Sondage stratifié

Antoine Rolland

DUT STID 2 - IUT Lumière

Nice - 2 mars 2022

Inégalité de Bienaymé- Tchebitcheff

Si X est une variable aléatoire de variance V_X et de moyenne $\mathbb{E}(X)$, alors

$$P(|\ \mathbb{E}(X) - X\ | \geq \alpha) \leq \frac{V_x}{\alpha^2}$$

On sait que cette inégalité est généralement sous-optimale. Par exemple pour la loi normale centrée réduite,

- $P(|\ \mathbb{E}(X) - X\ | \geq 2) = 0,05$ (et non 0,25)
- $P(|\ \mathbb{E}(X) - X\ | \geq 3) = 0,01$ (et non 0,11)

Application aux sondages

Si X est une variable aléatoire de variance V_X et de moyenne m, alors la moyenne $\overline{X_n}$ d'un échantillon de n réalisations de X est une V.A. de moyenne m et de variance V_X/n.

$$P(\mid m - \overline{X_n} \mid \geq \alpha) \leq \frac{V_x}{\alpha^2 n}$$

Principe du sondage stratifié

Si la population se divise en plusieurs sous-population de variances inférieures à la variance de la population, on peut améliorer la précision du sondage en effectuant des sondages à l'intérieur de chaque sous-population, de manière proportionnelle.

Démonstration

Démonstration avec 2 sous-populations

- de proportions p_1 et p_2 dans la population totale avec $p_1 + p_2 = 1$,
- de moyennes m_1 et m_2 (et donc $p_1 m_1 + p_2 m_2 = m$)
- de variances V_1 et V_2.
- et deux échantillons de taille n_1 et n_2 avec $n_1 = p_1 n$ et $n_2 = p_2 n$.

Démonstration

- $\overline{X}_{1,n}$ est la moyenne de l'échantillon de la sous-population 1
- $\overline{X}_{2,n}$ est la moyenne de l'échantillon de la sous-population 2
- $\overline{X}_{1,2,n} = p_1 \overline{X}_{1,n} + p_2 \overline{X}_{2,n}$ est la moyenne des deux moyennes pondérées par leur importance
- V_1/n_1 est la variance de $\overline{X}_{1,n}$
- V_2/n_2 est la variance de $\overline{X}_{2,n}$
- $\left(p_1^2 \frac{V_1}{n_1} + p_2^2 \frac{V_2}{n_2}\right)$ est la variance de $\overline{X}_{1,2,n}$

Démonstration

Il reste à montrer que

$$\left(p_1^2 \frac{V_1}{n_1} + p_2^2 \frac{V_2}{n_2}\right) \leq \frac{V}{n}$$

Or $p_1^2 \frac{V_1}{n_1} = p_1 \frac{V_1}{n}$ et $p_2^2 \frac{V_2}{n_2} = p_2 \frac{V_2}{n}$,

Comme $V_1 \leq V$ et $V_2 \leq V$,

$$p_1 \frac{V_1}{n} \leq p_1 \frac{V}{n}$$

et

$$p_2 \frac{V_2}{n} \leq p_2 \frac{V}{n}$$

D'où $\left(p_1^2 \frac{V_1}{n_1} + p_2^2 \frac{V_2}{n_2}\right) \leq (p_1 + p_2)\frac{V}{n} = \frac{V}{n}$

Exemple avec deux strates équilibrées

Prenons par exemple la population française simplifiée

- taille moyenne $m = 171$cm avec un écart-type $\sigma = 9,2$
- 50% de femmes de taille moyenne $m_F = 164$cm et d'écart-type $\sigma_F = 6$
- 50% d'hommes de taille moyenne $m_H = 178$cm et d'écart-type $\sigma_H = 6$

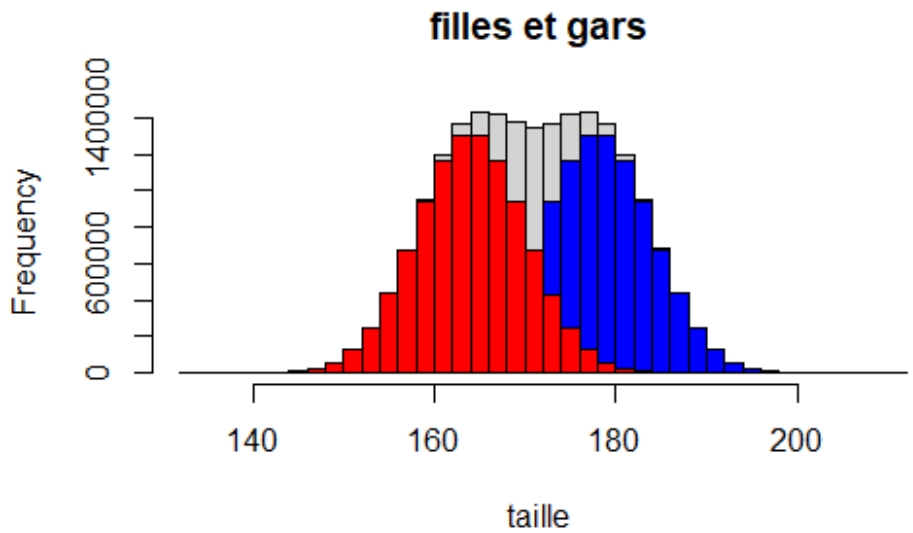
CALCULS THÉORIQUES

Avec l'inégalité de BT regardons la probabilité d'être à plus d'1 cm de m avec un échantillon de taille 1000

- sondage simple : $\frac{V}{n} = \frac{9{,}2^2}{1000}$
  $P(\mid m - \overline{X_n} \mid \geq 1) \leq 0.846$
- sondage stratifié : $\left(p_1^2 \frac{V_1}{n_1} + p_2^2 \frac{V_2}{n_2}\right) = \frac{1}{4} \times \frac{36}{500} + \frac{1}{4} \times \frac{36}{500}$
  $P(\mid m - \overline{X_n} \mid \geq 1) \leq 0.72$

SIMULATIONS AVEC DEUX STRATES ÉQUILIBRÉES

```
filles<-rnorm(10e6, mean=164, sd=6)
gars<-rnorm(10e6, mean=178, sd=6 )
tous<-c(filles, gars)
```



Simulations avec deux strates équilibrées

```
ech_tous<-replicate(1000,mean(sample(tous, 100,
replace = T)))
var(ech_tous)
0.7932108
sum(abs(ech_tous-mean(tous))>=1)
275 (846 au plus d'après BT)
```

SIMULATIONS AVEC DEUX STRATES ÉQUILIBRÉES

```
ech_filles<-replicate(1000,mean(sample(filles,
50, replace = T)))
ech_gars<-replicate(1000,mean(sample(gars, 50,
replace = T)))
moyennes<-(ech_filles+ech_gars)/2
var(ech_filles)
0.7290293
var(ech_gars)
0.7274294
var(moyennes)
0.3621479
sum(abs(moyennes-mean(tous))>=1)
93 (720 au plus d'après BT)
```

PERSPECTIVES

- si V_1 et V_2 sont petits devant V on gagne beaucoup de précision
- à la limite si V_1 et V_2 sont nuls, $n = 2$ suffit (avec la connaissance de p_1 et p_2)
- c'est l'intuition derrière la méthode des quotas : dans chaque tranche identifiée les variances sont nulles (ou faibles) : le vote est 100% dicté par les variables socio-démo.

SIMULATION DANS LE CADRE DU VOTE

Population

```
votants_a<-c(rep(0,900000), rep(1,100000))
votants_b<-c(rep(0,120000), rep(1,880000))
votants<-c(votants_a,votants_b)
mean(votants)
0.49
```

Echantillon pris dans toute la population

```
ech_tout<-replicate(1000,mean(sample(votants,1000)))
sum(ech_tout>0.5)
```

236 soit 23,6% de mauvaise prédiction

Echantillons pris dans les deux sous-populations

```
ech_a<-replicate(1000,mean(sample(votants_a,500)))
ech_b<-replicate(1000,mean(sample(votants_b,500)))
resultat<-(ech_a+ech_b)/2
sum(resultat>0.5)
141 soit 14,1% de mauvaise prédiction
```