TRAVAUX DIRIGES DE METHODES DE SONDAGE

THEME : RAPPELS DE STATISTIQUE MATHEMATIQUE

Exercice 1 :

Considérons deux variables aléatoires indépendantes discrètes *X* et *Y*.

X_i	$Prob(X_i)$	Y_i	$Proba(Y_j)$
0	0.50	0	0.40
1	0.20	2	0.60
2	0.30		**1**
	1		

1- Calculer l'espérance mathématique de *X* et de *Y*.
2- On multiplie la variable aléatoire *X* par 3. Vérifier que *E(3X)=3 E(X)*.
3- Calculer E(X+Y), et vérifier que E(X+Y)=E(X)+E(Y).
4- Calculer la variance de *X* et de *Y*.
5- Déterminer V(3Y) et V(X+Y). Conclure.

Exercice 2 :

Considérons une population composée de quatre entreprises {1,2,3,4}. On s'intéresse au chiffre d'affaire mensuel moyen $\overline{CA}$ de cette population d'entreprises. Le chiffre d'affaire des différentes entreprises est : $CA_1 = 6000$ €, $CA_2 = 12000$ €, $CA_3 = 8000$ € et $CA_4 = 6000$.

La contrainte de budget est telle que l'on ne peut interroger que deux entreprises sur quatre.

1- Combien peut-on constituer d'échantillon de taille 2 dans une population de taille 4 ?
2- Etablir la liste de tous les échantillons possibles.

Parce qu'on juge que l'entreprise 1 est particulièrement coopérative sur ce sujet, on veut lui donner une probabilité de tirage supérieure aux trois autres, si bien que les trois échantillons s_1, s_2, s_3 sont un « peu plus probables » que les autres. On a donc les probabilités de tirage suivantes : $p(s_1) = 0.25$, $p(s_2) = 0.25$, $p(s_3) = 0.2$, $p(s_4) = 0.1$, $p(s_5) = 0.1$ et $p(s_6) = 0.1$. On vérifie que : $\sum_{k=1}^{6} p(s_k) = 1$.

Par souci de simplicité, on choisit comme estimateur la moyenne simple dans l'échantillon.

3- Pour chaque échantillon, calculer l'estimateur du chiffre d'affaire moyen.

4- Déterminer le biais, la variance, le coefficient de variation et l'erreur quadratique moyenne de cet estimateur. Commenter.

Exercice 3 :

Soit *X* une variable aléatoire dont la loi dépend d'un paramètre réel θ. $\hat{T}_1$ et $\hat{T}_2$ sont deux estimateurs indépendants de θ, sans biais de variances respectives V_1 et V_2. On considère l'estimateur $\hat{T}_3 = a\hat{T}_1 + (1-a)\hat{T}_2$, *a* étant un nombre réel quelconque.

1- Montrer que $\hat{T}_3$ est un estimateur sans bais de θ.
2- Déterminer *a* de telle sorte que l'estimateur $\hat{T}_3$ soit de variance minimale.

Exercice 4:

Soit $(X_1, X_2, ..., X_n)$ un échantillon issu d'une loi normale de moyenne θ et de variance $\theta(1-\theta)$, où θ est un paramètre inconnu appartenant à l'intervalle $]0.1[$. On considère les estimateurs :

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{et} \qquad \hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

1- Montrer que $\hat{\theta}_1$ et $\hat{\theta}_2$ sont sans biais et convergents.
2- Quel estimateur a-t-on intérêt à choisir ?

NB : on donne $V(X_1^2) = 2\theta^2(1-\theta^2)$

Exercice 5:

Soit $\hat{\theta}$ un estimateur du paramètre θ de biais= $E(\hat{\theta}) - \theta$. Pour comparer deux estimateurs quelconques de θ, on peut utiliser l'erreur quadratique moyenne :

EQM= $E(\hat{\theta} - \theta)^2$

1- Montrer que l'erreur quadratique moyenne est la somme du bais au carré et de la variance.

TRAVAUX DIRIGES DE METHODES DE SONDAGE

THEME : LE SONDAGE ALEATOIRE SIMPLE

Exercice 1 :
On désire estimer la surface moyenne cultivée dans les fermes d'un canton rural. Sur N=2010 fermes que comprend ce canton, on en tire 100 par sondage aléatoire simple. On mesure alors Y_i la surface cultivée dans la ferme i en hectares et on trouve les résultats suivants :

$$\sum_{i \in s} Y_i = 2907 \text{ ha} \qquad \text{et} \qquad \sum_{i \in s} Y_i^2 = 154593 \text{ ha}^2$$

1- Donner la valeur de l'estimateur sans biais de la moyenne dans la population.
2- Donner un intervalle de confiance à 95% pour $\overline{Y}$.

Exercice 2 :
145 ménages de touristes ayant séjourné en France dans une région donnée, ont dépensé en moyenne journalière 126.53 €. L'écart-type estimé de ces 145 dépenses journalières s'élève à 32.01 €. Sachant que dans la région où a été effectuée l'enquête il est venu 50000 ménages de touristes, que peut-on dire de la dépense totale journalière de ces ménages ?

Exercice 3 :
Un sondage portant sur l'opinion relative à une personnalité politique donne un pourcentage d'opinions favorables égal à $\hat{P} = 30\%$. Combien de personnes ont-elles été interrogées pour que l'on puisse dire, avec un degré de confiance de 95%, que le vrai pourcentage d'opinions favorables dans la population ne s'écarte pas de $\hat{P}$ de plus de deux points ?
Même question lorsque $\hat{P} = 50\%$? Commenter.

Exercice 4 :
Quelle taille d'échantillon retenir pour connaître à deux points de pourcentage près et avec 95 chances sur 100, la proportion de parisiens qui portent des lunettes ?

Exercice 5 :
On souhaite estimer le nombre d'ecclésiastiques dans la population française. Pour ce faire, on choisit d'échantillonner n individus. Si la véritable proportion (inconnue) d'ecclésiastiques parmi la population est de 0.1%, combien, faut-il tirer de personnes pour obtenir un coefficient de variation de 5% ?

Exercice 6 : Tirer un échantillon au 1/10 dans la population suivante.

Iden.	Nom
1	A
2	B
3	C
4	D
5	E
6	F
7	G
8	H
9	I
10	J
11	K
12	L
13	M
14	N
15	O
16	P
17	Q
18	R
19	S
20	T
21	U
22	V
23	W
24	X
25	Y

Iden.	Nom
26	Z
27	AA
28	AB
29	AC
30	AD
31	AE
32	AF
33	AG
34	AH
35	AI
36	AJ
37	AK
38	AL
39	AM
40	AN
41	AO
42	AP
43	AQ
44	AR
45	AS
46	AT
47	AU
48	AV
49	AW
50	AX

Iden.	Nom
51	AY
52	AZ
53	AAA
54	AAB
55	AAC
56	AAD
57	AAE
58	AAF
59	AAG
60	AAH
61	AAI
62	AAJ
63	AAK
64	AAL
65	AAM
66	AAN
67	AAO
68	AAP
69	AAQ
70	AAR
71	AAS
72	AAT
73	AAU
74	AAV
75	AAW

Iden.	Nom
76	AAX
77	AAY
78	AAZ
79	AAAA
80	AAAB
81	AAAC
82	AAAD
83	AAAE
84	AAAF
85	AAAG
86	AAAH
87	AAAI
88	AAAJ
89	AAAK
90	AAAL
91	AAAM
92	AAAN
93	AAAO
94	AAAP
95	AAAQ
96	AAAR
97	AAAS
98	AAAT
99	AAAU
100	AAAV

TRAVAUX DIRIGES DE METHODES DE SONDAGE

THEME : LE SONDAGE STRATIFIE

Exercice 1 :

Une grande entreprise veut réaliser une enquête auprès de son personnel qui comprend 10000 personnes. Des études préliminaires ont montré que :

- Les variables que l'on cherche à analyser dans l'enquête sont très contrastées selon les catégories de personnel, et qu'il y a donc intérêt à stratifier selon ces catégories. Pour simplifier, on considère qu'il y a trois grandes catégories de personnel qui formeront les strates ;
- Les variables sont également très fortement liées à l'âge des individus.

On va donc proposer des plans d'échantillonnage comme si on voulait étudier l'âge des individus : si une stratégie est meilleure qu'une autre pour estimer l'âge moyen, on a aussi de bonnes raisons de penser qu'elle sera aussi la meilleure pour les vrais variables d'intérêt. Comme on connaît l'âge des membres du personnel, on peut raisonner en effectuant les comparaisons exactes. On dispose donc des renseignements suivants :

Catégorie	Poids dans l'ensemble du personnel	Ecart-type des âges
1	20%	18
2	30%	12
3	50%	3.6
Ensemble	100%	16

1- Soit $\overline{Y}$ l'âge moyen et $\hat{\overline{Y}}$ l'estimateur issu d'un échantillon aléatoire simple de taille n=100 individus. Quelle est la précision de l'estimateur $\hat{\overline{Y}}$?

2- On décide que l'échantillon de 100 individus doit être stratifié selon les trois catégories de personnel. Déterminer l'allocation proportionnelle ? Quelle est la précision de l'estimateur de $\overline{Y}$ qui en découle ? Comparer avec les résultats de la question 1.

3- Quelle serait la répartition optimale de l'échantillon ? Quelle est la précision de l'estimateur de $\overline{Y}$ qui en découle ? Comparer avec les résultats de la question 2.

Exercice 2 :

Un directeur de cirque possède 100 éléphants classés en 2 catégories : « mâles et femelles ». Le directeur veut estimer le poids total de son troupeau car il veut traverser un fleuve en bateau. Toutefois, l'année précédente, ce même directeur de cirque avait fait peser tous les éléphants de son troupeau et avait obtenu les résultats présentés dans le tableau ci-dessous :

	Effectifs	Poids moyen (tonnes)	Dispersions
Mâles	60	6	4
Femelles	40	4	2.25

1- Calculez la dispersion dans la population de la variable « poids de l'éléphant » pour l'année précédente.
2- Le directeur suppose désormais que les dispersions de poids n'évoluent pas sensiblement d'une année sur l'autre. Si le directeur procède à un tirage aléatoire simple sans remise de 10 éléphants, quelle est la variance de l'estimateur du poids total du troupeau ?
3- Si le directeur procède à un tirage stratifié avec allocation proportionnelle de 10 éléphants, quelle est la variance de l'estimateur du poids total du troupeau ?
4- Si le directeur procède à un tirage stratifié optimal de 10 éléphants, quels sont les effectifs de l'échantillon dans chacune des deux strates et quelle est la variance de l'estimateur du total ?

Exercice 3 :

Dans une grande ville, on considère le nombre moyen de patients que peut avoir un médecin pendant une journée de travail. On part de l'idée *a priori* que plus le médecin a d'expérience, plus il a de patients. Cela nous conduit à classer la population de médecins en 3 groupes : les « débutants » (classe 1), les « confirmés » (classe 2) et les « très expérimentés » (classe 3). Par ailleurs, on suppose que l'on connaît, dans la base de sondage des médecins, la classe de chacun d'entre eux (1 ou 2 ou 3 = information auxiliaire). Ainsi, on dénombre 500 médecins en classe 1, 1000 en classe 2 et 2500 en classe 3. Par sondage aléatoire simple, on sélectionne 200 médecins dans chaque classe. On calcule alors, dans chaque classe, le nombre moyen de patients par jour et médecins échantillonné : 10 en classe 1, 15 en classe 2 et 20 en classe 3. On calcule enfin les dispersions des nombres de patients par médecin dans chacun des 3 échantillons et on trouve respectivement 4 (classe 1), 7 (classe 2) et 10 (classe 3).

1- Comment s'appelle ce plan de sondage ? Justifiez a priori son utilisation.
2- Estimez le nombre moyen de patients soignés par jour et par médecin.
3- Donnez un intervalle de confiance à 95% pour le « vrai » nombre moyen de patients soignés par médecin et par jour.
4- Si vous n'aviez comme contrainte que le nombre total de médecins à enquêter (soit 600), procéderiez-vous comme ci-dessus ?
5- Quel est le gain de variance estimée obtenu avec une allocation proportionnelle par rapport au sondage aléatoire simple (de taille 600) ?

Exercice 4 :

Un chargé d'étude doit réaliser une enquête dans deux régions A et B auprès de 500 exploitants agricoles, afin d'évaluer le nombre moyen de bovins par exploitation agricole. Le nombre total d'exploitations est de 50000, soit 40000 pour la région A et

10000 pour la région B. Le chargé d'études peut obtenir la base de sondage, dans laquelle chaque exploitation figure avec son adresse.
Un recensement agricole récent a montré que l'écart-type du nombre de bovins par exploitation était de 20 dans la région A et 40 dans la région B. Le sondage a pour but d'actualiser les chiffres du recensement, mais on admet que les écarts-types ont gardé le même ordre de grandeur. Pour tout l'exercice, on tire les exploitations avec des probabilités égales.

1- Décrire précisément comment peut être constitué un échantillon proportionnel avec la région comme critère de stratification. En quoi cet échantillon est-il différent de celui qu'on aurait constitué par un tirage sans stratification dans l'ensemble du territoire ?
2- Avec quelle précision seront estimés, grâce à cet échantillon, les nombres moyens par exploitation pour chaque région et pour le territoire dans son ensemble (degré de confiance de 95%) ?
3- Quelle serait la répartition à envisager si on souhaitait obtenir une estimation du nombre moyen de bovins par exploitation avec la même précision pour chaque région ?
4- Avec cette répartition, comment obtiendrait-on une estimation du nombre moyen de bovins par exploitation sur l'ensemble du territoire et quelle serait la précision de cet estimateur (degré de confiance de 95%) ?
5- Quelle serait la répartition qui donnerait une précision optimale pour cet estimateur et quelle serait cette précision optimale (degré de confiance de 95%) ?
6- En évaluant le budget de l'enquête, le chargé d'études s'aperçoit que les coûts unitaires d'enquête d'une exploitation de la région B sont plus élevés qu'en A ($c_A = 200$€ et $c_B = 300$€). Il se demande alors s'il lui serait possible sans fixer *a priori* la taille de l'échantillon, de déterminer le nombre d'interviews n_A et n_B, de sorte que le coût global de l'enquête soit minimum pour une précision donnée. Le coût global est donné par la formule suivante : $C = c_A n_A + c_B n_B$.
    a- Pour une précision donnée, fournir l'expression des rapports $\frac{n_A}{n}$ et $\frac{n_B}{n}$ qui minimisent *C*.
    b- Calculer la taille d'échantillon *n* nécessaire pour que, avec cette répartition, on ait $V(\hat{\bar{Y}}) = 1.139$.
    c- En déduire le coût global d'enquête *C*.
    d- Comparer avec le coût global d'enquête qui résulterait des tailles n_A et n_B obtenues à la question 5.

TRAVAUX DIRIGES DE METHODES DE SONDAGE

THEME : SONDAGE A PLUSIEURS DEGRES ET SONDAGE EN GRAPPES

Exercice 1 :

Sur un disque dur d'ordinateur, on compte 400 fichiers, chacun comprenant exactement 50 enregistrements. Afin d'estimer le nombre moyen de caractères par enregistrement, on décide de tirer par sondage aléatoire simple 80 fichiers, puis 5 enregistrements dans chaque fichier. On note : $m = 80$ et $\overline{n} = 5$

. On mesure après tirage :

- La dispersion des estimateurs du nombre total de caractères par fichier, soit $s_T^2 = 905000$;
- La moyenne des *m* dispersions $s_{2,i}^2$ est égale à 805, où $s_{2,i}^2$ représente la dispersion du nombre de caractère par enregistrement dans le fichier *i*.

1- Comment estimer sans biais le nombre moyen $\overline{Y}$ de caractères par enregistrement ?
2- Comment estimer sans biais la précision de l'estimateur précédent ?
3- Donner un intervalle de confiance à 95% pour $\overline{Y}$.

Exercice 2 :

Une banque a 39800 clients dans ses fichiers informatiques, répartis dans 3980 agences gérant chacune exactement 10 clients. On souhaite estimer la proportion des clients à qui la banque a accordé un prêt. Pour cela, on échantillonne, par sondage aléatoire simple, 40 agences (échantillon *s*) et on dénombre dans chaque agence *i*, T_i clients bénéficiaires d'un prêt. Les données issues de l'enquête sont :

$$\sum_{i \in s} T_i = 185 \qquad \text{et} \qquad \sum_{i \in s} T_i^2 = 1263$$

1- Comment appelle-t-on ce type de sondage ?
2- Donner l'expression du paramètre à estimer et son estimateur sans biais.
3- Estimer sans biais la variance de cet estimateur, et fournir un intervalle de confiance approché à 95%.
4- Calculer l'effet de sondage (DEFF).
5- Calculer le coefficient de corrélation intra-grappes ρ .
6- Estimer la précision que l'on obtiendrait en échantillonnant 80 agences et 5 clients par agence sélectionnée.

Exercice 3 :

Un journal a 40000 abonnés, desservis par transporteurs. Il y a une carte pour chaque abonné, et le fichier des cartes est trié par ordre géographique de sorte que les zones géographiques se suivent les unes les autres. Le but de l'enquête est d'estimer combien

d’abonnés sont propriétaires de leur résidence principale, en vue de cibler certaines publicités. La direction du journal passe commande d’une enquête par interviews auprès de 800 abonnés, pris par grappes de 10. Le responsable du plan de sondage considère donc les N=40000 cartes comme une base de sondage composée de M=4000 grappes de N_0 unités chacune. On sélectionne selon un procédé de tirage systématique, assimilable à un tirage à probabilités égales, 80 grappes au sein des 4000.
Soit T_i le nombre total d’abonnés de la grappe i propriétaires de leur résidence principale ($0 \le Y_i \le 10$). On a les résultats suivants :

$$\sum_{i=1}^{80} T_i = 370 \qquad \text{et} \qquad \sum_{i=1}^{80} T_i^2 = 2536$$

1- Trouver un intervalle de confiance à 95% pour le nombre total de propriétaires parmi la population des abonnés au journal.