REVUE DE STATISTIQUE APPLIQUÉE

ANNE-MARIE DUSSAIX

Détermination de la taille d'échantillon pour la mesure d'évolutions

Revue de statistique appliquée, tome 35, n° 4 (1987), p. 25-35

<http://www.numdam.org/item?id=RSA_1987__35_4_25_0>

Rev. Statistique Appliquée, 1987, **XXXV** (4), 25-36

DÉTERMINATION DE LA TAILLE D'ÉCHANTILLON POUR LA MESURE D'ÉVOLUTIONS

Anne-Marie DUSSAIX

Professeur à l'E.S.S.E.C.
Chargée de cours à l'I.S.U.P.

Dans la conception des enquêtes répétées dans le temps, une étape essentielle est la détermination de la taille d'échantillon nécessaire à chaque vague d'enquête. Lorsque l'objectif des enquêtes répétées est essentiellement la mesure d'évolution, la détermination de la taille d'échantillon se fait généralement par une méthode classique utilisée dans la conception des expérimentations.

Dans tout ce qui suit, on supposera que les échantillons sont des échantillons aléatoires simples.

1. Détermination des tailles d'échantillon nécessaires pour mettre en évidence une évolution significative — Approche classique

On trouvera une présentation détaillée de ces résultats dans COCHRAN et COX (1957) dans le contexte des expérimentations et dans DEROO et DUSSAIX (1980) dans le contexte des enquêtes par sondage.

Rappelons-en les principaux résultats en ce qui concerne la comparaison de moyennes et la comparaison de proportions.

1.1. Comparaison de moyennes

Soit $\mathcal{Y}$ la variable étudiée, m_1 et m_2 les moyennes de $\mathcal{Y}$ aux temps t_1 et t_2 dans une population que l'on supposera invariante entre t_1 et t_2 et de taille suffisamment importante pour que l'on puisse négliger les facteurs d'exhaustivité. σ_1^2 et σ_2^2 sont les variances de la variable étudiée dans la population. On les suppose égales pour simplifier; ceci constitue souvent dans la pratique une hypothèse acceptable.

a) Cas d'échantillons indépendants

Lorsque les variances sont supposées égales ainsi que les coûts d'enquête en t_1 et t_2, la meilleure stratégie est de tirer en t_1 et en t_2 des échantillons de tailles n_1 et n_2 égales entr'elles : $n_1 = n_2 = n$.

La taille d'échantillon commune n est alors fonction :

— des risques de première et deuxième espèce α et β que l'on accepte de prendre. Dans le cas d'un test bilatéral par exemple :

$$H_0 : m_1 = m_2$$

$$H_1 : m_1 \neq m_2$$

Mots clés : *Enquêtes répétées dans le temps, Panels, Sondages, Taille d'échantillon.*

α est le risque de conclure à tort à une évolution, quel que soit son sens, alors qu'elle n'a pas eu lieu.
β est le risque de conclure à tort à pas d'évolution, alors qu'une évolution a eu lieu.

— de la variance commune $\sigma^2 = \sigma_1^2 = \sigma_2^2$

— de la plus petite différence $\Delta = |m_1 - m_2|$ que l'on souhaite mettre en évidence, si elle se produit, en limitant les risques α et β aux valeurs définies plus haut.

La taille d'échantillon commune n est en effet donnée :

— dans le cas d'un test *unilatéral* :

(on teste $H_0 : m_1 = m_2$ contre $H_1' : m_2 > m_1$ ou $H_1'' : m_2 < m_1$)

par : $$n = (t_{1-\alpha} + t_{1-\beta})^2 \times 2\sigma^2/\Delta^2 \qquad (1)$$

où $t_{1-\alpha}$ (resp. $t_{1-\beta}$) est le fractile de la loi normale centrée réduite correspondant à une probabilité $(1 - \alpha)$ (resp. $(1 - \beta)$).

— dans le cas d'un test *bilatéral*

(on teste $H_0 : m_1 = m_2$ contre $H_1 : m_1 \neq m_2$)

par : $$n = (t_{1-\alpha/2} + t_{1-\beta})^2 \times 2\sigma^2/\Delta^2 \qquad (2)$$

où $t_{1-\alpha/2}$ est le fractile de la loi normale centrée réduite correspondant à une probabilité $(1-\alpha/2)$.

b) Cas du panel (échantillons identiques en t_1 et en t_2)

Test unilatéral ($H_0 : m_1 = m_2$ contre $H_1 : m_2 > m_1$)

Si ρ désigne le coefficient de corrélation linéaire dans la population entre la variable $\mathcal{Y}$ au temps t_1 et la variable $\mathcal{Y}$ au temps t_2, la taille d'échantillon nécessaire pour mettre en évidence une évolution $\Delta = m_2 - m_1$ si elle se produit, en limitant à α et à β les risques de première et deuxième espèce, est :

$$n' = n(1 - \rho) \text{ où n est donné par (1)} \qquad (3)$$

Test bilatéral

De la même façon, pour mettre en évidence une différence $\Delta = |m_2 - m_1|$, la taille d'échantillon nécessaire est :

$$n' = n(1 - \rho) \text{ où n est donné par (2)} \qquad (4)$$

c) Cas d'échantillons renouvelés partiellement

n individus sont interrogés en t_1. Sur ces n individus, n_1 sont réinterrogés en t_1 (la proportion $k = n_1/n$ est fixée à l'avance) et $n_2 = n - n_1$ individus sont remplacés en t_2 par n_2 nouveaux interviewés.

Dans ce cas, la taille d'échantillon nécessaire est

$$n' = n(1 - k\rho) \qquad (5)$$

où n est donnée par (1) ou (2) selon qu'il s'agit d'un test unilatéral ou bilatéral.

Remarque

Dans ce qui précède, on a supposé que le taux de renouvellement était fixé a priori — Un des critères pour le déterminer est la recherche du taux de renouvellement minimisant la variance de l'estimateur de l'évolution, si l'objectif est d'estimer une évolution (cf. GOURIEROUX et ROY (1978) pour la détermination d'un taux de renouvellement optimal).

1.2. Comparaison de proportions

Dans le cas de comparaison de proportions P_1 et P_2, on obtient les tailles d'échantillon suivantes en fonction des risques α et β choisis (échantillons de taille égale n et indépendants).

Cas du test *unilatéral*

$$n = (t_{1-\alpha} + t_{1-\beta})^2/2\,(\arcsin\sqrt{p_2} - \arcsin\sqrt{p_1})^2 \qquad (6)$$

Cas du test *bilatéral*

$$n = (t_{1-\alpha/2} + t_{1-\beta})^2/2\,(\arcsin\sqrt{p_2} - \arcsin\sqrt{p_1})^2 \qquad (7)$$

Ces formules (6) et (7) nécessitent, contrairement aux formules (1) et (2) de faire des hypothèses non seulement sur la différence minimale $\Delta = |p_2 - p_1|$ que l'on souhaite mettre en évidence, mais aussi sur la valeur de départ « inconnue » p_1.

Fleiss (1981) propose des résultats un peu différents prenant en compte en particulier une correction de continuité.

Dans le cas d'un panel ou d'échantillons renouvelés partiellement, on déduit de (6) et (7) les tailles d'échantillon nécessaires de la même façon qu'au paragraphe 1.1.

L'approche rappelée dans les § 1.1 et 1.2 se révèle très utile dans la conception des enquêtes par sondage répétées dans le temps. Elle permet en particulier de contribuer à la détermination des objectifs de l'enquête par la détermination des paramètres α, β et surtout Δ.

Elle est tout à fait adaptée au cas d'une enquête répétée seulement deux fois (enquête avant et après campagne par exemple).

Cependant, dans le cas d'une enquête répétée en $t = 1, \ldots T\,(T > 2)$, on peut accepter d'avoir des résultats dont l'évolution soit non significative d'une période à la suivante mais significative au bout d'un certain nombre de périodes.

Une solution est alors d'adopter la méthode du paragraphe § 1 pour la comparaison des moyennes observées au temps 1 et au temps T. Il est clair que, dans cette démarche, on néglige les T-2 informations intermédiaires qui devraient permettre pour un risque α donné, de diminuer la taille d'échantillon à interroger à chaque période.

L'idée de la méthode proposée est de déterminer une taille d'échantillon nécessaire pour déterminer une évolution significative au bout de T périodes d'enquête, en tenant compte de toutes les informations intermédiaires. Cette méthode fera l'objet du § 3.

Une méthode répondant à un objectif un peu différent a été proposée par LAYCOCK et AL-KASSAB (1982). Elle fera l'objet du § 2.

2. Détection d'un changement de tendance dans l'analyse d'enquêtes répétées

Chaque enquête est constituée d'un échantillon aléatoire simple de taille n; la variable à laquelle on s'intéresse est $\mathcal{Y}$. LAYCOCK et AL-KASSAB (1982) introduisent le modèle à effet aléatoire suivant :

$$y_{ti} = \mu_t + e_{ti} \quad (t = 1, \ldots, T ; \quad i = 1, \ldots n)$$

où t indice les enquêtes, i indice l'individu; μ_t et e_{ti} sont des variables aléatoires telles que :

$$\mu_t \sim \mathcal{N}(\mu, \sigma_\mu^2) ; \quad e_{ti} \sim \mathcal{N}(0, \sigma^2) \; (t = 1, \ldots, T ; \quad i = 1, \ldots, n) \qquad (8)$$

Les μ_t et e_{ti} sont mutuellement indépendantes

d'où $\overline{y_t}. = \mu_t + \overline{e_t}.$, où $\overline{y_t}. = \sum_{i=1}^{n} y_{ti}/n$, $\overline{e_t}. = \sum_{i=1}^{n} e_{ti}/n$

et $\overline{e_t}. \sim \mathcal{N}(0, \sigma^2/n)$.

Définissant alors les événements

$$A = (\overline{y_{t+1}}, . > \overline{y_t}, .) \quad \text{et} \quad B = (\mu_{t+1} > \mu_t)$$

ils montrent que la probabilité p de détection correcte d'un changement de tendance est donnée par :

$$p = P(A \cap B) + P(\overline{A} \cap \overline{B}) \qquad (9)$$

ce qui, tous calculs faits, donne

$$p = \frac{1}{\Pi} \arctan \left\{ \left(\frac{\sigma^2}{n\sigma_\mu^2} \right)^{1/2} \right\} \qquad (10)$$

d'où

$$n = \frac{\sigma^2}{\sigma_\mu^2} \cot^2 \{\Pi(1 - p)\} \qquad (11)$$

Pour p proche de 1 :

$$n \simeq \frac{\sigma^2}{\Pi^2 \sigma_\mu^2 (1 - p)^2}$$

Cette approche admet, à notre avis, deux critiques :

1) Le modèle (8) est mal adapté à la conception des enquêtes répétées où l'objectif est généralement de mettre en évidence des évolutions et non pas des oscillations autour d'un niveau moyen μ.

2) La formule (11) nécessite une estimation de σ^2 et de σ_μ^2 et pratiquement, un certain nombre d'enquêtes déjà faites. L'approche nous semble plus utile pour la détermination de la probabilité p de détection correcte d'un changement de tendance donnée par (9) étant donné le modèle (8).

LAYCOK et KASSAB (1982) développent ensuite leur approche dans le cas où les μ_t suivent un modèle autorégressif du premier ordre, c'est-à-dire :

$\mu_t = \lambda\mu_{t-1} + \varepsilon_t$ où $\varepsilon_t \sim \mathcal{N}$ $(\mu_\varepsilon, \sigma_\varepsilon^2)$, ε_t indépendant de e_t et $|\lambda| < 1$. (12)

Ce modèle a été suggéré pour la première fois dans ce contexte par BLIGHT et SCOTT (1973).

La détermination de la taille d'échantillon nécessaire dépend alors du paramètre λ.

3. Détermination de la taille d'échantillon dans le cas d'une tendance linéaire

3.1. Comparaison de moyennes

3.1.1. Le modèle

Soit $\overline{y}_t$ l'estimation au temps t de la moyenne m_t de la variable dans la population ($t = 1, \ldots, T$). Cette estimation est faite à partir d'un échantillon aléatoire. L'échantillon est renouvelé de période en période.

On a donc, en confondant dans les notations estimation et estimateur :

$$\overline{y}_t = m_t + e_t \quad (t = 1, \ldots, T) \tag{13}$$

où e_t désigne l'erreur d'échantillonnage.

Dans la théorie des sondages classiques, les m_t sont des valeurs inconnues, mais fixes, de telle sorte que la connaissance des estimations précédentes $\overline{y}_1, \ldots, \overline{y}_{t-1}$ n'apporte aucune information sur l'estimation faite au temps t si les échantillons sont indépendants.

En fait, il peut être souvent plus réaliste de modéliser le comportement des m_t au cours du temps comme le font LAYCOCK et AL-KASSAB (1982). Un modèle simple est de considérer que les m_t sont les réalisations de variables aléatoires μ_t suivant le modèle :

$$\begin{aligned} &\mu_t = at + b + u_t \qquad (t = 1, \ldots, T) \\ &\mathscr{E}(u_t) = 0 \qquad \mathscr{V}(u_t) = \sigma_u^2 \\ &\mathscr{C}ov(u_t, u_{t'}) = 0 \qquad \forall t \neq t' \end{aligned} \tag{14}$$

Les opérateurs $\mathscr{E}$, $\mathscr{V}$ et $\mathscr{C}ov$ désignent respectivement les opérateurs espérance, variance et covariance associés à la distribution de probabilité jointe ξ des variables aléatoires $\mu_1, \ldots, \mu_T$.

Le modèle (13) se réécrit donc :

$$\overline{y}_t = \mu_t + e_t = at + b + u_t + e_t \qquad (t = 1, \ldots, T)$$

où e_t et μ_t (et donc e_t et u_t) sont indépendantes, d'où, en posant $\eta_t = u_t + e_t$

$$\overline{y}_t = at + b + \eta_t \qquad (t = 1, \ldots, T) \tag{15}$$

On supposera que l'espérance de e_t par rapport au plan de sondage est nulle :

$$E_p(e_t \mid m_t) = 0 \qquad \text{i.e. } E_p(\overline{y}_t \mid m_t) = m_t$$

et que

$$V_p(e_t \mid m_t) = V_p(e_t) = \frac{\sigma^2}{n}$$

où σ^2 est la variance de la variable étudiée dans la population et où E_p et V_p désignent respectivement l'espérance et la variance par rapport au plan de sondage supposé aléatoire

d'où
$$\text{Var}(\eta_t) = \mathscr{V} E_p(e_t + u_t/m_t) + \mathscr{E} V_p(e_t + u_t/m_t)$$
$$\text{Var}(\eta_t) = \mathscr{V}(u_t) + \mathscr{E}\left(\frac{\sigma^2}{n}\right)$$
$$\text{Var}(\eta_t) = \sigma_u^2 + \frac{\sigma^2}{n}$$

On pose
$$\sigma_\eta^2 = \sigma_u^2 + \frac{\sigma^2}{n}$$

3.1.2. Le test

En supposant $e_t \sim \mathscr{N}\left(0, \frac{\sigma^2}{n}\right)$, $u_t \sim \mathscr{N}(0, \sigma_u^2)$, on en déduit que

$$\eta_t \sim \mathscr{N}\left(0, \sigma_u^2 + \frac{\sigma^2}{n}\right)$$

a) Test unilatéral

Le test de significativité de l'évolution est équivalent au test

$$\begin{cases} H_0 : a = 0 \\ H_1 : a > 0 \end{cases}$$

Sous l'hypothèse H_0 et sous l'hypothèse de normalité des résidus, on a, si â est l'estimateur des moindres carrés de a (calculé à partir de (15)) et si

$$\bar{t} = \sum_{t=1}^{T} t/T :$$

$$\frac{\hat{a}}{\sqrt{\text{Var}(\hat{a})}} = \frac{\hat{a}}{\sigma_\eta}\sqrt{\sum_{t=1}^{T}(t-\bar{t})^2} \sim \mathscr{N}(0,1)$$

La zone de rejet de H_0 pour un risque de première espèce α fixé est donnée par :

$$\frac{\hat{a}}{\sigma_\eta}\sqrt{\Sigma(t-\bar{t})^2} > t_{1-\alpha}$$

où $t_{1-\alpha}$ est le fractile de la loi normale centrée réduite correspondant à la probabilité $(1 - \alpha)$.

Soit Δ la plus petite évolution des moyennes que l'on souhaite mettre en évidence au bout de T périodes.

Autrement dit, on souhaite mettre en évidence une pente minimale

$$a_0 = \frac{\Delta}{(T-1)}$$

avec un risque de ne pas la déceler par l'enquête au bout de T périodes si elle se produit, inférieur à β

d'où la condition

$$t_{1-\alpha} + t_{1-\beta} < \frac{\Delta/(T-1)}{\sigma_\eta/\sqrt{\Sigma (t-\bar{t})^2}}$$

soit :

$$\frac{\sigma^2}{n} < \frac{\Delta^2}{(T-1)^2} \times \frac{\sum_{t=1}^{T} (t-\bar{t})^2}{(t_{1-\alpha} + t_{1-\beta})^2} - \sigma_u^2 \qquad (16)$$

Si le deuxième membre de l'inégalité (16) est négatif, ce qui se produit

— pour une variance σ_u^2 par rapport au modèle trop forte

et/ou

— pour une évolution à mettre en évidence trop faible et/ou sur un nombre de périodes trop faible

on ne peut pas, quelle que soit la taille de l'échantillon, assurer un risque de deuxième espèce inférieur à β.

Par contre, si :

$$\frac{\Delta^2}{(T-1)^2} \times \frac{\Sigma (t-\bar{t})^2}{(t_{1-\alpha} + t_{1-\beta})^2} > \sigma_u^2$$

on obtient :

$$n > \frac{\sigma^2}{\dfrac{\Delta^2}{(T-1)^2} \times \dfrac{\Sigma (t-\bar{t})^2}{(t_{1-\alpha} + t_{1-\beta})^2} - \sigma_u^2}$$

ou

$$n > \frac{\sigma^2}{\dfrac{\Delta^2}{(T-1)} \dfrac{T(T+1)}{12\,(t_{1-\alpha} + t_{1-\beta})^2} - \sigma_u^2}$$

Remarque

Si $\sigma_u = 0$ et $T = 2$, on retrouve la formule (1).

Pratiquement, pour $\sigma_u = 0$, pour une évolution minimale Δ donnée à mettre en évidence avec des risques α et β donnés, il faut environ 10 périodes d'observation pour que la taille d'échantillon nécessaire par cette approche soit deux fois plus faible que celle obtenue dans l'approche du §1 (dans ce cas, on déterminerait n pour mettre en évidence une augmentation Δ en ne comparant que les résultats en période 1 et en période 10).

b) Test bilatéral

On teste

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$$

Dans ce cas, on obtient :

$$n > \frac{\sigma^2}{\frac{\Delta^2}{(T-1)^2} \times \frac{\Sigma(t-\bar{t})^2}{(t_{1-\alpha/2} + t_{1-\beta})^2} - \sigma_u^2}$$

3.2. Comparaison de proportions

Dans ce cas, le modèle est plus complexe car les variables aléatoires e_t et μ_t ne sont plus indépendantes. On se limitera dans ce document au modèle linéaire. D'autres modèles comme le modèle logistique pourraient être envisagés.

On change les notations par rapport au §3.1 pour reprendre les notations concernant les proportions.

a) Le modèle

Considérons que la proportion p_t dans la population au temps t est la réalisation d'une variable aléatoire P_t. Le modèle linéaire s'écrit :

$$\begin{aligned} &P_t = at + b + u_t \qquad (t = 1, \ldots, T) \\ &\mathscr{E}(u_t) = 0 \qquad \mathscr{V}(u_t) = \sigma_u^2 \\ &\mathscr{C}ov(u_t, u_{t'}) = 0 \qquad \forall t \neq t' \end{aligned} \tag{17}$$

et le modèle complet en désignant par f_t l'estimation ou l'estimateur de P_t

$$\begin{aligned} &f_t = P_t + e_t \qquad (t = 1, \ldots, T) \\ &E_p(e_t/p_t) = 0 \qquad V_p(e_t/p_t) = \frac{p_t(1-p_t)}{n} \\ &P_t = at + b + u_t \\ &\mathscr{E}(u_t) = 0 \qquad \mathscr{V}(u_t) = \sigma_u^2 \qquad \mathscr{C}ov(u_t, u_{t'}) = 0 \\ &\text{Les } e_t \text{ et les } u_t \text{ sont des v.a. indépendantes.} \end{aligned} \tag{18}$$

Soit, en posant

$$e_t + u_t = \eta_t \qquad (t = 1, \ldots, T)$$

$$f_t = at + b + \eta_t \qquad (t = 1, \ldots, T)$$

où

$$\begin{aligned} E(\eta_t) &= \mathscr{E}E_p(e_t/p_t) + \mathscr{E}E_p(u_t) = 0 \\ V(\eta_t) &= V(e_t) + V(u_t) \\ &= \mathscr{E}V_p(e_t/p_t) + \mathscr{V}E_p(e_t/p_t) \\ &\quad + V_p\mathscr{E}(u_t) + E_p\mathscr{V}(u_t) \\ &= \frac{1}{n}\mathscr{E}(at + b + u_t)(1 - at - b - u_t) + \sigma_u^2 \\ &= \frac{1}{n}(at + b)[1 - (at + b)] - \frac{\sigma_u^2}{n} + \sigma_u^2 \end{aligned}$$

Posons

$$V(\eta_t) = K_t^2 \quad \text{avec} \quad K_t^2 = \frac{1}{n}(at + b)[1 - (at + b)] + \sigma_u^2\left(1 - \frac{1}{n}\right)$$

Le modèle (18) se réécrit donc :

$$f_t = at + b + \eta_t \qquad (t = 1, \ldots, T)$$
$$E(\eta_t) = 0$$
$$V(\eta_t) = K_t^2$$

b) Estimation et test du modèle

L'hypothèse d'homoscedasticité n'étant pas satisfaite, la solution correcte est d'employer les moindres carrés généralisés, mais dans la variance K_t^2 interviennent les paramètres inconnus a et b.

Dans l'approche qui suit, on suppose que la pente a est faible de telle sorte que

$$K_1^2 \simeq \cdots \simeq K_T^2 \simeq K^2$$

et on utilise les m.c.o. pour estimer a et b :

$$\hat{a} = \frac{\sum_{t=1}^{T} (f_t - \overline{f})(t - \overline{t})}{\sum_{t=1}^{T} (t - \overline{t})^2}$$

où

$$\overline{f} = \sum_{t=1}^{T} f_t/T \qquad \text{et} \qquad \overline{t} = \sum_{t=1}^{T} t/T$$
$$\hat{b} = \overline{f} - \hat{a}\overline{t}$$
$$\hat{K}^2 = \sum_{t=1}^{T} (f_t - \hat{a}t - \hat{b})^2 / (T - 2)$$

Pour calculer la taille d'échantillon nécessaire, on utilisera l'approximation suivante :

$$(at + b)[1 - (at + b)] \simeq (a\overline{t} + b)[1 - (a\overline{t} + b)]$$

et on posera :

$$K^2 = \sigma_u^2 \left(1 - \frac{1}{n}\right) + \frac{1}{n}(a\overline{t} + b)(1 - a\overline{t} - b)$$

c) Détermination de la taille d'échantillon dans le cas d'un test unilatéral

On teste :

$$\begin{cases} H_0 : a = 0 \\ H_1 : a > 0 \end{cases}$$

Sous l'hypothèse H_0 et sous l'hypothèse de normalité de M_t,

$$\frac{\hat{a}}{\dfrac{K'}{\sqrt{\Sigma(t - \overline{t})^2}}} \sim \mathcal{N}(0, 1)$$

avec

$$K' = \sqrt{\sigma_u^2 \left(1 - \frac{1}{n}\right) + \frac{1}{n} b(1 - b)}$$

La zone de rejet de H_0 pour un risque de première espèce α fixé est donnée par :

$$\frac{\hat{a}}{K'} \sqrt{\Sigma(t - \bar{t})^2} > t_{1-\alpha}$$

où $t_{1-\alpha}$ est le fractile de la loi normale centrée réduite correspondant à la probabilité $(1 - \alpha)$.

Soit Δ la plus petite évolution des proportions que l'on souhaite mettre en évidence au bout de T périodes.

On souhaite donc mettre en évidence une pente minimale $a_0 = \frac{\Delta}{(T - 1)}$ avec un risque de ne pas la déceler par l'enquête au bout de T périodes, si elle se produit, inférieur à β, d'où la condition :

$$t_{1-\alpha} + t_{1-\beta} < \frac{\Delta/(T - 1)}{K'/\sqrt{\Sigma(t - \bar{t})^2}}$$

soit

$$\sigma_u^2 \left(1 - \frac{1}{n}\right) + \frac{1}{n} b(1 - b) < \frac{\Delta^2}{(T - 1)^2} \times \frac{\Sigma(t - \bar{t})^2}{(t_{1-\alpha} + t_{1-\beta})^2}$$

Soit, encore :

$$n > \frac{b(1 - b) - \sigma_u^2}{\frac{\Delta^2 T(T + 1)}{12(T - 1)(t_{1-\alpha} + t_{1-\beta})^2} - \sigma_u^2} \qquad (19)$$

Par la formule (19), on voit que n dépend :

— des valeurs moyennes de la proportion étudiée par l'intermédiaire de $b(1 - b)$ qui, dans l'hypothèse $H_0 : a = 0$, sera estimée par $\bar{f}(1 - \bar{f})$;

— de la pente minimale $\Delta/(T - 1)$ que l'on souhaite mettre en évidence ;

— du nombre de périodes étudié T ;

— de la variance σ_u^2 (traduisant l'écart par rapport à la linéarité).

Le tableau I donne le calcul de différentes tailles d'échantillon dans le cas d'un test unilatéral et avec

- $\sigma_u = 0$ i.e. les proportions suivent exactement le modèle linéaire
- $\alpha = \beta = 10\,\%$

et pour différentes valeurs du nombre de périodes et des proportions moyennes sur les périodes considérées.

Ce tableau indique que, pour une évolution donnée à mettre en évidence, il faut environ dix enquêtes successives pour que la taille nécessaire à chaque vague soit divisée par deux par rapport à la taille nécessaire pour mettre en évidence la même évolution par comparaison de deux enquêtes.

TABLEAU 1

Détermination de tailles d'échantillon pour $\sigma_u = 0$ et pour $\alpha = \beta = 10\,\%$
d'où $(t_{1-\alpha} + t_{1-\beta})^2 = 6{,}57$

	Proportions moyennes sur les périodes considérées de l'ordre de :						
	1,5 %	5 %		15 %		30 %	
	$\Delta = 1\,\%$	$\Delta = 1\,\%$	$\Delta = 2\,\%$	$\Delta = 2\,\%$	$\Delta = 3\,\%$	$\Delta = 3\,\%$	$\Delta = 5\,\%$
T = 2*	n = 1 941 (1 871)	n = 6 241 (6 157)	n = 1 560 (1 539)	n = 4 188 (4 162)	n = 1 861 (1 845)	n = 3 066 (3 056)	n = 1 104 (1 097)
T = 5	n = 1 553	n = 4 993	n = 1 248	n = 3 351	n = 1 489	n = 2 453	n = 883
T = 10	n = 953	n = 3 064	n = 766	n = 2 056	n = 914	n = 1 505	n = 542
T = 20	n = 527	n = 1 694	n = 424	n = 1 137	n = 505	n = 832	n = 300

* Le premier nombre indique la taille d'échantillon calculée par la formule (19), le nombre entre parenthèses la taille donnée par la formule (6).

La critique évidente de cette approche est que la variance par rapport au modèle linéaire, $\mathscr{V}(u_t) = \sigma_u^2$ n'est pas connue lorsque l'on essaie de déterminer une taille d'échantillon pour estimer des évolutions.

On peut répondre à cette critique que l'approche plus classique du §1 nécessite aussi de faire des hypothèses sur les proportions théoriques p_1 en période $t = 1$ et p_2 en période t_2.

Une autre façon indirecte de répondre à cette objection serait d'examiner la sensibilité des tailles d'échantillon lorsque le modèle de fluctuation est un autre modèle comme, par exemple, le modèle logistique.

Bibliographie

[1] BERKSON J. — « Why I prefer logits to probits », *Biometrics,* 1951, 7, pp. 327-339.

[2] BLIGHT B.J.N. et SCOTT A.J. — « A Stochastic Model for Repeated Surveys, *JRSS,* ser. B, 1973, 35, pp. 61-66.

[3] COCHRAN W.G. et COX G.M. — *Experimental Designs,* Wiley, 2e éd. 1957.

[4] COX D.R. — *Analyse des données binaires,* Dunod, Paris, 1972.

[5] DEROO M. et DUSSAIX A.M. — « Pratique et analyse des enquêtes par sondage », *Presses Universitaires de France,* 1980.

[6] FLEISS J.L. — *Statistical Methods for Rates and Proportions,* 2e édition, Wiley, 1981.

[7] GOURIEROUX Ch. et ROY G. — « Enquête en deux vagues, renouvellement de l'échantillon », *Annales de l'INSEE,* 1978, Vol. 29, pp. 115-134.

[8] LAYCOCK P.J. et AL KASSAB M.M.T. — « Optimum sample size and swing detection for repeated surveys », *Utilitas Mathematica,* 21, Vol. B, 1982, pp. 227/237.