Université Paul Valéry
Arts - Lettres - Langues
Sciences humaines & sociales

Cours 6
Étude des fluctuations d'échantillonnage par simulation

Indépendance et fluctuation d'échantillonnage

1 Dans le cours précédent (§14), nous avons dit que l'observation d'un échantillon ne produisait jamais l'égalité stricte des distributions conditionnelles, même lorsque les variables sont indépendantes dans la population (D n'est jamais égale à $\tilde{D}$, la distance $\chi^2(D)$ n'est jamais nulle) ; cela est dû au phénomène des fluctuations d'échantillonnage, les variations qu'on remarque quand on fait la même observation sur différents échantillons.

Ce phénomène inévitable rend impossible l'observation de l'indépendance à partir de la définition ; alors, plutôt que de tester l'égalité stricte de D et $\tilde{D}$, nous allons nous fonder sur leur proximité : l'idée est que si la différence observée entre D et $\tilde{D}$ est *conforme* aux fluctuations d'échantillonnage, celles-ci expliquent cette différence et nous permettent de conclure à l'indépendance des variables, et que si elle ne l'est pas, si la différence ne semble pas s'expliquer par les fluctuations d'échantillonnage, c'est qu'elle est probablement due à une liaison entre les variables.

Toute la question revient ainsi à préciser ce qu'on entend par « différence entre D et $\tilde{D}$ *conforme* aux fluctuations d'échantillonnage ».

Pour cela nous allons étudier ces variations à partir d'observations que nous allons *simuler* ; nous commencerons par décrire une méthode pour simuler la mesure d'une variable X de distribution connue sur un individu quelconque de la population ; nous appliquerons ensuite cette méthode pour simuler la mesure de cette variable sur des échantillons de taille n et observer les variations ; enfin, nous la généraliserons pour simuler l'observation conjointe de deux variables supposée indépendantes et de distribution connue, et étudier les différences entre D et $\tilde{D}$.

Simulation de la mesure de X sur un individu

2 déf Soit X une variable de distribution D sur la population P. Simuler la mesure de X sur un individu quelconque de P consiste à sélectionner une modalité de X :

- de manière aléatoire,
- les chances de sortie des modalités étant leur fréquence dans la population [1].

La première condition traduit le terme « quelconque » : la procédure de simulation ne doit pas permettre à l'expérimentateur de déterminer *à l'avance* la modalité qui sera sélectionnée, même s'il connaît celles qui peuvent l'être.

La seconde condition rend compte de l'importance relative des modalités dans la population : plus une modalité est fréquente, plus cette modalité doit avoir de chance d'être sélectionnée ; en d'autres termes, il faut qu'en répétant un grand nombre de fois la simulation, la fréquence de sortie d'une modalité se rapproche de sa fréquence dans P.

Le principe de la procédure de simulation retenue ici est simple : mettre dans une urne des boules dont les couleurs représentent les modalités, dans les mêmes proportions que dans la population, puis tirer une boule au hasard : la modalité sélectionnée est celle qui est représentée par la couleur. Cette procédure assure évidemment les deux conditions. Comme en pratique nous allons utiliser une urne imaginaire, nous devons rendre réaliste l'action de tirer une boule au hasard : il suffit de numéroter les boules de 1 à t (en supposant qu'il y a t boules), et

1. Dans tout ce cours, on suppose le nombre de modalités fini.

de tirer un de ces numéros au hasard, ce qui est relativement facile avec des dés, une table de nombres au hasard ou une calculatrice pourvue de la fonction Hasard (Rand, Random). En inscrivant directement (par l'imagination) les modalités sur les boules, plutôt que d'en passer par l'intermédiaire d'une couleur, on peut décrire le procédé de simulation de la manière suivante :

1. composer une urne de simulation U_X composée de t boules imaginaires numérotés, portant les modalités de X dans les mêmes proportions que dans P,
2. tirer au hasard un de ces numéros, et noter la modalité de la boule qui porte ce numéro.

3 Composition de l'urne de simulation U_X. On note D la distribution de X dans la population P, et f_i ses fréquences ; composer l'urne de simulation U_X consiste à déterminer le nombre de boules portant chaque modalité, t_1 pour la première modalité, t_2 pour la seconde modalité, ..., t_k pour la dernière, puis à numéroter les boules ; la somme des nombres t_1, ..., t_k doit être égale à la taille t de U ; d'autre part, comme les distributions de X doivent être identiques dans U et P, la fréquence $\frac{t_i}{t}$ de la modalité m_i dans U doit être égale à sa fréquence f_i dans P ; cette égalité $\frac{t_i}{t} = f_i$ donne $t_i = t * f_i$; et pour que ce produit soit un entier il faut choisir t assez grand :

- si les proportions f_i ont une seule décimale (exprimées en pourcentage ce sont des dizaines), il faut prendre $t = 10$ (ou un multiple de 10) puisque $10 * f_i$ est alors un entier ;
- si les proportions f_i ont 2 décimales (les pourcentages sont des entiers), il faut prendre $t = 100$ (ou un multiple de 100) puisque $100 * f_i$ est alors un entier ;
- si les proportions f_i ont 3 décimales (les pourcentages ont une décimale), il faut prendre $t = 1000$ (ou un multiple de 1000) puisque $1000 * f_i$ est alors un entier ;
- et ainsi de suite.

Pour résumer, on choisit d'abord $t = 10^p$, p étant le nombre de décimales des proportions f_i (si $p = 1$ $t = 10$, si $p = 2$ $t = 100$, etc.) ; puis on calcule les nombres $t_i = f_i * t$ qui sont ainsi des entiers ; enfin on numérote les t boules, modalité par modalité : les t_1 premières sont numérotées de 1 à t_1 et portent la modalité m_1 ; les t_2 suivantes numérotées de $t_1 + 1$ à $t_1 + t_2$ portent la modalité m_2, et ainsi de suite jusqu'aux t_k dernières, numérotées de $t_1 + ... + t_{k-1} + 1$ à $t_1 + t_2 + ... + t_k = t$ qui portent la modalité m_k. Cette construction garantit l'égalité des distributions de X dans U et P : m_i a les mêmes chances de sortir de U que d'être sélectionnée dans P : t_i chances sur t pour U identiques aux $100 * f_i$ chances sur 100 pour P.

4 Exemple. Composons l'urne de simulation U_X pour la variable X de distribution D dans P :

X	m_1	m_2	m_3
%	55,6	22,2	22,2

Comme ces pourcentages ont une seule décimale, les fréquences en proportion en ont 3 (55,6% équivaut à la proportion 0,556) ; on choisit donc $10^3 = 1000$ boules pour l'urne U_X ; on en déduit $t_1 = 0{,}556 * 1000 = 556$, $t_2 = 222$ et $t_3 = 222$; les 556 boules numérotées de 1 à 556 portent la modalité m_1 (leur proportion 556/1000 est bien égale à $f_1 = 0{,}556$), les 222 boules suivantes, numérotées de 557 à 556+222=778, portent la modalité m_2, et les 222 dernières, numérotées de 779 à 778+222=1000, la modalité m_3 :

U	m_1	m_2	m_3	total
t_i	556	222	222	1000
n°	1-556	557-778	779-1000	
%	55,6	22,2	22,2	100

On vérifie bien que la distribution de X dans U est la distribution D de X dans P.

5 Tirage d'un numéro au hasard. Il consiste à sélectionner un nombre compris entre 1 et t :

- de manière aléatoire,
- de sorte que tous les nombres ont la même chance d'être sélectionné.

Plusieurs procédures sont possibles :

a On peut lancer p fois un dé non pipé à 10 faces numérotées de 0 à 9 ; la suite des chiffres donne un nombre compris entre 0 et $10^p - 1$, auquel on ajoute donc 1 pour aller de 1 à 10^p.

b On peut utiliser une table de nombre au hasard ; c'est un tableau de chiffres qui permet de simuler le lancer d'un ou plusieurs dés à 10 faces : à partir d'une cellule initiale et dans une direction est, sud, ouest, nord, nord-est etc. on lit autant de chiffres successifs qu'il y a de dés ; pour le tirage suivant, on lit les chiffres suivants, dans la même direction. Par exemple, dans la table donnée en annexe, à partir de la cellule $c_{11,26}$ et dans la direction *est* (gauche-droite) on lit la suite des nombres à 3 chiffres (p=3) 404 907 778 768 545 027 957 267 676 926 108 etc. ; il faut ici aussi ajouter 1 pour obtenir des nombres compris entre 1 et 1000.

c On peut utiliser la fonction random (ou l'équivalent) d'une calculatrice ; cette fonction s'amorce avec un nombre-graine et donne des nombres pseudo-aléatoires (à partir d'une « graine » donnée, on obtient toujours la même suite) ; elle donne un nombre décimal compris entre 0 et 1 exclu : on prend les p premières décimales qui donnent un nombre compris entre 0 et $10^p - 1$, auquel on ajoute 1 pour obtenir un nombre compris entre 1 à $10^p = t$.

d Enfin, on peut utiliser un site web prévu à cet effet : http://www.math-info.univ-paris5.fr/\verb+~+smel/lexique/generateur/generateur.html donne une série de nombres au hasard compris entre 0 et 1 exclu ; http://www.randomnumbergenerator.com/ les fournit un à un, en cliquant sur « More random numbers » après avoir éventuellement paramétré (« Customize... ») le générateur.

Pour l'exemple précédent, supposons que le procédé c donne le nombre pseudo-aléatoire 0,58620... ; on retient donc 586 auquel on ajoute 1 : la boule 587 portant la modalité m_2, celle-ci est la mesure simulée de X sur un individu quelconque de P.

Simulation de la mesure de X sur un échantillon.

6 Pour simuler la mesure de X sur un échantillon aléatoire de P, il suffit de répéter n fois la procédure précédente :

1. on compose une urne de simulation U_X,
2. on tire une série de n numéros au hasard compris entre 1 et t, en notant à chaque fois la modalité de la boule portant le numéro.

7 **Exemple** : simulation de l'observation de la variable Y « absentéisme » dans un échantillon de taille 27. On suppose que la distribution de l'absentéisme dans la population des élèves est identique à la distribution de l'absentéisme dans l'échantillon, c'est à dire à la distribution marginale ; comme cette distribution en fréquence est celle de l'exemple précédent, l'urne de simulation U_Y est la même ; on tire 27 nombres au hasard entre 1 et 1000 en notant la modalité associée ; en utilisant la table donnée donnée en annexe, et en choisissant la direction *est* à partir de la cellule (11,26), on trouve (la modalité associée est entre parenthèses) : 404+1(R), 907+1(F), 778+1(F), 769(M), 546(R), 28(R), 958(F), 268(R), 677(M), 927(F), 109(R), 664(M), 974(F), 451(R), 513(R), 42(R), 182(R), 781(F), 128(R), 871(F), 587(M), 195(R), 284(R), 716(M), 454(R), 507(R), 613(M). Ce qui donne la distribution suivante :

Y	Rare	Moyen	Fréquent
Effectif	14	6	7
%	51,9	22,2	25,9

Cette distribution est un peu différente de la distribution dans la population en raison des fluctuations d'échantillonnage ; celles-ci ont un effet d'autant plus grand que la taille de l'échantillon est petite ; une simulation (par ordinateur) sur un échantillon de taille 2000 donne par exemple la distribution suivante, globalement plus proche de la distribution dans P :

Y	Rare	Moyen	Fréquent
Effectif	1100	462	438
%	55,0	23,1	21,9

Simulation d'une observation conjointe de deux variables indépendantes.

8 Pour simuler l'observation conjointe de X et Y sur un échantillon, on commence par simuler la mesure x_e de X à partir de la distribution D de X dans P ; il faudrait ensuite utiliser la distribution conditionnelle Y_{x_e} pour simuler y_e. Dans le cas présent où X et Y sont supposées indépendantes, les distributions conditionnelles de Y sont toutes égales à la distribution D' de Y dans la population P : pour simuler l'observation conjointe de X et Y sur un individu e, il suffit donc de simuler x_e à partir de D puis de simuler y_e à partir de D' ; cela nécessite évidemment de construire deux urnes de simulation, U_X et U'_Y.

9 **Exemple « Niveau scolaire et absentéisme ».** Supposons la distribution de X dans la population égale en fréquence à (le choix de pourcentage sans décimale est volontaire) :

X	A	B
%	56	44
proportion	0,56	0,44

Les proportion ayant 2 décimales, il suffit de prendre 100 boules (10^2) pour construire U_X :

U_X	A	B	total
t_i	56	44	100
n°	1-56	57-100	
%	56	44	100

Pour Y, nous reprenons l'urne de simulation construite au §4 puisque la marge de Y est la distribution de l'exemple :

U'_Y	Rare	Moyen	Fréquent	total
t'_i	556	222	222	1000
n°	1-556	557-778	779-1000	
%	55,6	22,2	22,2	100

Pour la mesure conjointe d'un premier individu, on tire au hasard un nombre compris entre 1 et 100 pour X, puis un nombre compris entre 1 et 1000 pour Y ; si les nombres sont 97 et 374 on attribuera à l'individu la modalité conjointe (B,Rare) et il participera à l'effectif n_{21} ; pour le second individu, on tire une seconde paire, 29 et 32, ce qui lui attribue la modalité conjointe (A,Rare) ; et ainsi de suite : la série de 27 paires 97 374, 29 32, 2 781, 59 193, 95 60, 57 785, 16 451, 55 236, 25 447, 11 87, 38 560, 58 787, 71 554, 35 485, 22 587, 96 641, 7 621, 49 616, 60 993, 54 436, 21 177, 25 230, 12 990, 11 720, 21 767, 96 118 et 96 720 donne la distribution suivante, à comparer avec la distribution théorique calculée dans le cours précédent :

X / Y	Rare (1 à 556)	Moyen (557 à 778)	Fréq. (779 à 1000)	
A (1 à 56)	9	6	2	17
B (57 à 100)	5	2	3	10
	14	8	5	27

Simulation du χ^2 sous hypothèse d'indépendance

10 La distribution conjointe précédente est une simulation d'une observation conjointe de deux variables indépendantes sur un échantillon de taille 27 ; en théorie le χ^2 de cette distribution devrait être nul, mais il est en réalité un peu supérieur en raison des fluctuations d'échantillonnage : $\chi^2 = 1{,}64$.

11 On peut étudier expérimentalement la variation du χ^2 sous hypothèse d'indépendance en répétant un grand nombre de fois la simulation précédente, 1000 par exemple, et en notant la série des 1000 χ^2 obtenus ; la situation statistique est la suivante :

- la population P' est l'ensemble de tous les échantillons de taille 27 imaginables de la population P des élèves (un individu de P' est un échantillon de taille 27) ;
- l'échantillon est le sous-ensemble des 1000 individu-échantillons (de taille 27) simulés ;
- la variable Z est le χ^2 (on mesure chaque individu-échantillon par son χ^2), qui est une variable continue ;
- les modalités sont (par exemple) des intervalles de largeur 2 : [0 2[, [2 4[, [4 6[etc. ;
- pour chacun des 1000 échantillons, on simule l'observation conjointe du niveau et de l'absentéisme en les supposants indépendants, et en note le χ^2 de la distribution conjointe obtenue ($z_e = \chi^2(D)$.

On obtiendrait alors une distribution de Z de la forme :

Z	[0 2[	[2 4[	[4 6[	[6 8[	[8 10[	[10 12[	[12 14[	[14 16[	total
n_i	600	243	109	34	8	3	3	0	1000
f_i en %	60,0	24,3	10,9	3,4	0,8	0,3	0,3	0,0	100

Cette distribution montre que pour 84% des échantillons sur lesquels on a fait l'observation conjointes de X et Y en les supposant indépendantes, les fluctuations d'échantillonnage conduisent à un χ^2 inférieur à 4 (en théorie ils devraient tous être nuls), et dans 0,6% seulement des cas à un χ^2 supérieur à 10. Supposons alors que dans une autre population d'élèves pour laquelle on ignore si X et Y sont indépendantes, une observation conjointe sur un échantillon de taille 27 donne un χ^2 égal à 18,2 : on pourrait à juste titre en conclure que, puisque cette valeur 18,2 est très peu probable si X et Y étaient indépendantes, les variables ne semblent pas l'être dans ce cas ; si on avait trouvé un χ^2 égal à 3,8 on aurait pu conclure que tout se passe comme si X et Y étaient indépendantes. Cette démarche est celle du test du χ^2 qui est l'objet du cours suivant.

Questions de cours

1. Qu'est-ce qu'on peut appeler fluctuations d'échantillonnage ?
2. Quelles sont les deux conditions que doit respecter la simulation d'une mesure de X de distribution D sur un individu ?
3. Qu'est-ce que signifie tirer au hasard un nombre compris entre 1 et 100 ?
4. Quelle est la taille minimale d'une urne de simulation quand les fréquences en pourcentage sont indiquées avec une décimale ?
5. Quelle sera l'effectif de la modalité m_i dans l'urne de simulation si sa fréquence dans P est 12,7% ?
6. Dans l'exemple 4, quelle est la modalité sélectionnée si le nombre tiré est 900 ? 555 ? 778 ? 779 ?
7. Combien faut-il construire d'urne de simulation pour simuler l'observation conjointe de deux variables indépendantes ?
8. Comment simule-t-on l'observation conjointe de deux variables indépendantes ?
9. Dans la situation exposée à la fin du §11, quelle conclusion peut-on tirer si le χ^2 de l'échantillon est 10,5 ? 9,1 ? 6,3 ? 4,7 ? 2,4 ? 1,64 ?

Annexe : table de chiffres au hasard

Cette table a 40 lignes et 50 colonnes regroupées 5 par 5 pour faciliter la lecture. On choisit *au hasard* une cellule initiale et une direction (est, sud, ouest, nord, nord-est, etc.) et on lit autant de chiffres qu'il est nécessaire.

83760	31255	71609	89887	00940	54355	44351	89781	58054	65813
66280	56046	50526	33649	87067	02697	06577	16707	96368	47678
70218	28376	98535	34190	96911	81578	97312	20500	48030	27256
02349	88955	52760	73696	91510	38633	38883	90419	26716	98215
93606	21415	34843	12969	84847	06280	95916	12991	08262	58385
24274	18747	37327	06780	08032	98544	24902	81607	87914	22721
67778	70496	57588	89813	71211	83848	93494	27946	79722	70315
89134	06458	40897	73025	04191	77144	49340	89446	71852	80854
83625	00097	71092	12009	63223	37993	50067	25688	98179	34628
03324	68196	72460	55616	27006	50790	28629	88726	97143	63218
84392	36623	91964	03505	46525	40490	77787	68545	02795	72676
76926	10866	39734	50512	04181	78012	78705	86194	28371	54535
06612	60200	49085	85108	71438	10099	99027	65081	82492	77584
76721	02889	95600	07984	31925	59685	91510	40039	43205	37149
64599	51953	55612	89088	58436	21501	86219	74528	59805	65020
79440	99677	49530	55291	34867	54774	52449	23294	94815	95124
35839	00177	57742	09502	42624	29017	94284	81409	36904	54329
83013	94568	75490	12138	24067	86954	00910	61171	82982	87191
19980	47085	46064	19102	26297	79745	99611	04555	52501	32088
55716	10350	67645	62922	81919	47925	91448	36025	20611	38939
36624	03992	27656	33092	22252	54461	83386	55340	11313	23290
50678	33814	07643	81452	60689	48745	49894	27285	90420	31188
17932	27351	34623	55864	58659	06992	88558	45742	56792	71027
76795	23022	20409	60100	59507	40596	16971	96490	47676	49129
20654	64916	59927	62495	81133	29095	64024	02792	39809	85302
73601	60099	50404	41700	53664	54397	49600	46980	13882	54275
59678	14528	96293	12957	68229	95753	15727	75113	09892	71487
92132	51012	09399	30175	73025	99849	34334	20089	19323	95149
76143	16802	32819	34057	94227	25779	93959	89810	47627	70561
99617	64239	13967	90188	60291	38478	09723	10697	78020	51388
02841	25077	02368	75931	42679	70900	33040	08871	46696	18647
57979	28621	03155	03704	98473	25894	26753	62390	54746	84189
41233	68027	17036	28310	50551	84295	80793	93235	78902	18351
48049	09367	15040	29166	64290	16439	67192	16681	46304	68190
10984	97394	23070	90585	53139	96998	39834	27678	42288	33778
59531	76937	15645	70938	00036	72773	25984	06507	27933	46779
36874	61476	74611	74476	48713	36124	98549	70465	58742	28707
49377	53222	14506	80260	59070	47101	02248	99520	08803	79772
59707	00510	29216	53012	47115	39798	79797	06491	72669	05055
63469	49151	35960	88792	43961	62352	78114	77810	95638	84227