STATISTIQUES

I. PRÉSENTATION

STATISTIQUES MATHÉMATIQUES

La partie authentiquement mathématique des statistiques consiste pour l'essentiel en des probabilités. Quelques exemples de situations qui relèvent de ce domaine :

- Une compagnie d'assurance a besoin de savoir quel montant elle doit facturer un contrat donné, par exemple une assurance décès.
- Des climatologues, constatant le nombre de cyclones survenus pendant un an se demandent si la variation par rapport aux années précédentes est significative, si elle traduit une tendance générale du climat.
- Un institut de sondage veut savoir combien de personnes il doit interroger pour que les réponses de l'échantillon aient de bonnes chances de ressembler à celles de la population tout entière.
- Le ministère de la santé essaye d'évaluer l'intérêt de procéder à une généralisation d'un vaccin donné.

Dans les cas les plus subtils, on voudra, à partir d'observations, essayer de remonter jusqu'aux lois de probabilité qui pourraient se cacher derrière, de façon à faire des prévisions ou des choix. L'idée est de déduire quelque chose de général à partir d'une connaissance partielle. Soit qu'on n'ait qu'une partie des données (sondage,

échantillon), soit qu'on ait un ensemble complet de données, mais qui n'est que la manifestation d'un phénomène sous-jacent (…). On essaye alors de deviner une loi cachée ou des corrélations. On parle alors de **statistiques inférentielles**[1]. Plus généralement, la partie probabiliste des statistiques, s'appelle aujourd'hui **statistiques mathématiques**.

STATISTIQUES DESCRIPTIVES

Mais les statistiques, à l'origine, ne sont pas une discipline mathématique. Le mot vient du latin *statisticus* : relatif à l'État. (Penser à l'anglais *state*, par exemple.) Les statistiques sont au départ des études de faits sociaux, par des dénombrements, des catalogues, des inventaires, des recensements, etc. Les rapports pouvaient être donnés sous formes de textes, de tableaux dont les données n'étaient pas forcément chiffrées. En tout cas, si aujourd'hui, son champ d'application est vaste, en ses origine, la statistique s'occupe des *hommes*. Les statistiques ont pris leur essor au XIX^{e} s., dans des sociétés devenues suffisamment atomisées pour que les hommes puissent y être considérés comme équivalents[2].

---

[1] *Inférer*, c'est passer de principes à une conclusion, par une opération de l'esprit. Une inférence peut être une *déduction* (comme en math.), mais cela peut aussi être une *induction* qui est une sorte de généralisation : par exemple, à force de voir tomber les objets, j'en *induis* que les objets tombent. L'induction est exclue en mathématiques : ce n'est pas parce qu'une loi fonctionne pour mille exemples qu'elle sera toujours vraie. En revanche, l'induction est utilisée en physique. Les statistiques « inférentielles » devraient en réalité s'appeler statistiques inductives. Il s'agit quand même d'une branche des mathématiques car elles sont fondées sur de vrais théorèmes de probabilités. C'est l'application à la réalité qui sort du domaine proprement mathématique.

[2] Ceux que l'histoire des statistiques intéressent pourront lire *Quand le Monde s'est fait Nombre*, d'Olivier Rey.

Aujourd'hui, la partie des statistiques qui n'est pas proprement mathématique (puisqu'elle ne contient ni théorèmes ni démonstrations, mais seulement des formules, des définitions et de zolis dessins), s'appelle les **statistiques descriptives**.

II. Vocabulaire

Population

Une **population** statistique est un *ensemble* dont les éléments sont appelés **individus**. Mathématiquement, ces individus sont comme de simples étiquettes. Mais elles sont là pour représenter des « individus » réels. Ces « individus » doivent être de nature assez homogène. On peut considérer à titre d'exemple que ces individus sont des êtres humains, dans une société où leurs statuts sont suffisamment indifférenciés pour que chacun puisse « compter pour un » (mais ce pourrait aussi bien être des huîtres ou des chaises produites par une même usine). Pour population, prenons donc comme exemple la population française actuelle, constituée de l'ensemble des individus de nationalité française.

Une **classe** est un regroupement d'*individus* (c'est un sous-ensemble de la *population*). Par exemple dans la population française, on peut considérer la classe des femmes, celle des retraités, celle des lycéens, des individus de moins de 25 ans…

Le nombre d'individus dans une classe s'appelle l'**effectif** de cette *classe*. L'*effectif* de la *population* est appelé **effectif total**.

L'adjectif *effectif* signifie : qui produit un effet. Dans l'armée, au

XVII[e] s., les soldats *effectifs* étaient ceux qui, réellement présents, étaient en état se de battre. Cet adjectif s'est substantivé[3].

Fréquence

La **fréquence** d'une classe est le rapport de l'effectif de cette classe sur l'effectif total. (Un *rapport* est un *quotient* de deux grandeurs de même nature.)

Par exemple, au premier janvier 2017, l'*effectif total* de la population française était de 66 990 826 individus. L'effectif des moins de 25 ans était de 20 172 427. Alors, la *fréquence* des moins de 25 ans dans la population française est de : $\frac{20\,172\,427}{66\,990\,826} \approx 0{,}3$

Ce rapport, cette « proportion », peut se dire en pourcentage :

$$0{,}3 = \frac{3}{10} = \frac{30}{100} = 30\,\%$$

La *fréquence* des moins de 25 ans dans la population française est de 30%.

Pourquoi appeler cela *fréquence* ?

Il faut peut-être, pour comprendre ce mot, lier les statistiques aux probabilités. Imaginons qu'on tire au sort, plusieurs fois de suite, un individu parmi tous ceux de la population (avec possibilité de tirer plusieurs fois le même). Alors, il arrivera plus ou moins *fréquemment* qu'on tombe sur une personne de moins de 25 ans. Si l'on fait un très grand nombre de tirages au sort, cela arrivera à peu près 30 fois sur 100. Lorsque le nombre de tirages croît, la

[3] Un *substantif* est un nom commun. Se *substantiver*, c'est se changer en nom commun.

fréquence se rapproche de la probabilité.[4]

(En physique, la *fréquence* est le nombre de fois qu'un phénomène arrive par unité de temps. Elle se donne généralement en Hertz, abrégé Hz, nombre de fois par seconde.)

CARACTÈRE

Pour chaque individu, on considère un **caractère**, qui est une propriété particulière[5]. La fonction qui, à chaque individu, associe son caractère s'appelle une **variable statistique**. On retrouvera cette particularité de vocabulaire consistant appeler « variable » une fonction, en probabilités (en classe de première) avec la notion de « variable aléatoire ». Le caractère associé à un individu donné peut alors s'appeler une **valeur**.

Le caractère peut être **qualitatif** (la couleur des yeux) ou **quantitatif** (l'âge). S'il est quantitatif, il peut être **discret** ou **continu**. Discret : il ne peut prendre que certaines valeurs bien séparées les unes des autres. (Le préfixe *dis-* marque la séparation.) Continu : il peut prendre toutes les valeurs réelles sur un certain intervalle. Exemple de caractère *discret* : le nombre d'enfants. Exemple de caractère *continu* : la taille. Un caractère continu peut être regroupé par intervalles. Par exemple, pour relever le salaire

[4] Il est possible aussi que le mot *fréquence*, en statistiques, vienne d'un sens plus ancien du mot. Fréquence signifiait *foule*, affluence (on parle d'un magasin plus ou moins *fréquenté*). En ce sens, *fréquence* serait proche de *effectif*. Mais le Dictionnaire Historique de la Langue Française d'Alain Rey semble plutôt considérer que le sens du mot en statistiques part du sens temporel pris par l'adjectif.

[5] Il est possible, pour une même population, d'étudier plusieurs caractères simultanément.

mensuel, on pourra grouper par tranches de 100 euros. On regroupera tous les individus dont le salaire mensuel est compris entre 1500 euros et 1600 euros.

L'**amplitude** d'une classe ou d'un intervalle, c'est l'écart entre la plus petite valeur et la plus grande. Ici, les intervalles ont des amplitudes de 100 euros. L'amplitude de la population totale, s'appelle son **étendue**.

SÉRIE STATISTIQUE

Une **série statistique** est la liste des valeurs associées aux individus. Si une valeur apparaît plusieurs fois, ce nombre de fois qu'elle apparaît doit être une information contenu d'une façon ou d'une autre dans la série. Par exemple, dans une classe de 25 élèves, on s'intéresse aux notes obtenues par les élèves à un contrôle donné. Ces notes sont :

16 ; 9 ; 9 ; 16 ; 10 ; 12 ; 18 ; 9 ; 15 ; 15 ; 13 ; 9 ; 6 ; 12 ; 15 ; 6 ; 4 ; 13 ; 9 ; 8 9 ; 9 ;10 ; 8 ; 14

Elles ont ici été regroupées par paquets de cinq, simplement pour plus de lisibilité. L'ordre dans lequel elles sont données n'a pas d'importance (ici, c'est l'ordre alphabétique du nom des élèves). On aurait aussi bien pu les classer par ordre croissant :

4 ; 6 ; 6 ; 8 ; 8 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 9 ; 10 ; 10 ; 12 ; 12 ; 13 ; 13 ; 14 ; 15 ; 15 ; 15 ; 16 ; 16 ; 18

On ne perdrait pas d'information si l'on donnait la série sous la forme suivante, appelée **tableau d'effectifs** :

Valeur	Effectif
4	1
6	2
8	2
9	7
10	2
12	2
13	2
14	1
15	3
16	2
18	1
Total :	25

Cette façon de mémoriser les données est économe en place si la population est importante et qu'il y a peu de valeurs.

À chaque valeur (ou à chaque intervalle), on peut ainsi associer une classe, formée de l'ensemble des individus dont le caractère considéré prend cette valeur. La *fréquence* d'une valeur donnée est la fréquence de la classe associée à cette valeur. Ainsi, en reprenant le tableau qui précède, on peut constater que la note 9 est assez fréquente. Sa fréquence est de $\frac{7}{25} = 0{,}28 = \underline{28\,\%}$.

III. Statistiques déscriptives

Dans ce paragraphe, nous prendrons encore comme exemple des notes obtenues par des élèves à un contrôle. Mais, pour simplifier les expressions numériques, on imaginera un groupe de seulement treize élèves. Les notes sont les suivantes :

7 ; **11** ; **14** ; **11** ; **12** ; **15** ; **4** ; **14** ; **11** ; **11** ; **15** ; **14** ; **14**.

Moyenne

On obtient la moyenne d'une série statistique en additionnant toutes les valeurs et en divisant par le nombre de valeurs.

Ici :

$$\frac{7+11+14+11+12+15+4+14+11+11+15+14+14}{10}$$

Si une valeur apparaît plusieurs fois, il faut la compter autant de fois qu'elle apparaît. Ainsi, si la série statistique est présentée en *tableau d'effectifs* :

Valeur	Effectif
4	1
7	1
11	4
12	1
14	4
15	2

Alors, il faut tenir compte des effectifs pour exprimer la moyenne :

$$\frac{1\times 4+1\times 7+4\times 11+1\times 12+4\times 14+2\times 15}{1+1+4+1+4+2}$$

Les effectifs sont, au numérateur, des *coefficients*.

Médiane

La médiane est « la » valeur qui sépare la série en deux classes de même effectif : il doit y avoir autant d'individus au dessus qu'en dessous.

Ainsi, le salaire[6] médian, en France, en 2016 était de 1772 euros. Cela signifie qu'il y avait autant de salariés touchant un salaire

[6] Salaire mensuel net équivalent temps plein des salariés d'une entreprise publique ou privée.

inférieur à cette valeur que de salariés touchant un salaire supérieur. Le salaire moyen, lui, était de 2202 euros.

Pour obtenir la médiane, on peut ordonner les valeurs par ordre croissant :

rang	1	2	3	4	5	6	7	8	9	10	11	12	13
valeur	**4**	**7**	**11**	**11**	**11**	**11**	**12**	**14**	**14**	**14**	**14**	**15**	**15**

On prend alors la valeur centrale. Ici, la médiane est de **12**. C'est assez simple car il y a un rang central, le rang 7, qu'on obtient en faisant la moyenne des rangs extrêmes : $\frac{1+13}{2}=7$. Lorsque la série contient un nombre impair de termes, c'est un tantinet plus délicat. Reprenons, la même série, mais amputée de son terme central :

rang	1	2	3	4	5	6	7	8	9	10	11	12
valeur	**4**	**7**	**11**	**11**	**11**	**11**	**14**	**14**	**14**	**14**	**15**	**15**

$\frac{1+12}{2}=7{,}5$. Il n'y a pas de rang central, mais deux rangs qui entourent ce rang central absent : le rang 6 et le rang 7, dont les valeurs sont respectivement **11** et **14**. Alors, on pourrait prendre pour médiane toute valeur strictement comprise entre **11** et **14**. Nous déciderons de prendre la moyenne de ces deux valeurs (certains cours, ou logiciels, procèdent différemment). Dans le cas présent, la médiane sera donc égale à **12,5**. Nous pouvons proposer en résumé la définition suivante :

Soit n l'effectif total d'une série statistique.
Si n est pair, la médiane est la valeur de rang $\frac{1+n}{2}$.
Si n est impair, nous prenons pour médiane la moyenne des deux valeurs centrales, c'est-à-dire de celles dont les rangs encadrent le nombre (pas entier) $\frac{1+n}{2}$.

QUARTILES

Ce sont les trois valeurs qui partagent la série en quatre classes de même effectif. Notons-les, par ordre croissant, Q_1, Q_2 et Q_3.
Le quartile du milieu, Q_2 est la médiane. Les deux autres ont des définitions qui ressemblent à celle de la médiane, mais simplifiées :

Le quartile Q_1 est la plus petite valeur telle que $\frac{1}{4}$ des données lui soient inférieures ou égales. Le quartile Q_3 est la plus petite valeur telle que $\frac{3}{4}$ des données lui soient inférieures ou égales.

Autrement dit, Q_1 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{1}{4}n$ et Q_3 est la valeur dont le rang est le premier entier supérieur ou égal à $\frac{3}{4}n$. (n étant là encore l'effectif total.)

Par exemple, en reprenant notre groupe de 13 élèves :

rang	1	2	3	4	5	6	7	8	9	10	11	12	13
valeur	**4**	**7**	**11**	**11**	**11**	**11**	**12**	**14**	**14**	**14**	**14**	**15**	**15**

$\frac{13}{4}=3{,}25$. Donc Q_1 est la valeur du terme de rang 4. Q_1 = **11**.

Q_2 est la médiane. Q_2 = **12**.

$\frac{3\times13}{4}=9{,}75$. Donc Q_3 est la valeur du terme de rang 10. Q_3 = **14**.

Effectifs cumulés, fréquences cumulées

Voici un tableau fictif donnant le nombre d'enfants par femme d'un groupe (non représentatif) de 205 mères :

Nombre d'enfants	Effectif
1	25
2	75
3	54
4	26
5	18
6	4
7	2
8	1

Si l'on souhaitait calculer la médiane de cette série, il faudrait trouver la mère de rang $\frac{1+205}{2} = 103$. Pour cela, on peut rajouter une colonne donnant les *effectifs cumulés*. **L'effectif cumulé** d'une valeur donnée, c'est le nombre d'individus dont la valeur est inférieure ou égale à cette valeur donnée. On calcul cela très facilement à l'aide des effectifs :

Nombre d'enfants	Effectif	Effectif cumulé
1	25	25
2	75	100
3	54	**154**
4	26	180
5	18	198
6	4	202
7	2	204
8	1	205

154 = 25+75+54

On voit ainsi facilement que la mère de rang 103 a 3 enfants.

De la même façon, les **fréquences cumulées** s'obtiennent en additionnant les fréquences correspondant aux valeurs inférieures ou égales à la valeur considérée :

Nombre d'enfants	Effectif	Effectif cumulé	Fréquences	Fréquences cumulées
1	25	25	0,12	0,12
2	75	100	0,37	0,49
3	54	154	0,26	**0,75**
4	26	180	0,13	0,88
5	18	198	0,09	0,97
6	4	202	0,02	0,99
7	2	204	0,01	1
8	1	205	0	1

0,75 = 0,12+0,37+0,26

Représentations graphiques

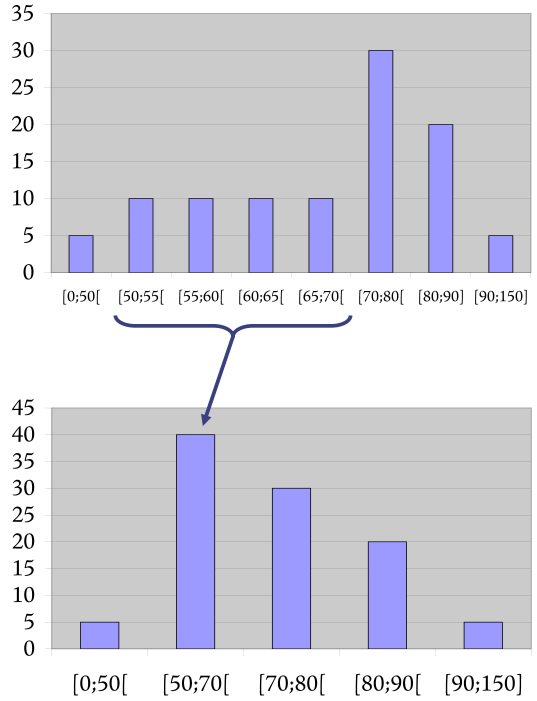
Nous n'allons pas passer en revue toutes les formes de représentations graphiques. Certaines — comme les « toiles d'araignées » — mériteraient une critique précise dont nous confions l'exécution au bon sens du lecteur. Contentons-nous de parler des *histogrammes*. (Le mot vient du grec *histos*, tissus, pris au sens de réseau, graphe et de *gramma*, lettre, écriture ; il n'a pas la même origine que le mot *histoire*.)

L'**histogramme** est particulièrement utile lorsqu'on représente un *caractère continu*, regroupé par classes d'*amplitudes* inégales. Supposons par exemple qu'on cherche à représenter graphiquement la façon dont se répartissent les poids d'un groupe de 100 personnes et que les données soient regroupées ainsi par classes d'amplitudes inégales :

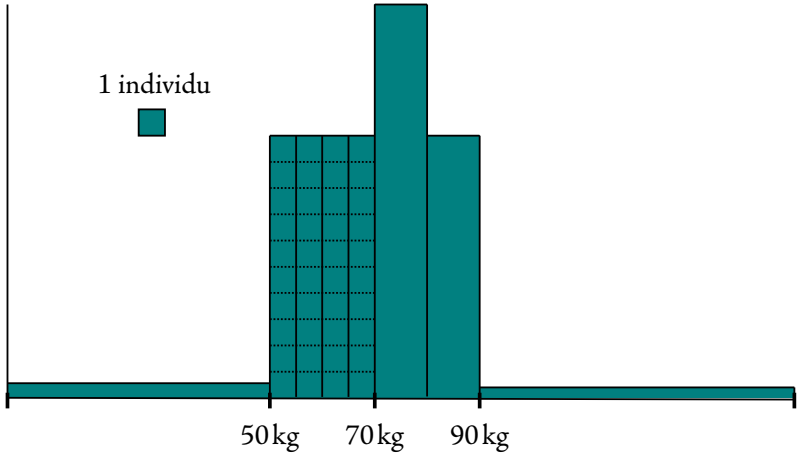
Poids en Kg	Effectif
[0 ; 50[	5
[50 ; 55[	10
[55 ; 60[	10
[60 ; 65[	10
[65 ; 70[	10
[70 ; 80[	30
[80 ; 90[	20
[90 ; 150]	5

Poids en Kg	Effectif
[0 ; 50[	5
[50 ; 70[	40
[70 ; 80[	30
[80 ; 90[	20
[90 ; 150]	5

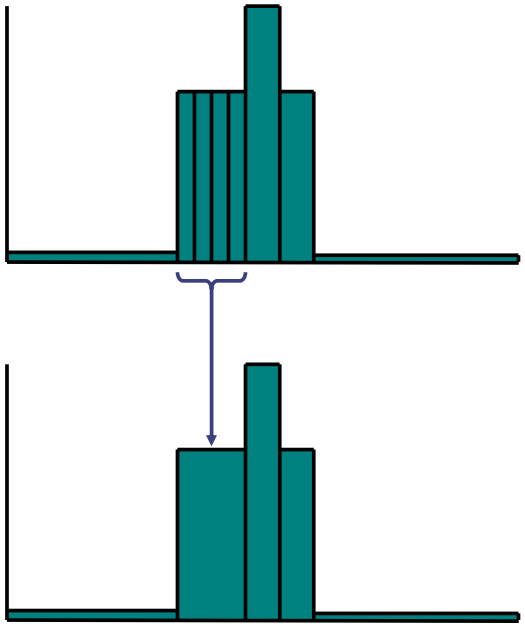
Alors, si l'on représente ces données par un **diagramme en bâton**, où seule la hauteur des barres varie, on ne prend pas en compte l'amplitude des classes. Dans ce cas, selon la façon dont on regroupe les valeurs en intervalles, la représentation pourra considérablement changer d'aspect.



On évite cela avec l'**histogramme**, dont le principe est de représenter l'effectif non pas par la hauteur des rectangles (en ordonnée), mais par leur aire. La largeur des rectangles étant, elle, proportionnelle à l'amplitude de l'intervalle.



De cette façon, si par exemple on réunit plusieurs rectangles adjacents de même hauteur, cela forme un nouveau rectangle (plus large) dont la hauteur ne change pas. Et l'aspect de la représentation n'est pas modifié.



III. Statistiques inférentielles

Intervalle de fluctuation

Exemple

Prenons comme *population* les français inscrits sur les listes électorales[7]. On est à quelques jours du premier tour d'élections présidentielles. Nous supposerons que tout le monde a choisi son candidat, que personne ne mentira aux sondeurs ni ne refusera de répondre, ni ne changera d'avis, etc. (hypothèses peu plausibles, mais il faut bien simplifier). Il y a donc une proportion p, bien déterminée, d'électeurs qui va voter pour le candidat Gudule. Disons que c'est pile-poil 30%. Autrement dit : $p = 0{,}3$. Ce nombre

[7] Ils sont environ 46 millions, mais cela n'a pas vraiment d'importance ici.

p est aussi la *probabilité* qu'on a de tomber sur un électeur de Gudule lorsqu'on tire un électeur au sort.

Un institut de sondage interroge 100 personnes, parmi les inscrits, en les tirant au sort[8]. Ces personnes, qui constituent un sous-ensemble de la population totale, forment ce que l'on appelle une **échantillon**. Notons f la fréquence des personnes votant Gudule dans cet échantillon. Intuitivement, f devrait être « proche » de p. Mais rien n'est sûr : si les sondeurs n'ont pas de bol, ils peuvent tomber sur un échantillon qui n'est pas représentatif, où, par exemple, personne ne vote Gudule ; ou encore un échantillon où tout le monde vote Gudule. Plus l'échantillon est gros, moins les sondeurs ont de risque de se planter trop gravement.

Simulation du tirage d'un échantillon

On peut très facilement simuler le sondage à la calculatrice ou avec un tableur : il suffit de faire 100 tirages au sort avec à chaque fois une probabilité 0,3 d'obtenir « Gudule » et donc 0,7 d'obtenir « pas Gudule ».

```
PROGRAM:GUDULE
:0→N
:For(I,1,100)
:If NbrAléat≤0.3
:N+1→N
:End
:Disp N/100
```

```
prgmGUDULE
                .28
                Fait
                .25
                Fait
                .28
                Fait
```

[8] Ce n'est pas comme ça qu'on pratique, du moins en France, mais peu importe. En outre, il faudrait préciser qu'on fait ici un tirage *avec remise*, c'est-à-dire que théoriquement, on pourrait interroger plusieurs fois la même personne. De cette façon, la probabilité de tirer un électeur de Gudule ne change pas, à chaque tirage d'un nouvel électeur.

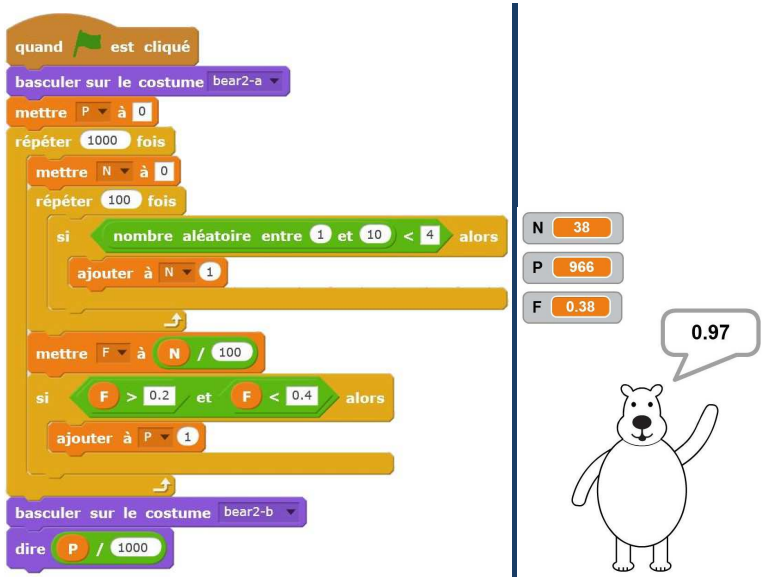
En lançant ce programme plusieurs fois, on obtient, par exemple les valeurs de *f* suivantes :

0,28 ; 0,25 ; 0,28 ; 0,34 ; 0,31 ; 0,33 ; 0,29 ; 0,28 ; 0,3.

Ce que l'on souhaiterait avoir, c'est un intervalle (si possible de centre *p* et par trop large) dans lequel *f* a au moins 95% de chances de se trouver. Pourquoi 95% ? Comme ça. On aurait pu choisir par exemple 99%, mais on va prendre systématiquement 95%, pour ne pas embrouiller nos esprits. Disons qu'« on » considère souvent que 5% d'erreur est un seuil tolérable[9]. Ici, l'intervalle **[0,2 ; 0,4]** satisfait à ces conditions. Nous n'expliquons pas pour l'instant comment nous avons trouvé cet intervalle et nous ne donnons pas de preuve de notre affirmation. En revanche, on peut vérifier la plausibilité de la chose par une simulation. On peut lancer par exemple 1000 fois le programme précédent et regarder, sur les 1000 fois, combien de fois *f* tombe en effet dans l'intervalle **[0,2 ; 0,4]**. Comme la calculatrice risque d'être un peu longue et que nous éprouvons une certaines nostalgie du collège, nous allons faire appel au bon vieux Scratch, en mode turbo.

[9] Ce laxisme semble préjudiciable à la recherche scientifique : http://passeurdesciences.blog.lemonde.fr/2013/11/13/une-etude-ebranle-un-pan-de-la-methode-scientifique/.

SIMULATION DES TIRAGES DE 1000 ÉCHANTILLONS



Que nous dit le nounours ?

Après avoir réalisé à toute berzingue 1000 fois la simulation d'un sondage auprès de 100 électeurs (donc 1000 fois 100 tirages au sort), il nous dit que dans 97% des cas, *f* se trouvait bien dans l'intervalle **[0,2 ; 0,4]**. Il semblerait[10] que cet intervalle **[0,2 ; 0,4]** soit en effet un intervalle dans lequel *f* fluctue dans au moins 95% des cas. On appelle cela un **intervalle de fluctuation au seuil de** 95%.

[10] Si l'on y réfléchit bien, il nous faudrait ici un « méta intervalle de fluctuation ». Cela montre la limite de ces présentations « pédagogiques ». En définitive, on a beau faire des expériences et des simulations, la connaissance des choses reste probabiliste. C'est la théorie qui prime : elle permet de parler de l'expérience et sans elle, l'expérience ne dit rien.

Définition

On répète n fois une même expérience aléatoire à deux issues A et B, dans laquelle la probabilité d'obtenir l'issue A est toujours la même. Notons p cette probabilité. On note alors f la fréquence de l'issue A dans les n tirages. On l'appelle la **fréquence observée**. On appelle **intervalle de fluctuation au seuil de** 95% tout intervalle (inclus dans [0 ;1]) dans lequel f a au moins 95% de chances de se trouver.

Commentaires

Le plus souvent, on donnera pour intervalles de fluctuation des intervalles de centre p.

L'intervalle [0 ; 1] est un intervalle de fluctuation toujours valable, mais qui n'a aucun intérêt : on essaye d'obtenir des intervalles de fluctuation les moins larges possibles. Ce qu'on souhaiterait avoir, plus précisément, ce sont des formules qui donneraient, en fonction de n et de p un intervalle de fluctuation le plus serré possible.

Répétons qu'au lieu de 95%, on aurait pu choisir une autre valeur, par exemple 99%.

Quel intérêt ?

Déjà, cela nous donne une idée des erreurs qu'on risque de commettre en faisant un sondage. Mais nous verrons cet aspect plus en détail en abordant les intervalles de confiance. Sinon, donnons un exemple. On a une pièce de monnaie que l'on soupçonne d'être truquée au profit du pile. Théoriquement, la probabilité d'obtenir

pile est de $\frac{1}{2}$. Donc $p = 0{,}5$. C'est du moins notre hypothèse. On lance cette pièce 50 fois et l'on obtient 30 *pile*. Est-ce que l'écart par rapport aux 25 *pile* théoriques est assez significatif pour renforcer nos soupçons ? Est-il dû aux légitimes fluctuations du hasard ? On ne pourra pas répondre à coup sûr à cette question, mais il sera utile de savoir si l'on reste dans un intervalle de fluctuation raisonnable.

On peut étendre la notion d'intervalle de fluctuation à des cas où l'expérience aléatoire n'est pas une répétition de la même expérience. On peut l'étendre à toute loi de probabilité.

Formule

Lorsque $0{,}2 \leq p \leq 0{,}8$ et que $n \geq 25$, on obtient (approximativement) un intervalle de fluctuation au seuil de 95% par la formule : $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$.

Exemples

Reprenons l'exemple du vote pour le candidat Gudule. p est la probabilité de tomber sur un électeur de Gudule lorsqu'on tire un électeur au sort. On avait : $p = 0{,}3$. L'entier n, lui, est le nombre de tirages au sort. Il correspond ici à la taille de l'échantillon : $n = 100$. On vérifie aisément que p et n satisfont aux conditions exigées par le théorème. Il ne reste plus qu'à appliquer la formule. Un intervalle de fluctuation au seuil de 95% est : $\left[0{,}3 - \frac{1}{\sqrt{100}} \ ; \ 0{,}3 + \frac{1}{\sqrt{100}} \right]$, c'est-à-dire : **[0,2 ; 0,4]**. C'est l'intervalle que nous avions proposé

en exemple sans expliquer d'où il sortait. Nous l'avions obtenu grâce à cette formule.

Reprenons à présent l'exemple de la pièce que nous soupçonnions d'être truquée. On a $p=0{,}5$ et $n=50$. Les conditions sont satisfaites, on peut donc appliquer la formule, qui nous donne un intervalle à 95% de fluctuation de $\left[0{,}5-\frac{1}{\sqrt{50}}\ ;\ 0{,}5+\frac{1}{\sqrt{50}}\right]$, soit environ $[0{,}35\ ;\ 0{,}65]$. Avec une fréquence observée de 0,3 on sort de cet intervalle. On a de bonne raisons d'avoir des soupçons. (Mais bon, je trouve tout cela un peu vaseux, si vous voulez mon opinion personnelle.)

Commentaires

D'abord, cette formule n'est pas parfaitement vraie. En effet, pour certaines valeurs de n et de p, la probabilité d'être dans l'intervalle en question est légèrement inférieure à 95%. Par exemple, lorsque $n=201$ et $p=0{,}45$, cette probabilité n'est que de 0,948. D'où le « approximativement » du théorème.

Ensuite, on peut se demander où est passé le « 95% » dans cette formule. Ben justement, par souci de produire une formule très simple, on a un intervalle qui est parfois un peu trop large (d'autres formules un peu plus compliquées, donnent un intervalle plus serré) et parfois, ce qui est beaucoup plus grave, un peu trop étroit.

La démonstration de cette formule n'est pas vue en seconde.

L'avantage de cette formule un peu grossière est qu'elle donne un intervalle (centré en p et) dont l'amplitude ne dépend pas de p.

Résumé

Lorsqu'on répète suffisamment une même expérience aléatoire à deux issues A et B, la fréquence f réellement observée de l'événement A se rapproche de sa probabilité p. Les formules sur les intervalles de fluctuation nous renseignent sur la rapidité de ce rapprochement.

Intervalle de confiance

La notion d'**intervalle de confiance** est très proche de celle d'*intervalle de fluctuation* et au niveau de la seconde, on a toutes les raisons de confondre les deux. Essayons tout de même de les distinguer dès à présent.

Il y a deux valeurs. L'une, théorique, fixe, p. Et l'autre, obtenue par un processus aléatoire, la fréquence observée f. Si l'on connaît p et qu'on veut un intervalle où f a de bonnes chances de « fluctuer », on cherche un *intervalle de fluctuation*.

Mais c'est souvent le problème inverse, qui se pose : en général, on ne connaît pas p : on a juste un sondage et l'on veut estimer les chances qu'on a que la valeur de f obtenue par ce sondage ne s'éloigne pas trop de la valeur réelle p qu'on aimerait bien connaître (mais qu'on ne connaîtra que le jour du vote).

Un **intervalle de confiance au seuil de** 95% est un intervalle dépendant l'échantillon obtenu (imaginons-le centré en f pour nous aider à comprendre) et qui a au moins 95% de chances de contenir p. (Certes, p est fixe, ne résulte d'aucun tirage au sort, mais on ne le connaît pas.)

Théoriquement, si l'on dispose d'une formule donnant des intervalles de fluctuation, on doit pouvoir inverser cette formule pour avoir des intervalles de confiance.

La formule que nous avons vu s'inverse si facilement qu'on pourrait se demander l'intérêt de créer un concept distinct venant se surajouter à celui d'intervalle de fluctuation. En effet, $f \in \left[p-\frac{1}{\sqrt{n}} \ ; \ p+\frac{1}{\sqrt{n}}\right] \Leftrightarrow p \in \left[f-\frac{1}{\sqrt{n}} \ ; \ f+\frac{1}{\sqrt{n}}\right]$. En effet, dans les deux cas, l'appartenance dit que l'écart entre p et f est inférieur à $\frac{1}{\sqrt{n}}$. Donc, à la condition qu'on admette que p est compris entre 0,2 et 0,8 et à la condition que n soit supérieur à 25, l'intervalle $\left[f-\frac{1}{\sqrt{n}} \ ; \ f+\frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance au seuil de 95%.

Mais il existe des formules à la fois plus fines et plus compliquées, dans lesquelles le « rayon » de l'intervalle de fluctuation dépend de p. Par exemple, on voit en terminale la formule suivante (valable lorsque n est « assez grand », que $np>5$ et $n(1-p)>5$:
$\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \ ; \ p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

Le problème, c'est que la formule donnant un intervalle de fluctuation, il est difficile de l'inverser pour produire une formule donnant un intervalle de confiance. En effet : dans le second cas, on ne connaît pas p.

Il est vrai qu'un intervalle de fluctuation centré en p donne automatiquement un intervalle de confiance centré en f. Mais ce qu'on demande à l'intervalle de confiance, c'est qu'on soit capable d'en trouver un en ne connaissant pas p.