le cnam

STA108 : Enquêtes et sondages

Pratique des redressements

Philippe Périé, janvier 2014

Pratique des redressements

- Utilisez au mieux les sources d'information auxiliaire
- Principe du redressent, les effets
- Une pratique à ne pas banaliser
- Objectifs
  - Corriger les biais
  - Diminuer la variance
- Exemples- marketing- soirées electorales
- Conditions pour qu'un redressement soit efficace (exemple avec la post stratification simple)
- Mise en œuvre pratique
- Peut on se fier aux redressements ?
- Bibliographie

(rappel) Utilisez au mieux les sources d'information auxiliaire

Il y a un principe fondamental à retenir en enquête : lorsqu'on dispose d'une information auxiliaire, il faut chercher à l'utiliser pour obtenir des estimations plus précises.

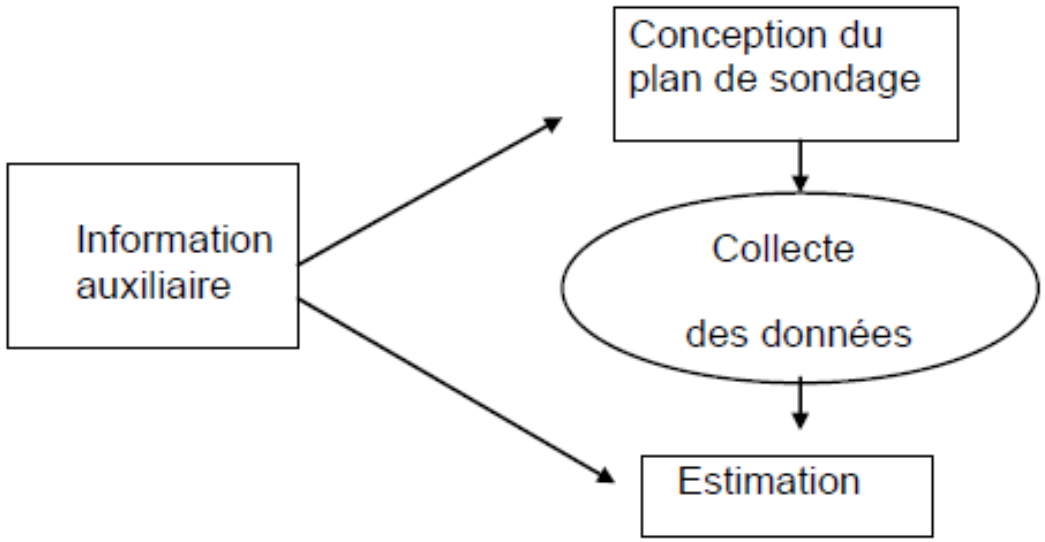
Cette information peut être utilisée à deux moments : au moment du tirage de l'échantillon ou au moment du calcul de l'estimateur.

- 1. Au moment du tirage, on utilise des techniques de stratification, de tirage proportionnel à un critère de taille, ou de tirage équilibré
- 2. Au niveau du calcul de l'estimateur, on utilise des techniques de redressement, d'estimation par ratio ou par régression. Le gain d'efficacité réalisé au niveau du calcul de l'estimateur est particulièrement intéressant car il est obtenu à un coût très faible relativement à ceux générés par la mise en place de techniques de tirage telles que la stratification, ou le tirage équilibré.

(rappel) Utilisez au mieux les sources d'information auxiliaire

Information auxiliaire intégrée avant ou après la collecte des données

FIG. 1 - *Les deux étapes de l'utilisation de l'information auxiliaire*



Principe du redressement

Le redressement est un procédé destiné à améliorer la précision des estimations. Il consiste

- à vérifier dans l'échantillon la distribution de quelques variables qui présentent une relation plausible avec la variable que l'on veut connaître et dont les valeurs réelles sont connues au niveau de l'ensemble de la population
- puis à en rétablir les distributions exactes par un jeu de pondérations quand l'échantillon s'en écarte

Il s'agit de prendre en compte des informations sur la population (post-stratification) afin de corriger les distorsions dues à des erreurs de non-observation (erreurs de couverture et/ou de non-réponse)

Ce type de correction est plus courant pour les échantillons non-probabilistes (eg quota), ou dans les échantillons probabilistes entachés d'importants erreurs de non-observation

Redresser pour quoi faire ?

Le redressement est donc une des techniques permettant d'intégrer de l'information auxiliaire dans le but d'améliorer la précision de l'enquête.

- Souvent, on perçoit le redressement comme un passage obligé permettant de rattraper un plan de sondage déficient, idéalement, on préfèrerait s'en passer.
- C'est faux : bien utilisée, c'est une technique d'amélioration de la qualité globale de l'enquête.

C'est la méthode la plus utilisée en études de marché, mais il en existe d'autres, telles que l'estimation par la méthode du ratio ou par la régression

Une pratique qu'il ne faut pas banaliser

Le redressement est trop souvent considéré comme une simple étape « informatique », permettant de caler mécaniquement la structure de l'échantillon sur celle de la population étudiée

Cela fini par devenir une pratique de « maquillage d'échantillon », ayant pour but de corriger les écarts entre quotas demandés et quotas réalisés

Comme toute autre phase de l'enquête, le redressement doit être préparé en amont : il faut penser à poser les bonnes questions, codées de façon homogène aux données de référence les plus récentes, en prenant garde aux unités statistiques (ménages vs individus, entreprises vs établissements, …)

Le choix des critères

Afin de réduire la variance des estimateurs, les variables de redressement doivent être :

- le plus corrélées possible aux thématiques de l'enquête. Leur sélection dépend donc du sujet traité : nombre de personnes au ménage, présence d'enfants, type et équipement du logement, « restitution » du vote à une élection antérieure, …
- peu nombreuses et agrégées de façon pertinente (afin d'éviter des effets mal maîtrisés)

Les non-répondants aux questions utilisées dans le redressement doivent être éliminés ou laissés à leur poids (si l'on veut éviter des hypothèses trop fortes à leur égard)

Objectifs du redressement

1. **Corriger les biais,** assurer la meilleure représentativité possible d ’un échantillon par rapport à la population cible, selon des critères définis par les objectifs de l ’enquête, en tenant compte des sur- et sous-représentations décidées a priori (sur- et sous-échantillons raisonnés) et des biais induits par :
   - le terrain
   - les non réponses (complètes ou partielles)
   - aléas de l’échantillonnage
   - soucis de comparabilité

2. **Améliorer la précision**

1 - Correction des biais : Correction de la non réponse (1/8)

Ne pas redresser revient à attribuer aux non-répondants le comportement moyen de l'ensemble des répondants, ce qui constitue souvent une grossière erreur

Il est bien connu que les non-répondants se trouvent plus particulièrement dans des catégories sociales spécifiques (personnes âgées, femmes, personnes à faible niveau d'instruction, …)

D'habitude il est préférable attribuer aux non-répondants le comportement moyen des répondants appartenant aux mêmes catégories sociales

1 - Correction des biais : études marketing (2/8)

… Ne pas prendre de critères par défaut choisir ceux qui sont liés à la variable d'intérêt

- Un exemple : Analyse de la consommation dans le segment ultra frais (yaourts, desserts chocolatés, crèmes desserts, etc …).
- Il faut s'assurer du bon calage de l'échantillon sur le revenu, la taille du foyer, la présence et le nombre d'enfants de moins de 6 ans, 15 ans, etc …

1 - Correction des biais - Opération d'estimation en soirée électorale (3/8)

L'échantillon est constitué de bureaux de vote. Les résultats ne sont pas issus de sondages 'sorties des urnes', mais bien des dépouillements des bureaux. Les résultats de la soirée sont 'remontés' bureau par bureau, et recalculés en permanence.

Les premiers bureaux analysés sont ceux qui ferment à 18h et 19h : donc si l'échantillon total de tous les bureaux entrants dans l'estimation est représentatif, donc susceptible de donner de bonnes estimations, l'échantillon analysé à un temps t a toutes les chances d'être décalé. En particulier on fait l'annonce a 20h avec un échantillon de bureaux de 18h et 19h …

A chaque instant, il faut le caler sur les résultats nationaux des élections de référence selon les tendances politiques

Correction du biais (4/8)

Estimateurs bruts et calés en soirée électorale : 19h25

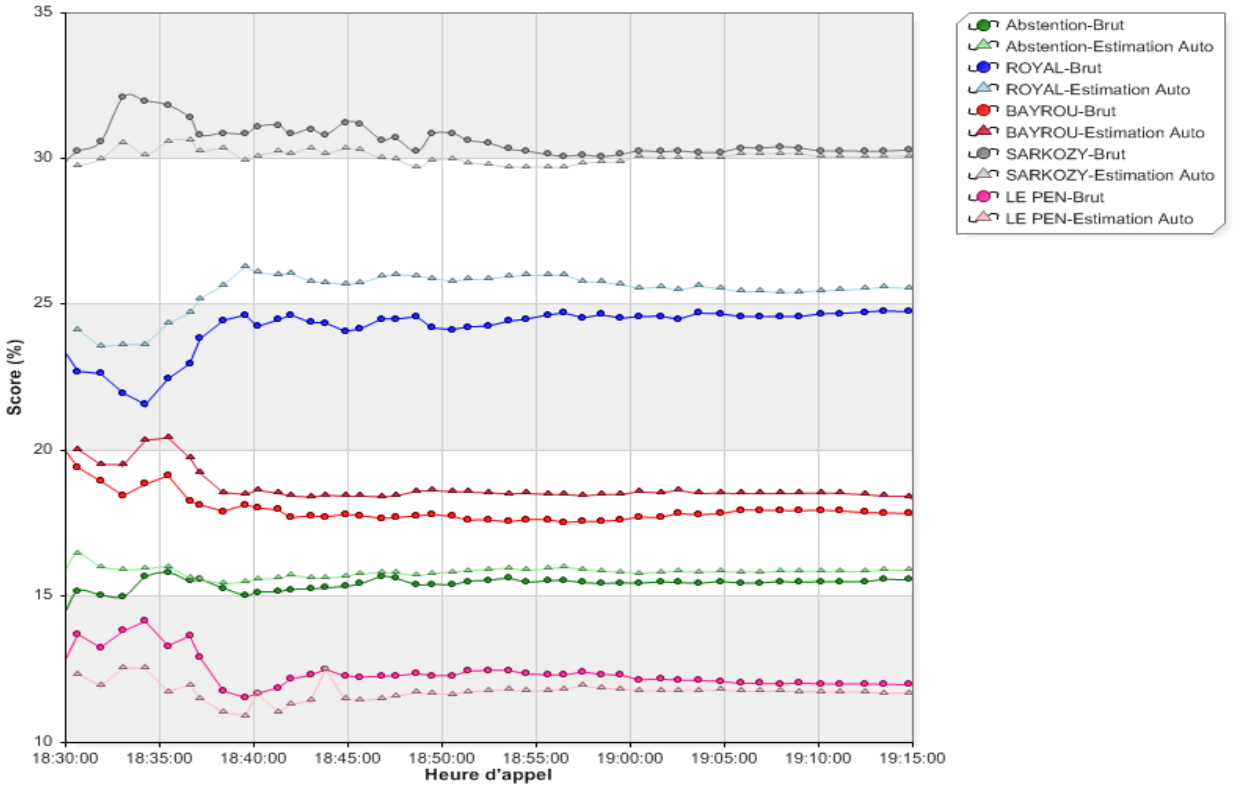
Visu estimation voix

19:25:31	0	1	2	3	4	Total	Echantillon
Nb bureaux	0	0	95	0	63	158	219

Echantillon NATIONAL						
Candidats	Scénarios	INS/EXP	BRUT	CORREL.	ESTIM	INCERT.
Abstention	ABSTENTION auto P02.1	INS	15.68	0.75	15.95	0.43
LAGUILLER	auto P02.1	EXP	1.56	0.42	1.49	0.14
BESANCENOT	auto P02.1	EXP	4.58	0.59	4.43	0.25
SCHIVARDI	auto P02.1	EXP	0.44	0.44	0.43	0.07
BOVE	auto P02.1	EXP	1.24	0.29	1.26	0.13
BUFFET	auto P02.1	EXP	1.81	0.89	1.82	0.13
ROYAL	JOSPIN+TAU+CHE P02.1	EXP	24.75	0.77	25.39	0.73
ROYAL	auto P02.1	EXP	24.75	0.81	25.44	0.68
ROYAL	auto R04	EXP	24.75	0.72	24.82	0.80
VOYNET	auto P02.1	EXP	1.44	0.44	1.52	0.12
BAYROU	BAYROU P02.1	EXP	17.81	0.65	18.13	0.61
Candidats	Scénarios	INS/EXP	BRUT	CORREL.	ESTIM	INCERT.
BAYROU	auto P02.1	EXP	17.81	0.77	18.31	0.51
SARKOZY	auto R04	EXP	30.28	0.72	30.20	0.84
SARKOZY	auto P02.1	EXP	30.28	0.82	30.19	0.71
SARKOZY	CHIRAC+DVD P02.1	EXP	30.28	0.61	30.06	1.03
VILLIERS	auto P02.1	EXP	2.63	0.24	2.58	0.32
LE PEN	auto P02.1	EXP	12.10	0.81	11.80	0.47
LE PEN	LE PEN+MEG P02.1	EXP	12.10	0.73	11.85	0.75
LE PEN	auto R04	EXP	12.10	0.73	11.80	0.52
NIHOUS	auto P02.1	EXP	1.36	0.76	1.23	0.12

Correction du biais (5/8)

graphiques entre 18h30 et 19h15



Correction du biais (6/8)

- L'échantillon entre 18h30 et 19h15 était biaisé (bureaux qui fermaient à 18h):
  - Il surestimait le score de J.M. Le Pen
  - Il sous estimait le score de Ségolène Royal
  - Il sous estimait le score de François Bayrou
  - Il est bon sur le score de Nicolas Sarkozy
  - Il est bon sur l'abstention
- Sur ce graphique, le calage se faisait sur les résultats du premier tour de l'élection présidentielle 2002. Lors de l'opération, il y avait le choix entre 3 élections de référence en calage possible (en 2002 présidentielle et législatives 1er tour, en 2004 régionales)
- Notez que le biais diminue au cours du temps (l'échantillon se complète et donc revient à sa structure initiale)

Correction du biais (7/8)

Estimateurs bruts et calés en soirée electorale : 20h25

Visu estimation voix

20:24:02	0	1	2	3	4	Total	Echantillon
Nb bureaux	0	0	10	0	162	172	219

Echantillon NATIONAL

Candidats	Scénarios	INS/EXP	BRUT	CORREL.	ESTIM	INCERT.
Abstention	ABSTENTION auto P02.1	INS	15.60	0.72	15.76	0.44
LAGUILLER	auto P02.1	EXP	1.44	0.55	1.38	0.09
BESANCENOT	auto P02.1	EXP	4.53	0.71	4.35	0.17
SCHIVARDI	auto P02.1	EXP	0.41	0.44	0.40	0.05
BOVE	auto P02.1	EXP	1.30	0.34	1.32	0.09
BUFFET	auto P02.1	EXP	1.94	0.94	1.94	0.11
ROYAL	JOSPIN+TAU+CHE P02.1	EXP	25.42	0.82	25.81	0.61
ROYAL	auto R04	EXP	25.42	0.78	25.36	0.67
ROYAL	auto P02.1	EXP	25.42	0.86	25.83	0.56
VOYNET	auto P02.1	EXP	1.46	0.61	1.53	0.07
BAYROU	auto P02.1	EXP	18.01	0.81	18.44	0.40
Candidats	**Scénarios**	**INS/EXP**	**BRUT**	**CORREL.**	**ESTIM**	**INCERT.**
BAYROU	BAYROU P02.1	EXP	18.01	0.68	18.32	0.50
SARKOZY	auto R04	EXP	29.97	0.78	30.13	0.71
SARKOZY	CHIRAC+DVD P02.1	EXP	29.97	0.71	29.93	0.86
SARKOZY	auto P02.1	EXP	29.97	0.88	30.04	0.57
VILLIERS	auto P02.1	EXP	2.55	0.24	2.51	0.28
LE PEN	auto R04	EXP	11.63	0.76	11.32	0.46
LE PEN	LE PEN+MEG P02.1	EXP	11.63	0.78	11.49	0.67
LE PEN	auto P02.1	EXP	11.63	0.85	11.44	0.38
NIHOUS	auto P02.1	EXP	1.34	0.84	1.24	0.10

Correction du biais (8/8)

Estimateurs bruts et calés en soirée électorale : 23h et fin

Visu estimation voix

23:18:21	0	1	2	3	4	Total	Echantillon
Nb bureaux	0	0	0	0	219	219	219

Echantillon NATIONAL

Candidats	Scénarios	INS/EXP	BRUT	CORREL.	ESTIM	INCERT.
Abstention	ABSTENTION auto P02.1	INS	15.44	0.62	15.08	0.43
LAGUILLER	auto P02.1	EXP	1.32	0.63	1.33	0.08
BESANCENOT	auto P02.1	EXP	4.23	0.79	4.24	0.14
SCHIVARDI	auto P02.1	EXP	0.37	0.50	0.39	0.04
BOVE	auto P02.1	EXP	1.27	0.44	1.30	0.07
BUFFET	auto P02.1	EXP	1.99	0.94	1.95	0.10
ROYAL	JOSPIN+TAU+CHE P02.1	EXP	26.28	0.86	25.73	0.55
ROYAL	auto P02.1	EXP	26.28	0.90	25.82	0.48
ROYAL	auto R04	EXP	26.28	0.81	26.05	0.62
VOYNET	auto P02.1	EXP	1.47	0.58	1.49	0.06
BAYROU	BAYROU P02.1	EXP	18.41	0.70	18.57	0.43
Candidats	**Scénarios**	**INS/EXP**	**BRUT**	**CORREL.**	**ESTIM**	**INCERT.**
BAYROU	auto P02.1	EXP	18.41	0.83	18.60	0.33
SARKOZY	auto R04	EXP	30.70	0.82	30.93	0.67
SARKOZY	CHIRAC+DVD P02.1	EXP	30.70	0.80	30.25	0.72
SARKOZY	auto P02.1	EXP	30.70	0.91	30.64	0.52
VILLIERS	auto P02.1	EXP	2.31	0.29	2.33	0.22
LE PEN	LE PEN+MEG P02.1	EXP	10.58	0.81	11.08	0.57
LE PEN	auto P02.1	EXP	10.58	0.88	10.88	0.33
LE PEN	auto R04	EXP	10.58	0.79	10.67	0.39
NIHOUS	auto P02.1	EXP	1.07	0.86	1.15	0.07

2- Diminution de la variance (1/5)

… Si les critères de calage sont liés à la variable d'intérêt (là aussi !)

- Si on reprend l'exemple de l'Analyse de la consommation dans le segment ultra frais le redressement permet de fixer les totaux sur les variables de calage. Ces éléments ne sont alors plus source de variabilité (ils ne sont plus aléatoires)
- La part de variabilité qui leur est directement due est donc éliminée
- Sur l'exemple des soirées électorales, on visualise ce résultat sur le graphique des estimations au cours du temps : les courbes estimées 'bougent moins' : elles sont plus stables

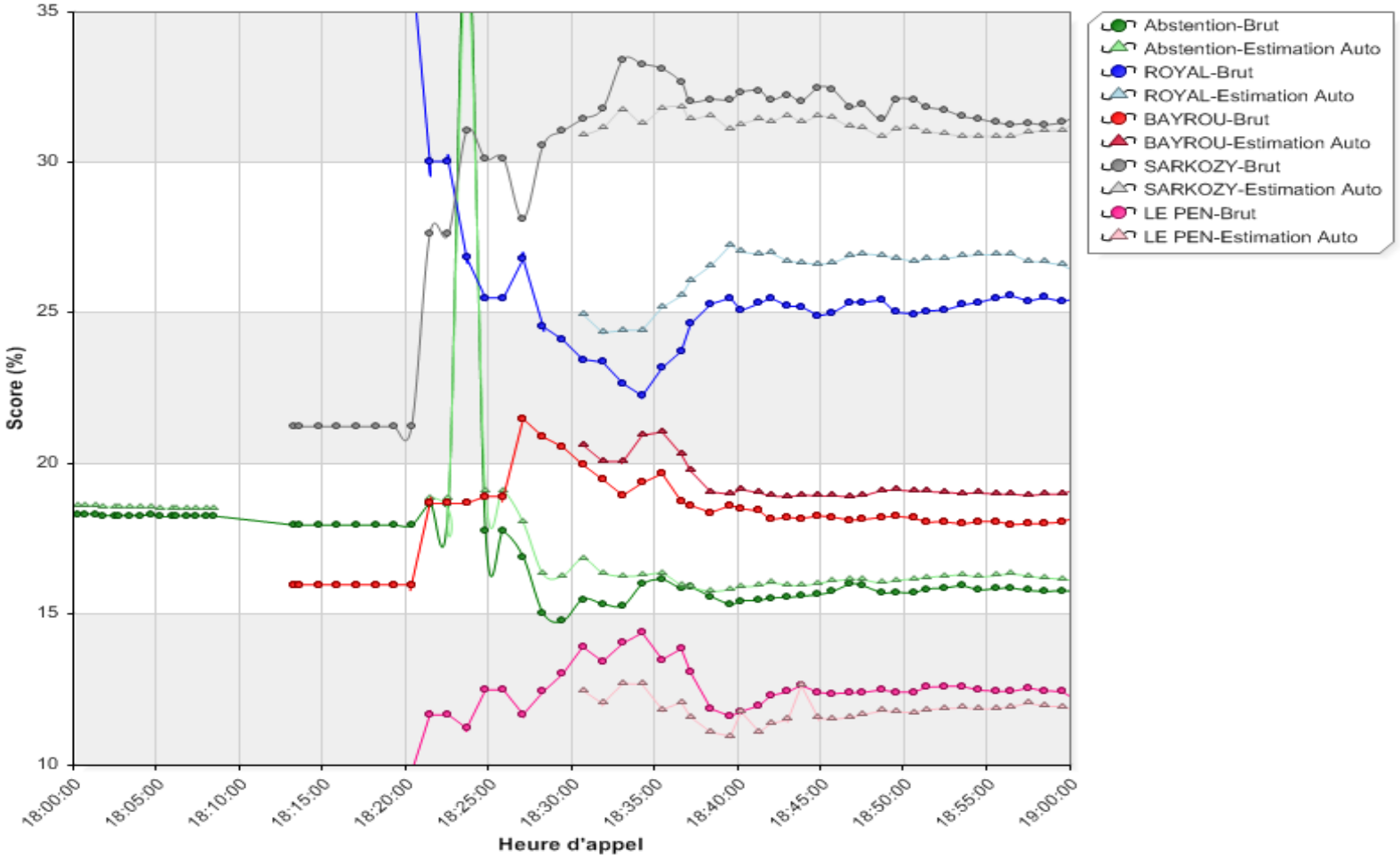
Un exemple d'estimateur par le ratio : soirées électorales (2/5)

Lors des soirées électorales, on utilise aussi un estimateur par le ratio, ainsi que des estimateurs par la régression).

Lors de chaque soirée, on dispose des informations sur les élections précédentes, ce qui fait que l'incertitude n'est plus sur l'étendue possible des scores (0 à 100) mais sur le RATIO d'évolution entre les deux élections considérées.

Cette incertitude est plus forte entre deux élections séparée par des années, qu'entre deux tours séparés par quelques jours

Diminution de la variance : les courbes estimées sont plus lisses (3/5)



2- Diminution de la variance
Présidentielles 2007 – 2ème tour (4/5)

La corrélation est de 0.98, donc le R^2 est de 0.96 : l'incertitude sur le score de Royal n'est plus que de 4% par rapport au cas où l'on n'aurait pas utilisé l'information du 1er tour !

Imprimer | Recharge | Recharge A | 5 sec

20:03:55	0	1	2	3	4	Total	E
Nb bureaux	0	0	12	0	160	172	

Echantillon NATIONAL									
Tendance	Scénarios	INS/EXP	Résultat brut	Corrélation	Ecart	Indice	Brut-Ecart	Brut-Indice	Incertitude
Abstention	ABSTENTION P07.1	INS	15.40	0.94	14.5	14.5	0.9	0.9	0.22
Abstention	ABS+ 2BAY+.30LP P07.1	INS	15.40	0.93	14.4	14.7	1.0	0.7	0.23
Abstention	ABSTENTION auto P07.1	INS	15.40	0.94	14.5	14.5	0.9	0.9	0.22
Abstention	auto tout	INS	15.40	0.94	14.6	14.6	0.8	0.8	0.20
Blancs nuls	BLANCS+NULS P07.1	INS	3.74	0.36	3.7	3.5	0.1	0.2	0.13
Blancs nuls	BN auto P07.1	INS	3.74	0.36	3.7	3.5	0.1	0.2	0.13
SARKOZY	SARKOZY P07.1	EXP	52.89	0.90	53.8	54.5	-0.9	-1.6	0.69
SARKOZY	SARKOZY+.4bay P07.1	EXP	52.89	0.87	54.1	54.6	-1.2	-1.7	0.75
SARKOZY	Sarkozy Compo P07.1	EXP	52.89	0.98	53.3	53.3	-0.4	-0.4	0.43
SARKOZY	auto P07.1	EXP	52.89	0.98	53.1	53.1	-0.2	-0.2	0.38
ROYAL	ROYAL P07.1	EXP	47.11	0.91	47.2	47.3	-0.1	-0.2	0.71
ROYAL	ROYAL+.9exg P07.1	EXP	47.11	0.96	46.8	46.7	0.3	0.4	0.44
ROYAL	ROYAL+exg+.4bay P07.1	EXP	47.11	0.97	47.1	47.1	0	0	0.40
ROYAL	ROYAL composite P07.1	EXP	47.11	0.98	46.9	46.8	0.3	0.3	0.41
ROYAL	auto P07.1	EXP	47.11	0.98	46.9	46.9	0.2	0.2	0.38

2- Diminution de la variance - Présidentielles 2007 – 2ème tour (5/5)

Regressions (scenario)

tns sofres — Unilog, a logica CMG company

Imprimer | Recharge | ☑ Recharge Auto | 5 sec

20:48:35	0	1	2	3	4	Total	Echantillon
Nb bureaux	0	0	38	0	171	209	219

Strate Total / Total - Election Présidentielles 2007 - 1 Tour (NATIONAL)

Entité politique courante	Scénario	Abstention	Blancs nuls	laguil+schiv+Buffet	besancenot	bove+voynet	Royal	bayrou	sarkozy	villiers	lepen+nhious
Abstention	ABSTENTION auto P07.1	0.99	-	-	-	-	-	-	-	-	-
Abstention	auto tout	0.91	0	0.01	0	0	0	0.08	0	0	0.01
Blancs nuls	BN auto P07.1	-	1.00	-	-	-	-	-	-	-	-
SARKOZY	auto P07.1	-	-	0	0	0	0	0.48	1.00	1.00	0.89
ROYAL	auto P07.1	-	-	1.00	1.00	1.00	1.00	0.52	0	0	0.11

Le score de S. Royal au deuxième tour est constitué de :

- 100% de [Laguiller + Schivardi + Buffet] + 100% de Besancenot + 100% de [Bove+Voynet] + 100% de Royal 1er tour + 48% de Bayrou + 11% de LePen…

Quand un redressement est il efficace du point de vue de la précision ?

Il est préférable de :

- Travailler sur un gros échantillon
- Découper les strates à partir de critères liés aux questions clés
- D'avoir suffisamment de strates pour que chaque strate soit homogène, mais pas trop : il faut que les strates soient suffisamment grandes

Une condition essentielle : Il faut que l'information exogène soit fiable (précise et d'actualité), sinon, on introduit un biais dans l'estimation des proportions et moyennes. Sources INSEE (enquête emploi, recensement, SIRENE …), …

➔ **Nous le montrons dans ce qui suit pour le cas d'une postratification simple**

Précision d'un estimateur poststratifié
Post Stratification simple (1/7)

L'estimateur d'une moyenne de la population est donné par :

$$\hat{\mu}_{post} = \sum_{n=1}^{h} \frac{N_h}{N} \mu_h$$

Où représente l'effectif des strates a posteriori et pour chaque $\hat{\mu}_h$:

$$\hat{\mu}_h = \frac{1}{n_h} \sum_{n=1}^{h} X_{hi}$$

Précision d'un estimateur poststratifié
Post Stratification simple (2/7)

C'est une formule en apparence identique à celle d'un estimateur stratifié $\hat{\mu}_{st}$

Mais il y a une différence importante :

- Dans le calcul de $\hat{\mu}_{st}$ les $\hat{\mu}_h$ sont basées sur des tailles fixées à l'avance
- Dans le calcul de $\hat{\mu}_{post}$ les $\hat{\mu}_h$ sont basées sur des tailles qui ne sont pas fixées à l'avance mais des résultats constatés sur l'échantillon, les n_h sont aléatoires

Précision d'un estimateur poststratifié
Post Stratification simple (3/7)

Nous ne détaillons pas les calculs assez complexes, mais on montre que la variance de l'estimateur post stratifié est donné par :

$$Var(\hat{\mu}_{post}) \cong \frac{(1-f)}{n}\sum_{h=1}^{k}\frac{N_h}{N}S^2{}_{h,c} + \frac{(1-f)}{n^2}\sum_{h=1}^{k}\frac{N-N_h}{N}S^2{}_{h,c}$$

C'est une somme de 2 termes :

(1) Terme classique d'un estimateur stratifié

(2) Terme supplémentaire 'pénalité' correspondant au caractère aléatoire des strates, ce terme tend vers 0 quand la taille des post strates augmente

Précision d'un estimateur poststratifié
Post Stratification simple (4/7)

Pour un total, on montre de manière analogue que :

$$Var\left(\hat{T}_{post}\right) \cong N\left(\frac{(1-f)}{n}\sum_{h=1}^{k} N_h {S^2}_{h,c} + \frac{(1-f)}{n^2}\sum_{h=1}^{k} N - N_h {S^2}_{h,c}\right)$$

Précision d'un estimateur poststratifié
Post Stratification simple (5/7)

On rappelle que :

$$Var(\hat{\mu}_{sas}) \cong \frac{(1-f)}{n}\sum_{h=1}^{k}\frac{N_h}{N}S^2{}_{h,c} + \frac{(1-f)}{n^2}\sum_{h=1}^{k}\frac{N_h}{N}(\overline{X_h}-\mu)^2$$

et :

$$Var(\hat{\mu}_{post}) \cong \frac{(1-f)}{n}\sum_{h=1}^{k}\frac{N_h}{N}S^2{}_{h,c} + \frac{(1-f)}{n^2}\sum_{h=1}^{k}\frac{N-N_h}{N}S^2{}_{h,c}$$

Précision d'un estimateur poststratifié
Post Stratification simple (6/7)

En comparant les deux termes, on a :

$$\frac{(1-f)}{n}\left(Var(\hat{\mu}_{sas}) - Var\left(\hat{\mu}_{post}\right)\right)$$
$$= \sum_{h=1}^{k} \frac{N_h}{N}(\overline{X_h} - \mu)^2 - \frac{1}{n}\sum_{h=1}^{k} \frac{N - N_h}{N} {S^2}_{h,c}$$

La stratification est bénéfique quand cette quantité est positive

Précision d'un estimateur poststratifié
Post Stratification simple (7/7)

On vérifie donc bien que :

1. La variable étudiée doit être corrélée avec le critère de stratification, c'est-à-dire avoir une valeur faible de variance intra strate
2. La taille n de l'échantillon doit être importante, pour que 1/n tende var 0 : inutile de redresser de petit échantillons
3. Le rapport $\frac{N-N_h}{N}$ doit être petit, donc $\frac{N_h}{N}$ grand : il est inutile d'avoir beaucoup de petites strates

Les contrôles à opérer

- Il est important d'opérer une validation préalable de la structure brute d'échantillon, sur un ensemble de variables critiques, qu'elles aient fait l'objet de quotas ou qu'elles soient utilisées comme simples variables de contrôle

- Après redressement, il faut vérifier la distribution des poids générés : min, max, quantiles et courbes de fréquence, indicateurs de forme du type

$$100 * (\sum \text{poids})^2 / n \sum \text{poids}^2$$

(*) Cela vaut 100 si tous les poids sont égaux, entre 50 et 70 s'il y a une forte dispersion; à moins de 50 le redressement est à revoir …

Programmes de redressement
La macro SAS CALMAR

L'INSEE met à disposition des utilisateurs sur son site un macro programme SAS permettant de réaliser des redressements http://www.insee.fr/fr/methodes/default.asp?page=outils/calmar/accueil_calmar.htm

« La macro SAS CALMAR (CALage sur MARges) permet de redresser un échantillon provenant d'une enquête par sondage, par re pondération des individus, en utilisant une information auxiliaire disponible sur un certain nombre de variables, appelées variables de calage. Le redressement consiste à remplacer les pondérations initiales (ou "poids de sondage") par de nouvelles pondérations telles que :

pour une variable de calage catégorielle (ou "qualitative"), les effectifs des modalités de la variable estimés dans l'échantillon, après redressement, seront égaux aux effectifs connus sur la population ;

pour une variable numérique (ou "quantitative"), le total de la variable estimé dans l'échantillon, après redressement, sera égal au total connu sur la population.

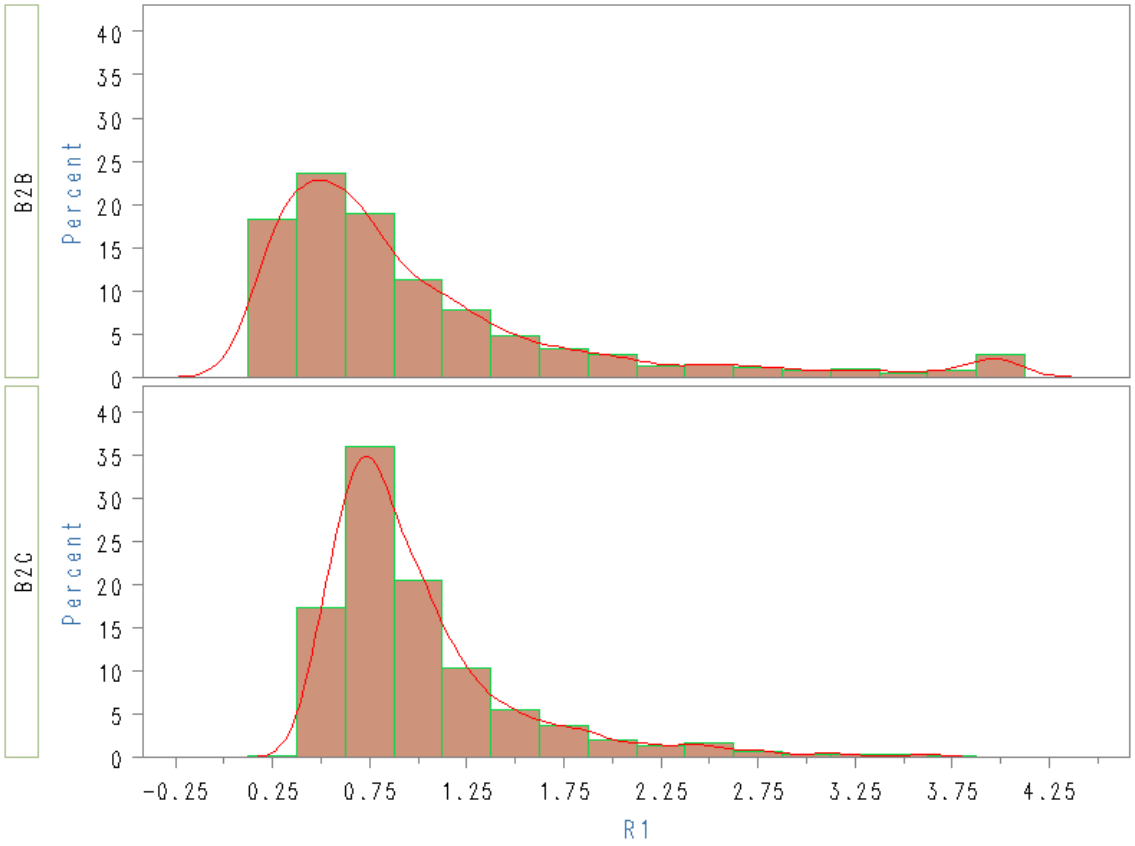
Cette méthode de redressement permet de réduire la variance d'échantillonnage, et, dans certains cas, de réduire le biais dû à la non réponse totale »

Programmes de redressement RIM, Raking Ratio

D'autres programmes utilisent la méthode beaucoup plus simple et robuste dite RIM (Reweight Iterative Method, Deming, Stephan, 1940) C'est le cas du logiciel Quantum utilisé par pratiquement tous les instituts d'étude de marchés. La méthode consiste dans le cas d'une postratification à réaliser des calculs de règles de 3 pour caler sur les différentes marges en itérant jusqu'à convergence. On parle aussi de Raking Ratio ('to rake' en anglais veut dire ratisser, ce qui fait référence au fonctionnement de l'algorithme).

CE Särndal et JC Deville ont montré que la méthode RIM est un cas particulier de la méthode CALMAR, que l'on peut d'ailleurs réaliser avec certaines options du programme SAS %Calmar

Distribution des poids



Vérification des sorties, attention si :

Non convergence (moins de 8-10 itérations avec RIM) : il faut revoir les critères de redressement (objectifs irréalistes) ou modifier les critères d'arrêt

Rapports de poids élevés (> 4)

Efficacité < 75% dans chaque strate.

$$Efficacité = \frac{100\left(\sum_j p_j\right)^2}{n_{brut} * \sum_j p_j^{\,2}}$$

Reflète la déformation de l'échantillon

Efficacité d'un redressement et notion de 'Base effective'

Cet indice compris entre 0 et 1 reflète la déformation des probabilités d'inclusion due au redressement. Il est d'autant plus faible que la dispersion de poids dans la strate considérée est importante

Interprétation : Si l'efficacité sur un redressement est de 75% sur un éch. de 1000, alors, la base ne « vaut » en fait que 750 individus (base effective).

Une remarque importante : cet indice est valide pour mesurer une dispersion des poids au sein d'une strate dans laquelle tous les individus ont le même taux de sondage. Si on veut vraiment mesurer un indice d'efficacité dans ce cas, on ne doit pas le calculer globalement, mais strate par strate, puis les agréger ensuite

Efficacité d'un redressement et 'Base effective'

L'inverse est le 'Design Effect Weight'

$$deff = \frac{n\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2}$$

$$deff \geq 1$$

Le 'Design Effect' est intégré dans toutes les formules de test et d'intervalle de confiance ; il représente l'accroissement de variance du à la déformation de l'échantillon durant le redressement. Un IC à 95% pour une proportion est donc :

$$\hat{p} \pm \left(\sqrt{deff} \times 1.96 \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right)$$

La pratique des redressements dans les études marketing

Région croisée par tranche d'unité urbaine, sexe, âge, CSP de la « personne de référence » ou de l'individu, présence d'enfants, niveau d'études, pratiques médias … sont les variables le plus souvent utilisées dans les études marketing

Le plus important c'est de :

- toujours utiliser des données de référence fiables et à jour
- veiller à redresser en plusieurs étapes s'il le faut : d'abord une première pondération - par ex. ménage ou pays -, ensuite un calage sur marges portant sur les variables « individu »
- rester aussi critiques que possible sur les éventuels erreurs de mesure commises

La pratique des redressements dans les études politiques (1/3)

Dans le domaine politique, l'utilisation des variables de redressement socio-démographiques est insuffisante. Elle ne permet pas de combattre le biais le plus important, qui est la sous- représentation chronique des certaines couches de l'électorat (notamment à l'extrême droite), entraînant automatiquement la sur- représentation d'autres groupes

On s'appuie donc sur la reconstitution d'un ou plusieurs scrutins intérieurs (en demandant par exemple aux interviewés comment ils avaient voté aux précédentes présidentielles) pour obtenir des distributions que l'on compare à la réalité

On calcule alors des coefficients de pondération, en espérant que le rétablissement des familles politiques à leur niveau réel du scrutin antérieur permettra leur représentation exacte dans les intentions de vote pour le scrutin à venir

La pratique des redressements dans les études politiques (2/3)

Le redressement fondé sur la « restitution de vote » constitue un correctif très puissant (les poids générés peuvent aller jusqu'à multiplier par trois le poids de certains individus !)

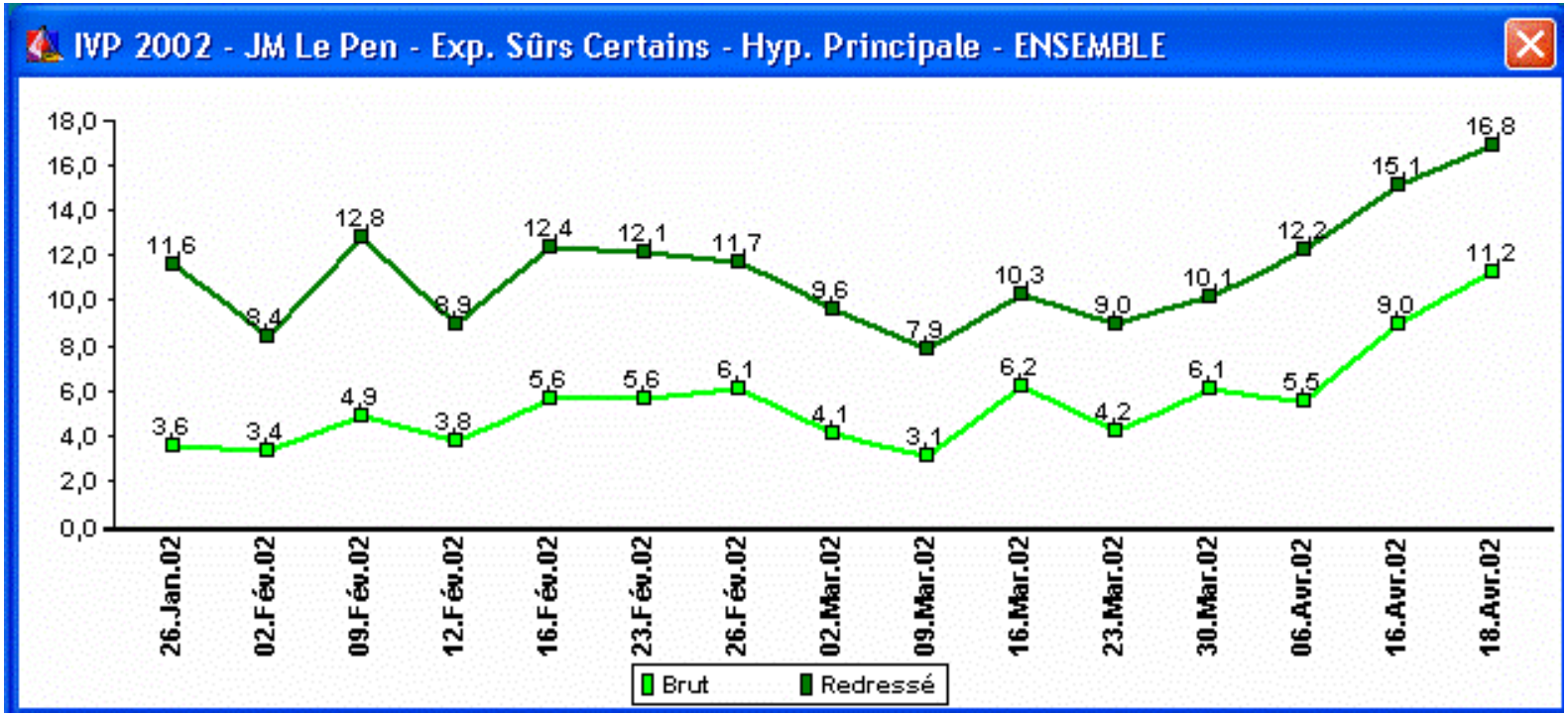
Son utilisation est cependant à manier avec précaution

- L'interviewé peut avoir oublié comment il a voté à l'élection restituée, qui a pu avoir eu lieu cinq ans auparavant
- Il peut confondre ses dispositions actuelles avec celles d'alors
- Il peut également dissimuler le vote précédent, afin de le rendre cohérent avec ses déclarations actuelles

On risque donc de sur-corriger les intentions enregistrées en voulant rétablir la « vraie » distribution des votes passés. L'inverse (sous-correction) est aussi possible …

La pratique des redressements dans les études politiques (3/3)

- Indications, contre-indications, précautions d'emploi …
- Pour le reste, un exemple vaut mieux que mille discours …



Peut-on se fier aux redressements ?

- Le redressement est indispensable
  - Correction des erreurs de non-observation
  - Standardisation des structures à des fins de comparaison
- Le redressement ne peut pas corriger les erreurs de mesure
  - Déclarations de revenus
  - Restitutions et intentions de vote
  - …
- Le redressement peut augmenter les biais
  - Poids délirants > limitation des poids (eg. de 0,25 à 4,00)
  - Disponibilité de données de référence fiables et récentes, codées de façon homogène

Bibliographie

- W. Edwards Deming and Frederick F. Stephan, “On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known” Ann. Math. Statist. Volume 11, Number 4 (1940), 427-444.http://projecteuclid.org/DPubS?service=UI&version=1.0&verb=Display&handle=euclid.aoms/1177731829
- Ardilly, P. (*2ème édition actualisée et augmentée,* 2006), Les techniques de sondage, Editions Technip, Paris
  - Chapitre III. Amélioration des estimateurs (redressements)
- Lejeune, M., éd. (2001), Traitements des fichiers d’enquêtes. Redressements, injections de réponses, fusions, PUG, Grenoble
- Brossier, G., Dussaix,A.-M., éd. (1999), Enquêtes et sondages, Dunod, Paris
  - Chapitre 5. Méthodes de redressement et de calage
- Brulé, M., « Peut-on se fier aux sondages? »,*Sociétal*, Paris, n° 37, 3e trimestre 2002, pp. 30-33