Statistiques

Table des matières

I Vocabulaire **1**

II Représentations graphiques **2**
II.1 Séries à caractère quantitatif discret . . . 2
II 1.1 Diagramme en b,tons . . . 2
II 1.2 Nuage de points . . . 3
II.2 Séries à caractère quantitatif continu . . . 3
II 2.1 Histogramme . . . 3
II 2.2 Polygone d'effectifs ou de fréquences cumulés . . . 4

III Paramètres statistiques **5**
III.1 Paramètres de position . . . 5
III 1.1 Mode . . . 5
III 1.2 Moyenne . . . 6
III 1.3 Médiane . . . 6
III 1.4 Quartiles . . . 7
III.2 Paramètres de dispersion . . . 7
III 2.1 Étendue . . . 7
III 2.2 Écart interquartile . . . 8

IV Fluctuation d'échantillonnage **9**
IV.1 Intervalle de fluctuation . . . 9
IV.2 Estimation d'une proportion . . . 10

I Vocabulaire

Une étude statistique commence par un recueil de données. On utilise le vocabulaire suivant pour décrire cette étude :

- **Série statistique** : Ensemble des valeurs collectées.
- **Population** : Ensemble sur lequel porte l'étude statistique.
- **Individus** : Éléments qui composent la population.
- **échantillon** : Partie de la population.
- **Caractère étudié** : Propriété que l'on observe sur les individus. Les différentes valeurs obtenues sont appelées **valeurs du caractère** ou **modalités**, souvent notées x_1, x_2,... x_p. On distingue deux types de caractères.
  - ◇ Un caractère peut être **qualitatif** (situation de famille, sexe...).
  - ◇ Un caractère peut être **quantitatif**. Il est dit **discret** lorsqu'il ne prend que des valeurs isolées (nombre d'enfants, notes dans une classe...). Il est dit **continu** lorsqu'il peut prendre théoriquement toutes les valeurs d'un intervalle (taille, temps d'écoute...) ; dans ce cas, les valeurs sont regroupées en sous-intervalles appelés des **classes**.

- **Effectif** : Pour une valeur du caractère (modalité ou classe), on appelle effectif le nombre d'individus de la population ayant cette valeur. On note souvent n_1, n_2,… n_p les effectifs respectifs des modalités x_1, x_2,… x_p.
- **Effectif total** : Nombre total d'individus de la population (ou de l'échantillon). Il est égal à $n_1 + n_2 + \ldots + n_p$, souvent noté N.
- **Fréquence** : Pour une valeur du caractère (modalité ou classe), on appelle fréquence le quotient de l'effectif de cette valeur par l'effectif total. On note souvent f_1, f_2,…, f_p les fréquences respectives des modalités x_1, x_2,… x_p, donc :

  $$f_1 = \frac{n_1}{N},\ f_2 = \frac{n_2}{N}, \ldots,\ f_p = \frac{n_p}{N}.$$

  On en déduit que $0 \leq f_1 \leq 1$, $0 \leq f_2 \leq 1$,…, $0 \leq f_p \leq 1$, et $f_1 + f_2 + \ldots f_p = 1$.
- **Valeurs extrêmes** : Valeurs minimales et maximales d'un caractère quantitatif.
- **Effectif cumulé** : Pour une valeur x d'une série statistique quantitative, l'effectif cumulé croissant (respectivement décroissant) de x est la somme des effectifs des valeurs inférieures (respectivement supérieures) ou égales à x.
- **Fréquence cumulée** : Pour une valeur x d'une série statistique quantitative, la fréquence cumulée croissante (respectivement décroissante) de x est la somme des fréquences des valeurs inférieures (respectivement supérieures) ou égales à x.

Exemple avec des notes :
Dans le tableau suivant sont regroupées les notes obtenues par les élèves d'une seconde lors du contrôle n° 1 (éventuellement arrondies pour simplifier l'étude) :

4	5	6	6	6	8	8	9	10	11	11	11	12	12	12	12	13
13	14	14	16	16	16	16	16	16	16	17	17	17	17	19	19	19

Dans cet exemple :

- La série statistique est l'ensemble des notes collectées.
- La population est l'ensemble des élèves de seconde .
- Les individus sont chacun des élèves de seconde Turner.
- Le caractère étudié est le résultat obtenu au contrà¥le n¬1.
- Les modalités sont les valeurs chiffrées des notes obtenues au contrà¥le n¬1.
- L'effectif total est le nombre d'élèves de la classe, à savoir 34.
- Les valeurs extrêmes sont 4 et 19.

Exemple avec des notes : Pour une meilleure lisibilité et pour simplifier l'étude, on peut commencer par compter le nombre d'individus ayant obtenu chaque note :

Note	4	5	6	8	9	10	11	12	13	14	16	17	19
Effectif	1	1	3	2	1	1	3	4	2	2	7	4	3
Fréquence à 10^{-2} près	0,03	0,03	0,09	0,06	0,03	0,03	0,09	0,12	0,06	0,06	0,21	0,12	0,09

Remarque

Dans le tableau précédent, la somme des fréquences est supérieure à 1 à cause des arrondis.

Exemple : série continue On a interrogé en 2008 un échantillon de 4812 Français concernant la durée hebdomadaire d'écoute de la télévision (en heures) :

Durée	[0 ; 0[	[10 ; 15[	[15 ; 20[	20 ; 30[	30 ; 50]
Effectif	972	924	826	1069	1021

Le caractère étudié, à savoir la durée d'écoute, est quantitatif continu : il peut prendre théoriquement toutes les valeurs de l'intervalle [0 ; 50]. Les données sont regroupées en classes [0 ; 10], [10 ; 15[, [15 ; 20[, [20 ; 30[et [30 ; 50].

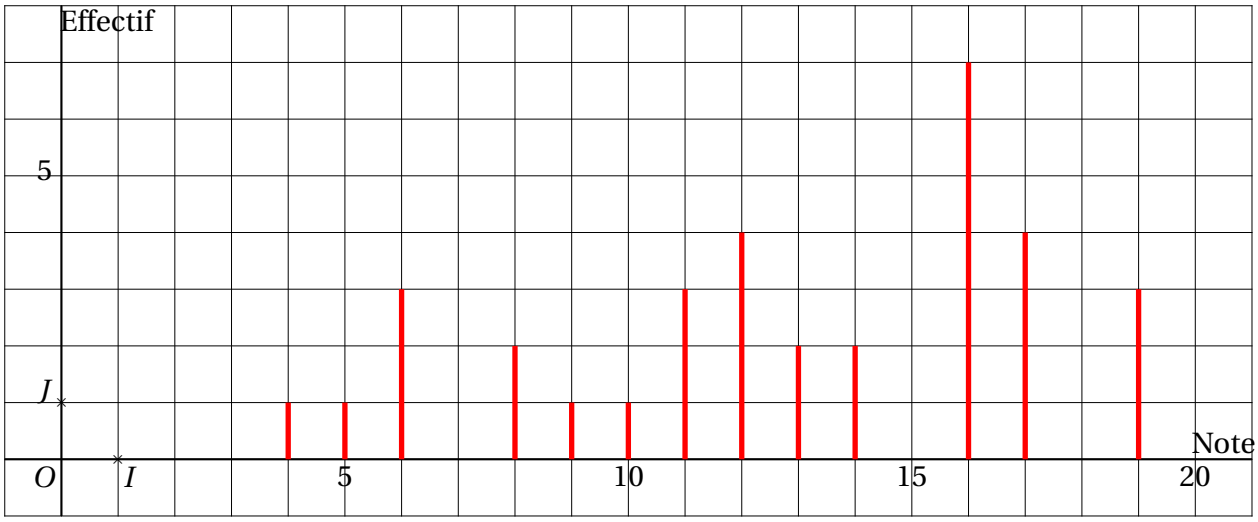
II Représentations graphiques

II.1 Séries à caractère quantitatif discret

II 1.1 Diagramme en b,tons

Dans un **diagramme en bâtons**, on représente une série statistique discrète par des segments dont la hauteur est proportionnelle à l'effectif de la valeur qu'ils représentent.

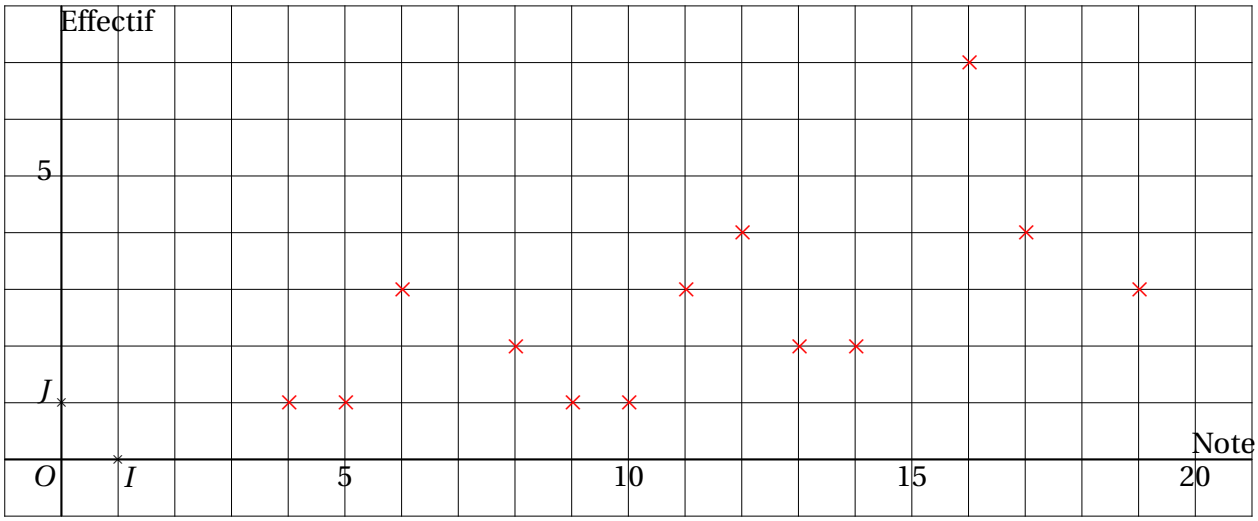
Exemple On continue à travailler avec les données de l'exemple sur les notes. Voici le diagramme en bâtons de cette série :



II 1.2 Nuage de points

Dans un **nuage de points**, on représente une série statistique discrète par des points dont les abscisses sont les valeurs du caractère, et les ordonnées sont les effectifs correspondants, parfois reliés par des segments.

Exemple On travaille toujours avec les données de l'exemple sur les notes. Voici le nuage de points de cette série :



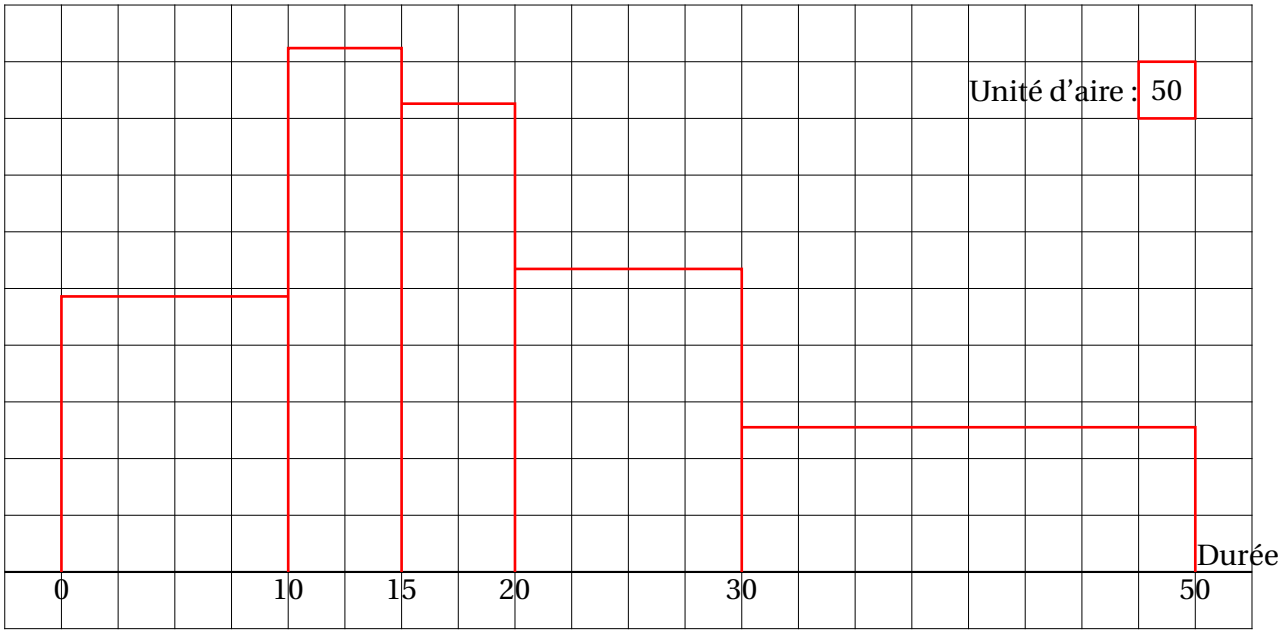
II.2 Séries à caractère quantitatif continu

II 2.1 Histogramme

Dans un **histogramme**, on représente une série statistique continue par des rectangles dont la largeur correspond à l'amplitude de chaque classe et dont l'aire est proportionnelle à l'effectif de la classe.

Exemple

On travaille avec les données de l'exemple sur la durée d'écoute de la télévision. Voici l'histogramme de cette série :



Remarque

Lorsque les classes ont toutes la même amplitude, la hauteur de chaque rectangle est proportionnelle à l'effectif de la classe qu'il représente. On dit alors que l'histogramme est à **pas constant**.

II 2.2 Polygone d'effectifs ou de fréquences cumulés

- Le **polygone des effectifs cumulés croissants** (respectivement **décroissants**) d'une série statistique continue est la ligne brisée qui joint les points du plan dont les abscisses sont les bornes de chaque classe et dont les ordonnées sont les effectifs cumulés croissants (respectivement décroissants) de ces valeurs.
- Le **polygone des fréquences cumulées croissantes** (respectivement **décroissantes**) d'une série statistique continue est la ligne brisée qui joint les points du plan dont les abscisses sont les bornes de chaque classe et dont les ordonnées sont les fréquences cumulées croissantes (respectivement décroissantes) de ces valeurs.

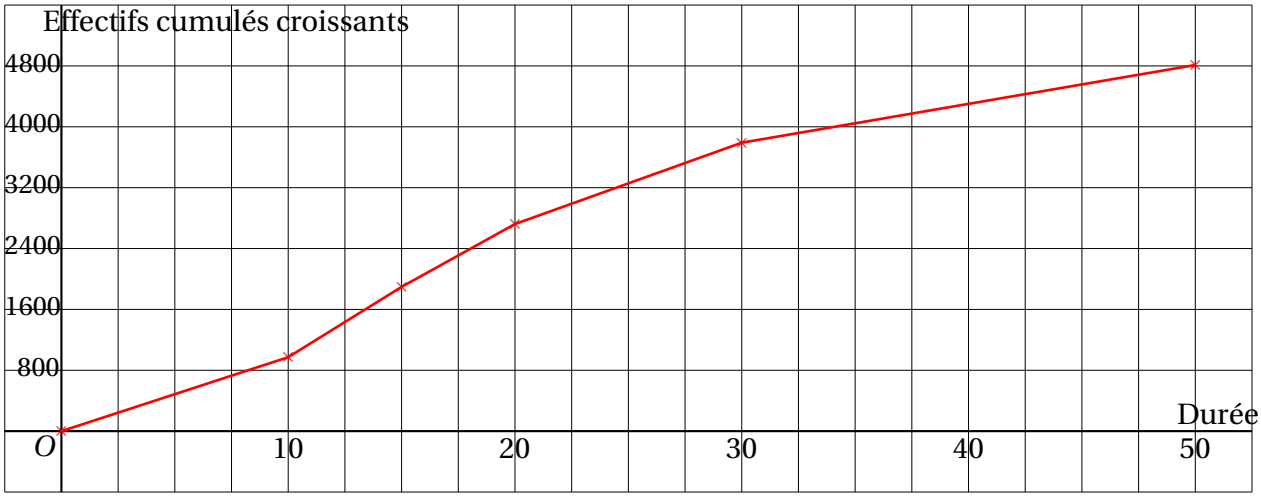
Ces représentations donnent l'allure de la répartition des valeurs de la série.

Exemple

La situation est toujours celle de l'exemple I. Le tableau des effectifs cumulés croissants est le suivant :

Durée <	0	10	15	20	30	50
Effectif	0	972	1896	2722	3791	4812

D'où le polygone des effectifs cumulés croissants :



Traitons à présent le cas des fréquences cumulées décroissantes :

Durée ≥	0	10	15	20	30	50
Effectif	4812	3840	2916	2090	1021	0
Fréquence	1	0,8	0,61	0,43	0,21	0

D'où le polygone des fréquences cumulées décroissantes :

III Paramètres statistiques

III.1 Paramètres de position

III 1.1 Mode

Définition :
Un **mode** d'une série statistique est une valeur de la série dont l'effectif est strictement supérieur à celui des autres valeurs.

Remarque

Dans une série statistique, il peut y avoir plusieurs modes.

Exemple

Dans l'exemple sur les notes, le mode est 16.

III 1.2 Moyenne

On considère une série statistique donnée par le tableau suivant :

Valeur	x_1	x_2	x_3	$\ldots$	x_{p-1}	x_p
Effectif	n_1	n_2	n_3	$\ldots$	n_{p-1}	n_p
Fréquence	f_1	f_2	f_3	$\ldots$	f_{p-1}	f_p

Définition

La **moyenne** de cette série statistique est le réel noté $\overline{x}$ défini par

$$\overline{x} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{n_1 + n_2 + \ldots n_p} = \frac{n_1x_1 + n_2x_2 + \ldots + n_px_p}{N}$$

en notant $N = n_1 + n_2 + \ldots + n_p$ l'effectif total de la série.

Propriété

On peut également calculer la moyenne à l'aide des fréquences :

$$\overline{x} = x1\,f_1 + x_2 f_2 + \ldots + x_p f_p.$$

Exemple

Avec les données de l'exemple I, la moyenne de la classe est :

$$\overline{x} = \frac{1\times 4+1\times 5+3\times 6+2\times 8+1\times 9+1\times 10+3\times 11+4\times 12+2\times 13+2\times 14+7\times 16+4\times 17+3\times 19}{34} = \frac{434}{34}$$

Une valeur approchée de $\overline{x}$ à 10^{-2} près est 12,76.

III 1.3 Médiane

Définition

La **médiane** M d'une série statistique est un réel qui partage cette série en deux parties telles que :

- Au moins 50 % des valeurs sont inférieures ou égales à la médiane ;
- Au moins 50 % des valeurs sont supérieures ou égales à la médiane.

Propriété

En pratique, on adopte la démarche suivante pour déterminer la médiane M d'une série statistiques d'effectif total N :

- On range d'abord les N valeurs du caractère par ordre croissant.
- Si N est pair, M est la moyenne des deux valeurs « centrales » de la série.
- Si N est impair, M est la valeur centrale de la série.

Exemple

Dans l'exemple I, l'effectif total est 34, c'est-à -dire pair. La médiane est donc la moyenne des deux valeurs centrales de la série, à savoir les 17^{e} et 18^{e} valeurs. Donc $M = \frac{13+13}{2} = 13$, ce qui signifie qu'au moins la moitié des notes est inférieure ou égale à 12 (en réalité 18 notes), et qu'au moins la moitié des notes est supérieure ou égale à 12 (en réalité 18 notes également).

Exemple

En France, en 2005, dans le secteur privé et semi-public (SNCF, Poste, Caisse d'épargne...), le salaire net mensuel médian est de 1528 €, alors que le salaire net mensuel moyen est de 1903,5 €.

Source : .

III 1.4 Quartiles

Définition

On considère une série statistique.

- Le premier **quartile** Q_1 est la plus petite valeur de la série telle qu'au moins 25 % des données soient inférieures ou égales à Q_1.
- Le troisième **quartile** Q_3 est la plus petite valeur de la série telle qu'au moins 75 % des données soient inférieures ou égales à Q_3.

Exemple

On considère toujours les données de l'exemple I.

- $34 \times \dfrac{25}{100} = 8,5$ donc le premier quartile Q_1 de la série est la 9evaleur, d'où $Q_1 = 10$, ce qui signifie qu'au moins un quart des notes sont inférieures ou égales à 10 (en réalité 9 notes, soit environ 26 %).
- $34 \times \dfrac{75}{100} = 25,5$ donc le troisième quartile Q_3 de la série est la 26e valeur, d'où $Q_3 = 16$, ce qui signifie qu'au moins trois quarts des notes sont inférieures ou égales à 16 (en réalité 27 notes, soit environ 79 %).

Remarque

Le fait que le partage théorique en 25 %, 50 % et 75 % de la série statistique à l'aide des indicateurs Q_1, M et Q_3 ne soit pas tout à fait exact provient du fait que la série comporte des valeurs identiques. Ce phénomène a tendance à s'amoindrir lors d'une étude sur une population plus importante avec un caractère dont les modalités sont plus disparates.

III.2 Paramètres de dispersion

III 2.1 Étendue

Définition L'**étendue** d'une série statistique est la différence entre sa plus grande et sa plus petite valeur.

Exemple

Dans l'exemple sur les notes, l'étendue est égale à $19 - 4 = 15$.

III 2.2 Écart interquartile

Définition

On considère une série statistique de premier quartile Q_1 et de troisième quartile Q_3.

- On appelle **intervalle interquartile** l'intervalle $[Q_1 \,;\, Q_3]$.
- On appelle **écart interquartile** la différence $Q_3 - Q_1$.

Exemple

Dans l'exemple I, l'intervalle interquartile est $[10 \,;\, 16]$ et l'écart interquartile est $Q_3 - Q_1 = 16 - 10 = 6$.

IV Fluctuation d'échantillonnage

IV.1 Intervalle de fluctuation

Lorsqu'on étudie un caractère d'une population, la connaissance de la population entière n'est pas toujours envisageable.
On doit alors se contenter de la connaissance d'une partie, convenablement choisie, de cette population qu'on appelle alors échantillon.

Définition

On appelle échantillon de taille n un sous-ensemble de la population, c'est-à-dire la liste de n résultats obtenus par répétition de n expériences indépendantes.

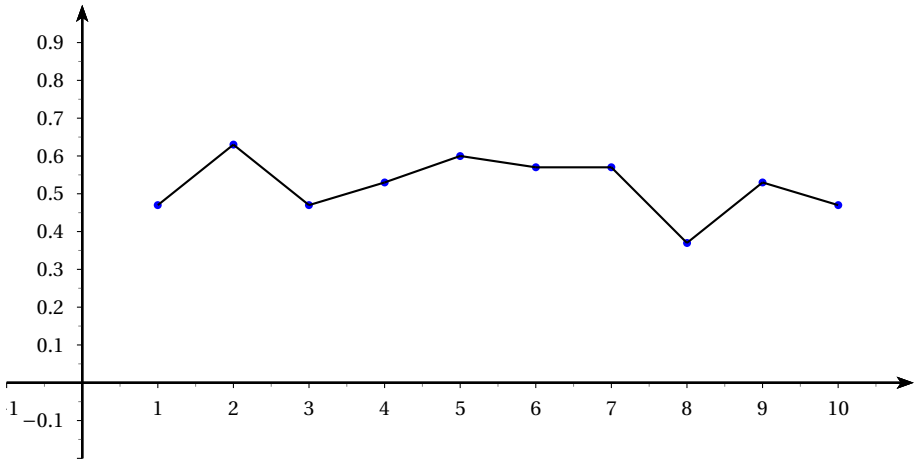
Exemples :

- On lance 100 fois un dé et on note la liste des 100 résultats.
- On prélève 100 ampoules d'une chaîne de fabrication et on note si elles sont conformes ou pas.
- On interroge 100 personnes et on note leur boisson au petit-déjeuner.
- Une urne contient des boules vertes, rouges et bleues.
  On tire au hasard une boule de l'urne, on note sa couleur puis on la remet, et on répète ceci 4 fois de suite.
  On obtient alors un échantillon de taille 4, par exemple V-R-B- V.

Remarque : Si on réalise plusieurs fois de suite une même expérience, on n'obtiendra pas à chaque fois la même liste de résultats.
Imaginons que chaque élève de la classe lance trente fois une pièce de monnaie, regarde les résultats Pile ou Fac et calcule la fréquence d'apparition du résultats « Pile ».
Les résultats risquent d'être différents d'un élève à l'autre.
C'est ce qu'on appelle fluctuation d'échantillonnage. Pour dix échantillons, on pourrait par exemple obtenir :

Échantillon	1	21314	5	6	7	8	9	10
Fréquence	0.47	0.63	0.47	0.53	0.60	0.57	0.37	0.53

Diagramme :



Définition

On appelle épreuve de Bernoulli une expérience aléatoire (dont le résultat survient au hasard) qui n'a que deux issues possibles.
(Jakob Bernoulli, mathématicien suisse, 1654-1705)

Intervalle de fluctuation

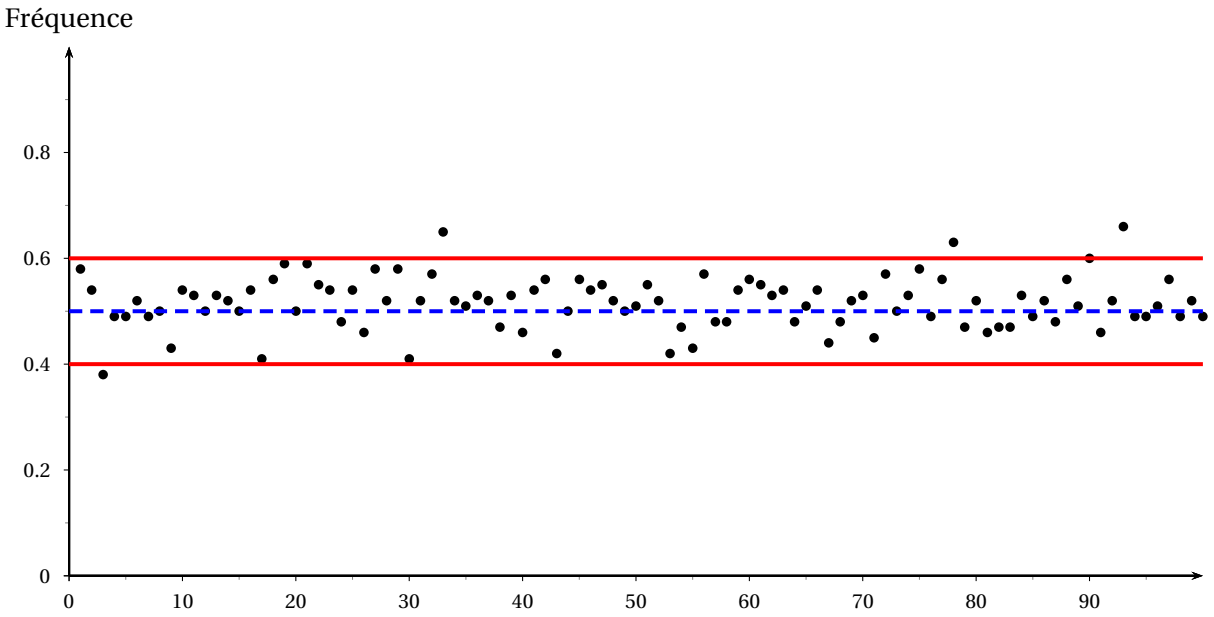
On étudie un échantillon d'épreuves de Bernoulli et on s'intéresse à l'une des deux issues (par exemple, obtenir Pile dans un tirage de Pile ou Face).
On note p la probabilité que cette issue se réalise.

Propriété

Si on analyse un grand nombre d'échantillons de taille n ($n \geqslant 25$) et que l'on observe à chaque fois la fréquence d'apparition f de l'issue choisie, si la probabilité p est comprise entre 0,2 et 0,8, au moins 95 % des fréquences se situent dans un intervalle $I = \left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right)$.

Cela signifie qu'il est rare (5 %) des cas de trouver un échantillon dans lequel la fréquence n'est pas dans l'intervalle de fluctuation. On peut alors considérer que cet échantillon n'est pas « normal » avec un risque de se tromper assez faible (5 % des cas).

Exemple : on lance 100 fois 100 dés et on note pour chaque série de 100 lancers la fréquence d'apparition de Pile.
$f = 0,5$, $\sqrt{100} = 10$ donc l'intervalle de fluctuation est $[0,4 \ ; \ 0,6]$



IV.2 Estimation d'une proportion

On cherche une estimation d'une proportion dans une population statistiquement inconnue, à partir d'un échantillon (c'est le principe des sondages).

Exemple : On s'intéresse à la proportion P d'un caractère dans une population.
Cette proportion P est supposée inconnue mais on sait que dans un échantillon de taille 100, la fréquence mesurée du caractère est $f = 0.58$.

Question : Quelle est une estimation de P ?

Réponse :

On peut penser que P est du même ordre de grandeur que f ;

Comme $0,2 \leqslant p \leqslant 0,8$ et $n = 100 \geqslant 25$ donc dans 95 % des cas au moins, $f \in \left[p - \frac{1}{\sqrt{100}} \ ; \ p + \frac{1}{\sqrt{100}}\right]$ donc il vient :

$p - \frac{1}{\sqrt{100}} \leqslant 0,58 \leqslant p + \frac{1}{\sqrt{100}}$

$p - \frac{1}{\sqrt{100}} \leqslant 0,58$ donne $p \leqslant 0,58 + \frac{1}{\sqrt{100}} = 0,68$

$0,58 \leqslant p + \frac{1}{\sqrt{100}}$ donne $0,58 - \frac{1}{\sqrt{100}} \leqslant p$ donc $0,48 \leqslant p$.

Ainsi $\boxed{p \in [0,48 \ ; \ 0,58]}$.

Cas général :

Si $0,2 \leqslant p \leqslant 0,8 et n \geqslant 25$, alors, dans 95 % des cas, $p \in \left[f - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$.

Cet intervalle est appelé **intervalle de confiance** au seuil de 95 %.

Résumé

Intervalle de fluctuation au seuilvde 95 %	Intervalle de confiance au seuil de 95 %
On connaît : la proportion p des individus ayant un caractère donné au sein d'une population	**On connaît** : la proportion f des individus ayant un caractère donné au sein d'un échantillon de taille n de la population.
On cherche : à savoir si un échantillon de taille n dans lequel la fréquence f de ce caractère est , est normal.	**On cherche** : à estimer la proportion p des individus ayant ce caractère donné au sein de la population totale.
Probablement si $f \in \left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right]$	Probablement que $p \in \left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right]$

Exemple : Le parti d'un candidat commande un sondage réalisé à partir de 1 600 personnes à l'issue duquel il est donné gagnant avec 52 % des voix.
A-t-il des raisons d'être confiant ?

Réponse :

Soit p la proportion de gens votant pour lui ; il est élu si $p \geqslant 0,5$.

D'près l'intervalle de confiance, $p \in \left[f - \frac{1}{\sqrt{1\,600}} \ ; \ f + \frac{1}{\sqrt{1\,600}}\right] = [0,495 \ ; \ 0,545]$ au seuil de 95%, donc il n'est pas sûr d'être élu.