Eléments de statistique

Répétition 2

16 octobre 2018

Notion d'échantillon i.i.d.

Soit $\mathcal{X}$ un caractère (ex : taille, poids, couleur des yeux, ...) étudié dans une population de taille N. Un échantillon i.i.d. est un ensemble de n individus tirés à partir de la population, de manière à ce que les n valeurs observées soient des réalisations de n variables aléatoires, indépendantes et identiquement distribuées selon la même loi que la variable $\mathcal{X}$.

Exercice

2.1. Un bowling met à disposition des clients cinq boules respectivement numérotées 6, 8, 10, 12 et 15. Les numéros indiqués sur les boules de bowling représentent leurs poids $\mathcal{X}$ en livres. On réalise des échantillons de 2 boules par tirage i.i.d. dans cet ensemble muni d'une loi uniforme (donc avec remise, en supposant qu'à chaque tirage chaque boule a une probabilité de 1/5 d'être choisie).

a) Ecrire tous les échantillons possibles.

b) Ecrire la moyenne m_x du poids x des boules de chaque échantillon.

c) Calculer l'espérance et la variance de la variable m_x.

d) Calculer la moyenne $\mu_{\mathcal{X}}$ et l'écart-type $\sigma_{\mathcal{X}}$ de la population, puis vérifier les formules $E\{m_x\} = \mu_{\mathcal{X}}$ et $V\{m_x\} = \frac{\sigma^2_{\mathcal{X}}}{n}$

e) Les m_x auraient-ils été plus proches de la moyenne de la population si on avait pris des échantillons de taille 3 ?

f) Calculez s^2 pour chaque échantillon et vérifiez que $E\{s^2\} \neq \sigma^2_{\mathcal{X}}$

g) Vérifiez que $E\{s^2_{n-1}\} = \sigma^2_{\mathcal{X}}$

2.2. Le poids de 10000 étudiants d'une université suit une loi de moyenne $\mu_{\mathcal{X}} = 68kg$ et d'écart-type $\sigma_{\mathcal{X}} = 3kg$.

a) Sous l'hypothèse d'une loi normale pour la variable $\mathcal{X}$, déterminez $P(66,5 \leq$ poids étudiant $\omega < 69,5)$.

b) On extrait de façon indépendante 80 échantillons i.i.d. de 25 étudiants chacun.

i. Sous l'hypothèse d'une loi normale pour la variable $\mathcal{X}$, pour combien d'échantillons peut-on s'attendre à trouver une moyenne d'échantillon m_x comprise entre 66,5 et 69,5 kg ?

ii. Sans faire d'hypothèse sur la distribution de la variable $\mathcal{X}$, pour combien d'échantillons peut-on s'attendre à trouver une moyenne comprise entre 66,5 et 69,5 kg ?

iii. Que peut-on en conclure ?

2.3. Une élection entre un candidat A et un candidat B va avoir lieu dans un pays dont le corps électoral est constitué de 35 millions de personnes. On sait que 17 millions de personnes soutiennent le candidat A, tandis que 18 millions soutiennent le candidat B.

a) On construit un échantillon i.i.d. de 1400 personnes, en tirant avec remise et de façon équiprobable dans la population globale.

i. Calculez l'espérance et l'écart-type de la proportion f_A et du nombre n_A de personnes de l'échantillon qui voteraient pour A.

b) Le même pays est "stratifié" en deux parties P_1 et P_2, dont le corps électoral est composé de respectivement 10 millions et 25 millions de personnes. On sait que la proportion p_{A_1} de personnes soutenant A dans la partie P_1 est de $\frac{6}{10}$. On tire un échantillon i.i.d. de 400 personnes habitant dans la partie P_1, et de façon indépendante un autre échantillon i.i.d. de 1000 personnes habitant dans la partie P_2.

i. Déterminez la proportion p_{A_2} de personnes de P_2 qui voteront pour A.

ii. Calculez pour chaque échantillon l'espérance et l'écart-type de la proportion et du nombre de personnes qui voteraient pour A.

iii. Si f_{A_1} et f_{A_2} désignent respectivement les deux fréquences relatives, définissions $f'_A = \alpha f_{A,1} + (1-\alpha) f_{A,2}$ une nouvelle statistique fonction des deux échantillons. Déterminez α pour que $E\{f'_A\} = E\{f_A\}$. Calculer ensuite la variance et l'écart-type de cette statistique f'_A.

iv. Que pouvez vous conclure ?

v. Que peut on affirmer au sujet des distributions d'échantillonnage des deux statistiques f_A et f'_A ?

Exercices suggérés

2.4. On s'intéresse à la population des femmes wallonnes. Une étude précise sur la population entière nous indique que la taille de ces femmes (en nombre fini, bien que très grand) suit de très près une loi normale de moyenne $\mu_{\mathcal{X}} = 160cm$ et d'écart-type $\sigma_{\mathcal{X}} = 20cm$.

Sous cette hypothèse :

a) Déterminez $P(140 \leq$ taille de madame $\omega < 170)$

b) Déterminez $P(140 \leq$ taille moyenne de 25 ω i.i.d. $< 170)$

2.5. Dans un entrepôt, on utilise un élévateur dont la charge utile est de 8200 kg. Il est destiné à charger 25 paquets à la fois. Le poids d'un paquet suit une loi normale dont la moyenne est de 300 kg avec un écart-type de 50 kg.

a) Quelle est la probabilité pour qu'il y ait surcharge ?

b) Si l'élévateur effectue 100 transports par jour, à quel nombre de surcharges peut-on s'attendre par semaine de 5 jours ouvrables ?

2.6. EXAMEN AOUT 2014

Une élection entre trois candidats A, B et C va avoir lieu dans un pays dont le corps électoral est constitué de 61 millions de personnes. On sait que 20 millions de personnes soutiennent le candidat A, 18 millions le candidat B et le reste soutient le candidat C. On construit un échantillon i.i.d. de 2500 personnes, en tirant avec remise et de façon équiprobable dans la population globale.

a) Calculez l'espérance et l'écart-type de la proportion f_C et du nombre n_C de personnes de l'échantillon qui voteraient pour C. (**2 points**)

Le même pays est "stratifié" en deux parties P_1 et P_2, dont le corps électoral est composé de respectivement 11 millions et 50 millions de personnes. On sait que dans la partie P_1, la proportion p_{A_1} de personnes soutenant A est de $\frac{1}{2}$ et la proportion p_{B_1} de personnes soutenant B est de $\frac{3}{10}$. On tire un échantillon i.i.d. de 1000 personnes habitant dans la partie P_1, et de façon indépendante un autre échantillon i.i.d. de 1500 personnes habitant dans la partie P_2.

b) Calculez pour chaque échantillon l'espérance et l'écart-type de la proportion et du nombre de personnes qui voteraient pour C. (**4 points**)

c) Si f_{C_1} et f_{C_2} désignent respectivement les deux fréquences relatives de vote pour le candidat C dans les parties P_1 et P_2, définissions $f'_C = \alpha f_{C_1} + (1 - \alpha) f_{C_2}$ une nouvelle statistique fonction des deux échantillons. Déterminez α pour que $E\{f'_C\} = E\{f_C\}$. (**2 points**)

d) Calculez la variance et l'écart-type de f'_C. (**4 points**)

e) L'écart-type de f'_C est-il plus grand que celui de f_C ? A quoi cela est-il dû ? (**2 points**)

f) Si au lieu de tirer 1000 personnes dans la partie P_1 et 1500 dans la partie P_2, on en avait tiré respectivement 451 et 2049, aurions-nous forcément obtenu un écart-type de f'_C plus petit que celui de f_C ? Justifier. (**2 points**)

g) En préservant un nombre total de 2500 personnes, y-a-t-il moyen d'obtenir une plus grande réduction de variance qu'en tirant 451 personnes dans la partie P_1 et 2049 personnes dans la partie P_2 ? Justifier. (**2 points**)

h) Que peut on affirmer au sujet des distributions d'échantillonnage des deux statistiques f_C et f'_C ? (**2 points**)

2.7. EXAMEN JANVIER 2015

La quantité de cannabis absorbée mensuellement par un consommateur régulier suit une loi de moyenne $\mu_{\mathcal{X}} = 149g$ et d'écart-type $\sigma_{\mathcal{X}} = 40,3g$. On tire un échantillon i.i.d. de 15 consommateurs réguliers de cannabis.

a) Sous l'hypothèse d'une loi normale pour la variable $\mathcal{X}$, estimer la probabilité que la moyenne de l'échantillon soit comprise entre 134g et 164g ? (**4 points**)

b) Sans faire d'hypothèse sur la distribution de la variable $\mathcal{X}$, que peut-on dire sur la probabilité que la moyenne de l'échantillon soit comprise entre 134g et 164g ? (**4 points**)

c) La probabilité calculée au point b) devait-elle forcément être plus petite que celle calculée au point a) ? (**2 points**)

Supposons désormais que les consommateurs achètent des paquets de 25g et que le nombre théorique de sachets achetés suit la loi indiquée à la table 1. On tire toujours un échantillon i.i.d. de 15 consommateurs réguliers de cannabis.

TABLE 1 – Quantité achetée mensuellement par un consommateur de cannabis.

Nombre de sachets de 25g	Probabilité
3	5%
4	15%
5	20%
6	25%
7	18%
8	10%
9	5%
10	2%

d) Que valent l'espérance et la variance de la moyenne d'échantillon ? (**2 points**)

e) Que valent les espérances de s_x^2 et de s_{n-1}^2 ? (**2 points**)

f) Peut-on utiliser les formules du formulaire pour calculer les variances de s_x^2 et de s_{n-1}^2 ? Justifier. (**1 point**)

g) Déterminez le rapport entre les variances de s_x^2 et de s_{n-1}^2. (**1 point**)

h) Quel est l'intérêt de corriger la variance ? Quel est le désavantage ? (**2 points**) Suggestion : repartez de vos réponses aux sous-questions e) et g).

i) Si vous deviez calculer la probabilité que la moyenne de l'échantillon soit comprise entre 134g et 164g dans les conditions qui nous occupent, pensez-vous qu'elle serait forcément plus grande que celle calculée au point b) ? (**2 points**)

2.8. EXAMEN AOUT 2015

La quantité de bière absorbée par un supporter du standard de Liège (les Rouches) lors d'un match à domicile suit une loi de moyenne $\mu_{\mathcal{X}} = 1898$ ml et d'écart-type $\sigma_{\mathcal{X}} = 980,3$ ml. On tire un échantillon i.i.d. de 12 matchs à domicile et on tire à chaque fois un supporter des Rouches au hasard.

a) Sous l'hypothèse d'une loi normale pour la variable $\mathcal{X}$, estimer la probabilité que la moyenne de la consommation des 12 personnes soit comprise entre 1498 et 2298 ml ? (**4 points**)

b) Sans faire d'hypothèse sur la distribution de la variable $\mathcal{X}$, que peut-on dire sur la probabilité que cette moyenne soit comprise entre 1498 et 2298 ml ? (**4 points**)

c) La probabilité calculée au point b) devait-elle forcément être plus petite que celle calculée au point a) ? (**2 points**)

Supposons désormais que les supporters des Rouches achètent des bières en gobelets de 250ml et que le nombre théorique de bières achetées lors d'un match à domicile suive la loi indiquée à la table 2. Notez que le code d'honneur du supporteur des Rouches veut que toute bière achetée soit entièrement consommée ! On tire toujours un échantillon i.i.d. de 12 matchs à domicile et à chaque fois un supporter du standard de Liège au hasard.

TABLE 2 – Quantité de bières achetées par un supporter lors d'un match à domicile

Nombre de gobelets de 250ml	Probabilité
0	4,9%
2	4,1%
3	6%
4	8%
5	9%
6	8%
7	9%
8	10%
9	8%
10	11%
11	8%
12	5%
13	3%
14	2%
15	2%
18	1%
20	1%

d) Que valent l'espérance et la variance de la moyenne d'échantillon ? (**2 points**)

e) Que valent les espérances de s_x^2 et de s_{n-1}^2 ? (**2 points**)

f) Peut-on utiliser les formules du formulaire pour calculer les variances de s_x^2 et de s_{n-1}^2 ? Justifier. (**1 point**)

g) Déterminez le rapport entre les variances de s_x^2 et de s_{n-1}^2. (**1 point**)

h) Quel est l'intérêt de corriger la variance ? Quel est le désavantage ? (**2 points**) Suggestion : repartez de vos réponses aux sous-questions e) et g).

i) Si vous deviez calculer la probabilité que la moyenne consommée par les 12 personnes soit comprise entre 1,498 et 2,298 litre dans les conditions qui nous occupent, pensez-vous qu'elle serait forcément plus grande que celle calculée au point b) ? (**2 points**)

2.9. Les lecteurs MP3 produits par un fabricant A ont une durée de vie moyenne de 1400 heures avec un écart-type de 200 heures. Ceux du concurent B ont une durée de vie moyenne de 1200 heures avec un écart-type de 100 heures. On tire un échantillon i.i.d. de 125 lecteurs MP3 de chaque marque, indépendamment les uns des autres. Quelle est la probabilité que l'échantillon de la marque A ait une durée de vie moyenne supérieure d'au moins 160 heures à celle de B.

Suggestion : commencez par trouver les formules de distribution d'échantillonnage de la différence entre deux moyennes (qui n'ont pas été vues en répétition).

Solutions des exercices suggérés

2.4. a) 0,5328 b) 0,99379

2.5. a) 0,00256 b) 1,28

2.6. a) $E\{f_C\} = 0,377$, $E\{n_C\} = 942,62$, $\sigma_{f_C} = 9,69\text{x}10^{-3}$, $\sigma_{n_C} = 24,23$.

b) $E\{f_{C_1}\} = \frac{1}{5}$, $E\{n_{C_1}\} = 200$, $\sigma_{f_{C_1}} = 0,01265$, $\sigma_{n_{C_1}} = 12,65$.
$E\{f_{C_2}\} = 0,416$, $E\{n_{C_2}\} = 624$, $\sigma_{f_{C_2}} = 0,0127$, $\sigma_{n_{C_2}} = 19$.

c) $\alpha = \frac{11}{61}$.

d) $V\{f'_C\} = 1,14\text{x}10^{-4}$, $\sigma_{f'_C} = 1,0678\text{x}10^{-2}$.

e) Oui, nous n'avions donc aucune garantie de réduction de variance car la taille des échantillons n'a pas été choisie de manière proportionnelle à la taille de P_1 et P_2.

f) Cette taille garantit une réduction de variance, mais pas la réduction maximale. Lorsqu'il y a une grosse différence entre les parties, ce qui est le cas ici, alors il existe forcément une meilleure répartition.

h) Les conditions de tailles d'échantillon sont remplies pour que f_C, f_{C_1} et f_{C_2} soient quasi-gaussiennes. De plus, f'_C étant une combinaison linéaire de deux variables quasi-gaussiennes indépendantes, elle suivra

aussi une loi quasi-gaussienne.

2.7. a) 85%

b) L'inégalité de Bienaymé-Tchebycheff garanti une probabilité $\geq 51,88\%$.

c) Oui, la méthode qui fait le moins d'hypothèses conduit toujours à une garantie moins forte.

d) $E\{m_x\} = 149\text{g} \quad V\{m_x\} = 108,27\ \text{g}^2$.

e) $E\{s_x^2\} = 1515,73\text{g}^2 \quad E\{s_{n-1}^2\} = 1624\text{g}^2$

f) Non car on est pas dans le cas gaussien.

g) $\frac{(n-1)^2}{n^2}$.

h) Intérêt : estimateur non-biaisé. Désavantage : variance de l'estimateur plus élevée.

i) Oui. L'inégalité de Bienaymé-Tchebyshev est garantie quelque soit la distribution. On trouverait donc une probabilité $\geq 51,88\%$.

2.8. a) 84%

b) L'inégalité de Bienaymé-Tchebycheff garanti une probabilité $\geq 50\%$.

c) Oui, la méthode qui fait le moins d'hypothèses conduit toujours à une garantie moins forte.

d) $E\{m_x\} = 1898\text{ml} \quad V\{m_x\} = 80082,34\ \text{ml}^2$.

e) $E\{s_x^2\} = 880890\text{ml}^2 \quad E\{s_{n-1}^2\} = 960971\text{ml}^2$

f) Non car on est pas dans le cas gaussien.

g) $\frac{(n-1)^2}{n^2}$.

h) Intérêt : estimateur non-biaisé. Désavantage : variance de l'estimateur plus élevée.

i) Oui. L'inégalité de Bienaymé-Tchebyshev est garantie quelque soit la distribution. On trouverait donc une probabilité $\geq 50\%$.

2.9. 0,97725

TABLE 3 – Table de Gauss (exemple de lecture : $P(\mathcal{Z} \leq 1,46) = 0,92785$)

	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,50000	0,50399	0,50798	0,51197	0,51595	0,51994	0,52392	0,52790	0,53188	0,53586
0,1	0,53983	0,54380	0,54776	0,55172	0,55567	0,55962	0,56356	0,56749	0,57142	0,57535
0,2	0,57926	0,58317	0,58706	0,59095	0,59483	0,59871	0,60257	0,60642	0,61026	0,61409
0,3	0,61791	0,62172	0,62552	0,62930	0,63307	0,63683	0,64058	0,64431	0,64803	0,65173
0,4	0,65542	0,65910	0,66276	0,66640	0,67003	0,67364	0,67724	0,68082	0,68439	0,68793
0,5	0,69146	0,69497	0,69847	0,70194	0,70540	0,70884	0,71226	0,71566	0,71904	0,72240
0,6	0,72575	0,72907	0,73237	0,73565	0,73891	0,74215	0,74537	0,74857	0,75175	0,75490
0,7	0,75804	0,76115	0,76424	0,76730	0,77035	0,77337	0,77637	0,77935	0,78230	0,78524
0,8	0,78814	0,79103	0,79389	0,79673	0,79955	0,80234	0,80511	0,80785	0,81057	0,81327
0,9	0,81594	0,81859	0,82121	0,82381	0,82639	0,82894	0,83147	0,83398	0,83646	0,83891
1,0	0,84134	0,84375	0,84614	0,84849	0,85083	0,85314	0,85543	0,85769	0,85993	0,86214
1,1	0,86433	0,86650	0,86864	0,87076	0,87286	0,87493	0,87698	0,87900	0,88100	0,88298
1,2	0,88493	0,88686	0,88877	0,89065	0,89251	0,89435	0,89617	0,89796	0,89973	0,90147
1,3	0,90320	0,90490	0,90658	0,90824	0,90988	0,91149	0,91309	0,91466	0,91621	0,91774
1,4	0,91924	0,92073	0,92220	0,92364	0,92507	0,92647	0,92785	0,92922	0,93056	0,93189
1,5	0,93319	0,93448	0,93574	0,93699	0,93822	0,93943	0,94062	0,94179	0,94295	0,94408
1,6	0,94520	0,94630	0,94738	0,94845	0,94950	0,95053	0,95154	0,95254	0,95352	0,95449
1,7	0,95543	0,95637	0,95728	0,95818	0,95907	0,95994	0,96080	0,96164	0,96246	0,96327
1,8	0,96407	0,96485	0,96562	0,96638	0,96712	0,96784	0,96856	0,96926	0,96995	0,97062
1,9	0,97128	0,97193	0,97257	0,97320	0,97381	0,97441	0,97500	0,97558	0,97615	0,97670
2,0	0,97725	0,97778	0,97831	0,97882	0,97932	0,97982	0,98030	0,98077	0,98124	0,98169
2,1	0,98214	0,98257	0,98300	0,98341	0,98382	0,98422	0,98461	0,98500	0,98537	0,98574
2,2	0,98610	0,98645	0,98679	0,98713	0,98745	0,98778	0,98809	0,98840	0,98870	0,98899
2,3	0,98928	0,98956	0,98983	0,99010	0,99036	0,99061	0,99086	0,99111	0,99134	0,99158
2,4	0,99180	0,99202	0,99224	0,99245	0,99266	0,99286	0,99305	0,99324	0,99343	0,99361
2,5	0,99379	0,99396	0,99413	0,99430	0,99446	0,99461	0,99477	0,99492	0,99506	0,99520
2,6	0,99534	0,99547	0,99560	0,99573	0,99585	0,99598	0,99609	0,99621	0,99632	0,99643
2,7	0,99653	0,99664	0,99674	0,99683	0,99693	0,99702	0,99711	0,99720	0,99728	0,99736
2,8	0,99744	0,99752	0,99760	0,99767	0,99774	0,99781	0,99788	0,99795	0,99801	0,99807
2,9	0,99813	0,99819	0,99825	0,99831	0,99836	0,99841	0,99846	0,99851	0,99856	0,99861
3,0	0,99865	0,99869	0,99874	0,99878	0,99882	0,99886	0,99889	0,99893	0,99896	0,99900
3,1	0,99903	0,99906	0,99910	0,99913	0,99916	0,99918	0,99921	0,99924	0,99926	0,99929
3,2	0,99931	0,99934	0,99936	0,99938	0,99940	0,99942	0,99944	0,99946	0,99948	0,99950
3,3	0,99952	0,99953	0,99955	0,99957	0,99958	0,99960	0,99961	0,99962	0,99964	0,99965
3,4	0,99966	0,99968	0,99969	0,99970	0,99971	0,99972	0,99973	0,99974	0,99975	0,99976
3,5	0,99977	0,99978	0,99978	0,99979	0,99980	0,99981	0,99981	0,99982	0,99983	0,99983
3,6	0,99984	0,99985	0,99985	0,99986	0,99986	0,99987	0,99987	0,99988	0,99988	0,99989
3,7	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999
3,8	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999	0,99999
3,9	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000	1,00000