MÉTHODES D'IMPUTATION ALÉATOIRES ÉQUILIBRÉES

David Haziza

Université de Montréal

Travail joint avec Guillaume Chauvet et Jean-Claude Deville

PLAN

- INTRODUCTION
- IMPUTATION DÉTERMINISTE VS. ALÉATOIRE
- VARIANCE DUE À L'IMPUTATION
- IMPUTATION ÉQUILIBRÉE
- ÉTUDE PAR SIMULATION
- TRAVAUX FUTURS

CONTEXTE

- Population finie U de taille N
- L'objectif est d'estimer le total dans la population

$$Y = \sum_{i \in U} y_i ,$$

pour une variable d'intérêt y.

- On tire un échantillon aléatoire, s, de taille n, selon un plan de sondage $p(s)$.

ESTIMATION: RÉPONSE COMPLÈTE

- Un estimateur de Y est l'estimateur de Horvitz-Thompson

$$\hat{Y}_\pi = \sum_{i \in s} d_i y_i$$

- $d_i = 1/\pi_i$ désigne le poids de sondage de l'unité i
- π_i désigne la probabilité d'inclusion de l'unité i dans l'échantillon s $i = 1,..., N$.

- L'estimateur de HT est sans biais sous le plan de sondage:

$$E_p\left(\hat{Y}_\pi\right) = Y.$$

ESTIMATEUR IMPUTÉ

- En présence de non-réponse à la variable y, on définit un estimateur imputé de Y

$$\hat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i \left(1 - r_i\right) y_i^*,$$

où r_i est une variable indicatrice de réponse telle que

$$r_i = \begin{cases} 1 & \text{si l'unité } i \text{ a répondu à la variable } y \\ 0 & \text{sinon} \end{cases}$$

et y_i^* désigne la valeur imputée utilisée pour remplacer la valeur manquante y_i

MÉTHODES D'IMPUTATION

Les méthodes d'imputation peuvent être classées en deux groupes:

• Les méthodes dites déterministes: Méthodes qui fournissent une valeur fixe étant donné l'échantillon

• Les méthodes dites stochastiques ou aléatoires: Méthodes d'imputation ayant une composante aléatoire (et donc qui ne donnent pas nécessairement la même valeur étant donné l'échantillon si la méthode est répétée)

MÉTHODES D'IMPUTATION

• La plupart des méthodes d'imputation peut être représentée par le modèle suivant:

$$y_i = f\left(\mathbf{z_i}\right) + \sigma\sqrt{\nu_i}\,\varepsilon_i$$

• Les erreurs ε_i sont des variables aléatoires indépendantes de moyenne 0 et de variance 1 avec une distribution commune (L)

• Les quantités ν_i sont supposées connues

• Le modèle peut également servir à motiver une imputation à l'intérieur de classes d'imputation

MÉTHODES D'IMPUTATION

- Imputation déterministe:

$$y_i^* = \hat{f}_r(\mathbf{z_i})$$

- Imputation aléatoire: c'est une imputation déterministe à laquelle on ajoute un résidu aléatoire:

$$y_i^* = \hat{f}_r(\mathbf{z}_i) + \hat{\sigma}\sqrt{v_i}\,\varepsilon_i^*$$

  - Les résidus ε_i^* sont tirés aléatoirement parmi l'ensemble des résidus « standardisés » observés chez les répondants:

$$\tilde{e}_j = \frac{1}{\hat{\sigma}\sqrt{v_j}}\left(y_j - \hat{f}_r(\mathbf{z}_i)\right) - \bar{e}_r, \quad j \in s_r$$

MÉTHODES D'IMPUTATION

• Les résidus aléatoires sont tirés de manière à respecter les probabilités suivantes:

$$P\left(\varepsilon_i^* = \tilde{e}_j\right) = \omega_j \Big/ \sum_{l \in s} \omega_l r_l$$

• Lorsque $\omega_j = 1$, on est en présence d'imputation aléatoire non-pondérée

• Lorsque $\omega_j = d_j$, on est en présence d'imputation aléatoire pondérée

• D'autres choix de pondération sont possibles (e.g., Haziza, 2009)

MÉTHODES DÉTERMINISTES

- Imputation par la régression:

$$f(\mathbf{z}_i) = \mathbf{z}_i'\boldsymbol{\beta} \text{ et } \nu_i = \boldsymbol{\lambda}'\mathbf{z}_i \Rightarrow y_i^* = \mathbf{z}_i'\hat{\mathbf{B}}_r$$

- Imputation par le ratio:

$$f(\mathbf{z}_i) = \beta z_i \quad \text{et} \quad \nu_i = z_i \Rightarrow y_i^* = \hat{B}_r z_i = \frac{\bar{y}_r}{\bar{z}_r} z_i$$

- Imputation par la moyenne:

$$z_i = 1 \,\forall i, \quad f(z_i) = \beta \text{ et } \nu_i = 1 \Rightarrow y_i^* = \hat{B}_r = \bar{y}_r$$

MÉTHODES ALÉATOIRES

- Imputation par la régression aléatoire:

$$y_i^* = \mathbf{z}_i' \hat{\mathbf{B}}_r + \hat{\sigma} \sqrt{\boldsymbol{\lambda}' \mathbf{z}_i} \varepsilon_i^* .$$

- Imputation par hot-deck aléatoire:
  - On remplace la valeur manquante d'un receveur par celle d'un répondant (donneur) tiré au hasard dans l'ensemble des répondants

  - Peut être vue comme de l'imputation par la moyenne à laquelle on a rajouté un résidu:

$$y_i^* = \bar{y}_r + \underbrace{\left(y_j - \bar{y}_r \right)}_{\tilde{e}_j}, \qquad j \in s_r$$

PROPRIÉTÉS DES MÉTHODES D'IMPUTATION

- Estimation d'un total (ou moyenne): Les méthodes déterministes mènent à des estimateurs asymptotiquement sans biais
- Estimation d'un quantile (e.g., médiane): Les méthodes déterministes (exception: PPV) tendent à distordre la distribution des variables que l'on impute

Estimateurs de quantiles biaisés !

- Les méthodes aléatoires préservent les distributions au prix d'un terme de variance additionnel (variance due à l'imputation), due à la sélection aléatoire des résidus.

VARIANCE TOTALE

- Variance totale de l'estimateur imputé: Imputation aléatoire

$$V\left(\hat{Y}_I\right) = V_p E_q E_I\left(\hat{Y}_I \mid s\right) + E_p V_q E_I\left(\hat{Y}_I \mid s\right) + E_p E_q V_I\left(\hat{Y}_I \mid s\right)$$
$$= V_{SAM} + V_{NR} + V_{IMP}$$

- Les trois termes sont du même ordre de grandeur

- En présence d'imputation aléatoire, l'estimateur imputé s'écrit comme:

$$\hat{Y}_I = \sum_{i \in s} d_i r_i y_i + \sum_{i \in s} d_i \left(1 - r_i\right) \hat{f}_r\left(\mathbf{z}_i\right) + \sum_{i \in s} d_i \left(1 - r_i\right) \hat{\sigma} \sqrt{v_i} \varepsilon_i^*$$

- La variance due à l'imputation provient du troisième terme

VARIANCE DUE À L'IMPUTATION

$$V_{IMP} = E_p E_q \left\{ \frac{\sum_{j \in s} d_j^2 \left(1 - r_j\right) v_j}{\sum_{j \in s} \omega_j r_j} \sum_{j \in s} \omega_j r_j \tilde{e}_j^2 \right\}$$

- V_{IMP} est petite lorsque:
  - le taux de réponse est élevé
  - le taux de réponse est faible
  - les résidus $\tilde{e}_j$ sont petits (le modèle ajuste bien les données)

UN EXEMPLE

- Imputation par la régression linéaire simple aléatoire:

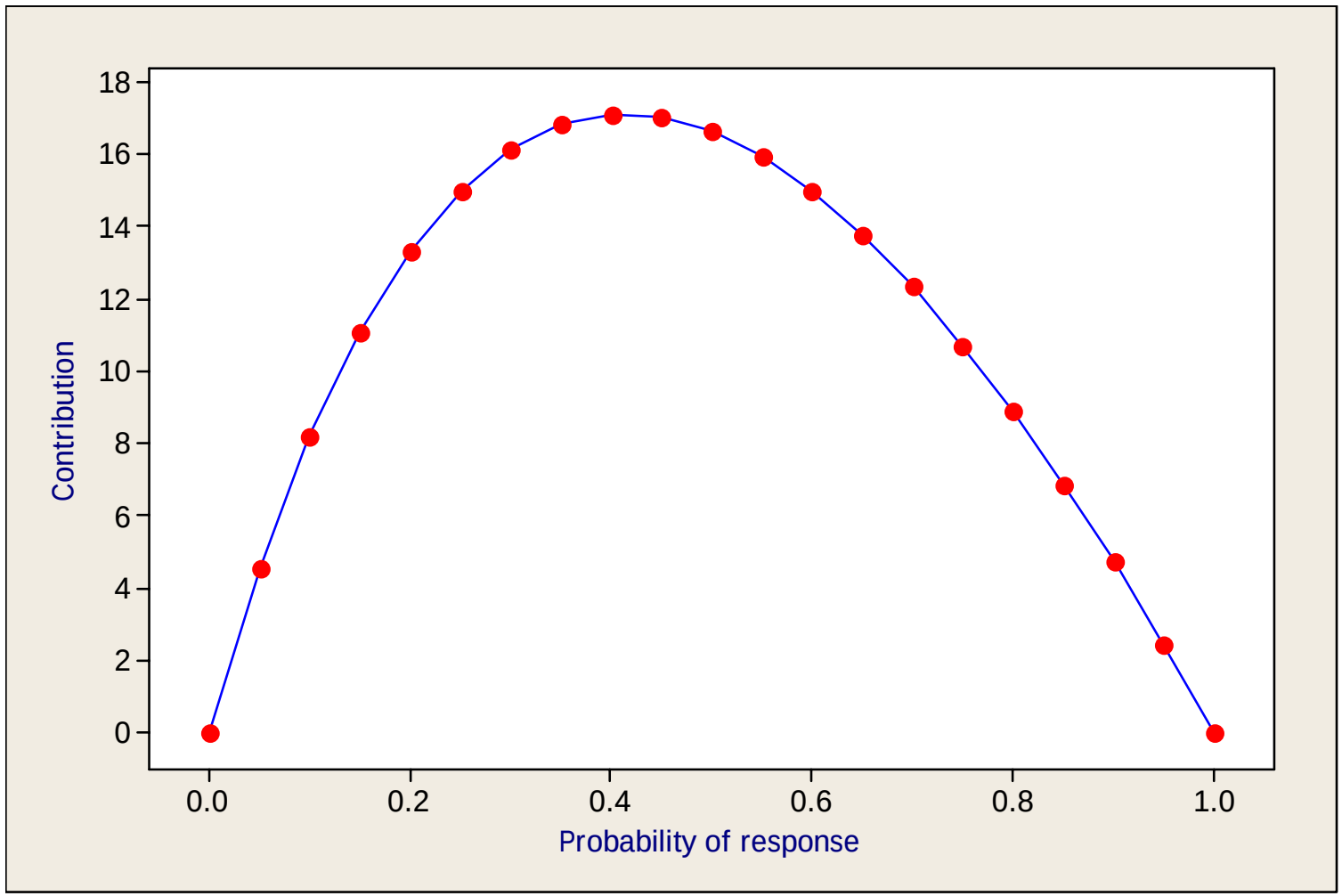
$$\mathbf{z}_i = (1, z_i)' \text{ et } \nu_i = 1.$$

- Échantillonnage aléatoire simple sans remise et mécanisme de non-réponse uniforme.

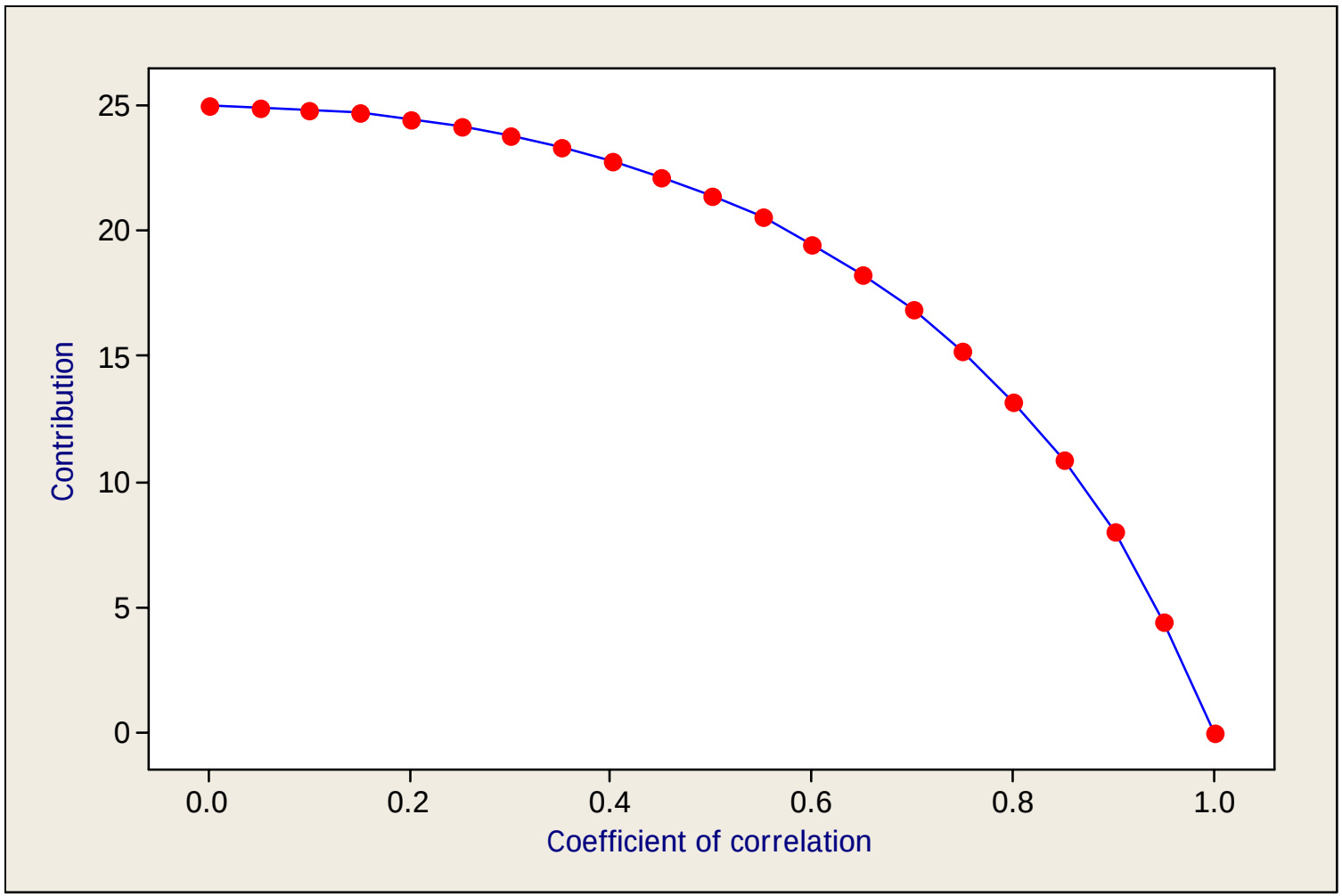
- Variance additionnelle (en %) relative due à l'imputation aléatoire:

$$\frac{V_A\left(\hat{Y}_I\right) - V_D\left(\hat{Y}_I\right)}{V_D\left(\hat{Y}_I\right)} \approx \frac{p(1-p)\left(1-\rho_{yz}^2\right)}{1-(1-p)\rho_{yz}^2}.$$

Variance additionnelle (en %) avec $\rho_{xy} = 0.8$.



Variance additionnelle (en %) avec $p = 0.5$



RÉDUIRE LA VARIANCE DUE À L'IMPUTATION

• Trois approches ont été proposées pour réduire/éliminer la variance due à l'imputation:

(1) L'imputation fractionnelle: (Kalton et Kish, 1981;1984; Fay, 1996; Kim et Fuller (2002); Fuller et Kim (2005)

- consiste à imputer M valeurs (méthode aléatoire)
- chaque valeur se voit attribuer un poids (habituellement, $1/M$)
- similaire à l'imputation multiple mais les procédures d'estimation sont différentes
- la variance due à l'imputation décroit à mesure que M croit.

RÉDUIRE LA VARIANCE DUE À L'IMPUTATION

(2) Procédure d'ajustement des valeurs imputées: Chen, Rao et Sitter (2000)

- on commence par utiliser l'imputation par hot-deck aléatoire
- ensuite, on ajuste les valeurs imputées de manière à éliminer le troisième terme. Autrement dit, on veut satisfaire:

$$\sum_{i \in s} d_i \left(1 - r_i\right) \hat{\sigma} \sqrt{\nu_i} \varepsilon_i^* = 0 \quad \Leftrightarrow \quad \bar{y}_m^* = \bar{y}_r$$

- Chen, Rao et Sitter (2000) ont montré que cette procédure permet de préserver la distribution des variables que l'on impute

RÉDUIRE LA VARIANCE DUE À L'IMPUTATION

(3) Imputation équilibrée:

- on sélectionne les résidus aléatoires de manière à satisfaire (au moins approximativement)

$$\sum_{i \in s} d_i \left(1 - r_i\right) \hat{\sigma} \sqrt{\nu_i} \varepsilon_i^* = 0$$

- cette idée n'est pas nouvelle: Kalton et Kish (1981, 1984) et Deville (2006).

- on utilise la méthode du Cube (Deville et Tillé, 2004) proposée dans le contexte de l'échantillonnage équilibrée.

AVANTAGES DE L'IMPUTATION ÉQUILIBRÉE

• Simplicité pour les utilisateurs car un seul fichier est crée (la majorité des enquêtes utilisent l'imputation simple)

• Aucun ajustement des valeurs imputées n'est nécessaire. Les procédures d'estimation habituelles peuvent être appliquées.

• Peut être utilisée pour des variables continues ou catégorielles

• La distribution des variables que l'on impute est préservée:

$$\hat{F}_I(t) \xrightarrow{p} F_N(t)$$

ÉTUDE PAR SIMULATION

- Population extraite d'un échantillon de l'Enquête Canadienne sur la Santé de taille $N = 10000$
- La population a été stratifiée par province (11 strates)
- De la population, 1000 échantillons de taille $n = 500$ ont été tirés selon un plan stratifié aléatoire simple sans remise avec répartition proportionnelle
- Dans chaque échantillon, la non-réponse a été générée selon un mécanisme uniforme à l'intérieur des strates ($p_h = 0.6$ ou $p_h = 0.7$)
- 3 méthodes d'imputation à l'intérieur des classes: imputation par la moyenne, imputation par hot-deck aléatoire et imputation équilibrée

ÉTUDE PAR SIMULATION

- Note: l'imputation équilibrée consiste à tirer des donneurs pour imputer les receveurs tel que $\bar{y}^{*}_{mh} = \bar{y}_{rh}$ dans la strate h
- On s'intéresse à deux variables:
  - Le poids en kg d'un individu (y_1)
  - La présence ou absence d'asthme (y_2): $y_2 = 1$ si l'individu souffre d'asthme et $y_2 = 0$, sinon
- On cherche à estimer :
  - $\bar{Y}_1$ et $F_1(t)$ (fonction de répartition de la variable « poids »)
  - $\bar{Y}_2$: la proportion d'individus souffrant d'asthme

MESURES MONTE CARLO

- Biais relatif monte carlo (en %)

$$RB_{MC}\left(\hat{\theta}\right) = \frac{E_{MC}\left(\hat{\theta}\right) - \theta}{\theta} \times 100$$

- Erreur quadratique moyenne monte carlo:

$$EQM_{MC}\left(\hat{\theta}\right) = E_{MC}\left(\hat{\theta} - \theta\right)^2$$

- Efficacité relative (imputation par hot-deck aléatoire utilisé comme référence):

$$ER = \frac{EQM_{MC}\left(\hat{\theta}_{(.)}\right)}{EQM_{MC}\left(\hat{\theta}_{(HD)}\right)}$$

EFFICACITÉ RELATIVE MONTE CARLO

• Le biais relatif des estimateurs imputés de $\bar{Y}_1$ et $\bar{Y}_2$ est négligeable dans tous les scénarios

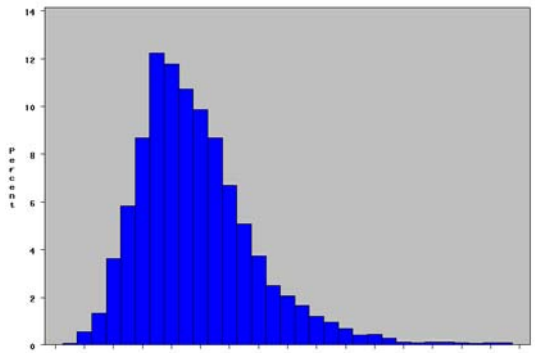
ER	Moyenne	Hot-deck aléatoire	Equilibrée
y_1	0.80	1	0.82
y_2	0.82	1	0.86

FONCTION DE RÉPARTITION

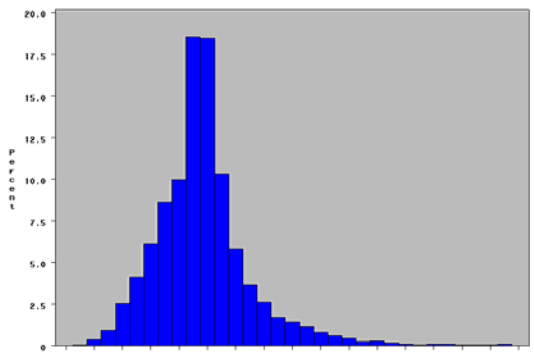
Biais relatif et EQM pour $F_1(t)$

$F_1(t)$	Biais relatif (en %)			*ER*	
	Moyenne	Hot-deck	Équilibrée	Moyenne	Équilibrée
0.25	-29.03	0.18	0.25	9.52	0.90
0.5	-16.30	-0.11	-0.02	11.02	0.90
0.75	8.94	0.03	0.01	7.90	0.87

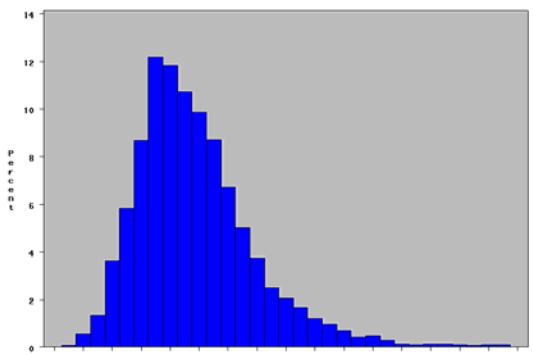
FDR dans la population



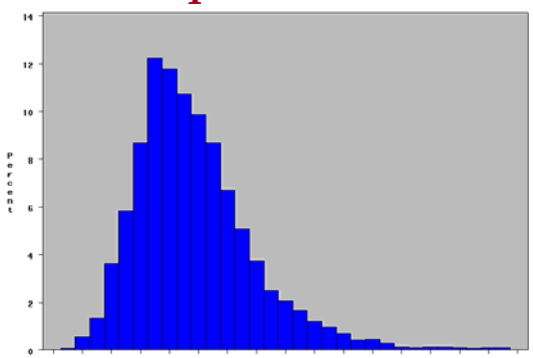
FDR: imputation par la moyenne



FDR: imputation par hot deck aléatoire



FDR: imputation équilibrée



TRAVAUX FUTURS

- Estimation de la variance:
  - Si les contraintes d'équilibrage sont exactement satisfaites, on peut utiliser n'importe quelle méthode d'estimation de la variance
  - Sinon, il y a une certaine variance due à la phase d'atterrissage

- L'imputation équilibrée peut être utilisée pour imputer dans le cas de coefficients bivariés (exemple: coefficient de corrélation)