summary of the information contained in $\boldsymbol{X}^{(c)}$, obtained through an aggregation function $\pi(\cdot)$ which may contain some deterministic elements ϑ

For observations $\boldsymbol{x}^{(c)}$, the information is summarised in $s_c = \{n_c, \Upsilon_c, \alpha_c\}$, where $\Upsilon_c = \mathcal{D}(\boldsymbol{X}^{(c)})$ and α_c contains the summary statistics specific to $\pi(\cdot)$.

Estimating Equations for Symbolic Data (1)

- Let $\phi_c := f_{X|S_c=s_c}$ denote the density of X given a summary s_c.
- Let $\phi := f_{X|\boldsymbol{S}=\boldsymbol{s}}$ denote the density of X given the set of all summaries $\boldsymbol{s}$.

- These densities are linked through

$$\phi_c(x) = \frac{\mathbb{1}(x \in s_c)\phi(x)}{\mathbb{P}(s_c)},$$

where the indicator restricts to the c-th symbol and the denominator corresponds its probability of occurrence $\mathbb{P}(s_c) = \int \phi(y)\mathbb{1}\{y \in s_c\}\mathrm{d}y$.

- **Example:** If classes/symbols are independent then we can take $\mathbb{P}(s_c) = \frac{n_c}{N}$ such that $\sum_{c=1}^{C} \mathbb{P}(s_c) = 1$ which yields

$$\phi(x) = \sum_{c=1}^{C} \frac{n_c}{N}\phi_c(x).$$

Estimating Equations for Symbolic Data (2)

We define the **estimating equations for symbolic inputs** as

$$\mathbb{E}_{F_S}[g'_r(S, \theta, \vartheta)] = 0, \text{ for all } r = 1, \ldots, R$$

where g'_r is a transformation of g_r due to the aggregation and F_S is the symbolic distribution function.

1. Need to define an empirical alternative for F_S
2. Need to derive g'_r (from g_r)

Estimating Equations for Symbolic Data (3)

The **symbolic empirical likelihood** associated with the observed symbol $\boldsymbol{s}$ is

$$L(F_S; \boldsymbol{s}) = \prod_{c=1}^{C} \mathbb{P}(S = s_c) = \prod_{c=1}^{C} \prod_{n=1}^{n_c} \mathbb{P}(X = x_n^{(c)}),$$

F_S can be seen as a discrete distribution on $s_1, \ldots, s_C$ with probabilities $p_1, \ldots, p_C$, such that:

$$\begin{cases} \textbf{(C3)} : p_c = \mathbb{P}(S_c = s_c) > 0, c = 1, \ldots, C \\ \textbf{(C4)} : \sum_{c=1}^{C} p_c = 1 \end{cases}$$

Each observation $x_n^{(c)}$, $n = 1, \ldots, n_c$, $c = 1, \ldots, C$ has probability $q_n^{(c)} = \frac{p_c}{n_c}$, such that $\sum_{c=1}^{C} \sum_{n=1}^{n_c} q_n^{(c)} = 1$.
If there are no other conditions on $\boldsymbol{s}$ other than **(C3)** and **(C4)** then

$$\hat{q}_n^{(c)} = \underset{q_n^{(c)} | C3, C4}{\arg\max} \prod_{c=1}^{C} \prod_{n=1}^{n_c} q_n^{(c)} = \frac{1}{N}, \quad \Longrightarrow \quad \hat{p}_c = \frac{n_c}{N}$$

and the estimating equation becomes:

$$\sum_{c=1}^{C} \frac{n_c}{N} g_r'(s_c, \hat{\theta}, \vartheta) = 0, \text{ for all } r = 1, \ldots, R.$$

Estimating Equations for Symbolic Data (4)

The symbolic estimating equations are obtained by integrating the classical estimating equation over all the data points $\boldsymbol{x}$ from which the symbols $\boldsymbol{s}$ are produced with their corresponding weights, i.e.

$$\mathbb{E}_{F_S}[g'_r(S,\theta,\vartheta)] = \int_{\mathcal{D}_X^N} \mathbb{E}_{F_X}[g_r(X,\theta)] f_{S|\boldsymbol{x}}(\boldsymbol{s}|\boldsymbol{x},\vartheta)\phi(\boldsymbol{x})d\boldsymbol{x}.$$

$$\begin{aligned}
\mathbb{E}_{F_S}[g'_r(S,\theta,\vartheta)] &= \int_{\mathcal{D}_X^N} \mathbb{1}\left\{\pi\left(\boldsymbol{x}^{(c)}\right) = s_c; c = 1,\ldots,C\right\} \\
&\qquad \times \left\{\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \mathbb{1}\{x_n \in \boldsymbol{x}^{(c)}\} g_r(x_n;\theta)\right\} \phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} \int_{\mathcal{D}_X^N} \mathbb{1}\{x_n \in \boldsymbol{x}^{(c)}\} \mathbb{1}\left\{\pi\left(\boldsymbol{x}^{(c)}\right) = s_c\right\} g_r(x_n;\theta)\phi(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\
&= \sum_{c=1}^{C} \frac{n_c}{N} \int_{\Upsilon_c} \phi_c(x) g_r(x;\theta)\mathrm{d}x,
\end{aligned}$$

And we conclude

$$g'_r(s_c,\theta,\vartheta) = \int_{\Upsilon_c} \phi_c(x) g_r(x,\theta)\mathrm{d}x.$$

Estimating the within-symbol density ϕ_c (1)

Back to the Example: $\theta = (\mu, \sigma^2)$, we can rewrite

$$\mathbb{E}_{F_X}\left[g\left(\boldsymbol{X}, \hat{\theta}\right)\right] = \left(\frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu}), \frac{1}{N}\sum_{n=1}^{N}\left((x_n - \hat{\mu})^2 - \hat{\sigma}^2\right)\right)$$
$$= \left(\sum_{c=1}^{C}\frac{n_c}{N}(\mu_c - \hat{\mu}), \sum_{c=1}^{C}\frac{n_c}{N}\left((\mu_c - \hat{\mu})^2 + \sigma_c^2 - \hat{\sigma}^2\right)\right) = 0$$

$\Rightarrow$ Need $\hat{\mu}_c = \mathbb{E}_{\phi_c}(X)$ and $\hat{\sigma}_b^2 = \mathbb{V}_{\phi_c}(X)$.

For *skewness*: $\gamma_c = \mathbb{E}_{F_{X^c}}[((X^{(c)} - \mu_b)/\sigma_b)^3]$

For *correlation*: $\rho_{cde} = \mathbb{E}_{F_{X^c}}[(X_{[d]}^{(c)} - \mu_{c[d]})(X_{[e]}^{(c)} - \mu_{c[e]})]\sigma_{c[d]}^{-2}\sigma_{c[e]}^{-2}$

We first need to estimate ϕ

Common approach: (Bertrand and Goupil, 2000; Billard and Diday, 2003)

- Assume uniformity within classes: ϕ_c is uniform density;
- **Example**: $\alpha_c = (\alpha_{c,l}, \alpha_{c,u})$ the upper and lower bounds of an interval
  - $\hat{\mu}_c = \frac{\alpha_{c,l}+\alpha_{c,u}}{2}$, $\hat{\sigma}_c^2 = \frac{(\alpha_{c,u}-\alpha_{c,l})^2}{12}$

Estimating the within-symbol density ϕ_c (2)

Our approach: **Borrowing information from adjacent symbols**

- Attribution of a realisation to a class is **random**:
  $\Lambda_x \in \{1, \ldots, C\}$ with d.f. H_x, indicates the list of symbols in which an observation x could have been aggregated in.
- Redefine ϕ (density of X given $\boldsymbol{S}$) as

$$\phi(x) = \int_{\mathcal{D}(\Lambda_x)} f_{X|\boldsymbol{S}=\boldsymbol{s}'}(x) \mathrm{d}H_x(\lambda)\mathrm{d}\lambda$$

  where $\boldsymbol{s}'$ **denotes the set of symbols given that the observation x is assumed to be grouped in the λ-th class**.

  If an observation x can only be associated to a **unique** *symbol s_c then H_x is a Dirac delta function and $\phi(x) = f_{X|\boldsymbol{S}=\boldsymbol{s}'}(x)$.*

- The density of an observation given a symbol s_c is given by

$$\phi_c(x) = \frac{\mathbb{1}\{x \in s_c\}\phi(x)}{P_H^{s_c}}, \quad P_H^{s_c} = \int_{\Upsilon_c} \phi(x)\mathrm{d}H_x(c)\mathrm{d}x.$$

ϕ_c for interval-valued data

- s_c: D-dimensional intervals with $\alpha_c = (\alpha_{c,l}, \alpha_{c,u})$ and $\Upsilon_c = [\alpha_{c,l}, \alpha_{c,u}]$
- If x is in a region where some intervals from the set $\boldsymbol{s}$ overlap, then Λ_x has a discrete outcome λ, a subset of $\{1, \ldots, C\}$

$$H_x(\lambda, c) = \frac{n_c}{\sum_{a \in \lambda} n_a}, \text{ for all } c \in \lambda.$$

- We can derive that:

$$\phi(x) = \sum_{c=1}^{C} \frac{n_c}{N|\Upsilon_c|} \mathbb{1}\{x \in \Upsilon_c\},$$

- Split the range of $\boldsymbol{x}$ into a grid of subintervals denoted by $\upsilon_{\boldsymbol{b}}$ with $\boldsymbol{b} = (b_1, \ldots, b_D)$ and $b_d = 1, \ldots, (2C+1); d = 1, \ldots D$

  The normalising term $P_H^{s_c} = m_c(1)$ where

$$\begin{aligned} m_c(f(x)) &= \int_{\Upsilon_c} f(x)\phi(x)\mathrm{d}H_x(c)\mathrm{d}x \\ &= \frac{1}{N} \sum_{c'=1}^{C} \frac{n_{c'}}{|\Upsilon_{c'}|} \left(\sum_{\boldsymbol{b}} \mathbb{1}\{\upsilon_{\boldsymbol{b}} \subset \Upsilon_{c'}\} H_{\boldsymbol{b}}(\lambda, c) \int_{\upsilon_{\boldsymbol{b}}} f(y)\mathrm{d}y \right), \end{aligned}$$

ϕ_c for interval-valued data

The estimates of the mean, variance, skewness and correlation within an interval c are obtained using the density ϕ_c and are given as follows

$$\hat{\mu}_{cd} = \frac{m_c\left(x_{[d]}\right)}{m_c(1)}, \qquad \hat{\sigma}^2_{cd} = \frac{m_c\left(x^2_{[d]}\right)}{m_c(1)} - \hat{\mu}^2_{cd},$$

$$\hat{\gamma}_{cd} = \frac{m_c\left(x^3_{[d]}\right)}{m_c(1)\hat{\sigma}^3_{cd}} - \frac{\hat{\mu}^3_{cd}}{\hat{\sigma}^3_{cd}} - 3\frac{\hat{\mu}_{cd}}{\hat{\sigma}_{cd}}, \qquad \hat{\rho}_{cde} = \frac{m_c\left(x_{[d]}x_{[e]}\right)}{m_c(1)\hat{\sigma}_{cd}\hat{\sigma}_{ce}} - \frac{\hat{\mu}_{cd}\hat{\mu}_{ce}}{\hat{\sigma}_{cd}\hat{\sigma}_{ce}},$$

for $d = 1, \ldots, D$.

ϕ_c for histogram-valued data

- s_c: D-dimensional histogram bins with $n_{\boldsymbol{c}} = \alpha_{\boldsymbol{c}} = \sum_{n=1}^{N} \mathbb{1}\{x_n \in \Upsilon_{\boldsymbol{c}}\}$ and $\Upsilon_{\boldsymbol{c}} = \left((y^1_{c_1-1}, y^1_{c_1}) \times \cdots \times (y^D_{c_D-1}, y^D_{c_D})\right)$.
- Assume the **marginal bin width to be equal** and given by $\delta_d = y^d_{c_d} - y^d_{c_d-1}\ \forall c_d, d$ s.t. the area of each bin is $|\Upsilon_{\boldsymbol{c}}| = \prod_{d=1}^{D} \delta_d$.
- Choice of the bin locations $\Upsilon_{\boldsymbol{c}}$ is arbitrary and *other histograms could have arisen by shifting the bin locations by up to half of their width to the left or to the right.*
- New bin locations $\Upsilon' = \Upsilon + \boldsymbol{u}$ where $\boldsymbol{u} = (u_1, \ldots, u_D)$, $u_d \sim \mathcal{U}\left(-\frac{\delta_d}{2}, \frac{\delta_d}{2}\right)$.
- Λ_x: the bin Υ in which x could have been aggregated to, is now a continuous as the set of outcomes contains all the shifted bins Υ' that will include x.

$$H_x(\Upsilon'_c) = \frac{1}{\prod_{d=1}^{D} \delta_d} \mathbb{1}\{x \in \Upsilon'_c\}$$

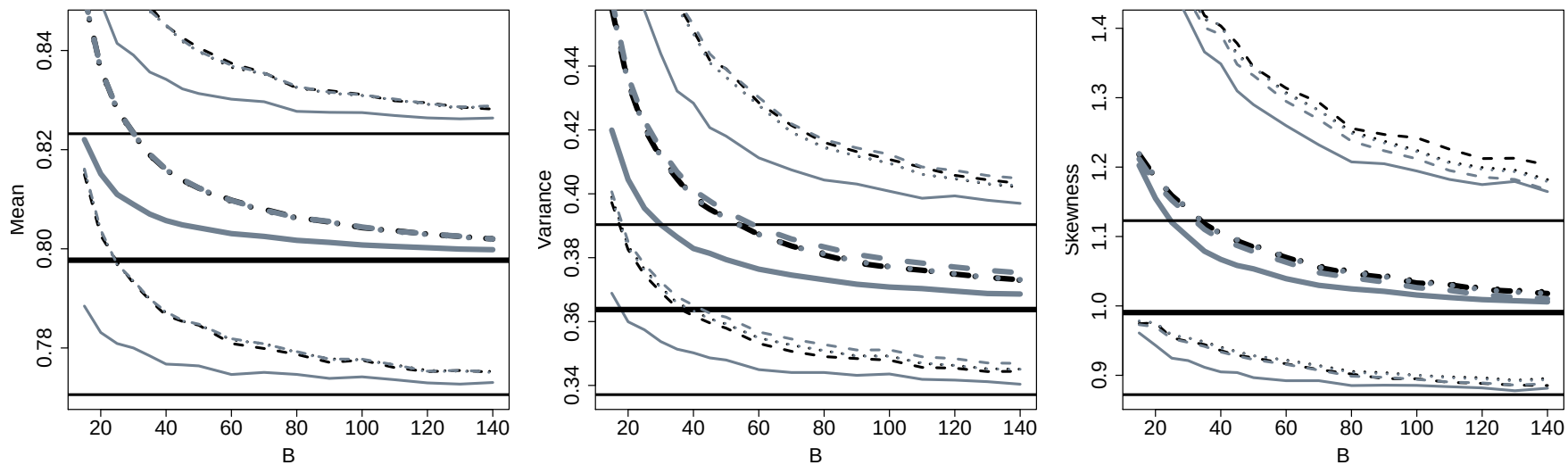
- Assuming that histogram bins can be shifted relates to the ideas of Heitjan and Rubin (1991)

Examples

$N = 2000$ draws from the Skew-Normal distribution.

- **C** min/max intervals

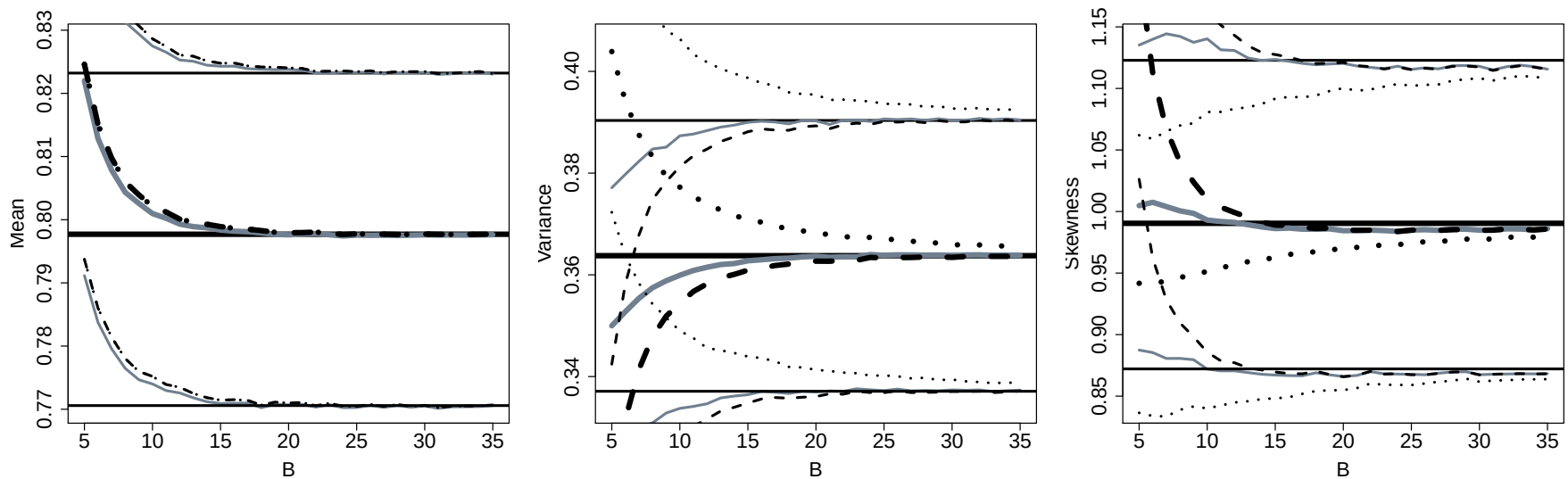
- Classical estimates and 95% CI (**Solid black**)
- Borrowing/Not borrowing information (**Solid grey**/**Dotted black**)
- Billard and Didday (2003) histogram estimator of intervals by partitioning, borrowing (**Dotted grey**) and not borrowing information (**Dashed grey**)



Examples

$N = 2000$ draws from the Skew-Normal distribution.
- **1** Histogram with **C** equal bins

- Classical estimates and 95% CI (**Solid black**)
- Borrowing/Not borrowing information (**Solid grey**/**Dotted black**)



Discussion

- A **new non-parametric approach to estimate statistics** presented in the general framework of **Estimating Equations**

- Proposal to **borrow information from neighbours** to get a sense of the **dependence**:
  - Borrow lots of information = low dependence
  - Borrow no information = dependence

- Not always necessary: e.g. mushroom dataset analysed in de A Lima Neto et al. (2011)

UNSW SYDNEY

ACEMS

THANK YOU

Manuscripts:

- Whitaker, Beranger & Sisson (2020). Estimating Equations for data summaries.
  Available very soon!
- Beranger, Lin & Sisson (2018). New models for symbolic data.
  `https://arxiv.org/pdf/1805.03316.pdf`
- Whitaker, Beranger & Sisson (2019). Composite likelihood methods for histogram-valued random variables.
  `https://arxiv.org/pdf/1908.11548.pdf`

Contact:

`B.Beranger@unsw.edu.au`