STT-4600 STT-6220

CHAPITRE 2

Section 2.2 **Échantillonnage probabiliste.**

Un plan de sondage aléatoire est une fonction de probabilité définie dans l'ensemble des sous-ensembles de la population U. Définir un plan de sondage est équivalent à spécifier une distribution conjointe pour les variable indicatrices Z_i, i=1,…,N, où Z_i=1 si l'unité i est dans l'échantillon et Z_i=0 sinon.

Pour caractériser les distributions échantillonnales de statistiques calculées à partir de l'échantillon on va utiliser les probabilités de sélection simple et conjointe:

$$\pi_i = \Pr(Z_i = 1) \qquad \pi_{ij} = \Pr(Z_i = 1, Z_j = 1) \text{ où } i \neq j$$

Exemple : PLAN DE SONDAGE POUR UNE POPULATION DE 6 UNITÉS

Soit une population de N=6 unités et une variable Y dont les valeurs sont

Unité	1	2	3	4	5	6
Y	20	10	8	3	12	4

La sélection de n=3 unités se fait selon la méthode suivante : la première unité (#1) est toujours tirée et les deux autres unités sont choisies au hasard parmi les 5 restantes. Il y a 10 façons de choisir les deux unités restantes. Le plan de sondage est donné par

$$P(s)=\begin{cases}1/10 & \text{si s contient 3 unités dont la 1ere}\\ 0 & \text{sinon}\end{cases}$$

Les 10 échantillons possibles et la valeur de la moyenne échantillonnale pour chacun sont donnés dans le tableau 1.

Tableau 1.Les 10 échantillons possibles selon le plan *P(.)*

Echant.	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6	*P(s)*	$\bar{y}_s$
1	1	1	1	0	0	0	1/10	38/3
2	1	1	0	1	0	0	1/10	11
3	1	1	0	0	1	0	1/10	14
4	1	1	0	0	0	1	1/10	34/3
5	1	0	1	1	0	0	1/10	31/3
6	1	0	1	0	1	0	1/10	40/3
7	1	0	1	0	0	1	1/10	32/3
8	1	0	0	1	1	0	1/10	35/3
9	1	0	0	1	0	1	1/10	9
10	1	0	0	0	1	1	1/10	12

où Z_i=1 si l'unité i fait partie de l'échantillon et 0 sinon.

La moyenne échantillonnale s'écrit $\bar{y}_s = \frac{1}{3}\sum_{i\in S} y_i = \frac{1}{3}\sum_{i=1}^{N} Z_i y_i$. On peut calculer l'espérance et la variance à partir de l'énumération de tous les échantillons possibles.

Exemple (suite): CALCUL DES DEUX PREMIERS MOMENTS DE LA MOYENNES ÉCHANTILLONNALE
Son espérance vaut

$$E[\bar{y}_s] = \frac{38+33+42+34+31+40+32+35+27+36}{3\times 10} = 11.6$$

Elle est différente de $\bar{y}_U = 9.5$, la moyenne de *Y* dans la population de taille 6. (Vérifier que si on exclût la première unité des calculs, l'espérance de la moyenne échantillonnale pour les deux unités sélectionnées au hasard dans la population de taille 5 est bien égale à la moyenne pour la population de taille 5.)

On peut également calculer la variance de $\bar{y}_s$ à l'aide de la formule usuelle,

$$\mathrm{Var}[\bar{y}_s] = \mathrm{E}\left[\{\bar{y}_s - E(\bar{y}_s)\}^2\right] = \frac{(12.67-11.6)^2 + ... + (12-11.6)^2}{10} = 1.90$$

Dans le plan précédent on a π_1=1 et π_i=2/5 pour i=2,3,…,6. De même π_{1i}=2/5 et π_{ij}=1/10 pour *i* et *j* supérieurs à 1. Pour calculer les probabilités simples et conjointes, il suffit de faire la somme des probabilités de tous les échantillons contenant les unités visées. Par exemple 4 échantillons contiennent les unités 1 et 2, donc π_{2i}=4/10=2/5.

Sachant que $\bar{y}_s = \frac{1}{3}\sum_{i=1}^{N} Z_i y_i$ on fait un calcul d'espérance direct,

$$E(\bar{y}_s) = \frac{1}{3}\sum_{i=1}^{N} E(Z_i) y_i = \frac{1}{3}\left(y_1 + \frac{2}{5}\left[y_2 + y_3 + ... + y_6 \right] \right) = 11.6$$

qui ne nécessite pas une énumération complète de tous les cas possibles. C'est donc en utilisant les Z_i que nous allons déterminer les propriétés échantillonnales des statistiques pour des plans de sondage dans de grandes populations où un dénombrement exhaustif est impossible.

CONCLUSION

1. L'espérance de la moyenne échantillonnale pour un plan de sondage aléatoire quelconque n'est pas toujours égale à la moyenne dans la population
2. Le calcul des deux premiers moments d'une statistique peut se faire en utilisant les probabilités de sélection simple et conjointe, sans faire une énumération de tous échantillons possibles.

Section 2.3 **Échantillonnage aléatoire simple sans remise.**

Si N et n représentent les tailles de la population et de l'échantillon, un plan de sondage aléatoire simple est défini par le plan de sondage suivant,

$$P(s)=\begin{cases} 1/\binom{N}{n} & \text{si } s \text{ contient } n \text{ unités} \\ 0 & \text{sinon} \end{cases}$$

Proposition. Pour un plan aléatoire simple sans remise, les probabilités de sélection simple et conjointe s'écrivent :

$$\pi_i = n/N \qquad \pi_{ij} = n(n-1)/[N(N-1)] \text{ où } i \neq j$$

Démonstration : La probabilité pour que l'unité i soit dans l'échantillon est

$$\pi_i = \frac{\#\text{ sous-ensembles de taille n contenant i}}{\#\text{ sous-ensembles de taille n}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N}$$

De même $\pi_{ij} = \dfrac{\#\text{ sous-ensembles de taille n contenant i et j}}{\#\text{ sous-ensembles de taille n}} = \binom{N-2}{n-2} / \binom{N}{n} = \dfrac{n(n-1)}{N(N-1)}$

♦

Pour sélectionner un tel échantillon dans une population on peut utiliser la procédure SURVEYSELECT de SAS ou bien la fonction `sample` de R . Par exemple pour tirer n=4 unités parmi N=20 unités on utilise :

```
sample(20,4)
[1] 10  1 11 17
```

L'échantillon de taille n=4 sera donc composé des unités numérotées 1, 10 11 et 17, s={1,10,11,17}.

Tableau 2. Nombres aléatoires produit avec Excel

6344	5415	0198	0508	2305	9691	5266
7195	2949	7886	2195	1433	6977	0035
3606	6544	0845	7871	4701	0567	2605
1089	**10**89	9330	9773	2290	3827	0194
3743	**01**27	0162	1956	5972	8041	7470
7094	3675	5542	5102	4242	3391	0627
1569	1539	2278	9006	0345	8837	5406
9911	1607	0053	4458	1195	6681	0476
3383	5399	7222	9556	0996	1378	5161
0845	8969	5673	4903	4441	9292	7733

On peut également utiliser des nombres aléatoires, comme ceux du tableau 2, et faire les *n* tirages une unité à la fois. Fixons la règle de lecture suivante : on lit les colonnes de gauche à droite en ne considérant que les deux premières chiffres de chaque nombre et en ne conservant que les nombres de 1 à 20. Ceci donne les unités suivantes 10, 15, 08, 10 (déjà tiré ne compte pas), 01. On obtient s={10,15,08,01}.

DISCUSSION : Pour tirer un échantillon aléatoire simple sans remise, il faut utiliser un mécanisme aléatoire qui fait en sorte que toutes les unités de la population aient des chances égales d'être tirées. Prendre les *n* premières unités rencontrées est acceptable dans la mesure où les unités de la population ont au préalable subies une permutation aléatoire.

Propriétés échantillonnales de la moyenne dans un plan aléatoire simple sans remise

Soit $\{y_i, i \in \mathcal{S}\}$ les valeurs de la variable *y* recueillies auprès des *n* unités de l'échantillon. On rappelle que les probabilités de sélections sont $\pi_i = n/N \qquad \pi_{ij} = n(n-1)/[N(N-1)]$ où $i \neq j$.
L'estimateur de la moyenne $\bar{y}_U$ de *y* est la moyenne échantillonnale

$$\bar{y}_s = \frac{1}{n}\sum_{i\in\mathcal{S}} y_i = \frac{1}{n}\sum_{i=1}^{N} Z_i y_i$$

Proposition 1 L'espérance et la variance de $\bar{y}_s$ par rapport au plan de sondage sont :

$$E(\bar{y}_s) = \bar{y}_U = \frac{1}{N}\sum_{i=1}^{N} y_i \quad \text{et} \quad \text{Var}(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1} = \left(\frac{1}{n} - \frac{1}{N}\right)S_y^2 = (1-f)\frac{S_y^2}{n}.$$

où $f=n/N$ est la fraction de sondage et S_y^2 est la variance de *y* dans la population,

$$S_y^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y}_U)^2}{N-1} = \frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i^2 - N\bar{y}_U^2\right\}.$$

Démonstration : Puisque $E(Z_i)=n/N$,

$$E(\bar{y}_s) = E\left(\frac{1}{n}\sum_{i=1}^{N} Z_i y_i\right) = \frac{1}{n}\sum_{i=1}^{N} E(Z_i) y_i = \frac{1}{N}\sum_{i=1}^{N} y_i$$

Pour la variance on utilise la formule $\mathrm{Var}(\sum X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{cov}(X_i, X_j)$. Puisque $\mathrm{Var}(Z_i)=(n/N)(1-n/N) = n(n-N)/N^2$ (variance d'une Bernoulli) et

$$\mathrm{Cov}(Z_i, Z_j) = E(Z_i Z_j) - E(Z_i)E(Z_j) = \Pr(Z_i = 1, Z_j = 1) - \left(\frac{n}{N}\right)^2$$

$$= \frac{n(n-1)}{N(N-1)} - \frac{n^2}{N^2} = -\frac{n(N-n)}{N^2(N-1)},$$

on peut faire les calculs suivants

$$\mathrm{Var}(\bar{y}_s) = \frac{1}{n^2}\left\{\sum_{i=1}^{N} \frac{n(N-n)}{N^2} y_i^2 - \sum_{i \neq j}^{N} \frac{n(N-n)}{N^2(N-1)} y_i y_j\right\}$$

$$= \frac{(N-n)}{nN^2}\left\{\sum_{i=1}^{N} y_i^2 - \frac{1}{(N-1)} \sum_{i \neq j}^{N} y_i y_j\right\} \text{ (on ajoute et on soustrait } \sum y_i^2/\text{(N-1) aux deux termes)}$$

$$= \frac{(N-n)}{nN^2}\left\{\sum_{i=1}^{N}\left(1 + \frac{1}{N-1}\right) y_i^2 - \frac{1}{(N-1)}\left(\sum_{i=1}^{N} y_i\right)^2\right\}$$

$$= \frac{(N-n)}{nN(N-1)}\left\{\sum_{i=1}^{N} y_i^2 - N\bar{y}_U^2\right\} = \frac{1-f}{n} S_y^2$$

♦

En statistique classique, la variance de la moyenne vaut σ^2/n. Lorsque la population est finie S_y^2 jour le rôle de σ^2 et il faut multiplier la variance par *1-f* , c'est la correction pour population finie. Souvent *N* est beaucoup plus grand que *n* et $f=n/N\approx 0$; on retrouve alors le résultat classique pour la variance de la moyenne échantillonnale.

Note : Si y est dichotomique et prend deux valeurs, 0 ou 1, on a alors

$$S_y^2 = \frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i^2 - N\bar{y}_U^2\right\} = \frac{1}{N-1}\left\{\sum_{i=1}^{N} y_i - N\bar{y}_U^2\right\} = \frac{N}{N-1}\bar{y}_U(1-\bar{y}_U) \approx \bar{y}_U(1-\bar{y}_U)$$

Proposition 2 Un estimateur non biaisé de $\mathrm{Var}(\bar{y}_s)$ est donné par :

$$v(\bar{y}_s) = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{\sum_{i\in\mathcal{S}}(y_i - \bar{y}_s)^2}{n-1} = (1-f)\frac{s_y^2}{n}.$$

<u>Démonstration</u> : On utilise le fait que $\frac{1}{2N(N-1)}\sum_{i,j=1}^{N}(x_i - x_j)^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$. En effet

$$\sum_{i,j=1}^{N}(x_i - x_j)^2 = \sum_{i=1}^{N}\sum_{j=1}^{N}(x_i^2 - 2x_i x_j + x_j^2) = N\sum_{i=1}^{N}x_i^2 + N\sum_{j=1}^{N}x_j^2 - 2\left(\sum_{i=1}^{N}x_i\right)\left(\sum_{j=1}^{N}x_j\right)$$

$$= 2N\left\{\sum_{i=1}^{N}x_i^2 - N\bar{x}_U^2\right\} = 2N\sum_{i=1}^{N}(x_i - \bar{x})^2.$$

On doit montrer que s_y^2 est un estimateur non biaisé de S_y^2 pour ce faire on écrit

$$s_y^2 = \frac{1}{2n(n-1)}\sum_{i,j\in U} Z_i Z_j (y_i - y_j)^2 .$$

On évalue cette espérance en utilisant le fait que E(Z_iZ_j)=Pr(Z_i=*1*, Z_j=*1*)=*n(n-1)/{N(N-1)}* :

$$E(s_y^2) = E\left(\frac{1}{2n(n-1)}\sum_{i,j\in U} Z_i Z_j (y_i - y_j)^2\right)$$

$$= \frac{1}{2N(N-1)}\sum_{i,j\in U}(y_i - y_j)^2$$

$$= S_y^2$$

♦`

Note : Si y est dichotomique et prend deux valeurs, 0 ou 1, on a alors $v(\bar{y}_s) = (1-f)\dfrac{\bar{y}_s(1-\bar{y}_s)}{n-1}$

Poids d'échantillonnage et estimation du total.

Le total T_y de y dans la population est $T_y = \sum_{i=1}^{N} y_i = N\bar{y}_U$. Le poids d'échantillonnage d'une unité de l'échantillon s est défini comme étant l'inverse de la probabilité de sélection. Pour le plan aléatoire simple sans remise $w_i = 1/\pi_i = N/n$. L'estimation du total de y, T_y, dans la population s'écrit

$$\hat{T}_y = N\bar{y}_s = \sum_{i \in S} w_i y_i$$

En fonction des poids d'échantillonnage, l'estimation de la variance du total s'écrit

$$v(\hat{T}_y) = (1 - f)\sum_{i \in S} \frac{n(w_i y_i - \hat{T}_y / n)^2}{n-1}$$

On va voir dans le cours que l'écriture $\hat{T}_y = \sum_{i \in S} w_i y_i$ est vraie quel que soit le plan de sondage utilisé pour recueillir l'échantillon. C'est un des résultats fondamentaux de ce cours.

Tirage avec remise

Les unités sont tirées avec remise s'il est possible de sélectionner une unité plus d'une fois. Un plan aléatoire simple avec remise pour tirer *n* unités consiste à tirer *n* échantillons aléatoires simples indépendants (sans remise) de taille 1. Au premier tirage toutes les unités ont une probabilité de 1/*N* d'être choisie. Le deuxième tirage est indépendant du résultat du premier tirage: toutes les unités ont une probabilité de 1/*N* d'être choisie et ainsi de suite. Il y a donc N^n échantillons ordonnés possibles. Quelles sont l'espérance et la variance de $\bar{y}_s$ pour ce plan?

Pour étudier cette question il est utile de définir une variable aléatoire *Y* comme étant la valeur de *y* d'une unité tirée au hasard dans la population. « Au hasard » veut dire que chaque unité a une probabilité 1/*N* d'être tirée. L'espérance et la variance de *Y* valent :

$$E(Y) = \sum_{i=1}^{N} y_i / N = \bar{y}_U \text{ et } \mathrm{Var}(Y) = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 / N = (N-1)S_y^2 / N .$$

Pour un tirage avec remise $\bar{y}_s$ la moyenne de *n* copies indépendantes de *Y*. Pour calculer l'espérance et la variance de $\bar{y}_s$ on utilise des résultats standards de statistique mathématique :

$$E_{ar}(\bar{y}_s) = E(Y) = \bar{y}_U \text{ et } \mathrm{Var}_{ar}(\bar{y}_s) = \frac{\mathrm{Var}(Y)}{n} = \frac{1}{n}\left(1 - \frac{1}{N}\right) S_y^2 .$$

De plus une estimation non biaisée de la variance de $\bar{y}_s$ s'écrit $v_{ar}(\bar{y}_s) = s_y^2 / n$, où s_y^2 est la variance échantillonnale, L'indice « *ar* » veut dire avec remise.

Discussion On note que la variance avec remise est plus grande que la variance sans remise : $\mathrm{Var}_{ar}(\bar{y}_s) > \mathrm{Var}(\bar{y}_s)$. Ainsi l'estimateur de variance avec remise $v_{ar}(\bar{y}_s) = s_y^2 / n$ est une approximation conservatrice de $v(\bar{y}_s) = (1 - f) s_y^2 / n$.

Si la taille d'échantillon *n* est beaucoup plus petite que la taille de la population *N*, c'est-à-dire si $f = n/N \approx 0$, les deux types d'échantillonnage sont, à toute fin pratique, équivalent. En effet il est très improbable qu'une unité soit tirée plus d'une fois dans le plan avec remise si $N \gg n$. On note que si $f \approx 0$, les formules pour la variance théoriques et l'estimateur de variance sont les mêmes pour les deux plans :

$$(1 - f)\frac{S_y^2}{n} \approx \frac{1}{n}\left(1 - \frac{1}{N}\right)S_y^2 \text{ et } (1 - f)\frac{s_y^2}{n} \approx \frac{s_y^2}{n}$$

En pratique l'échantillonnage est toujours sans remise. D'un point de vue technique le calcul et l'estimation des variances est cependant plus simple si on suppose que l'échantillonnage est avec remise. Dans la deuxième partie du cours, qui traite de plans compliqués, on va utiliser la stratégie d'estimer la variance pour un plan sans remise à l'aide de la variance avec remise correspondante. La variance avec remise devrait être, comme dans un plan aléatoire simple, une approximation conservatrice de la variance sans remise

Section 2.4 **Intervalle de confiance**

Le calcul d'un intervalle de confiance s'appuie sur une version « population finie » du théorème de la limite centrale due à Hajek en 1960. Son théorème porte suite une suite infinie de populations indexées par *N*. Si cette suite satisfait certaines hypothèses de régularité alors

$$\frac{\bar{y}_s^{(N)} - \bar{y}_U^{(N)}}{\sqrt{v(\bar{y}_s)^{(N)}}} \xrightarrow{N} N(0,1)$$

Les bornes d'un intervalle de confiance asymptotique à 100(1-α)% pour $\bar{y}_U$ sont données par $\bar{y}_s \pm z_{1-\alpha/2}\sqrt{v(\bar{y}_s)}$ où $z_{1-\alpha/2}$ est un percentile de la N(0,1) ($z_{.975}$=1.96). Lorsque *n* est petit, on remplace parfois la valeur critique normale par celle d'une *t* à *n-1* degrés de liberté.

Variable d'intérêt dichotomique :

Lorsque y_i prend les valeurs 0 ou 1, $\bar{y}_U$ représente la proportion d'unités de la population pour lesquelles *y*=1. On retrouve ce type de statistique dans les sondages d'opinion où *y*=1 si une personne appuie un certain parti et *y*=0 sinon. Dans ce cas $\bar{y}_s = \hat{p}$ est la proportion échantillonnale de 1 et

$$\text{Var}(\bar{y}_s) = (1-\frac{n}{N})\frac{N\bar{y}_U(1-\bar{y}_U)}{n(N-1)} \text{ et } v(\bar{y}_s) = (1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n-1};$$

la formule précédente pour l'intervalle de confiance asymptotique s'applique.

Tableau 3. Distribution échantillonnale de la moyenne dans une population de taille N=6 d'où on tire n=3 unités

Echant.	y_1=2	y_2=3	y_3=4	y_4=5	y_5=12	y_6=17	$\bar{y}_s$	$v(\bar{y}_s)$	IC-	IC+	Co.
1	1	1	1	0	0	0	3.00	0.17	1.24	4.76	0
2	1	1	0	1	0	0	3.33	0.39	0.65	6.01	0
3	1	1	0	0	1	0	5.67	5.06	-4.00	15.34	1
4	1	1	0	0	0	1	7.33	11.72	-7.39	22.06	1
5	1	0	1	1	0	0	3.67	0.39	0.99	6.35	0
6	1	0	1	0	1	0	6.00	4.67	-3.29	15.29	1
7	1	0	1	0	0	1	7.67	11.06	-6.63	21.96	1
8	1	0	0	1	1	0	6.33	4.39	-2.68	15.34	1
9	1	0	0	1	0	1	8.00	10.50	-5.93	21.93	1
10	1	0	0	0	1	1	10.33	9.72	-3.07	23.74	1
11	0	1	1	1	0	0	4.00	0.17	2.24	5.76	0
12	0	1	1	0	1	0	6.33	4.06	-2.33	14.99	1
13	0	1	1	0	0	1	8.00	10.17	-5.71	21.71	1
14	0	1	0	1	1	0	6.67	3.72	-1.63	14.96	1
15	0	1	0	1	0	1	8.33	9.56	-4.96	21.63	1
16	0	1	0	0	1	1	10.67	8.39	-1.79	23.12	1
17	0	0	1	1	1	0	7.00	3.17	-0.65	14.65	1
18	0	0	1	1	0	1	8.67	8.72	-4.03	21.37	1
19	0	0	1	0	1	1	11.00	7.17	-0.51	22.51	1
20	0	0	0	1	1	1	11.33	6.06	0.75	21.91	1
Moyenne de $\bar{y}_s$ pour les 20 échantillons							7.167	Taux de couverture moyen			0.8
Moyenne de $v(\bar{y}_s)$ pour les 20 échantillons							5.961				

On va étudier le taux de couverture réel de l'intervalle confiance à 95% pour $\bar{y}_U$ dans la population du tableau 3. Les valeurs de y pour les 6 unités sont 2, 3, 4, 5, 12, 17 ainsi $\bar{y}_U = 7.167$ S^2=35.77 et $\mathrm{Var}(\bar{y}_s) = 5.961$. On note au tableau 3 que $\bar{y}_s$ et $v(\bar{y}_s)$ s ont des estimateurs sans biais.

Le tableau 3 donne les 20 échantillons possibles et les bornes de l'intervalle de confiance pour chacun obtenu avec la valeur critique $t_{2,0.025} = 4.3$. On note que 16 des 20 intervalles de confiance recouvre la vraie valeur

$\bar{y}_U = 7.167$ ainsi le taux de couverture réel est de 80% ce qui est beaucoup plus faible que le taux de couverture nominal de 95%. Si on refaisait cette exercice avec une population où la distribution de *y* est voisine d'un normale on devait obtenir un taux nominal plus proche du vrai taux.

Conclusion : Les intervalles de confiance calculés en échantillonnage s'appuient sur des approximations asymptotiques qui sont susceptibles de donner des taux de couverture réels inférieurs aux taux nominaux suggérés. par l'approximation asymptotique.

Exemple : **Comparaison de deux proportions**
Dans un sondage d'opinion il y a souvent plusieurs choix de réponse mutuellement exclusifs pour une question. Par exemple à la question « Pour quel parti politique voteriez-vous s'il y avait une élection aujourd'hui? » il y a autant de choix de réponse que de partis politiques; c'est une expérience multinomiale.

S'il y a *J* choix de réponse, les variables dans la population s'écrivent y_{ij}, $i=1,\ldots,N$, $j=1,\ldots,J$ où y_{ij} =1 si l'unité *i* choisi la réponse *j* et 0 sinon; de plus $\Sigma_j y_{ij}=1$ pour tout *i*. L'estimation de la proportion de la population qui choisit la jième réponse est $\bar{y}_{sj}$ avec $v(\bar{y}_{sj}) = (1-f)\bar{y}_{sj}(1-\bar{y}_{sj})/(n-1)$. Si on veut comparer le soutien aux choix *j* et *k* on considère $\bar{y}_{sj} - \bar{y}_{sk}$. Comment estime-t-on la variance de cette différence ? Peut-on prendre

$v(\bar{y}_{sj} - \bar{y}_{sj}) = (1-f)\{\bar{y}_{sj}(1-\bar{y}_{sj}) + \bar{y}_{sk}(1-\bar{y}_{sk})\}/(n-1)$ comme si les deux estimations étaient indépendantes?

En fait $\bar{y}_{sj} - \bar{y}_{sk} = \bar{x}_s$ où x_i vaut 1 si le choix du sujet *i* est *j*, -1 si c'est *k* et 0 si ce n'est ni *j* ni *k*. De plus la variance échantillonnale des *x* est égale à

$$s_z^2 = \frac{1}{n-1}\{\sum_S x_i^2 - n(\bar{y}_{sj} - \bar{y}_{sk})^2\} = \frac{n}{n-1}\{\bar{y}_{sj} + \bar{y}_{sk} - (\bar{y}_{sj} - \bar{y}_{sk})^2\}$$

$$= \frac{n}{n-1}\{\bar{y}_{sj}(1-\bar{y}_{sj}) + \bar{y}_{sk}(1-\bar{y}_{sk}) + 2\bar{y}_{sj}\bar{y}_{sk}\}$$

Ainsi $v(\bar{y}_{sj} - \bar{y}_{sj}) = v(\bar{y}_{sj}) + v(\bar{y}_{sk}) + 2(1-f)\bar{y}_{sj}\bar{y}_{sk}/(n-1)$ et la somme des deux variances sous-estime la vraie variance.

Exemple (suite) : Dans un sondage auprès de n=301 jeunes de 25-35 ans de Québec on a demandé « Quelle personnalité sportive admirez-vous le plus ? Les 3 personnalités les plus choisies sont Alexandre Despaties, Chantal Petitclerc et Mario Lemieux. Les données sont

n=301	Despaties	Petitclerc	Lemieux	Autres
# personnes	57	30	15	204
P(chapeau)	0.189	0.100	0.050	0.678
e.t.	0.023	0.017	0.013	0.027

Peut-on conclure que Despaties est plus populaire que Petitclerc ? La différence des préférences est de 0.089 et la variance est de $v(\hat{p}_{De} - \hat{p}_{Pe}) = v(\hat{p}_{De}) + v(\hat{p}_{Pe}) + 2\hat{p}_{De}\hat{p}_{Pe}/(n-1) = .031^2$. La

statistique normale pour faire ce test est .089/.031=2.931 pour un seuil observé de $2\times Pr(N(0,1)>2.931)=0.34\%$; la différence est significative; Despaties est plus populaire que Petitclerc. L'erreur-type calculée sous l'hypothèse d'indépendance est de .028; elle sous-estime la vraie erreur-type.

Section 2.5 **Calcul de tailles d'échantillons et marge d'erreur**

Lorsque l'on planifie une enquête on choisit souvent la taille d'échantillon pour faire en sorte que l'estimateur de la moyenne ait une précision prédéterminée. Cet objectif peut s'exprimer en terme d'une marge d'erreur absolue e_a si, par exemple, on cherche une taille d'échantillon telle que

$$P(|\bar{y}_s - \bar{y}_U| \leq e_a) = 1-\alpha$$

Cette équation peut se traduire en une équation pour n_a, la taille d'échantillon pour une marge d'erreur absolue. En effet, en vertu du théorème de la limite centrale de Hajek,

$$P(|\bar{y}_s - \bar{y}_U| \leq e_a) = P\left(|\frac{\bar{y}_s - \bar{y}_U}{\sqrt{\mathrm{Var}(\bar{y}_s)}}| \leq \frac{e_a}{\sqrt{\mathrm{Var}(\bar{y}_s)}}\right) = 2\Phi\left(\frac{e_a}{\sqrt{\mathrm{Var}(\bar{y}_s)}}\right)-1.$$

Pour que cette probabilité vaille (1-α), il faut que

$$\frac{e_a}{\sqrt{\text{Var}(\bar{y}_s)}} = z_{1-\alpha/2} \text{ c'est-à-dire } e_a = z_{1-\alpha/2}\sqrt{\left(1-\frac{n_a}{N}\right)\frac{S_y^2}{n_a}} \text{ ou bien } n_a = \frac{z_{1-\alpha/2}^2 S_y^2 / e_a{}^2}{1+\frac{z_{1-\alpha/2}^2 S_y^2}{N e_a{}^2}}$$

Si *N* est grand la formule de taille d'échantillon se réduit à $n_a = z_{1-\alpha/2}^2 S_y^2 / e_a{}^2$.

Pour une marge d'erreur relative de e_r, on veut que

$$P\left(\frac{|\bar{y}_s - \bar{y}_U|}{\bar{y}_U} \le e_r\right) = 1-\alpha$$

La solution pour n_r (l'indice *r* veut dire relative) se calcule comme pour une erreur absolue. On obtient

$$n_r = \frac{z_{1-\alpha/2}^2 S_y^2 / (\bar{y}_U e_r)^2}{1+\frac{z_{1-\alpha/2}^2 S_y^2}{N(\bar{y}_U e_r)^2}}$$

Cette formule fait intervenir le carré du coefficient de variation de la variable *y*, CV= $\sqrt{S_y^2} / \bar{y}_U$.

Lorsque l'on estime une proportion la valeur maximale de la variance S^2 est ¼. Ainsi si on veut avoir une marge d'erreur absolue de e_a avec un taux de confiance de 95% (19 fois sur 20), on applique la formule pour n_a , avec une variance de ¼, ce qui donne

$$n_a = \frac{1.96^2 / (4e_a{}^2)}{1 + 1.96^2 / (4Ne_a{}^2)}$$ ou bien $n_a = 3.84/(4e_a{}^2)$ si $N=\infty$.

Si $e_a = 5\%$ on a n_a=278 si N=1000 et n_a=384 si $N=\infty$. Avec $e_a = 1\%$ il faut prendre n_a=906 et n_a=9600 dans ces deux cas.

MARGE D'ERREUR ESTIMÉE

La marge d'erreur (absolue) estimée est la demi-longueur de l'intervalle de confiance pour $\bar{y}_U$. Dans le cas d'une variable y dichotomique, on peut rapporter la marge d'erreur maximale obtenue avec $\hat{p} = 1/2$.

Exemple : Dans un sondage auprès de n=425 personnes de la région de Québec, 26.8% veulent voir Céline Dion au festival d'été. Calculer un intervalle de confiance à 95% pour la proportion réelle de la population de Québec qui veut voir un tel spectacle. On a $f \approx 0$ et les bornes de l'IC sont

$$.268 \pm 1.96\sqrt{\frac{.268 \times .732}{424}} = (.226, .310)$$

Ainsi la marge d'erreur est (.310-.226)/2=4.2%, 19 fois sur 20.

On rapporte souvent, dans les média, la marge d'erreur maximale obtenue pour une estimation de ½. Avec un échantillon de 425 personnes elle est donnée par

$$1.96\sqrt{\frac{.5\times.5}{425}}=\frac{1.96}{2\sqrt{425}}=4.76\%.$$

On utilise n et non pas n-1 dans la formule de variance parce qu'on travaille avec la variance exacte $\mathrm{Var}(\bar{y}_s)$ et non pas son estimateur $v(\bar{y}_s)$.

CONCLUSION

Voici les principaux résultats obtenus pour le plan aléatoire simple sans remise :
La moyenne de y dans la population est $\bar{y}_U$, N est la taille de la population et n celle de l'échantillon. Si $\bar{y}_s$ est la moyenne échantillonnale la variance théorique de cet estimateur est

$\mathrm{Var}(\bar{y}_s) = \frac{1-f}{n} S_y^2$ où $S_y^2 = \sum_{i=1}^{N} (y_i - \bar{y}_U)^2 / (N-1)$ et f=n/N est la fraction de sondage.

Un estimateur de cette variance est donné par $v(\bar{y}_s) = (1-f)s_y^2 / n$ où s_y^2 est la variance échantillonnale de y. Le poids d'échantillonnage est $w_i=N/n$ pour tout $i \in \mathcal{S}$

Lorsqu'on planifie une enquête, on peut obtenir les tailles d'échantillons nécessaires pour obtenir une marge d'erreur prédéterminée. Pour obtenir une marge d'erreur absolue de e avec un taux de confiance de 100(1-α)% on prend $n = \frac{z_{1-\alpha/2}^2 S_y^2 / e^2}{1 + \frac{z_{1-\alpha/2}^2 S_y^2}{Ne^2}}$ alors que $n = \frac{z_{1-\alpha/2}^2 S_y^2 / (\bar{y}_U e)^2}{1 + \frac{z_{1-\alpha/2}^2 S_y^2}{N(\bar{y}_U e)^2}}$ pour avoir une marge d'erreur relative de e .