3. Echantillonnage, distribution d'échantillonnage, estimation

Cette section présente des techniques d'échantillonnage avec Stata, qui sont utilisées pour les sondages ou avant l'estimation. Nous verrons ce que Ardilly (2006, p. 669) appelle l'approche modèle et l'approche traditionnelle. Les notions importantes ici sont essentiellement celles de tirage sans ou avec remise, échantillon aléatoire, échantillon *bootstrap*, pondération à la Horvitz-Thompson. Cette section souligne aussi les liens entre ces notions et celles de population et super population. Nous adoptons l'approche fréquentiste, qui est indissociable de l'approche modèle. Nous voyons toutes ces notions dans le cadre de l'estimation de l'indicateur de centralité d'une variable, son moment d'ordre 1.

3.1 Approches modèle et traditionnelle

Dans la **section 2**, nous avons présenté des variables aléatoires suivant des lois connues. Nous n'avions pas défini l'unité statistique ou unité d'observation d'une population (*item*, *statistical unit*, *individual*), i.e. l'individu (objet ou sujet) auquel chaque variable se rapportait. Car, il n'y avait pas de "population", au sens où l'on entend d'habitude ce terme. Nous avions un modèle d'échantillonnage, i.e. N variables aléatoires $Y_1, \dots, Y_N$. Chaque valeur y_i était produite par un mécanisme dit processus générateur des données (PGD ou *Data Generating Process*, DGP). Un exemple de PGD : $Y \sim N(10,3)$. Il suffit de faire `di rnormal(10,3)` pour faire un tirage.

L'approche modèle

Selon cette approche, la population $\{1, \dots, N\}$ est un "echantillon aléatoire" tiré (avec remise) dans une **superpopulation** (Ardilly, 2006, p. 669). Par exemple, la moyenne générale de chaque étudiant du M1 EE peut se modéliser de cette manière dans un cadre **semi-paramétrique** (on ne suppose pas de loi précise, mais seulement l'existence de paramètres) : $E(Y_i) \equiv \mu$ et $\sqrt{V(Y_i)} \equiv \sigma$, i.e. $Y_i \sim \text{i.i.d.}$, $i = 1, \dots, 19$. Par abus de langage, on appelle $(Y_1, \dots, Y_{19})$ un échantillon aléatoire (*random sample*) alors que ce sont les **unités d'échantillonnage** (*sample units*) $\{1, \dots, 19\}$ qui sont tirées aléatoirement.

Chaque Y_i est tiré avec remise, ce que sous-entend le "i.i.d." dans la notation précédente (on n'a qu'une superpopulation dans ce cas), et tous les échantillons sont équiprobables. Deux sources d'aléas (deux lois) peuvent se superposer (Ardilly, 2006, p. 556) : le mode de tirage de n unités parmi les N, et le modèle pour Y dans la superpopulation. Dans l'approche modèle, le problème est d'estimer sans biais le(s) paramètre(s) de la superpopulation à partir des valeurs observées de l'échantillon (superpopulation → échantillon). C'est généralement cette approche que l'on adopte quand on spécifie un modèle économétrique.

Dans cette section, on écartera la première source d'aléa (le mode de tirage). Par exemple, si on souhaite connaître la part Y des étudiant.e.s de l'ISEM ($N = 2500$) qui doivent travailler pour payer leurs études, la population de l'ISEM est considérée comme un échantillon aléatoire issu d'une superpopulation $Y \sim B(p)$, l'échantillon c'est le M1 EADE ($n = 19$). Le problème est d'estimer p dans l'échantillon aléatoire $\{1, \dots, 19 = n\}$, avec $Y_j \sim Y$ (les Y_j sont i.i.d.).

L'approche traditionnelle

L'autre approche, dite **traditionnelle**, fonde la nature aléatoire de l'échantillon dans le mode de tirage. L'échantillon est tiré dans la population (population → échantillon). Comme avant, le problème est d'estimer sans biais des paramètres, ceux de la population. La population d'intérêt $\{1, \ldots, N\}$ est notée $\mathcal{P}$, avec N le nombre d'unités d'observation distinctes. Pour Newbold et *alii* (2007), *a population is the complete set of all items (statistical units, individuals) that* ***interest*** *an investigator* (p. 80).

La population, c'est nous qui la définissons. L'ensemble $\mathcal{P}$ est aussi appelé la **base de sondage**. On parle d'**échantillonnage en population de taille finie** car $\mathcal{P}$ est dénombrable ($\text{Card}(\mathcal{P}) = N < \infty$). On se focalise sur la probabilité d'inclusion d'un individu dans un échantillon. Un échantillon est un sous-ensemble de $\mathcal{P}$. En notant cet échantillon $\mathcal{S}$, on peut écrire $\mathcal{S} \subseteq \mathcal{P}$. Toutes les unités d'observation peuvent être dans l'échantillon (on pourrait essayer de recenser la population française, par exemple). Autrement dit, on peut avoir $\mathcal{S} = \mathcal{P}$ (**mode de tirage trivial**), mais si Y_i est observé avec erreur (on observe $Y_i' = Y_i + \varepsilon_i$), on retombe dans l'approche modèle.

L'approche traditionnelle peut paraître plus simple car on ne présupose absolument rien sur les Y_i (Ardilly, 2006, p. 555). On y a recours plutôt dans les situations où l'on veut éviter un recensement trop long et coûteux.

Unité d'observation, unité d'échantillonnage

On distingue **unité d'observation** (*unit of interest*) et **unité d'échantillonnage** (*sample unit*). Dans l'approche traditionnelle, la première appartient à $\mathcal{P}$, la seconde à $\mathcal{S}$. Dans l'approche modèle, comme il n'y a pas forcément de $\mathcal{P}$ (nous avions dit que nous ne disposions que d'un échantillon aléatoire), on a seulement des *sample units* !

Quelle que soit l'approche, l'unité d'observation i est le plus souvent discrète en économie ; chaque i possède un identifiant (par exemple, le numéro SIREN si l'unité est une entreprise). Néanmoins, l'approche modèle a une certaine préférence en économétrie, notamment quand $\mathcal{P}$ n'est pas dénombrable. En effet, l'unité d'observation peut être un indice continu comme, une unité de temps. Par exemple, supposons que la variable qui nous intéresse soit le cours quotidien du CAC40 à 9h00 du matin (la valeur de quotation d'ouverture). Quelle est la population dans ce cas ? 9h00 tous les jours de l'année ?

Si c'est la saisonnalité *intra-day* qui m'intéresse, dois-je faire un relevé par heure, …, par seconde ou toutes les millisecondes ? La taille de l'échantillon serait d'environ 7×10^9 observations (245 jours, fois 8 heures de quotation/jour, fois 60 minutes/heure, fois 60 secondes/minute, fois 1000 millisecondes/seconde). Est-ce une population de taille suffisante ? Quel est l'intérêt de définir une population faite d'une quantité dénombrable d'unités d'observations ? Autant prendre t entre 0 et 1, et travailler sur la variation de cours de bourse, modélisée par des acroissements $dY(t) = Y(t + \delta t) - Y(t)$ indépendant $N(0, \sqrt{\delta t}$.

D'autres "populations" sont théoriquement infinies (coordonnées GPS, mots prononcés sur des réseaux sociaux), ou, au contraire, très petites (l'effectif de la classe). On s'en sort mieux avec l'approche modèle (Tassi, 2004, pp. 58-59). On se focalise sur l'échantillon aléatoire, et on

essaie d'identifier les paramètres de la superpopulation, qu'on appelle généralement un modèle.

3.2 L'approche fréquentiste

Il y a deux grandes approches de l'inférence en statistique : fréquentiste et bayésienne. Ce cours adopte la première, qui est aussi appelée classique. On rappelle au passage que l'**inférence statistique** d'un modèle c'est l'**estimation** des paramètres de ce modèle et des **tests** d'hypothèses sur ces paramètres.

Dans l'approche traditionnelle, dans laquelle les Y_i ne sont pas considérés aléatoires, la distribution d'une statistique dépend du mode de tirage, tirage sans remise (TSR) ou avec remise (TAR), donc de n et N. En revanche, dans le modèle d'échantillonnage, où le mode de tirage est un TAR, elle dépend de n et de la loi suivie par les Y_i (par ex., $Y_i \sim$ n.i.d.).

Approche modèle-fréquentiste

On considère un modèle d'échantillonnage : on a une distribution de probabilité F (partiellement connue) pour une variable aléatoire réelle Y. Supposons que le paramètre d'intérêt soit le premier moment (moment non-centré d'ordre 1) de Y, i.e. son espérance mathématique : $E_F(Y)$. Rappelons que c'est la valeur que nous obtiendrions en moyenne si nous tirions (avec remise) une nouvelle observation ("en moyenne" sous-entend que nous faisons un grand nombre de tirages). Cette valeur est inconnue dans la mesure où F n'est elle-même que partiellement connue. On a n observations de Y, que l'on note par le vecteur $\mathbf{y} \equiv (y_1, \dots, y_n)$. Les y_i sont des réalisations de variables aléatoires i.i.d. $Y_i \sim Y$. On note $\mathbf{Y} \equiv (Y_1, \dots, Y_n)$, et une **statistique** $t(\mathbf{Y})$.

C'est ici qu'intervient le concept de "fréquence" : comme pour les observations individuelles Y_i, nous avons $\mathbf{Y}^{(k)}$, et donc $t(\mathbf{Y}^{(k)})$, d'où la méthode consistant à prendre l'espérance $E_F(t(\mathbf{Y}))$. Or, $E_F(t(\mathbf{Y}))$ ne vaut pas forcément $E_F(Y)$, d'où un biais possible ! Dit autrement, si je pouvais prendre la moyenne de tous les $t(\mathbf{y})$, je n'aurais pas forcément $E_F(Y)$. Non seulement $t(\mathbf{y})$ peut ne pas être égal à $E_F(Y)$, mais $E_F(t(\mathbf{Y}))$ non plus. On calcule le **biais** comme suit :

$$E_F(t(\mathbf{Y})) - E_F(Y).$$

Une estimation naturelle de $E_F(Y)$ est obtenue en calculant la moyenne $t(\mathbf{y}) \equiv (y_1 + \cdots + y_n)/n$, notée $\overline{y}$. Dans ce cas, $t(\mathbf{Y})$ est sans biais. En effet, supposons le *modèle* $E(Y_i) \equiv \mu$. La statistique $t(\mathbf{Y}) \equiv n^{-1}\sum Y_i$, aussi notée $\hat{Y}$, est sans biais. C'est un résultat archiconnu :

$$E(\hat{Y}) = E(\sum Y_i/n) = \sum E(Y_i/n) = (1/n)\sum E(Y_i) = (1/n)\sum \mu = (1/n)n\mu = \mu.$$

Le principe de l'approche fréquentiste ici est de substituer $t(\mathbf{y})$ que l'on n'observe qu'une fois à $t(\mathbf{Y})$ que l'on n'observe pas. Considérons une mesure de précision de $\overline{y}$. Prenons par exemple l'erreur standard de $\hat{Y}$, qui vaut $(V_F(Y)/n)^{1/2}$. Le problème est qu'on ne connaît pas $V_F(Y)$. En revanche, on connaît une estimation sans biais, $\sum(y_i - \overline{y})^2/(n-1)$. C'est ce que Efron et Hastie (2015) appellent le **plug-in principle**. La correction $n/(n-1)$ n'est pas le sujet ici. Le sujet est que l'on remplace $V_F(Y)$ par une estimation fonction de $\mathbf{y}$ et $\overline{y}$, parce que l'on a remplacé $E_F(Y)$ par $\sum y_i/n$. La précision de $\overline{y}$ est la précision probabiliste de l'estimateur $t(\mathbf{Y})$.

Approche traditionnelle-fréquentiste

Revenons à l'approche traditionnelle. Le moment d'ordre 1 dans ce cas est la moyenne dans la population, i.e. la fonction (Ardilly et Tillé, 2003)

$$f(Y_1, \dots, Y_N) = \frac{\sum Y_i}{N} \equiv \overline{Y}.$$

Si Card($\mathcal{S}$) $\equiv n < N$, on a affaire à un **sondage** (le cas $n = N$ est le **recensement**). Le problème est de trouver le mode de tirage garantissant que, en moyenne, $\hat{Y}$ soit proche de $\overline{Y}$ ci-dessus. L'échantillonnage peut se faire selon un ou plusieurs critères (Ardilly, 2006, p. 94) afin de représenter la population de manière équilibrée (**sondages aléatoires simple**, **stratifié**, etc.).

$\hat{Y}$ est aussi une statistique (une variable aléatoire) dans cette approche. Supposons qu'après avoir choisi un mode de tirage, l'on puisse tirer autant de n-échantillons que l'on veut. On pourrait tracer une distribution des valeurs de $\hat{Y}$ (sa **distribution d'échantillonnage**). Cette distribution est dispersée autour de la valeur centrale $\overline{Y}$ (μ dans l'autre approche), la valeur que nous obtiendrions si nous prélevions un nombre infini d'échantillons. Plus les moyennes d'échantillons sont "proches" de ces valeurs, plus la précision de $\hat{Y}$ est grande. À chaque échantillon correspond un $\hat{Y}$, qui dépend de n (fini) et du mode de tirage.

Notons que "fini" ne veut pas dire petit. Pour une population de 1000000, un échantillon fini de taille 10000, c'est relativement petit ($1/100^e$), mais suffisamment grand dans l'absolu pour obtenir des statistiques précises. L'enquête Budget des Familles est un exemple. En revanche, pour une population de taille 1000, un échantillon fini de taille 100 pèse 10 fois plus que précédemment ($1/10^e$), mais 100 est trop "petit" pour obtenir des statistiques précises (cf. 3.3). Le mode de tirage est plus crucial dans le second cas.

[Donner l'exemple des élections américaines de 1936]

3.3 Précision de la moyenne dans le modèle d'échantillonnage

Comment mesure-t-on généralement la précision de la moyenne – dans la théorie fréquentiste ? Par l'**erreur type** $\sqrt{V(\hat{Y})}$. Calculons d'abord $V(\hat{Y})$:

$$V(\hat{Y}) = V(\sum Y_i/n) = \sum V(Y_i/n) = (1/n)^2 \sum V(Y_i) = (1/n)^2 \sum \sigma^2 = (1/n)^2 n\sigma^2 = \sigma^2/n.$$

L'erreur-type dépend donc de l'écart-type σ de la variable Y, en général inconnu, et de n : $\sqrt{\sigma^2/n} = \sigma/\sqrt{n}$. L'estimateur de σ^2 généralement employé est

$$\frac{1}{n-1} \sum_{i \leq n} (Y_i - \hat{Y})^2 \equiv \hat{\sigma}_c^2.$$

La précision est donc

$$\frac{\hat{\sigma}_c}{\sqrt{n}} = \left(\frac{1}{n(n-1)} \sum_{i \leq n} (Y_i - \hat{Y})^2 \right)^{1/2}.$$

Il suffit ensuite de remplacer les Y_i par les valeurs observées de notre unique échantillon, et $\hat{Y}$ par la moyenne de l'échantillon (***plug-in***).

Voyons un exemple avec $n = 3$ pour faire simple : $y_1 = 6$, $y_2 = 2$ et $y_3 = 8$.

```
. cls

. clear   all


. set     obs     3
Number of observations (_N) was 0, now 3.

. input   int var1

        var1
  1. 6
. 2
. 8

. sum     var1

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
        var1 |          3    5.333333     3.05505          2          8

. di              r(sd)/sqrt(r(N))
1.7638342

. ci              means var1                     // Stata >=17

    Variable |        Obs        Mean    Std. err.       [95% conf. interval]
-------------+------------------------------------------------------------
        var1 |          3    5.333333    1.763834       -2.255833     12.9225
```

La commande **`summarize var1`** renvoie "l'écart-type corrigé" de la variable **`var1`**. Il suffit de diviser par $\sqrt{3}$ pour avoir une estimation de l'erreur-type. La commande **`ci var1`** fournit une estimation de l'erreur-type directement, et un **intervalle de confiance non-asymptotique** à 95%. Afin de construire l'intervalle, nous avons besoin non pas du fractile d'une N(0,1), mais d'une Student du fait de la taille de l'échantillon. Il y a 97,5% de chances qu'une variable aléatoire distribuée selon un loi de Student à 2 degrés de liberté prenne une valeur plus petit que 4,302 env. Nous pouvons obtenir cette valeur avec **`di invttail(2,0.025)`**.

```
. display invttail(2,0.025)
4.3026527
```

La borne inférieure de l'intervalle est bien environ égale à $5{,}333 - 4{,}302 \times 1{,}763$.

Exercice. 1. Démontrer que $\hat{\sigma}_c^2/n$ est un estimateur sans biais de σ^2/n. 2. Peut-on en déduire que $\hat{\sigma}_c/\sqrt{n}$ est un estimateur sans biais de $\sigma/\sqrt{n}$? 3. Quelle méthode pourrions-nous utiliser pour calculer $E(\hat{\sigma}_c/\sqrt{n})$?

Correction.

- Pour la question 1, c'est la même chose que montrer que $\hat{\sigma}_c^2$ estime σ^2 sans biais.
- La question 2 attire l'attention sur le point sivant : montrer que $\hat{\sigma}_c^2/n$ est un estimateur sans biais de σ^2/n n'implique pas que $\hat{\sigma}_c/\sqrt{n}$ est un estimateur sans biais de l'erreur-type $\sigma/\sqrt{n}$. En effet, de $E(\hat{\sigma}_c^2/n) = \sigma^2/n$ on peut déduire $\sqrt{E(\hat{\sigma}_c^2/n)} = \sigma/\sqrt{n}$, mais ce n'est pas la question 2 ; la question 1 porte sur $E(\hat{\sigma}_c^2/n)$. Tandis que $E(\hat{\sigma}_c/\sqrt{n}) = E(\sqrt{\hat{\sigma}_c^2/n})$, est l'espérance de la racine de la statistique sans biais. Or, d'après l'**inégalité de Jensen**, si

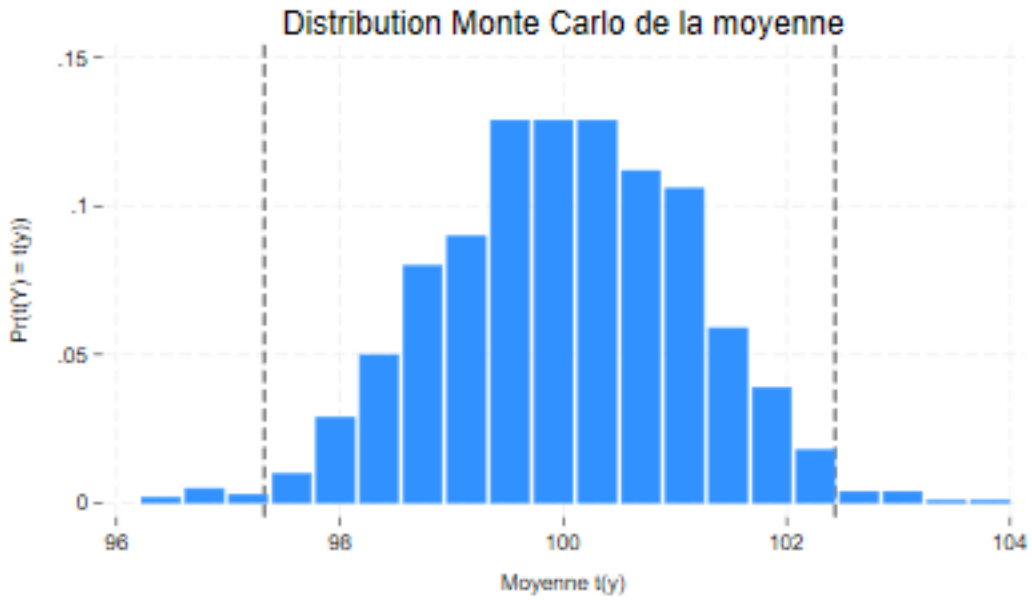
$E(|\hat{\sigma}_c|) < \infty$ et $\hat{\sigma}_c > 0$ (Wooldridge, 2010, p. 32), alors puisque $\sqrt{\cdot}$ est concave, on a $\sqrt{E(\hat{\sigma}_c^2/n)} \geq E(\sqrt{\hat{\sigma}_c^2/n})$. Par conséquent, l'estimateur proposé est plus grand, en moyenne, que $\hat{\sigma}_c/\sqrt{n}$.

- Pour la question 3, on pourrait regarder la distribution de $\hat{\sigma}_c/\sqrt{n}$.

On ne va pas simuler cette variable, mais la distribution de $\hat{Y}$. Une simulation Monte Carlo va nous permettre d'avoir une idée plus claire de ce que l'on entend par "précision". Le bootstrap, que nous verrons en fin de section, nous donnera un point de vue fréquentiste. La loi faible des grands nombres (LFGN, par la suite) garantit que – sous certaines conditions – la précision de la moyenne augmente assez vite quand n augmente. Combien faut-il d'observations pour que la précision double ? On peut répondre à cette question en utilisant la méthode Monte Carlo vue dans la **Section 2**.

```
. cls

. clear all


. set seed               21041971

. set more off

. set     obs                    10000
Number of observations (_N) was 0, now 10,000.

. local                  R=1000

. generate               Y=rnormal(100,10)       // mu = 100, sigma = 10

. generate               RANK=0

. matrix                 V1=J(`R',1,0)

. forvalues              I=1(1)1000 {
.  qui: replace   RANK=runiform()
.  sort                  RANK Y
.  quietly sum    Y in 1/80                     // Taille de l'echantillon
.  matrix define  V1[`I',1]=r(mean)       // La statistique dont on cherche la distr.
. }

. drop                   Y RANK

. qui: svmat             V1, names(MEAN_)

. qui: summarize  MEAN_1, d

. display                "r(p1) = " r(p1) " , r(p99) = " r(p99)
r(p1) = 97.32061 , r(p99) = 102.42999

. twoway hist            MEAN, bin(20) gap(10) fraction ///
                               title("Distribution Monte Carlo de la moyenne") ///
                               xline(`r(p1)' `r(p99)') xscale(noline titlegap(3)) ///
                               yscale(noline titlegap(3)) ///
                               xtitle("Moyenne t(y)") ytitle("Pr(t(Y) = t(y))")



. graph export "statainitiation_3_distributionofthemean.png", width(400) replace
file statainitiation_3_distributionofthemean.png saved as PNG format
```

Graphique 3.1. Distribution Monte Carlo de la moyenne.



En théorie, plus l'échantillon est grand, plus la dispersion Monte Carlo autour de la moyenne, $\sigma/\sqrt{n}$ tend vers zéro. Les deux barres verticales délimitent un intervalle de confiance à 98%. L'intervalle de fluctuation asymptotique est $100 \pm 2{,}326 \times 10/\sqrt{80}$, qui vaut [97,39 ; 102,60], où 2,3263… est obtenu par `invnormal(.99)`. On peut voir que si on quadruple la taille de l'échantillon (remplacer `quietly summarize Y in 1/80` par `quietly summarize Y in 1/320`, la précision double !, c'est-à-dire, la longueur de l'intervalle de confiance qui contient 98% des moyennes est divisée par 2 ; il passe d'environ [97,32 ; 102,43] à [98,69 ; 101,36]. En effet, $102{,}43 - 97{,}32$ est environ égal à 5,11 et $101{,}36 - 98{,}69$ est environ égal à 2,67. Or, $5{,}11/2{,}67 \approx 2$, la précision double !).

Ce résultat aurait pu être obtenu analytiquement, puisqu'en théorie, si n et $n' > n$ sont deux tailles d'échantillons, alors le ratio des erreurs types vaut :

$$\frac{\frac{\sigma}{\sqrt{n'}}}{\frac{\sigma}{\sqrt{n}}} = \sqrt{\frac{n}{n'}},$$

qui vaut 1/2 si $n' = 4n$.

Bibliographie

Ardilly, P., 2006. Les techniques de sondage, Technip.

Ardilly, P., Tillé, Y., 2003. Exercices corrigés de méthodes de sondage. Ellipses.

Imbens, G.W., Rubin, D.B. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences. Cambridge University Press, Cambridge, USA, 625 pp.

Newbold, P., Carlson, W.L., Betty, T., 2007. Statistics for Business and Economics. Pearson Prentice Hall, New Jersey, 984 p.

Tassi, P., 2004. Méthodes Statistiques, Economica, 3e édition, pp. 58-59.

```
. quietly log close
```