

Quelques définitions en statistique

1-Définition de la statistique :est la discipline qui étudie des phénomènes à travers la collecte de données, leur traitement, leur analyse, l'interprétation des résultats et leur présentation afin de rendre ces données compréhensibles par tous. C'est à la fois une branche des mathématiques appliquées², une méthode et un ensemble de techniques

2-Un ensemble statistique :est une abstraction qui consiste à considérer une collection de copies virtuelles (ou *répliques*) d'un système physique dans l'ensemble des états accessibles où il est susceptible de se trouver, compte tenu des contraintes extérieures qui lui sont imposées, telles le volume, le nombre de particules, l'énergie et la température. Cette notion, introduite par le physicien américain Josiah Willard Gibbs en 1902, est un concept central de la physique statistique.

3-Une unité statistique :est une **unité d'observation ou de mesure pour laquelle des données sont recueillies ou dérivées**. L'unité statistique est donc l'élément de base pour compiler et tabuler les données statistiques.

Selon le cas, l'*unité statistique* peut être un individu, un ménage, une entreprise, un établissement, une commune, un département, une région ou encore un pays.

Ex : un étudiant de la 2^e année

4-Caractère statistique (ou variables statistiques) : C'est ce qui est observé ou mesuré sur les individus d'une population statistique. Il peut s'agir d'une variable qualitative ou quantitative. Exemples : Taille, poids, salaire, sexe, profession d'un groupe donné d'individus.

5-On appelle série statistique : le résultat d'une étude menée sur une certaine population, visant à mesurer la présence d'un certain caractère au sein de cette population.

6-Classe : On regroupe plutôt ensemble les personnes ayant un revenu dans des intervalles donnés. Ces intervalles sont appelés classes de la série statistique. Plus généralement, une classe est donc **un intervalle de valeurs prises par un caractère quantitatif**.

7-En statistique, la modalité : est **l'une des différentes variantes que peut présenter un élément d'une population relativement à une variable qualitative**. Par exemple, une pièce est bonne ou défectueuse (2 modalités) et le mois de naissance est janvier, février, etc

8-Définition de la bio statistique : consiste à appliquer un traitement statistique à des données biologiques. C'est un champ interdisciplinaire qui associe la biologie, l'informatique et les mathématiques statistiques.

9-Fréquences absolues / relatives

Il est très important de distinguer :

• **La fréquence absolue** : c'est le résultat brut du comptage.

• **La fréquence relative** : c'est la fréquence absolue divisée par la taille de l'entité. C'est toujours un nombre inférieur à 1 (rappel : $1\% = 0,01$; $1\text{‰} = 0,001$; $1 \text{ ‰} = 0,0001\text{‰‰}$; $1\text{ppm} = 0,000001$; etc.).

Quand les entités sont de taille différente, il ne faut jamais comparer les fréquences absolues entre elles, car cela n'a aucun sens ; on peut uniquement comparer les fréquences relatives.

La statistique descriptive

1-Définition : est la branche des statistiques qui regroupe les nombreuses techniques utilisées pour *décrire* un ensemble *relativement important* de données.

2-Objectif de la statistique descriptive : est de décrire, c'est-à-dire de résumer ou représenter, par des statistiques, les données disponibles quand elles sont nombreuses.

Toute description d'un phénomène nécessite d'**observer** ou de **connaître** certaines choses sur ce phénomène.

- Les **observations disponibles** sont toujours constituées d'ensemble d'observations *synchrones*. Par exemple : une température, une pression et une mesure de densité à un instant donné dans une cuve précise. Ces trois variables synchrones peuvent être observées plusieurs fois (à plusieurs dates) en plusieurs lieux (dans plusieurs cuves).
- Les **connaissances disponibles** sont quant à elles constituées de *formules* qui relient certaines variables. Par exemple, la loi des gaz parfaits.

3-Description intrinsèque des observations

Sans aucun *a priori* sur la question que l'on se pose, quelques statistiques simples permettent de décrire les observations :

- la moyenne ;
- la médiane ;
- le mode ;
- le maximum ;
- le minimum ;
- l'écart type (et la variance) ;
- les quartiles.

Les trois premiers sont souvent nommés **critères de position**(Les caractéristiques de position sont les quartiles, les quintiles, les déciles et les centiles. Les quartiles partagent l'effectif total en quatre parties égales. Les quintiles partagent l'effectif total en cinq parties égales. Quant aux déciles, elles partagent l'effectif total en dix parties égales.), et les autres entrent plutôt dans la catégorie des **critères de dispersion**(On appelle dispersion statistique, la tendance qu'ont les valeurs de la distribution d'un caractère à s'étaler, à se disperser les unes par rapport aux autres ou de part et d'autre d'une valeur centrale).

Les quatre paramètres de dispersion absolue les plus courants sont l'étendue, l'intervalle interquartiles, l'écart absolu moyen et l'écart type

- **Moyenne.**

La moyenne arithmétique est la somme des valeurs de la variable divisée par le nombre d'individus :

$$m = \frac{\sum X}{n} \quad m = \text{moyenne arithmétique}$$

Σ = somme de

X = ème observation

n = nombre d'observations

Exemple.- La moyenne arithmétique des fréquences cardiaques des 29 filles de la troisième année: 80, 60, 70, 80, 50, 80, 80, 80, 80, 50, 60, 80, 90, 80, 50, 70, 80, 70, 80, 50, 80, 60, 90, 80, 80, 80, 70, 60, 80, 70.

- **Médiane (centre).**

La médiane est la valeur centrale qui partage l'échantillon en deux groupes de même effectif : 50 % au-dessus et 50 % en dessous. La médiane peut avoir une valeur différente de la moyenne. En France (et d'ailleurs dans la plupart des pays), le salaire médian est inférieur au salaire moyen : il y a beaucoup de smicards et peu de gros salaires. Cependant, les gros salaires tirent la moyenne vers le haut.

La position de la médiane (pos Me) correspond à l'observation que se trouve au milieu des observations rangées.

Elle se calcule par la formule suivante: $\text{Pos Me} = \frac{n + 1}{2}$

n = nombre des observations

Exemple.- pour le calcul de la médiane des 29 observations des fréquences cardiaques, il faut:

a.- ordonner les observations: 50, 50, 50, 50, 60, 60, 60, 70, 70, 70, 70, 70, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 80, 90, 90,

b.- La position de la médiane est donnée par la formule:

$$\text{Pos Me} = \frac{29 + 1}{2} = 15$$

NB.- Si au lieu de 29 observations, nous avons 28, $\text{Pos Me} = \frac{28 + 1}{2} = 14,5$

Pour avoir la médiane on prend la moyenne entre la valeur de l'observation No 14 et celle No 15.

La valeur de l'observation No 14 : 80 et celle du No 15: 80 soit $\text{Pos Me} = \frac{80 + 80}{2} = 80$

- **Mode (statistique).**

Le mode correspond à la réalisation la plus fréquente.

Le mode d'une série, ou dominante d'une loi, est la valeur de la variable (ou de l'unité statistique) qui revient le plus fréquemment dans la série. C'est la valeur centrale de la classe qui a le plus grand effectif.

Ex : Soit la série {8, 4, 4, 3, 4, 3, 8, 2,5} La valeur la plus fréquente de cette série est 4. Le mode est donc égal à 4. L'effectif associé à ce mode est 3.

Il est l'indice le plus simple à déterminer puisqu'il suffit de lire un graphique ou de regarder le tableau des effectifs.

Exemple.- dans l'exemple des fréquences cardiaques, nous classons les observations, puis on identifie la valeur la plus fréquente:

Observation	Nb
50	4
60	3
70	5
80	15
90	2

Ici, la classe modale ou le mode est la classe 80

- Variance (statistiques et probabilités).

La variance empirique corrigée pour le carré de l'écart type (ou variance) :

Attention : la variance (notion de statistique descriptive) égale est la simple moyenne arithmétique des carrés des écarts à la moyenne arithmétique observée, mais la variance *sans biais* (notion de statistique mathématique, qui signifie qu'en moyenne la valeur empirique est égale à la valeur théorique) est fois la variance observée. La variance sans biais est donc supérieure à la variance observée.

Formule

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

= S^2 = variance d'échantillon

= X_i = valeur d'une observation

= $\bar{x}$ [valeur moyenne de toutes les observations](#)

= N = nombre d'observations

- Écart type.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

c'est la racine carrée de la variance

- **Coefficient de variation** :
- **Minimum et maximum**
- **Étendue** : c'est l'intervalle entre la plus petite et la plus grande valeur. On dit d'un phénomène qu'il présente une « forte dynamique » lorsque l'étendue (ou la dispersion) est grande.

Intervalle de confiance

Le théorème central limite garantit que la moyenne estimée est à une distance plus petite que de la moyenne théorique avec une probabilité proche de 1 où suit une loi gaussienne standard. Cela veut aussi dire que (est le quantile correspondant à pour une gaussienne):

Par conséquent, lorsque la taille de l'échantillon augmente linéairement, la précision de l'estimateur de la moyenne augmente en uand l'ensemble de points ne constitue pas un *échantillon* de la population, mais la population totale, la variance sans biais n'a pas à être utilisée, puisque l'on n'est plus dans un contexte d'estimation mais de mesure.

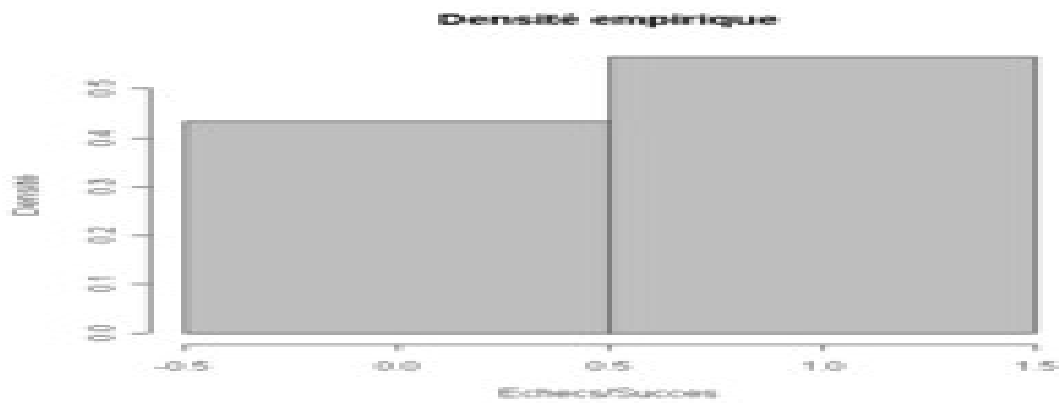
Quantiles

Les quantiles est une généralisation de la notion de médiane qui divise la loi en deux parties égales. On définit notamment les quartiles, déciles et centiles (ou percentiles) sur la population, ordonnée dans l'ordre croissant, que l'on divise en 4, 10 ou 100 parties de même effectif.

On parlera ainsi du « centile 90 » pour indiquer la valeur séparant les premiers 90 % de la population des 10 % restants. Ainsi, dans une population de jeunes enfants, un enfant dont la taille est au-delà du centile 90, ou en deçà du centile 10, peut être l'objet d'un suivi particulier.

- **Histogramme**

Même s'il est considéré souvent comme une représentation graphique, et qu'il a donc plus sa place dans une description des méthodes de Visualisation des données, l'histogramme est un hybride d'une représentation exhaustive des données et d'une description par recours à des lois statistiques.



histo

gramme de l'exemple des sportifs.

La densité empirique d'une variable à valeurs discrètes est simplement constituée de la proportion des observations prenant chaque valeur.

Dans l'exemple des sportifs, la densité empirique de notre population est 57 % de succès et 43 % d'échecs. L'histogramme associé est très simple (cf image à gauche).

On appelle **fonction de répartition empirique** associée une série d'observations à valeur réelles ayant les valeurs la fonction suivante:

Elle est une estimation de *la probabilité que la valeur d'un événement du phénomène observé ait une valeur supérieure ou égale à* .

Si on voulait en déduire la densité empirique associée aux observations, il faudrait dériver . Étant donné que la dérivée d'une indicatrice () est une loi de Dirac, le résultat ne serait pas très utilisable.

Plusieurs alternatives sont possibles :

- utiliser un estimateur par noyaux, il s'agit d'implémenter la densité suivante:

où est une fonction noyau (de masse égale à un).

- approximer la densité par une fonction en escalier.

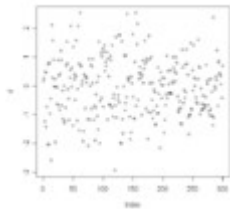
Un histogramme est *la meilleure estimation par une fonction en escalier* de la densité empirique. C'est-à-dire que l'intégrale de l'histogramme doit être la plus proche possible de . Remarquons que l'intégrale de l'histogramme est une fonction continue affine par morceaux. D'un certain point de vue :

trouver la fonction continue affine par morceaux qui approxime le mieux la fonction de répartition empirique revient à caractériser totalement l'histogramme.

Dans ce cadre, le nombre de morceaux (de *classes* ou de *barres*) est un paramètre très important. Il faut recourir à un critère supplémentaire si on veut trouver sa meilleure valeur

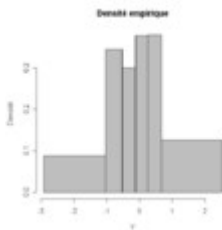
possible. On prend par exemple un critère d'information d'Akaike ou le critère d'information bayésien; il est aussi possible de recourir à un critère d'information ou d'entropie.

Par construction, les barres de l'histogramme ne sont donc pas nécessairement toutes de la même largeur.



•

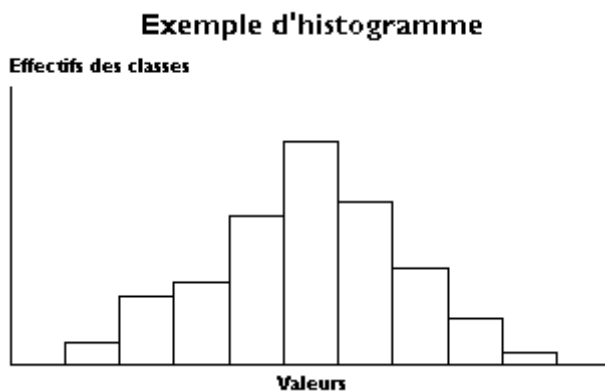
(A) Un tirage aléatoire de points



•

(B) Un histogramme associé

Construction d'un histogramme

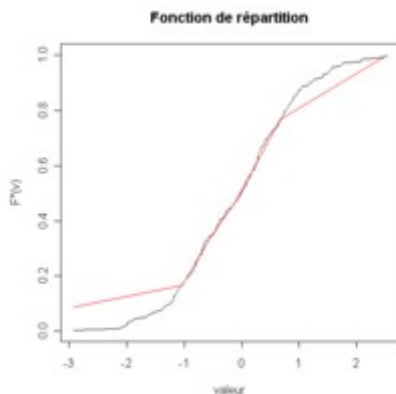


L'histogramme est une des nombreuses représentations graphiques de données statistiques possibles. Comme les quantiles, l'histogramme découpe la population en classes mais le point de vue est différent.

Avec les quantiles, le but est de localiser les frontières entre classes de même effectif. Ils sont souvent utilisés, par exemple en matière de revenus, pour comparer les deux classes extrêmes.

Pour les histogrammes, les largeurs de classes sont choisies afin de rendre le mieux possible compte de la loi réelle des observations. Il s'agit d'une tâche difficile.

Pour plus de simplicité, les classes des histogrammes sont parfois pris de même largeur et de hauteur variable : on appelle de tels histogrammes des **diagrammes en barres**. Ce ne sont pas de véritables histogrammes.



La fonction de répartition empirique (noir) et la fonction continue affine par morceaux associées à un histogramme

Il est possible de comparer la distance entre ces deux courbes.

- par exemple en utilisant le Test de Kolmogorov-Smirnov
- où en remarquant que la distance entre ces deux courbes (définie par la surface entre elles) suit une loi du.

En allant plus loin, ce genre de méthode de comparaison de fonctions de répartition (ici entre celles issues de l'histogramme et la loi empirique) peut être utilisé pour comparer la répartition empirique de nos observations à celle d'une loi connue (c'est par exemple le principe de la Droite de Henry). Cela permet de répondre à la question *ma répartition ressemble-t-elle à une loi connue ?*

Statistiques déductives

Statistiques déductives est le moteur de l'analyse des données, permettant aux chercheurs de faire des déductions sur une population plus large sur la base d'un échantillon. Cette colonne vertébrale analytique se différencie de statistiques descriptives en allant au-delà du simple résumé des données jusqu'aux prédictions et aux généralisations. Dans le cadre du commerce électronique, il est essentiel pour prise de décision et efficace tests d'hypothèses, alimentant des informations stratégiques allant du comportement des consommateurs aux prévisions de stocks.

I-Bases des tests d'hypothèses

Dans le monde des tests d'hypothèses, l'hypothèse nulle et l'hypothèse alternative sont le Yin et le Yang. Ils proposent des déclarations contrastées sur un paramètre de population, ouvrant la voie à une enquête. Comprendre les erreurs de type I (faux positifs) et de type II (faux négatifs) est crucial pour l'évaluation des risques. Le niveau de signification (α) choisi, souvent 0,05, et la valeur p travailler ensemble pour décider du sort des hypothèses, en guidant les acteurs du commerce électronique dans leurs projets, des ajustements marketing aux lancements de produits.

II-Méthodes d'échantillonnage et détermination de la taille de l'échantillon

A-Méthodes d'échantillonnage

Pour mettre les voiles dans la bonne direction, la sélection de l'échantillon optimal n'est pas négociable. Échantillonnage probabiliste (englobant les critères aléatoires simples, stratifiés, en grappes et systématiques) garantit que chaque membre de la population a une chance connue de sélection, ce qui évite les biais. En revanche, l'échantillonnage non probabiliste (comme les méthodes de commodité, de quota et de boule de neige) privilégie souvent l'aspect pratique plutôt que la perfection. La taille de l'échantillon est un exercice d'équilibre, affecté par le niveau de confiance souhaité, la marge d'erreur et la variabilité de la population, essentiels pour prendre le pouls du consommateur.

1. **Échantillonnage aléatoire simple**: une technique dans laquelle chaque élément de la population a une chance et une probabilité égales d'être sélectionné.
2. **Échantillonnage systématique**: est un processus symétrique dans lequel le chercheur choisit les échantillons après un intervalle spécifiquement défini
3. **Échantillonnage aléatoire stratifié**: est une méthode d'échantillonnage très productive dans les situations où le chercheur a l'intention de se concentrer uniquement sur des strates spécifiques de la population disponible. données disponibles sur la population
4. **Échantillonnage par grappes**: implique la division de la population en groupes ou en grappes, comme son nom l'indique. Suivant cette technique, un certain nombre de grappes est sélectionné au hasard, puis toutes les unités incluses à l'intérieur des grappes sélectionnées constituent l'échantillon.
5. **Échantillonnage par quotas**: méthode par laquelle les chercheurs créent un *échantillon* de commodité composé d'individus représentatifs d'une population. Les chercheurs choisissent
6. **Échantillonnage de convenance** : un échantillon choisi lors d'une étude pour des raisons pratiques d'accessibilité
7. **Échantillonnage en boule de neige**: Il s'agit d'une technique d'échantillonnage dans laquelle les sujets existants fournissent des références pour le recrutement.
8. **Échantillonnage raisonné**: les participants sont choisis par les chercheurs en fonction de : Leurs connaissances. Leur compréhension de la question de recherche. Ou leurs objectifs.

B-Formule de calcul de la taille de l'échantillon

Voilà la fameuse formule que vous attendez : $n = z^2 \times p (1 - p) / m^2$

n = taille de l'échantillon

z = niveau de confiance selon la [loi normale centrée réduite](#) (pour un niveau de confiance de 95%, $z = 1.96$, pour un niveau de confiance de 99%, $z = 2.575$)

p = proportion estimée de la population qui présente la caractéristique (lorsque inconnue, on utilise $p = 0.5$ ce qui correspond au cas le plus défavorable c'est-à-dire la dispersion la plus

grande)

m = marge d'erreur tolérée (par exemple on veut connaître la proportion réelle à 5% près)

Cette formule détermine le nombre de personnes **n** à interroger en fonction de la marge d'erreur **m** que l'on peut tolérer sur une proportion de réponses **p**.

Nous retenons ici le calcul d'une proportion qui est très fréquent dans les enquêtes en ligne et sur lequel vous appuierez souvent vos décisions stratégiques. Par exemple, si vous étudiez le marché pour un nouveau produit à lancer, vous investirez pour le mettre en marché si la proportion des répondants intéressés par ce produit est suffisamment grande.

Revenons à nos calculs pour vous montrer en exemples ce que cela donne :

1-Pour calculer une proportion avec un niveau de confiance de **95%** et une marge d'erreur à **5%** :

$$n = (1,96)^2 \times (0,5)(1-0,5) / (0,05)^2 = 384,16$$

2-Pour calculer une proportion avec un niveau de confiance de **95%** et une marge d'erreur à **7%** :

$$n = (1,96)^2 \times (0,5)(1-0,5) / (0,07)^2 = 196$$

3-Pour calculer une proportion avec un niveau de confiance de **99%** et une marge d'erreur à **2%** :

$$n = (2,575)^2 \times (0,5)(1-0,5) / (0,02)^2 = 4144,14$$