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

Cours 6
Simulation et fluctuation d'échantillonnage

Dans le cours précédent, nous avons dit que les distributions conditionnelles n'étaient jamais identiques, même lorsque les variables sont indépendantes ; cela est dû au phénomène des `fluctuations d'échantillonnage`, les variations qu'on remarque quand on fait la même observation sur différents échantillons, par exemple de deux variables conjointes.

C'est le même phénomène qui explique que le χ^2 d'une distribution n'est jamais nul, même lorsque les variables sont indépendantes ; pour comprendre comment une valeur de ce χ^2 peut être attribuée à une fluctuation d'échantillonnage plutôt qu'à une liaison entre les deux variables, nous allons mettre en évidence ce phénomène en simulant l'observation d'une variable, puis de deux variables conjointes indépendantes sur plusieurs échantillons ; les simulations supposent que les distributions de ces variables dans la population soient connues, ou estimées par les distributions marginales d'une observation conjointe effective.

Nous commencerons par décrire une méthode pour simuler la mesure sur un individu d'une variable X dont la distribution sur la population est connue.

Simulation de la mesure de X

1 **Simuler la mesure de X de distribution D sur un individu quelconque** consiste à sélectionner une modalité de X :
- de manière aléatoire,
- conformément à la distribution D.

La première condition revient à dire qu'on connaît à l'avance les modalités qui pourront être sélectionnées, et qu'on doit être incapable de déterminer *a priori* laquelle le sera.

La seconde condition signifie que la chance de sortie d'une modalité doit être proportionnelle à sa fréquence dans D : si, par exemple, la fréquence de m_1 est trois fois plus grande que celle de m_2 il faut que m_1 ait trois fois plus de chance d'être sélectionnée que m_2 ; en d'autres termes, il faut qu'en répétant un grand nombre de fois cette procédure, la fréquence de sélection d'une modalité soit égale (ou à peu près) à sa fréquence dans D.

Le procédé de simulation que nous utiliserons par la suite consiste à :

1. construire une `population de simulation` P_D composée de t individus imaginaires numérotés, ayant les modalités de X dans les mêmes proportions que D,
2. tirer au hasard un de ces numéros, et noter la modalité de l'individu qu'il identifie.

2 **Construction de la population de simulation P_D.** Il faut déterminer l'effectif de chaque modalité dans P_D ; ces nombres $t_1, t_2, ...,$ et t_k sont des entiers de somme t ; ils doivent vérifier $\frac{t_i}{t} = f_i$ pour que la distribution de X sur P_D soit égale à D, f_i étant la fréquence de m_i dans D ; pour que $t_i = f_i * t$ soit un entier, il faut donc choisir t assez grand : si p est le nombre de décimales des f_i, il faut que t soit au moins égal à 10^p :
- $p = 1$ (les proportions ont une seule décimale et les pourcentages sont des dizaines) : il faut au moins 10 individus dans la population de simulation, puisque $10 * f_i$ est un entier ;
- $p = 2$ (les proportions ont deux décimales et les pourcentages sont des entiers) : il en faut au moins 100, puisque $100 * f_i$ est un entier ;
- $p = 3$ (les proportions ont trois décimales et les pourcentages ont une décimale) : il en faut au moins 1000 ; et ainsi de suite.

Une fois t déterminé, on calcule les nombres $t_i = f_i * t$ qui sont ainsi des entiers, et, en les numérotant, on répartit les t individus dans les sous-échantillons induits : les t_1 premiers, numérotés de 1 à t_1 auront la modalité m_1 ; les t_2 suivants numérotés de t_1+1 à $t1+t2$ la modalité m_2, et ainsi de suite jusqu'aux t_k derniers numérotés de $t_1 + ... + t_{k-1} + 1$ à $t_1 + t_2 + ... + t_k = t$ qui seront supposés avoir la modalité m_k.

Cette construction garantit que la distribution de X sur P_D est exactement la distribution D.

3 Tirage d'un numéro au hasard. Il s'agit de tirer au hasard un nombre compris entre 1 et $t = 10^p$, c'est à dire de sélectionner un de ces nombres :
- de manière aléatoire,
- conformément à la distribution uniforme, les nombres ayant tous 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 donnera un nombre compris entre 0 et $10^p - 1$, auquel on ajoutera 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, verticale, horizontale, ou diagonale, 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.

c On peut utiliser la fonction random (ou l'équivalent) d'un calculateur ; 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) ; cette fonction donnant généralement un nombre décimal compris entre 0 et 1 exclu, on prendra les p premières décimales auquel on ajoutera 1.

d Enfin, on peut utiliser un site web prévu à cet effet : http://www.math-info.univ-paris5.fr/~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.

4 Exemple : simulation de la mesure de X de distribution :

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

1. Comme les proportions ont 3 décimales (un pourcentage à 1 décimale équivaut à une proportion à 3 décimales), on construit une population de simulation P_{S_X} de 1000 individus ; les 556 (0,556*1000) individus numérotés de 1 à 556 ont la modalité m_1 (leur proportion 556/1000 est bien égale à $f_1 = 0{,}556$), les 222 individus suivants, numérotés de 557 à 556+222=778, la modalité m_2, et les 222 derniers, numérotés de 779 à 778+222=1000, la modalité m_3 ; par construction, la proportion de X dans cette population est identique à D ;
2. on tire ensuite un nombre au hasard entre 1 et 1000, en utilisant par exemple le second site : 587 ; comme l'individu 587 possède la modalité m_2, c'est cette modalité qui est sélectionnée par la simulation.

Simulation de la mesure de X sur un échantillon.

5 Pour mesurer X sur un échantillon de taille n d'une population P , on commence en principe par sélectionner n individus par une méthode d'échantillonnage, puis on prend la mesure de X sur chacun de ces individus.

La méthode que nous retiendrons ici consiste à répéter n fois le procédé qu'on vient de décrire ; cela revient à choisir n individus de P au hasard, en supposant la taille de P[1] suffisamment importante pour que la distribution reste constante pendant la simulation : supposons en effet une modalité m_1 de fréquence 0,4 : si le premier individu choisi est de modalité m_1, elle diminue de 17% dans une population de taille 10, en passant à $3/9 = 0{,}333$, et seulement de 0,025% dans une population de 10000 ; mais pour l'usage que nous allons faire de la simulation, ces restrictions sont sans conséquence.

6 **En pratique,** pour simuler la mesure de X sur un échantillon de taille n, dans une population nombreuse P où X est de distribution D :

1. on construit une population de simulation P_D,
2. puis on tire une série de n numéros au hasard compris entre 1 et t, en notant à chaque fois la modalité de l'individu identifié.

7 **Exemple** : simulation de l'observation de la variable « absentéisme » dans un échantillon de taille 27. On suppose que la distribution dans la population des élèves est identique à la distribution marginale ; comme cette distribution en fréquence est la distribution de l'exemple précédent, la population de simulation est la même ; on tire 27 nombres au hasard entre 1 et 1000 en notant la modalité associée, ce qui donne par exemple (la modalité associée est entre parenthèse) : 911(3), 197(1), 335(1), 702(2), 277(1), 553(1), 477(1), 628(2), 364(1), 513(1), 952(3), 916(3), 637(2), 717(2), 141(1), 606(2), 242(1), 137(1), 804(3), 156(1), 400(1), 129(1), 108(1), 998(3), 218(1), 512(1), 839(3). Ce qui donne la distribution suivante :

Y	Rare	Moyen	Fréquent
Effectif	16	5	6
Proportion	0,593	0,185	0,222

Cette distribution est un peu différente de la distribution marginale du tableau de contingence, en raison des fluctuations d'échantillonnage ; elles sont causées ici par la composition aléatoire de l'échantillon et amplifiées par sa taille relativement réduite : les variations se réduisent en effet quand la taille des échantillons simulés augmente.

Simulation d'une observation conjointe de deux variables indépendantes.

8 Quand X et Y sont indépendantes, les distributions de X et Y sont les mêmes pour tous les individus (les distributions conditionnelles sont égales en fréquence), égales aux distributions de X et Y dans la population P ; pour simuler la mesure conjointe d'un individu, on peut donc simuler indépendamment la mesure de X et la mesure de Y, la première simulation sélectionnant une modalité m_i à partir d'une population de simulation P_{D_X}, la seconde une modalité m'_j à partir d'une autre population de simulation.

9 **La simulation d'une observation conjointe D sous hypothèse d'indépendance** peut se faire de cette manière, en prenant comme distribution de X et Y sur la population P les distributions marginales de D ; cette simulation de la distribution théorique d'indépendance $\tilde{D}$ permet d'étudier de quelle manière les fluctuations d'échantillonnage font varier le χ^2 de deux variables indépendantes, alors qu'en théorie il devrait être nul ; l'idée est de comparer le χ^2 de D à ces valeurs simulées : s'il peut être considéré comme une de ces valeurs, l'hypothèse de l'indépendance de X et Y sera plausible, dans le cas contraire non.

[1] À na pas confondre ici avec la population de simulation imaginaire P_D, qui sert uniquement à simuler la mesure d'une variable suivant une distribution donnée.

10 **Exemple « Niveau scolaire et absentéisme ».** Nous avons vu comment simuler l'absentéisme à l'aide d'une population de simulation de taille 1000.

Pour simuler le niveau scolaire, on suppose que la distribution marginale de X est la distribution dans la population, et on prend volontairement pour l'exemple une proportion à deux décimales : 0.56 pour A et 0.44 pour B ; on construit donc une population de simulation de taille 100, les 56 premiers individus imaginaires ayant le niveau A et les 44 derniers, numérotés de 57 à 100, le niveau B.

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 d'une statistique d'un échantillon.

11 En simulant plusieurs observations d'une variable sur un échantillon, on peut observer comment les fluctuations d'échantillonnage font varier une statistique comme la moyenne. Prenons l'exemple de la variable X de la situation « Revenus et situation géographique » (exemple 5 du premier cours) dont la distribution dans la population est estimée par la distribution marginale, de moyenne 11,7 :

X	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 30]	Total	Moyenne
Effectifs	72	120	108	100	400	11,7
%	18	30	27	25	100	

On construit la population de simulation avec 100 individus numérotés ; le revenu des 18 premiers (1 à 18) sera supposé être dans la première modalité (entre 0 et 5), celui des 30 suivants (19 à 48) entre 5 et 10, celui des 27 suivants (49 à 75) entre 10 et 15, et celui des 25 derniers (76 à 100) entre 15 et 30.

Voilà quelques simulations avec les moyennes associées, effectuées par un calculateur :

X	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 30]	Total	Moyenne
Simul-1	76	120	108	96	400	11,5
Simul-2	71	103	110	116	400	12,3
Simul-3	75	129	103	93	400	11,3
Simul-4	78	105	110	107	400	11,9
Simul-5	66	131	101	102	400	11,8

On peut observer que la moyenne de X varie autour de la moyenne dans la population ; ces différences sont également l'effet des fluctuations d'échantillonnage, et elles seraient d'autant plus réduites (les moyennes simulées seraient d'autant plus proches de la moyenne dans la population) que les tailles des échantillons seraient grandes .

D'une façon analogue on peut observer la variation du χ^2 en simulant de nombreuses fois une observation conjointe sous hypothèse d'indépendance ; c'est ce que nous verrons dans le prochain cours ; en attendant, remarquons que la simulation effectuée dans l'exemple du Niveau scolaire et absentéisme donne un χ^2 égal à 1,63 : puisque les variables sont supposées indépendantes dans cette simulation, on doit attribuer cette valeur aux fluctuations d'échantillonnage et aux erreurs d'arrondis ; que faut-il alors penser du χ^2 égal 1,05 de la situation effective, calculé dans le cours précédent ? Nous verrons dans le prochain cours comment un test permet de répondre à cette question..

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 population de simulation de D quand les fréquences en pourcentage sont indiquées avec une décimale ?
5. Quelle sera l'effectif de la modalité m_i dans la population de simulation si sa fréquence dans D 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 de population de simulation pour simuler l'observation conjointe d'une distribution théorique d'indépendance ?
8. Comment simule-t-on l'observation conjointe d'une distribution théorique d'indépendance ?
9. Autour de quelle valeur fluctue la moyenne d'un échantillon simulé à partir d'une distribution D ?