

SONDAGES ET PLANS D'EXPERIENCES

1. INTRODUCTION

Lorsque l'on réalise une enquête par sondage auprès d'une population, il semble naturel de construire un échantillon « représentatif » de la population. Or, bien souvent, l'objectif est en fait de comparer le comportement de sous-populations. On se trouve alors dans un cadre qui relève davantage du plan d'expériences que du sondage.

Au §7.2. nous présentons un exemple trivial de plan d'expériences, et au § 7.3. nous montrons sur un exemple les qualités respectives de deux échantillons, l'un optimal pour estimer une proportion dans une population, l'autre optimal pour comparer deux proportions relatives à deux sous-populations.

En conclusion :

même lorsqu'on réalise une enquête auprès d'une population, on peut être amené à construire un échantillon stratifié qui relève davantage du plan d'expériences que de l'échantillon « représentatif ».

2. QU'EST-CE-QU'UN PLAN D'EXPERIENCES ?

Prenons un exemple. On fabrique avec les mêmes ingrédients deux gâteaux dans chacune des conditions suivantes :

A : sans les blancs d'œufs (–) / avec les blancs d'œufs (+)

B : durée de cuisson 10 minutes (–) / durée de cuisson 20 minutes (+)

Les notes données par un jury de dégustation sont les suivantes :

A	B	Y
–	–	9.2
–	–	8.8
–	+	10.4
–	+	11.6
+	–	17.4
+	–	16.6
+	+	11.2
+	+	10.8

A et B sont les *facteurs* et ont deux *niveaux* chacun (–, +).

Y est la *réponse*.

On a ici un *plan complet avec répétition* (2 expériences par niveau des facteurs croisés).

La moyenne des notes est 12.

Par niveau de chaque facteur, les moyennes sont :

$$\begin{array}{ll} A - : 10 & A + : 14 \\ B - : 13 & B + : 11 \end{array}$$

L'effet du facteur A est +2 (par rapport à la moyenne) s'il est au niveau + (avec les blancs d'œufs) et -2 s'il est au niveau - (sans les blancs d'œufs).

De même l'effet du facteur B est +1 s'il est au niveau - (durée de cuisson 10 mn) et -1 s'il est au niveau + (durée de cuisson 20 mn).

Il semble donc que les conditions optimales de fabrication du gâteau sont A+ et B-.

Les moyennes par niveau des facteurs croisés confirment ce résultat :

A	B	$\bar{y}$
-	-	9
-	+	11
+	-	17
+	+	11

On peut même remarquer qu'au niveau A+ et B-, alors qu'on pouvait s'attendre à une note moyenne de 15 [(moyenne) + (effet du niveau + de A) + (effet du niveau - de B)], on a une moyenne de 17 ; on a donc une **interaction** de +2 lorsque le facteur A est au niveau + et le facteur B au niveau -.

On peut alors écrire, pour chacune des expériences, la relation suivante :

$$Y = \alpha + \alpha_A + \alpha_B + \alpha_{AB} + e$$

où α est l'effet moyen, α_A et α_B l'effet du facteur A et du facteur B respectivement, α_{AB} l'interaction des facteurs A et B et e une erreur.

On montre que les coefficients qui rendent minimum la somme des carrés des erreurs sont, pour chaque niveau des facteurs et des facteurs croisés :

$$\begin{aligned} \alpha &= \bar{Y} \\ \alpha_A &= \bar{Y}_A - \bar{Y} \\ \alpha_B &= \bar{Y}_B - \bar{Y} \\ \alpha_{AB} &= \bar{Y}_{AB} - \bar{Y}_A - \bar{Y}_B + \bar{Y} \end{aligned}$$

c'est-à-dire, dans notre exemple :

$$\alpha = 12$$

$$\alpha_A(-) = -2 \quad \alpha_A(+) = +2$$

$$\alpha_B(-) = +1 \quad \alpha_B(+) = -1$$

$$\alpha_{AB}(-,-) = -2 \quad \alpha_{AB}(-,+) = +2 \quad \alpha_{AB}(+,-) = +2 \quad \alpha_{AB}(+,+) = -2$$

L'estimation de ces paramètres relève de la méthode statistique connue sous le nom d'**Analyse de la Variance**.

Lorsqu'on peut réaliser un plan complet (c'est-à-dire au moins une expérience par niveau de facteurs croisés), il est relativement aisé de déterminer les **conditions expérimentales optimales**.

En revanche, lorsque le nombre de facteurs et les nombres de niveaux des facteurs augmentent, il est très vite impossible de réaliser un plan complet.

On a alors recours à des **plans incomplets**. Il existe une théorie utilisant des propriétés combinatoires et algébriques, pour la construction de plans d'expériences optimaux, c'est-à-dire des plans permettant, avec un nombre faible d'expériences par rapport au nombre de niveaux croisés, d'estimer avec précision des effets et de déterminer les conditions expérimentales optimales.

3. SONDAGES ET PLAN D'EXPERIENCES : A CHAQUE OBJECTIF SON ECHANTILLON

Dans le cas d'enquête socio-économique auprès d'une population d'individus en chair et en os, si l'on se trouve dans un contexte expérimental, il s'agira le plus souvent de plans complets avec répétition.

Prenons un exemple.

On réalise une enquête auprès des chefs d'entreprises françaises dans les secteurs secondaire et tertiaire.

La répartition des entreprises selon le secteur d'activité (Industrie, Commerce, Services) et la taille, mesurée en nombre de salariés et répartie en quatre classes (-50, 50 à 100, 100 à 500, +500), est donnée dans le tableau suivant :

Taille Secteur	−50	50 à 100	100 à 500	+500	Ens.
Industrie	38 200	6 000	4 800	1 000	50 000
Commerce	43 400	4 000	2 200	400	50 000
Services	39 300	5 000	5 000	700	50 000
Ens.	120 900	15 000	12 000	2 100	150 000

On se propose de réaliser un sondage stratifié selon ces deux variables, de taille globale 600.

Comment répartir l'échantillon selon les strates afin de répondre au mieux à chacun des objectifs suivants ?

Objectif 1 : estimer, avec le maximum de précision la proportion, p , de chefs d'entreprises satisfaits par une certaine mesure gouvernementale.

Objectif 2 : estimer, avec le maximum de précision, les différences de proportions de chefs d'entreprises satisfaits par la mesure gouvernementale selon les classes de taille.

On notera p_1 , p_2 , p_3 , p_4 les proportions de chefs d'entreprises satisfaits dans chaque classe.

Solution

On montre que, en l'absence d'information supplémentaire, l'objectif 1 est atteint avec un échantillon stratifié proportionnel et l'objectif 2 avec un échantillon équiréparti par strate.

- ***Objectif 1 : échantillon stratifié proportionnel***

Taille Secteur	−50	50 à 100	100 à 500	+500	Ens.
Industrie	153	24	19	4	200
Commerce	173	16	9	2	200
Services	157	20	20	3	200
Ens.	483	60	48	9	600

- ***Objectif 2 : échantillon équiréparti par strate***

Taille Secteur	−50	50 à 100	100 à 500	+500	Ens.
Industrie	50	50	50	50	200
Commerce	50	50	50	50	200
Services	50	50	50	50	200
Ens.	150	150	150	150	600

Supposons que les proportions observées sur les échantillons de chaque strate soient données dans le tableau suivant.

Dans les marges, on a reporté les proportions par secteur d'activité, par classe de taille et pour l'ensemble. Ces proportions sont calculées selon les formules d'estimation de proportion pour un échantillon stratifié. Par exemple, pour la dernière classe de taille, on a :

$$\frac{1}{2100} (1000 \times 0.9 + 400 \times 0.9 + 700 \times 0.6) = 0.8$$

Taille Secteur	−50	50 à 100	100 à 500	+500	Ens.
Industrie	0.60	0.70	0.80	0.90	0.6372
Commerce	0.30	0.50	0.70	0.90	0.3384
Services	0.30	0.40	0.50	0.60	0.3342
Ens.	0.3948	0.5467	0.6567	0.8000	0.4367

Dans la suite on se propose de répondre à chacun des deux objectifs pour chacun des deux échantillons.

- **Objectif 1 : estimation de p**

⇒ Échantillon stratifié proportionnel

$$\hat{p} = 0.4367$$

$$\hat{V}(\hat{p}) = \frac{1}{(150000)^2} \left[(38200)^2 \frac{0.6 \times 0.4}{152} + (43400)^2 \frac{0.3 \times 0.7}{172} + \dots + (700)^2 \frac{0.6 \times 0.4}{2} \right]$$

$$\hat{V}(\hat{p}) = (0.0193)^2$$

d'où un intervalle de confiance à 95% de p :

$$0.4367 \pm 1.96 \times 0.0193 \quad \text{soit} \quad [39.9\% ; 47.4\%]$$

⇒ Échantillon équiréparti par strate

$$\hat{p} = 0.4367$$

$$\hat{V}(\hat{p}) = \frac{1}{(150000)^2} \left[(38200)^2 \frac{0.6 \times 0.4}{49} + (43400)^2 \frac{0.3 \times 0.7}{49} + \dots + (700)^2 \frac{0.6 \times 0.4}{49} \right]$$

$$\hat{V}(\hat{p}) = (0.0316)^2$$

d'où un intervalle de confiance à 95% de p :

$$0.4367 \pm 1.96 \times 0.0316 \quad \text{soit} \quad [37.5\% ; 49.9\%]$$

Conclusion

L'échantillon stratifié proportionnel permet d'obtenir un intervalle de confiance à 95% de p plus étroit que l'échantillon équiréparti par strate.

• **Objectif 2 : estimation de $d = p_4 - p_3$**

$\Rightarrow$ Échantillon stratifié proportionnel

$$\hat{d} = \hat{p}_4 - \hat{p}_3 = 0.8 - 0.6567 \quad \hat{d} = 0.1433$$

$$\hat{V}(\hat{d}) = V(\hat{p}_4) + V(\hat{p}_3)$$

$$\begin{aligned} \hat{V}(\hat{d}) = & \frac{1}{(2100)^2} \left[(1000)^2 \frac{0.9 \times 0.1}{3} + (400)^2 \frac{0.9 \times 0.1}{1} + (700)^2 \frac{0.6 \times 0.4}{2} \right] \\ & + \frac{1}{(12000)^2} \left[(4800)^2 \frac{0.8 \times 0.2}{18} + (2200)^2 \frac{0.7 \times 0.3}{8} + (5000)^2 \frac{0.5 \times 0.5}{19} \right] \end{aligned}$$

$$\hat{V}(\hat{d}) = (0.167)^2$$

d'où un intervalle de confiance à 95% de d :

$$0.1433 \pm 1.96 \times 0.167 \quad \text{soit} \quad [-0.18 ; 0.47]$$

Conclusion

La différence entre p_3 et p_4 n'est pas significative au seuil de 5%. En fait on obtient une très mauvaise précision sur l'estimation de d .

$\Rightarrow$ Échantillon équiréparti par strate

$$\hat{d} = 0.1433$$

$$\begin{aligned} \hat{V}(\hat{d}) = & \frac{1}{(2100)^2} \left[(1000)^2 \frac{0.9 \times 0.1}{49} + (400)^2 \frac{0.9 \times 0.1}{49} + (700)^2 \frac{0.6 \times 0.4}{49} \right] \\ & + \frac{1}{(12000)^2} \left[(4800)^2 \frac{0.8 \times 0.2}{49} + (2200)^2 \frac{0.7 \times 0.3}{49} + (5000)^2 \frac{0.5 \times 0.5}{49} \right] \end{aligned}$$

$$\hat{V}(\hat{d}) = (0.051)^2$$

d'où un intervalle de confiance à 95% de d :

$$0.1433 \pm 1.96 \times 0.051 \quad \text{soit} \quad [0.04 ; 0.24]$$

Conclusion

La différence entre p_3 et p_4 est significative au seuil de 5%.