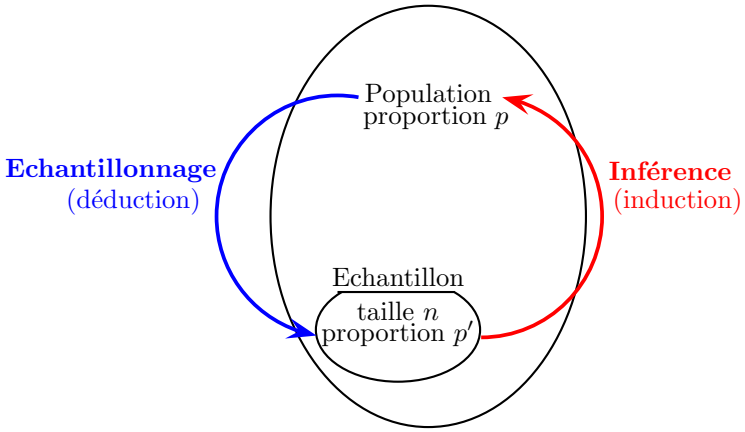
Fluctuation des échantillons - Estimation & sondage

I Inférence et échantillonnage statistiques



L'échantillonnage statistique consiste à prédire, à partir d'une population connue les caractéristiques des échantillons qui en seront prélevés.
On parle aussi de déduction des caractéristiques de l'échantillon.

L'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir de celles d'un échantillon.
On parle aussi d'induction, ou encore d'extrapolation des caractéristiques à l'ensemble de la population.

II Fluctuation d'échantillonnage

Définition: *Lorsqu'on répète n fois une expérience aléatoire on obtient une série de n résultats que l'on appelle* **échantillon de taille** n.
Si ces répétitions sont indentiques et indépendantes entre elles et que l'issue de chacune admet deux issues (0 ou 1, "Réussite" ou "Echec", ...), on dit que l'échantillon relève du modèle de Bernoulli.

Définition: *Si on réalise plusieurs échantillons, la distribution des proportions (ou fréquences) du nombre de "Réussite" varie d'un échantillon à l'autre.*
Ce phénomène s'appelle la **fluctuation d'échantillonnage**.

Exemple :
Il y a $p = 10\,\%$ de gauchers en France ($12\,\%$ exactement). Dans un échantillon de 30 élèves (la classe par exemple), on peut donc s'attendre à trouver 3 gauchers.

Dans certaines classes, il y a effectivement 3 gauchers, dans d'autes il y en a 2, soit $2/30 \simeq 7\,\%$, dans d'autres qu'un seul, soit $1/30 \simeq 3,5\,\%$, dans d'autres encore il peut y en a jusqu'à 6, soit ($6/30 = 20\,\%$), mais il semble très rare de trouver une classe avec 7 ou plus de gauchers.

La fréquence, ou proportion, de gauchers dans un échantillon de 30 personnes, varie, ou fluctue d'un échantillon à l'autre, mais reste sûrement dans l'intervalle $\left[3,5\,\% ; 20\,\%\right] = \left[0,35 ; 0,2\right]$.

III Intervalle de fluctuation au seuil de 95%

On sait néanmoins que plus la taille de l'échantillon est grande, plus la proportion observée se rapproche (se stabilise autour) d'une valeur limite donnée par la probabilité : c'est **la loi des grands nombres**.

Plus précisément, si on désigne par p la proportion dans la population complète, ou la probabilité s'il s'agit d'une expérience aléatoire, et qu'on réalise un échantillon de taille n, alors,

Prop.: *Si $n \geqslant 25$ et $0,2 \leqslant p \leqslant 0,8$, la proportion p' dans un échantillon de taille n est dans l'intervalle*

$$\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$$

avec une probabilité d'environ 95%.
Cet intervalle s'appelle l'intervalle de fluctuation à 95%, ou au seuil de 95%, ou encore au seuil d'erreur de 5%.

Cet intervalle quantifie la fluctuation due à l'aléatoire de la proportion p' que l'on peut observer en ne prélevant qu'un échantillon de taille n.
Même si, comme l'échantillon est constitué aléatoirement, la proportion p' peut a priori prendre n'importe quelle valeur entre 0 et 1 (ou 0% et 100%), on est "presque sûr" (en fait avec un "petit" risque d'erreur de 5%) que la variation de la proportion p' reste limitée dans cet intervalle.

Exercice 1 On réalise un échantillon de taille $n = 50$ lancers d'une pièce équilibrée ; on a alors équiprobabilité : les probabilités des issues "Pile" et "Face" sont $p = P("Pile") = P("Face") = 0,5$.

1. Combien de fois peut-on s'attendre à obtenir l'issue "Pile" ?
2. On obtient 30 fois l'issue "Pile". Calculer la fréquence correspondante.
   Est-ce que l'hypothèse "la pièce est équilibrée" est crédible ?
3. On réalise un échantillon de $n = 100$ lancers. Calculer la fréquence correspondante.
   On obtient 60 fois l'issue "Pile". Est-ce que l'hypothèse "la pièce est équilibrée" est crédible ?
4. On réalise un échantillon de $n = 200$ lancers. Calculer la fréquence correspondante.
   On obtient 120 fois l'issue "Pile". Est-ce que l'hypothèse "la pièce est équilibrée" est crédible ?

Exercice 2 **Influence d'une usine à proximité**
Une usine chimique est venue s'implanter près d'une ville il y a 3 ans. Pendant ces 3 ans sont nés dans cette ville 132 enfants dont 52 garçons.

1. Quelles sont les proportions de garçons et de filles nés dans cette ville ces 3 dernières années ?
2. Peut-on considérer que l'usine a eu un impact sur les naissances ?

Exercice 3 **Parité homme-femme**
Deux entreprises A et B recrutent leur personnel dans un bassin d'emploi où il y a autant d'hommes que de femmes. L'entreprise A emploie 60 personnes, dont 26 femmes, tandis que l'entreprise B emploie 1050 personnes, dont 480 femmes.

1. Calculer les proportions de femmes employées dans chaque entreprise.
   Laquelle semble au mieux respecter la parité ?
2. Déterminer pour chaque entreprise l'intervalle de fluctuation à 95% de la proportion de femmes.
3. Les deux entreprises respectent-elles la parité au seuil d'erreur de 5% ?

Exercice 4 **Conditionnement et commercialisation de pièces défectueuses**
La chaîne de production d'une usine produit des pièces commercialisables. 5% des pièces produites sont défectueuses. Ces pièces sont ensuites conditionnées et expédiées par carton de 50 pièces.

1. Quelle est la probabilité pour que dans un carton :
   a) aucune pièce ne soit défectueuse ?
   b) exactement une pièce soit défectueuse ?
   c) il y ait au plus une pièce défectueuse ?
2. Déterminer l'intervalle de fluctuation à 95% du nombre de pièces défectueuses. Interpréter.
3. Mêmes questions pour un conditionnement par cartons de 200 pièces.

Exercice 5 **Dimensionnement d'une cantine**
Dans un établissement de 3000 personnes, chaque personne peut ou non, librement, manger à la cantine chaque jour. En moyenne 65% des personnes y mangent.
On souhaite estimer le nombre de places assises dans la cantine telle manière que toutes les personnes aient une place pour manger. Par ailleurs, par soucis d'économie, on souhaite aussi que le nombre de places prévues soit minimal.
Combien de places doit-on prévoir ?

Exercice 6 **Texas contre Partida**
En novembre 1976, dans un comté du sud du Texas, Rodrigo Partida était condamné à 8 ans de prison pour cambriolage et tentative de viol. Il attaqua ce jugement au motif que la désignation des jurés était discriminante : alors que 79,1% de la population du comté était d'origine mexicaine, sur les 870 personnes convoquées pour être jurés lors d'une certaine période, il n'y eût que 339 personnes d'origine mexicaine.

1. Déterminer l'intervalle de fluctuation pour la proportion de personnes d'origine mexicaine dans un échantillon de 870 personnes.
2. Quelle est la proportion de personnes d'origine mexicaine constatée parmi les jurés convoqués ?
   Qu'en conclure ?

Exercice 7 Contrôle qualité
Dans une usine, le responsable de la fabrication affirme que la proportion de produits défectueux fabriqués est de 20%.
Sur la chaîne de fabrication on a prélevé au hasard 72 produits, et on a constaté que 24 d'entre eux étaient défectueux.

1. Quelle est la proportion de produits défectueux dans l'échantillon prélevé ?
2. Que penser de l'affirmation du responsable de la fabrication ?

Exercice 8 Influence du climat sur la couleur des yeux
En France, la proportion de personnes ayant les yeux bleus est de 31%.
Dans une grande ville française, au micro-climat particulièrement ensoleillé, sur 50 personnes rencontrées au hasard, on a recensé 10 personnes ayant les yeux bleus.

1. Déterminer l'intervalle de fluctuation à 95% de la proportion de personnes ayant les yeux bleus dans un échantillon de 50 personnes.
2. Peut-on attribuer au micro-climat une influence spécifique sur la couleur des yeux ?

IV Estimation d'une proportion inconnue à partir d'un échantillon

Exercice 9 Estimation du nombre de cyclistes citadins
Dans une grande ville de 200 000 habitants, la municipalité s'intéresse au nombre de cyclistes afin d'adapter au mieux les routes et pistes cyclables. Elle cherche donc à estimer la propotion p de cyclistes dans la ville, ou encore le nombre N de cyclistes.
La municipalité a ainsi effectué un sondage en interrogeant, au hasard, 400 personnes. Sur ces 400 personnes, 78 déclarent circuler régulièrement en vélo.

1. Quelle est la proportion p' de cyclistes dans l'échantillon interrogé ?
   Peut-on affirmer que $p = p'$?
2. Pour une proportion p de cyclistes dans la ville, donner l'intervalle à 95% de la proportion p' dans un échantillon de taille 400.
   Ecrire alors deux inégalités concernant les proportions p et p'.
3. Quelles inégalités doit alors satisfaire la proportion p ?
   En déduire l'intervalle contenant la proportion p, puis un encadrement de du nombre N de cylcistes dans toute la ville.

Définition: *Un échantillon est un sous-ensemble de la population.*
Un échantillon représentatif est un sous-ensemble choisi au hasard dans la population.

Connaissant la proportion (ou fréquence) p' d'un caractère pour un échantillon aléatoire de taille n, on peut estimer la proportion p du caractère dans la population complète de la façon suivante :

Prop.: *Lorsque* $n \geqslant 25$ *et* $0,2 \leqslant p \leqslant 0,8$*, l'intervalle*

$$\left[p' - \frac{1}{\sqrt{n}} \ ; \ p' + \frac{1}{\sqrt{n}} \right]$$

contient la proportion p avec une probabilité supérieure ou égale à 95%.
Cet intervalle s'appelle l'intervalle de confiance au seuil de 95%, ou l'intervalle au niveau de confiance de 95%.

Exercice 10 La semaine précédent une élection opposant deux candidats A et B, on a interrogé un échantillon de 200 électeurs supposé représentatif de l'ensemble des électeurs.
109 personnes de cet échantillon ont déclaré avoir l'intention de voter pour le candidat A.
Le candidat A, suite à ce sondage, affirme : « si les élections avaient eu lieu le jour du sondage j'aurais été élu ».
Qu'en pensez-vous ?

D'après ce qui précède, on peut (et doit !) maintenant traduire un résultat d'un sondage tel que :
« *Il y a 52% de personnes qui voteraient pour le candidat A (d'après un sondage réalisé auprès de 1000 personnes* », par : « *Il y a 95% de chances (ou une probabilité de 0,95=95%) pour que l'intervalle [49% ; 55%] contiennent le pourcentage de personnes prêts à voter pour le candidat A* ».

Les sondages sont réalisés en général sur des échantillons de $n = 1000$ personnes. Bien sûr, si ce n'est pas le cas, on adapte alors l'intervalle de confiance [49% ; 55%] grâce à la formule précédente.

V Dimensionnement des échantillons

En sondant un échantillon plus important, l'intervalle de confiance aurait été restreint. Deux élements sont alors en concurrence :
– si la taille de l'échantillon est faible, la fourchette obtenue est large, et l'information peut manquer de pertinence ;
– on ne souhaite pas par ailleurs à sonder des échantillons de taille trop importante, afion de diminuer le coût de l'étude.

Exercice 11 Avec les données de l'exercice précédent, en supposant que la proportion d'électeurs favorables au candidat A reste la même, de quelle taille devrait être l'échantillon des personnes interrogées pour pouvoir affirmer que A serait élu ?

Exercice 12 En 2002, avant le 1er tour des élections présidentielles, les sondages estimaient que L. Jospin allait l'emporter sur J.M. Le Pen avec 18% pour des votes contre 14%.
A la surprise générale, le jour de l'élection, J.M. Le Pen l'emporta avec 16,86% des votes contre 16,18% pour L. Jospin...

1. Voici un extrait d'un article publié dans le journal "Le Monde" par le statisticien Michel Lejeune après le premier tour de l'élection présidentielle de 2002 :

   "Pour les rares scientifiques qui savent comment sont produites les estimations, il est clair que l'écart des intentions de vote entre les candidats Le Pen et Jospin randait tout à fait plausible le scénario qui s'est réalisé. En effet, certains des derniers sondages indiquaient 18% pour Jospin et 14% pour Le Pen. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales [...], on obtient sur de tels pourcentages une incertitude de plus ou moins 3% étant donné la taille de l'échantillon [...]"

   Quel est la taille de l'échantillon auquel fait allusion Michel Lejeune dans son article publié dans "Le Monde" ?
2. Déterminer les intervalles de confiance à 95% après le sondage (effectué auprès de 1000 personnes).
   Le résultat de l'élection est-il si surprenant ?
3. Vérifier que les pourcentages de votes pour Jospin et Le Pen sont bien cohérents avec les résultats du sondage.
4. La taille de l'échantillon était ici trop faible, pour pouvoir tirer une conclusion, même avec un risque d'erreur de 5%.

   Quelle devrait être la taille de l'échantillon de personnes sondées pour que, si 18% des personnes votent pour Jospin et 14% pour Le Pen, on puisse conclure, avec un risque d'erreur de 5%, que Jospin l'emporterait bien sur Le Pen.