Plan de sondage informatif, distribution pondérée, et maximum de vraisemblance.

D. Bonnéry*, F. Coquet*, J. Breidt**

* Crest (Ensai)
** Colorado State University

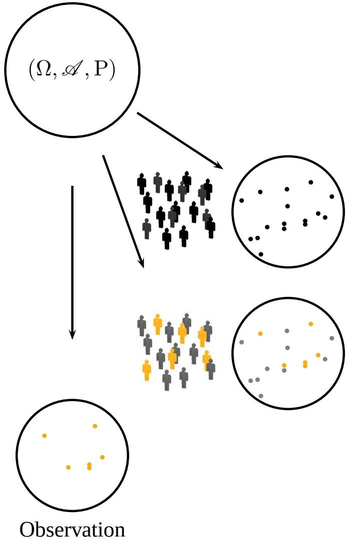
5 novembre 2012

1. Informative selection and asymptotic framework
    - Informative selection mechanism
    - Sample distribution
2. Parametric estimation
    - Maximum pseudo-likelihood estimator
    - Convergence
    - Simulations

An observation is the **outcome of two random processes**:

- The **population realization**: variables are generated for each element of a population U of size N.
- The **selection mechanism** : a **sample** is drawn from U.

The observation usually consists of the values of the study variables for each element in the sample.



Define:

- $(Y_1, \ldots, Y_N)$: **study variable**,
- $(Z_1, \ldots, Z_N)$: **design variables**,
- Π: **design measure** and symmetric function of $(Z_1, \ldots, Z_N)$,
- $(J_1, \ldots, J_N)$: **sample** (vector of $\mathbb{N}^N$),

$$\mathrm{P}^{\Pi,\mathcal{Y},\mathcal{Z}} - a.s. (\mathrm{p}, y, z), \mathrm{P}^{\mathcal{J}|\Pi=\mathrm{p},\mathcal{Z}=z,\mathcal{Y}=y} = \mathrm{p}.$$

- $\pi_k = \Pi\left(\{J_k \geqslant 1\}\right)$: **inclusion probability**,
- $n = \sum_{k \in U} J_k$: **sample size**,

Assume:

- $\lim_{N \to \infty} N^{-1} n > 0$.

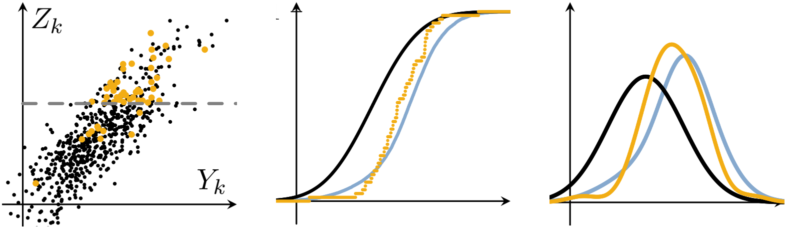
Definition

The weight function is:
$$\rho : y \mapsto \lim_{N\to\infty} \mathrm{E}\left[J_k | Y_k = y\right] / \mathrm{E}\left[J_k\right]$$

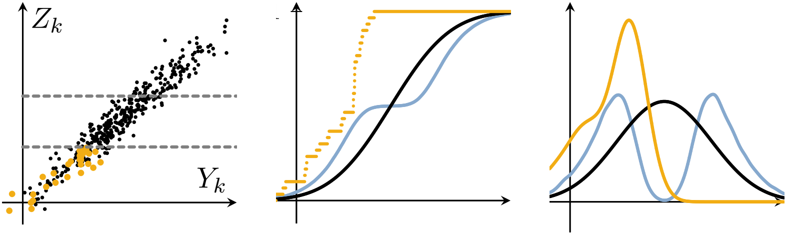
Denote (ℓ) the ℓth drawn element.

$$\text{Does } \mathrm{P}^{(Y_{(\ell)})_{\ell=1}^{n}} \sim \left(\rho . \mathrm{P}^{Y_1}\right)^{\otimes n}?$$

- Observations behave like iid $\rho.f$



- Observations do not behave like iid $\rho.f$



Population cdf, empirical sample cdf and sample cdf

f, $\rho.f$ and sample kernel density estimator

Results:

Under asymptotic independance of draws in the one-dimensinal case:

- the empirical sample cdf converges to $\alpha \mapsto (\rho.\, \mathrm{P}^Y)((-\infty, \alpha])$ (Bonnéry, Breidt and Coquet, 2011)
- a kernel density estimator (kde) of the pdf converges to

$$\rho.\frac{\mathrm{d}\,\mathrm{P}^Y}{\mathrm{d}\lambda}$$

Goal:

- Parametric estimation.

Inference on population model

Consider the population model

- $(Y_k)_{k\in\{1,\ldots,N\}} \sim (f_\theta.\lambda)^{\otimes N}$,
- $(Y_k, Z_k)_{k\in\{1,\ldots,N\}}$ are iid realizations,
- $\mathrm{P}^{Z_k|Y_k}$ is parametrized by $\xi \in \Xi$

The target of the inference is θ.

Definition

$$\rho_{\theta,\xi} : y \mapsto \lim_{N\to\infty} \mathrm{E}_{\theta\xi}\left[J_k|Y_k = y\right] / \mathrm{E}_{\theta\xi}\left[J_k\right].$$

Following Pfeffermann and Krieger (1992) we define:

$$\hat{\theta}(\xi) = \arg\max_{\theta \in \Theta} \left\{ \sum_{\ell=1}^{n} \ln\left(\rho_{\theta\xi} f_\theta \left(Y_{(\ell)} \right) \right) \right\}.$$

Assume $A0$. Let

- $\hat{\xi}$ be a consistent estimator of ξ,
- $\overline{\mathscr{L}}\left(\left(Y_{(\ell)} \right)_{\ell \in \{1,\dots,n\}}, \theta, \xi \right) = n^{-1} \sum_{k=1}^{n} \ln\left(\rho_{\theta\xi} f_\theta \right) \left(Y_{(\ell)}, \theta, \xi \right)$

Definition

The maximum pseudo-likelihood estimator associated to $\hat{\xi}$ is:

$$\hat{\theta} = \arg\max_{\theta \in \Theta} \left\{ \overline{\mathscr{L}}\left(\left(Y_{(\ell)} \right)_{\ell \in \{1,\dots,n\}}, \theta, \hat{\xi} \right) \right\},$$

Definition

Define

$$
\begin{aligned}
m(y) &= \mathrm{E}\left[J_1 | Y_1 = y\right] \\
m'(y_1, y_2) &= \mathrm{E}\left[J_2 | Y_1 = y_1, Y_2 = y_2\right] \\
v(y) &= \mathrm{Var}\left[J_1 | Y_1 = y\right] \\
c(y_1, y_2) &= \mathrm{Cov}\left[J_1, J_2 | Y_1 = y_1, Y_2 = y_2\right] \\
\delta(y_1, y_2) &= m'(y_1, y_2) m'(y_2, y_1) - m(y_1) m(y_2)
\end{aligned}
$$

$$
\begin{aligned}
Y^* &\sim \rho_{\theta\xi} f_\theta . \lambda \\
\mathscr{I}_{11} &= \mathrm{E}_{\theta,\xi}\left[\left(\frac{\partial \ln\left(\rho_{\theta\xi} f_\theta\right)}{\partial\theta}\left(Y^*, \theta, \xi\right)\right)^2\right] \\
\mathscr{I}_{12} &= \mathrm{E}_{\theta,\xi}\left[\left|\left(\frac{\partial}{\partial\theta}\ln\left(\rho_{\theta\xi} f_\theta\right)\frac{\partial}{\partial\xi}\ln\left(\rho_{\theta\xi} f_\theta\right)\right)\left(Y^*, \theta, \xi\right)\right|\right]
\end{aligned}
$$

Assumption

- *Standard conditions on* $\rho_{\theta,\xi} f$,
- *asymptotic independance of draws:*
  $$\forall g \in \left\{ 1\!\!1, \left(\frac{\partial \ln(\rho_{\theta\xi} f_\theta)}{\partial \theta}(.,\theta,\xi) \right)^2, \left(\frac{\partial \ln(\rho_{\theta\xi} f_\theta)}{\partial \theta}(.,\theta,\xi) \frac{\partial \ln(\rho_{\theta\xi} f_\theta)}{\partial \xi}(.,\theta,\xi) \right) \right\},$$
  - $\mathrm{E}_{\theta\xi}\left[|g(Y_1)\, g(Y_2)|\, c_{,\theta,\xi}(Y_1, Y_2) \right] = o_{N\to\infty}(1)$,
  - $\mathrm{E}_{\theta\xi}\left[|g(Y_1)\, g(Y_2)|\, \delta_{,\theta,\xi}(Y_1, Y_2) \right] = o_{N\to\infty}(1)$,
  - $\mathrm{E}_{\theta\xi}\left[\left(g^2 (v_{\theta,\xi} + m^2_{\theta,\xi}) \right)(Y_1) \right] = o(N)$.

Theorem

Suppose

$$\sqrt{n}\begin{bmatrix}\left(\frac{\partial}{\partial\theta}\bar{\mathscr{L}}\right)\left(\left(Y_{(\ell)}\right)_{\ell\in\{1,\ldots,n\}},\theta,\xi\right)\\ \hat{\xi}-\xi\end{bmatrix}\xrightarrow{\mathscr{L}}\mathscr{N}\left(0,\Sigma=\begin{bmatrix}\Sigma_{11} & \Sigma_{12}\\ & \Sigma_{22}\end{bmatrix}\right).$$

Then

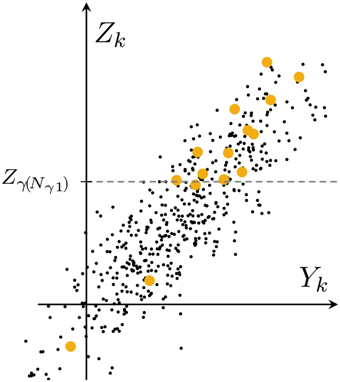
$$\sqrt{n}\left(\hat{\theta}-\theta\right)/\sigma\xrightarrow{\mathscr{L}}\mathscr{N}(0,1),$$

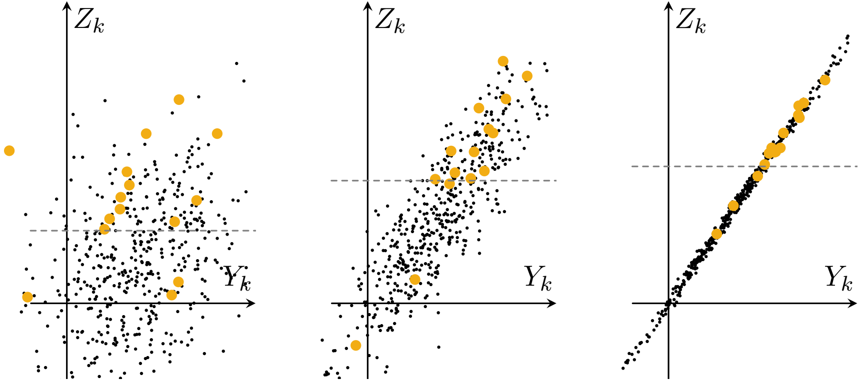
with

$$\sigma^2 = \frac{\Sigma_{11}}{\mathscr{I}_{11}^2}+\frac{\mathscr{I}_{12}}{\mathscr{I}_{11}^2}\left(\Sigma_{22}\mathscr{I}_{12}-2\Sigma_{12}\right).$$

Consider:

- $Y \sim \mathcal{N}(\theta.\mathbb{1}, Id_N)$,
- $\varepsilon \sim \mathcal{N}(0, Id_N)$, ε and Y are independent,
- $Z = \xi.Y + \eta\varepsilon$, η known,
- $N = 5000$,
- $\frac{N_1}{N} = 0.7$, $\frac{N_2}{N} = 0.3$,
- $\frac{n_1}{N_1} = 1/70$, $\frac{n_2}{N_2} = 4/30$.





$\eta = 10$ $\eta = 1$ $\eta = 0.1$

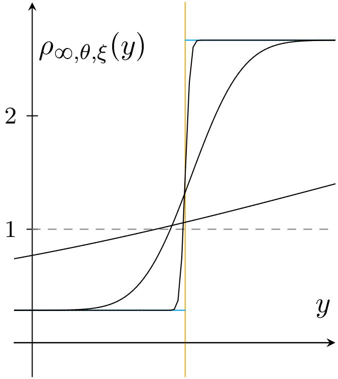
Let

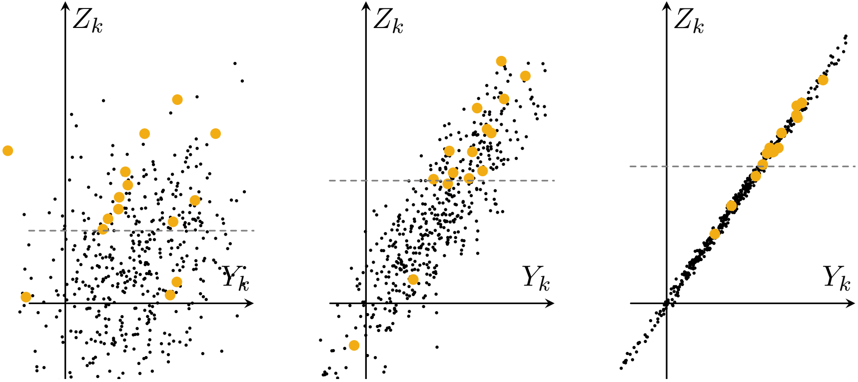
- $t_1 = \lim_{N\to\infty} \frac{N_1}{N} = 0.7$
- $\zeta = \phi^{-1}(t_1)$
- $\tau_h = \lim_{N\to\infty} \left(\frac{n_h}{N_h}\right)$,
  $\tau_1 = 1/70,\ \tau_2 = 4/30$
- $p(y) =$
  $\mathrm{P}\left(\varepsilon < \frac{\zeta\sqrt{\xi^2+\eta^2}+\xi(\theta-y)}{\eta}\right)$

We get:

$$\rho_{\theta,\xi}(y) = \frac{\tau_1 p(y) + \tau_2(1-p(y))}{\tau_1 t_1 + \tau_2(1-t_1)}$$

Figure: Plot of ρ for $\theta = 1.5$, $\xi = 2$, $\eta \in \{.1, 1, 10\}$





$\eta = 10$ $\eta = 1$ $\eta = 0.1$

To estimate ξ, we use

$$\hat{\xi} = \frac{\sum_{\ell=1}^{n} Z_{(\ell)} Y_{(\ell)} / \pi_{(\ell)}}{\sum_{\ell=1}^{n} Y_{(\ell)}^2 / \pi_{(\ell)}}.$$

Compare $\hat{\theta}$ to

- $\tilde{\theta} = \sum_{\ell=1}^{n} \frac{Y_{(\ell)}}{\pi_{(\ell)}} = \arg\max_{\theta \in \Theta} \left\{ \sum_{k=1}^{N} \frac{\ln(f_\theta(Y_k)) J_k}{\pi_k} \right\}.,$
- $\bar{\theta} = n^{-1} \sum_{\ell \in \{1,\ldots,n\}} Y_{(\ell)}.$

Table: Calculus of mean and mean square error on 1000 simulations

θ	ξ	σ		Mean[.]	MSE[.]	$\sqrt{\frac{\text{MSE}}{\text{MSE}(\hat{\theta})}}$	$\frac{1}{n_\gamma}\lim_{\gamma\to\infty} n_\gamma \text{ Var }[.]$
1.5	2	0.1	$\hat{\theta}$	1.502	$7.643\ 10^{-4}$	1	$6.962\ 10^{-4}$
			$\tilde{\theta}$	1.5	$4.811\ 10^{-3}$	2.509	$4.523\ 10^{-3}$
			$\bar{\theta}$	2.329	$6.887\ 10^{-1}$	30.02	$3.979\ 10^{-3}$
1.5	2	1	$\hat{\theta}$	1.5	$1.975\ 10^{-3}$	1	$2.975\ 10^{-3}$
			$\tilde{\theta}$	1.501	$5.583\ 10^{-3}$	1.681	$6.024\ 10^{-3}$
			$\bar{\theta}$	2.241	$5.509\ 10^{-1}$	16.7	$3.971\ 10^{-3}$
1.5	2	10	$\hat{\theta}$	1.497	$5.501\ 10^{-3}$	1	$2.943\ 10^{-3}$
			$\tilde{\theta}$	1.5	$1.030\ 10^{-2}$	1.368	$1.030\ 10^{-2}$
			$\bar{\theta}$	1.662	$2.999\ 10^{-2}$	2.335	$4.027\ 10^{-3}$

Summary:

- Definition of sample pdf and limit sample pdf, applicable to with or without replacement and fixed or random size samples,
- Simple and verifiable conditions on the sequence of sample schemes,
- Pertinence: The sample behaves like an independent sample (Uniform cdf convergence),
- The limit sample pdf can be used for inference (Convergence of the maximum pseudo likelihood estimator),
  - Asymptotic normality under stratified sampling and fixed number of strata.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3).

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, 71(3).

Bonnéry, D., Breidt, F. J., and Coquet, F. (2011). Uniform convergence of the empirical cumulative distribution function under informative selection from a finite population. *To appear in Bernoulli*.

Pfeffermann, D. and Krieger, A. M. (1992). Maximum likelihood estimation for complex sample surveys. *Survey Methodology*, 18(1):225–255.

Gong, G. and Samaniego, F. J. (1981). Pseudomaximum likelihood estimation: theory and applications. *The Annals of Statistics*, 9(4).

Pfeffermann, D., Krieger, A. M., and Rinott, Y. (1998a). Parametric distributions of complex survey data under informative probability sampling. *Statistica Sinica*, 8(4).

Thank you for your attention.