Statistiques

I. Présentation et Vocabulaire de base

Toute étude statistique s'appuie sur des données. Dans le cas ou ces données sont numériques (95% des cas), on distingue les données discrètes (qui prennent un nombre fini de valeurs : par ex, le nombre de voitures par famille en France) des données continues (qui prennent des valeurs quelconques : par ex, la taille des animaux d'un zoo).

- Dans le cas d'une série discrète, le nombre de fois ou l'on retrouve la même valeur s'appelle l'effectif de cette valeur. Si cet effectif est exprimé en pourcentage, on parle alors de fréquence de cette valeur.
- Dans le cas d'une série continue, on répartit souvent les données par classes.

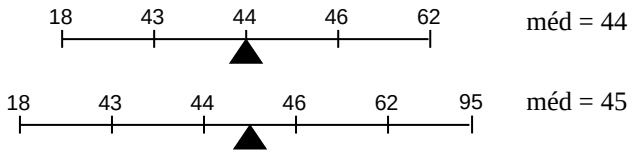
Le but des statistiques est d'analyser les données dont on dispose. Pour cela, on peut par exemple chercher à déterminer la moyenne ou la médiane de la série. De tels nombres permettent notamment de comparer plusieurs séries entre elles. On les appelle indicateurs statistiques ou paramètres statistiques. On distingue les indicateurs de position (qui proposent une valeur "centrale" de la série) et les indicateurs de dispersion (qui indiquent si la série est très regroupée autour de son "centre" ou non).
Ainsi, le mode d'une série (valeur qui a le plus grand effectif de la série) est un indicateur de position.
L'étendue de cette série (différence entre la plus grande et la plus petite valeur) est un indicateur de dispersion.
La moyenne et la médiane sont des indicateurs de position.

De plus, lorsque la série est trop importante (population d'un pays…), on est obligé de faire un sondage, c'est à dire de restreindre l'étude à un échantillon de cette série. Tout le problème est alors de choisir un échantillon vraiment représentatif (de taille suffisante et non biaisé) et d'évaluer l'erreur commise par rapport à une étude qui porterait sur l'ensemble de la série. (exemple des sondages électoraux)

II. Médiane

Définition
Soit une série statistique d'effectif total n, rangée par ordre croissant.
On appelle médiane la valeur "du milieu". On dit qu'elle partage la série en deux moitiés : il y a autant de valeurs en dessous qu'au dessus.
Pour déterminer son rang, il y a 2 cas :

- si n est impair : la médiane est la valeur de rang $\frac{n+1}{2}$
- si n est pair : nous prendrons la demi-somme des deux valeurs dont les rangs entourent le nombre $\frac{n+1}{2}$



Remarque :
Si les données ont été regroupées en classes, on ne peut déterminer la valeur exacte de la médiane. En revanche, on appellera classe médiane, la classe qui la contient (et permet donc d'en donner un encadrement).

Exemples
Données discrètes "en vrac"
10, 7, 12, 18, 16, 15, 5, 11, 11, 20, 15, 11, 18, 14
Ordonnons la série par ordre croissant : 5, 7, 10, 11, 11, 11, 12, 14, 15, 15, 16, 18, 18, 20
Il y a 14 termes or $\frac{14+1}{2} = 7{,}5$.

La médiane est donc la demi somme des $7^{ème}$ et $8^{ème}$ termes : méd $= \frac{12+14}{2} = 13$

Avec un tableau d'effectifs

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés	6	17	42	61	76	81

Attention, il faut bien interpréter cette dernière ligne : Les données qui valent 3 ont un rang compris entre 18 et 42 inclus

L'effectif total est de 81 or $\frac{81+1}{2} = 41$. La médiane est donc le $41^{ème}$ terme : méd = 3

Avec des données réparties par classes

classe	[0 ; 2[	[2 ; 4[	[4 ; 6[	[6 ; 8]
fréquence	10%	38%	45%	7%
fréquence cumulée	10%	48%	93%	100%

48% des valeurs sont strictement inférieures à 4
Et 93% des valeurs sont strictement inférieures à 6
La classe médiane est donc la classe [4 ; 6[
On peut donc en déduire l'encadrement suivant 4 < méd < 6

III. Quartiles

Définitions :
Le premier quartile Q_1 est la plus petite des valeurs de la série telle qu'au moins 25% de la population ait sa valeur inférieur ou égale à Q_1.
Le troisième quartile Q_3 est la plus petite des valeurs de la série telle qu'au moins 75% de la population ait sa valeur inférieure ou égale à Q_3.

Méthode pour trouver Q_1 et Q_3.
Pour une population d'effectif n, $\frac{n}{4}$ (ou $\frac{3n}{4}$) si il est un entiers nous donne le rang de Q_1 (ou Q_3), dans le cas contraire on prendra la valeur de l'élément du rang immédiatement supérieur.

Exemple
Avec un tableau d'effectifs

valeurs	1	2	3	4	5	6
effectifs	6	11	25	19	15	5
effectifs cumulés	6	17	42	61	76	81

Ici n = 81 Donc $\frac{n}{4} = 20{,}25$ il va donc falloir que je prenne la valeur de l'individu de rang 21, donc $Q_1 = 3$
$\frac{3n}{4} = 60{,}75$ L'individu de rang 61 a pour rang 5 donc $Q_3 = 4$

IV. Moyenne

Exemple :
Soit la série statistique ci-contre :

valeurs	0	1	2	3	4
effectifs	1	2	1	4	2

La moyenne est : $\bar{x} = \frac{0+1+1+2+3+3+3+3+4+4}{1+2+1+4+2} = \frac{24}{10} = 2{,}4$

On préférera écrire : $\bar{x} = \frac{0+2\times1+2+4\times3+2\times4}{1+2+1+4+2} = \frac{24}{10} = 2{,}4$

Notation.
Le symbole Σ (sigma) signifie que l'on ajoute les éléments, par exemple

$$\sum_{i=1}^{5} x_i = x_1 + x_2 + x_3 + x_4 + x_5$$

Formule pour le calcul de la moyenne
Soit la série statistique ci-contre :

valeurs	x_1	x_2	…	x_p
effectifs	n_1	n_2	…	n_p

La <u>moyenne</u> est : $\bar{x} = \frac{n_1x_1+n_2x_2+\ldots+n_{N-1}x_{N-1}+n_Nx_N}{n_1+n_2+\ldots+n_{N-1}+n_N} = \frac{\sum_{i=1}^{N} n_ix_i}{\sum_{i=1}^{N} n_i}$

Remarque :
Si les données ont été regroupées en classes, on ne peut calculer la valeur exacte de la moyenne. On peut toutefois en déterminer une bonne approximation en remplaçant chaque classe par son milieu.

Exemples :
a) Tableau de fréquences

valeurs	12	13	14	15	16
fréquences	0,05	0,17	0,43	0,30	0,05

$\bar{x}$ =

b) Données réparties en classes

Milieu de classe				
classes	[0 ; 5[	[5 ; 10[	[10 ; 15[	[15 ; 20]
effectifs	7	12	14	2

Remplaçons chaque classe par son milieu : $\frac{0+5}{2} =$ ……

$\bar{x} = \frac{7\times\ldots\ldots+12\times\ldots\ldots+14\times\ldots\ldots+2\times\ldots\ldots}{\ldots\ldots} \approx$

V. Propriétés de la moyenne

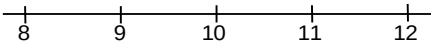
a) Addition ou Multiplication de toutes les données par un même nombre :
Ex : Soit la série : 10, 12, 14. $\bar{x}$ =
Ajoutons 2 : la nouvelle série est : 12, 14, 16. $\bar{x}$ =
Multiplions par $\frac{1}{2}$: la nouvelle série est : 6, 7, 8. $\bar{x}$ =

Théorème de linéarité de la moyenne.
Si toutes les valeurs d'une série sont multipliées par un nombre *k* alors la moyenne est aussi multipliée par *k*.
Si on ajoute un nombre *k* à toutes les valeurs d'une série alors la moyenne est augmentée de *k*.
b) Moyennes partielles
Ex : Sur les 5 premières interros, Paul a eu 12,5 de moyenne. Il vient d'avoir 15,5 à la 6ème interro.
Les notes ayant toutes le même coefficient, quelle est sa nouvelle moyenne ?
La somme des notes des 5 premières interros est : 5× 12,5
La somme des notes des 6 interros est donc : 5×12,5 + 15,5
La nouvelle moyenne est donc : $\bar{x} = \frac{5\times12,5+15,5}{6} = 13$

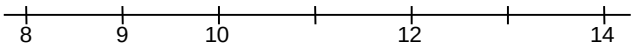
Cas général : Si on réunit deux groupes disjoints ayant respectivement pour moyennes et effectifs, $\bar{x}_1$ et n_1 d'une part, $\bar{x}_2$ et n_2 d'autre part, la moyenne de l'ensemble sera alors :

$$\bar{x} = \frac{n_1 \times \bar{x_1} + n_2 \times \bar{x_2}}{n_1 + n_2}$$

c) Lien entre la moyenne et la médiane
- Quand on modifie les valeurs extrêmes d'une série, la moyenne change contrairement à la médiane qui ne change pas. On dit que la moyenne est "sensible aux valeurs extrêmes".
  Il arrive que certaines de ces valeurs extrêmes soient douteuses ou influent de façon exagérée sur la moyenne. On peut alors, soit calculer une moyenne élaguée (c'est à dire recalculer la moyenne sans ces valeurs gênantes), soit utiliser la médiane.
- Comment interpréter un écart entre la moyenne et la médiane ?

  Soit la série suivante : 

  Ici la moyenne (10) et la médiane (10) sont identiques : la série est bien "centrée".

  Soit la nouvelle série : 

  Ici la moyenne (10,6) est plus importante que la médiane (10) : la série est plus "étalée à droite".

VI. Mode et classe modale

Définition
Un mode est une valeur du caractère ayant le plus grand effectif.
Une classe modale est une classe du caractère ayant le plus grand effectif.

Dans l'exemple précédent :
Le mode est 3 car l'effectif correspondant est 25 ce qui est le plus gros effectif

Dans le III exemple b) la classe modale est [10 ;15[car son effectif , 14, est le plus grand des effectifs.

VII. Fluctuation d'échantillonnage et simulation

Définition :
Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience.

Exemples :
On peut lancer une dizaine de fois un dès, et noter les résultats. L'indépendance est donnée par le fait qu'aucun tirage n'influence un autre.
On peut tirer des boules d'une urne, pour avoir l'indépendance il est impératif que l'on replace à chaque fois la boule tirée dans l'urne (sinon l'expérience ne cesse d'évoluer)
On peut sélectionner au hasard une partie d'une grande population, comme il n'y a pas de remise (on ne prends pas une personne au hasard, puis on la replace dans la population, puis on sélectionne une deuxièmes personne, on la replace etc …) on pourrait penser qu'il n'y a pas d'indépendance, cependant si l'échantillon est très petit par rapport à l'effectif de la population d'où on le tire, on pourra faire comme si.

Soit une population dont la proportion d'apparition d'une caractéristique donnée est p ($p\epsilon[0,1]$). Pour chaque échantillon la fréquence d'apparition sera notée f, on notera que pour chaque échantillon f aura une nouvelle valeur, on appelle ce phénomène **la fluctuation d'échantillonnage**.

La grande partie des f trouvées seront proches de p, et on pourra remarquer que plus l'effectif de l'échantillon est important plus les f trouvées sont proches de p.

Définition / Propriété
Lorsque l'on considère un échantillon de taille n d'une population l'intervalle $[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}]$ est appelé **intervalle de fluctuation de la fréquence f au seuil de 95%**.
L'échantillon est **représentatif (ou non biaisé)** si et seulement si f sa fréquence d'apparition de la caractéristique est dans cet intervalle.

Exemple
Les entreprises sont sensées ne pas faire de discrimination quant au sexe des personnes employées. Deux entreprises A et B ont respectivement 41 femmes pour 100 employés et 4 850 femmes sur 10 000 employés. On supposera que pour chaque poste il y avait autant de candidats que de candidates. Deviner si on peut vraisemblablement penser que la sélection s'est faite de manière équitable. Vérifiez par le calcul.
Pour l'entreprise A l'intervalle de fluctuation de la fréquence au seuil de 95% est :
$\left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right] = \left[0{,}5-\frac{1}{\sqrt{100}}; 0{,}5+\frac{1}{\sqrt{100}}\right] = [0{,}4; 0{,}6]$ or $f = \frac{41}{100} = 0{,}41$ et on a $0{,}41 \in [0{,}4; 0{,}6]$ donc l'échantillon est représentatif d'une situation de parité.
Pour l'entreprise B , l'intervalle de fluctuation est $\left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right] = \left[0{,}5-\frac{1}{\sqrt{10\,000}}; 0{,}5+\frac{1}{\sqrt{10\,000}}\right] \supset [0{,}49; 0{,}51]$
or ici $f = \frac{4850}{10000} = 0{,}485$ et 0,485 n'appartient pas à l'intervalle, donc l'échantillon n'est pas représentatif d'une situation de parité. C'est le contraire de ce que l'on pouvait penser

Inversement
Lors d'un sondage, on ne peut interroger toute la population, donc on ne peut connaître p (il faudra attendre le jour du vote pour cela). On se concentrera sur un échantillon non biaisé, on est sur à 95% que l'on a $p-\frac{1}{\sqrt{n}} \leq f \leq p+\frac{1}{\sqrt{n}} \Leftrightarrow p \leq f+\frac{1}{\sqrt{n}} \leq p+\frac{2}{\sqrt{n}}$
et $p-\frac{2}{\sqrt{n}} \leq f-\frac{1}{\sqrt{n}} \leq p \Leftrightarrow f-\frac{1}{\sqrt{n}} \leq p \leq f+\frac{1}{\sqrt{n}}$

Définition / Propriété
Soit une population dont on veut connaître p le pourcentage d'occurrence d'une propriété, et soit un échantillon représentatif de la population présentant la fréquence f. Alors l'intervalle $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ a une probabilité d'au moins 95% de contenir p.
Cet intervalle est appelé l'intervalle de confiance de p au niveau de confiance 0,95 (ou au risque de 5%)

Exemple
Un candidat Y est crédité 45% d'intention de vote lors d'un sondage fait sur 500 personnes.
En admettant que l'échantillon de personnes sondées est représentative de la population des votants dire donnez une fourchette contenant p sure à 95%
$f = 0{,}45$ n = 500 donc $\frac{1}{\sqrt{n}} \approx 0{,}04472$ $f-\frac{1}{\sqrt{n}} \approx 0{,}40528$ et $f+\frac{1}{\sqrt{n}} \approx 0{,}49472$
p sera donc compris entre 40,53% et 49,47%. Ces bornes sont des approximations, pour ne pas prendre de risque on préférera donner un intervalle légèrement plus petit qu'en donner un trop grand, on arrondira par excès la plus petite borne et par défaut la plus grande.

VII. Fluctuation d'échantillonnage et simulation

Définition :
Un échantillon de taille n est constitué des résultats de n répétitions indépendantes de la même expérience.

Exemples :
On peut lancer une dizaine de fois un dès, et noter les résultats. L'indépendance est donnée par le fait qu'aucun tirage n'influence un autre.
On peut tirer des boules d'une urne, pour avoir l'indépendance il est impératif que l'on replace à chaque fois la boule tirée dans l'urne (sinon l'expérience ne cesse d'évoluer)
On peut sélectionner au hasard une partie d'une grande population, comme il n'y a pas de remise (on ne prends pas une personne au hasard, puis on la replace dans la population, puis on sélectionne une deuxièmes personne, on la replace etc …) on pourrait penser qu'il n'y a pas d'indépendance, cependant si l'échantillon est très petit par rapport à l'effectif de la population d'où on le tire, on pourra faire comme si.

Soit une population dont la proportion d'apparition d'une caractéristique donnée est p ($p\epsilon[0,1]$). Pour chaque échantillon la fréquence d'apparition sera notée f, on notera que pour chaque échantillon f aura une nouvelle valeur, on appelle ce phénomène **la fluctuation d'échantillonnage**.

La grande partie des f trouvées seront proches de p, et on pourra remarquer que plus l'effectif de l'échantillon est important plus les f trouvées sont proches de p.

Définition / Propriété
Lorsque l'on considère un échantillon de taille n d'une population l'intervalle $[p-\frac{1}{\sqrt{n}}, p+\frac{1}{\sqrt{n}}]$ est appelé **intervalle de fluctuation de la fréquence f au seuil de 95%**.
L'échantillon est **représentatif (ou non biaisé)** si et seulement si f sa fréquence d'apparition de la caractéristique est dans cet intervalle.

Exemple
Les entreprises sont sensées ne pas faire de discrimination quant au sexe des personnes employées. Deux entreprises A et B ont respectivement 41 femmes pour 100 employés et 4 850 femmes sur 10 000 employés. On supposera que pour chaque poste il y avait autant de candidats que de candidates. Deviner si on peut vraisemblablement penser que la sélection s'est faite de manière équitable. Vérifiez par le calcul.
Pour l'entreprise A l'intervalle de fluctuation de la fréquence au seuil de 95% est :
$\left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right] = \left[0{,}5-\frac{1}{\sqrt{100}}; 0{,}5+\frac{1}{\sqrt{100}}\right] = [0{,}4; 0{,}6]$ or $f = \frac{41}{100} = 0{,}41$ et on a $0{,}41 \in [0{,}4; 0{,}6]$ donc l'échantillon est représentatif d'une situation de parité.
Pour l'entreprise B , l'intervalle de fluctuation est $\left[p-\frac{1}{\sqrt{n}}; p+\frac{1}{\sqrt{n}}\right] = \left[0{,}5-\frac{1}{\sqrt{10\,000}}; 0{,}5+\frac{1}{\sqrt{10\,000}}\right] \supset [0{,}49; 0{,}51]$
or ici $f = \frac{4850}{10000} = 0{,}485$ et 0,485 n'appartient pas à l'intervalle, donc l'échantillon n'est pas représentatif d'une situation de parité. C'est le contraire de ce que l'on pouvait penser

Inversement
Lors d'un sondage, on ne peut interroger toute la population, donc on ne peut connaître p (il faudra attendre le jour du vote pour cela). On se concentrera sur un échantillon non biaisé, on est sur à 95% que l'on a $p-\frac{1}{\sqrt{n}} \leq f \leq p+\frac{1}{\sqrt{n}} \Leftrightarrow p \leq f+\frac{1}{\sqrt{n}} \leq p+\frac{2}{\sqrt{n}}$
et $p-\frac{2}{\sqrt{n}} \leq f-\frac{1}{\sqrt{n}} \leq p \Leftrightarrow f-\frac{1}{\sqrt{n}} \leq p \leq f+\frac{1}{\sqrt{n}}$

Définition / Propriété
Soit une population dont on veut connaître p le pourcentage d'occurrence d'une propriété, et soit un échantillon représentatif de la population présentant la fréquence f. Alors l'intervalle $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ a une probabilité d'au moins 95% de contenir p.
Cet intervalle est appelé l'intervalle de confiance de p au niveau de confiance 0,95 (ou au risque de 5%)

Exemple
Un candidat Y est crédité 45% d'intention de vote lors d'un sondage fait sur 500 personnes.
En admettant que l'échantillon de personnes sondées est représentative de la population des votants dire donnez une fourchette contenant p sure à 95%
$f = 0{,}45$ n = 500 donc $\frac{1}{\sqrt{n}} \approx 0{,}04472$ $f-\frac{1}{\sqrt{n}} \approx 0{,}40528$ et $f+\frac{1}{\sqrt{n}} \approx 0{,}49472$
p sera donc compris entre 40,53% et 49,47%. Ces bornes sont des approximations, pour ne pas prendre de risque on préférera donner un intervalle légèrement plus petit qu'en donner un trop grand, on arrondira par excès la plus petite borne et par défaut la plus grande.