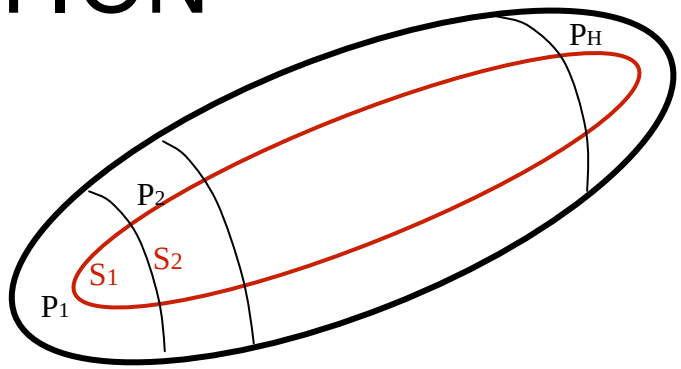
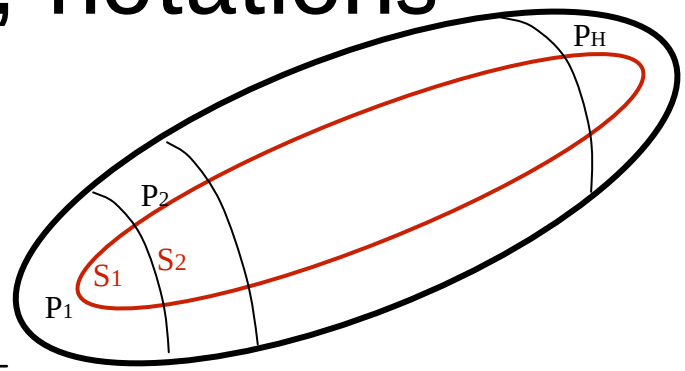
STRATIFICATION



- Utilisation d'une information auxiliaire qualitative
- Toujours efficace

STRATIFICATION, notations



- **Strates:**

$N_1, N_2 \ldots\ldots N_h \ldots\ldots N_H$ $\quad N = \sum N_h$

$\bar{Y}_1, \bar{Y}_2 \ldots\ldots \bar{Y}_h \ldots\ldots \bar{Y}_H$ $\quad \bar{Y} = \sum \frac{N_h}{N} \bar{Y}_h$

$\sigma_1^2, \sigma_2^2 \ldots\ldots \sigma_h^2 \ldots\ldots \sigma_H^2$ $\quad \sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2$

- **Échantillon:**

$n_1, n_2, \ldots\ldots n_h, \ldots\ldots n_H$ $\quad n = \sum n_h$

$\bar{y}_1, \bar{y}_2, \ldots\ldots \bar{y}_h, \ldots\ldots \bar{y}_H$ $\quad \bar{y} = \sum \frac{n_h}{n} \bar{y}_h$

$\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots\ldots \hat{\sigma}_h^2 \ldots\ldots \hat{\sigma}_H^2$

STRATIFICATION

$$\sigma^2 = \sum \frac{N_h}{N} \sigma_h^2 + \sum \frac{N_h}{N} \left(\bar{Y}_h - \bar{Y} \right)^2 = \sigma^2_W + \sigma^2_B$$

- Variance totale=

  moyenne des variances (*variance intra*)

  +variance des moyennes (*variance inter*)

STRATIFICATION

- Estimateur sans biais de $\bar{Y}$ (Horvitz Thomson)

$$\hat{\bar{Y}}_{str} = \sum \frac{N_h}{N} \bar{y}_h$$

- Variance:

$$V(\hat{\bar{Y}}_{str}) = \sum \left(\frac{N_h}{N} \right)^2 V(\bar{y}_h) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{\sigma_h^2}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$= \frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{S_h^2}{n_h}$$

STRATIFICATION, répartition proportionnelle

- Échantillon dit « représentatif »:

$$\frac{n_h}{n} = \frac{N_h}{N} \Rightarrow \tau_h = \frac{n_h}{N_H} = \frac{n}{N} = \tau$$

- Taux de sondage constant dans chaque strate

$$\hat{\bar{Y}}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{n_h}{n} \bar{y}_h = \bar{y} = \hat{\bar{Y}}_{prop}$$

STRATIFICATION, répartition proportionnelle

- variance :

$$V(\hat{\bar{Y}}_{prop}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_h)\frac{S_h^{\,2}}{n_h} = \frac{1}{N^2}\sum_{h=1}^{H}\frac{N_h - n_h}{n_h} N_h S_h^{\,2}$$

$$= \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N_h}{n_h} - 1\right)N_h S_h^{\,2} = \frac{1}{N^2}\sum_{h=1}^{H}\left(\frac{N}{n} - 1\right)N_h S_h^{\,2} = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2}$$

- Si N_h est grand:

$$V(\hat{\bar{Y}}_{prop}) = \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N} S_h^{\,2} \simeq \frac{N-n}{nN}\sum_{h=1}^{H}\frac{N_h}{N}\sigma_h^{\,2} = \frac{N-n}{N}\frac{\sigma^2_{\,w}}{n}$$

STRATIFICATION, répartition proportionnelle

- Variance de l'estimateur du SAS sans remise:

$$V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n} = \frac{N-n}{N}\frac{S^2}{n} \simeq \frac{N-n}{N}\frac{\sigma^2}{n}$$

- Avec les mêmes probabilités d'inclusion d'ordre 1, l'échantillon stratifié représentatif est plus efficace qu'un échantillon simple de même taille dès que les $\bar{Y}_h$ sont différents.

STRATIFICATION optimale

- Répartition optimale:

$$V(\widehat{\overline{Y}_{str}}) = \frac{1}{N^2}\sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2}$$

avec $S_h^{\,2} = \frac{N_h}{N_h - 1}\sigma_h^{\,2}$

c_h – coût unitaire d'une observation

$$\begin{cases} \min \sum \frac{N_h\left(N_h - n_h\right)}{n_h} S_h^{\,2} \\ \sum n_h c_h = c_0 \end{cases}$$

$$\sum \frac{N_h^{\,2}}{n_h} S_h^{\,2} - \underbrace{\sum N_h S_h^{\,2}}_{fixe}$$

STRATIFICATION optimale

- Solution:

$$\frac{N_h^2 S_h^2}{n_h^2}$$ proportionnel à c_h

$$\frac{n_h}{N_h} \propto \frac{S_h}{\sqrt{c_h}}$$

Si c_h constant:

$$n_h = n\frac{N_h S_h}{\sum N_h S_h}$$ - Répartition de Neyman

STRATIFICATION

- Exemple n° 1: présondage de 155 unités

Strates	**1**	**2**	**3**	**4**	
N_h	3750	3272	1387	2475	10 884
n_h	50	45	30	30	155
$\overline{y}_h$	12.6	14.5	18.6	13.8	
$\hat{\sigma}_h^2$	2.8	2.9	4.8	3.2	

STRATIFICATION

- Exemple n° 1:

$$\widehat{\overline{Y}} = \sum \left(\frac{N_h}{N} \right) \overline{y}_h = \frac{3750 \text{x} 12.6 + ... + 2475 \text{x} 13.8}{10884} = 14.21$$

$$\hat{V}(\widehat{\overline{Y}}) \simeq \sum \left(\frac{N_h}{N} \right)^2 \frac{\widehat{\sigma}_h^{\,2}}{n_h} = 0.02059 = (0.14)^2$$

Intervalle de confiance à 95% pour $\overline{Y}$:

$14.21 \pm 2 \text{x} 0.14$ soit: $\left[13.93 < \overline{Y} < 14.49 \right]$

Pour T: 154662 ± 3047

STRATIFICATION

- Exemple n° 1:

$$\sigma^2 = \sum \frac{N_h}{N} \sigma^2_h + \sum \frac{N_h}{N} \left(Y_h - Y \right)^2$$

On estime: σ^2_h par $\frac{n_h}{n_{h-1}} \hat{\bar{Y}}^2_{str} \sigma_h$

$\overline{Y}_h$ par $\overline{y}_h$

$\overline{\overline{Y}}$ par $\hat{\bar{Y}}_{str}$

$$\hat{\sigma}^2 = 6.06 = \left(2.46 \right)^2$$

STRATIFICATION

- Suite: Répartition de Neyman pour n=1000:

N_1S_1 = 6275 n_1 = 1000 x 6275/19 312 = 325

N_2S_2 = 5572 n_2 = 288

N_3S_3 = 3038 n_3 = 157

N_4S_4 = 4427 n_4 = 229

19 312

Variance: $\frac{1}{N^2}\sum\frac{N_h\left(N_h-n_h\right)}{n_h}S_h^{\,2}=0.0029=\left(0.0542\right)^2$

$\overline{Y}$ connu à $\pm$ 2 x 0.0542 soit $\pm$ 0.108

T connu à $\pm$ 1179

STRATIFICATION

- Échantillon simple à 1000:

$$\frac{\sigma^2}{n} \times \frac{N-n}{N-1} = 0.0055 = (0.0742)^2$$

$\overline{Y}$ connu à $\pm$ 0.15; T connu à $\pm$ 1615

- Échantillon stratifié représentatif:

$n_1 = 345$

$n_2 = 301$

$n_3 = 127$

$n_4 = 227$

STRATIFICATION

- Estimation d'une proportion p
- Même démarche: une proportion est une moyenne particulière

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} f_h$$

$$V(\hat{p}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{p_h(1-p_h)}{n_h} \frac{N_h - n_h}{N_h - 1}$$

$$\hat{V}(\hat{p}_{str}) \simeq \sum_{h=1}^{H} \left(\frac{N_h}{N} \right)^2 \frac{f_h(1-f_h)}{n_h} \left(1 - \frac{n_h}{N_h} \right)$$

STRATIFICATION

- Comment stratifier?
  - Remarque préalable: dans un sondage à probabilité inégale π_i proportionnel à Y_i annule la variance.
  - Nombre de strates: le maximum mais...
  - Limites de strates optimales:

    méthode de Dalenius et Hodges. Regrouper des classes selon le cumul de la racine des effectifs

STRATIFICATION

- Répartition dans les strates:
  - Si S_h inconnu : répartition proportionnelle
  - si S_h connu: Neyman
  - sinon, hypothèse fréquente $\frac{S_h}{\overline{Y}_h} = c$ d'où n_h proportionnel à la somme de la variable étudiée ou d'une variable corrélée.
  - Exemple: échantillon d'entreprises proportionnel au CA ou à l'effectif de la strate.

STRATIFICATION

- Variable de stratification: en théorie Y; sinon, variable bien corrélée avec Y.
- En pratique quand il y a plusieurs variables d'intérêt et une variable de stratification , on utilise la répartition proportionnelle