

Cours N°1 : Introduction, probabilités et variables aléatoires.

I. Introduction

En ce qui concerne les biostatistiques, on s'intéresse à des **caractères** ou à des **grandeurs biologiques**. Elles sont mesurées chez les êtres vivants sains et malades. Leur particularité principale est la **variabilité**.

A. Variabilités

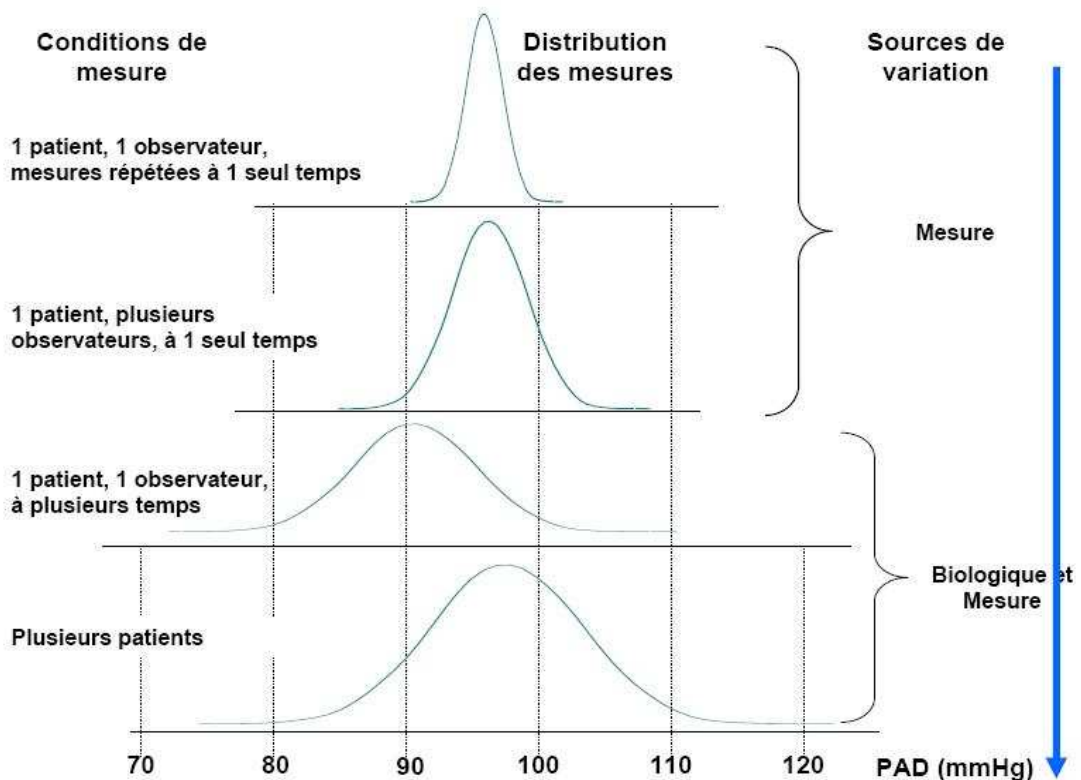
La **variabilité** est liée à la diversité biologique naturelle des sujets. On parle de :

- **Variabilité inter-individuelle**, représentant les différences entre les individus.
- **Variabilité intra-individuelle**, représentant les différences selon le moment et la situation, pour un même individu.

De plus, s'ajoutant à cette diversité biologique, la **variabilité de mesure** (du point de vue erreur), est aussi à prendre en compte. Elle est liée à :

- **L'instrument**, le moyen de faire la mesure.
- **L'observateur** qui fait la mesure.

Ces deux sources de variabilité s'additionnent !



B. Données quantitatives et qualitatives

Les caractères, ou grandeurs biologiques, sont « classées » en deux grands types :

- Les données **quantitatives**, qui font référence à un **nombre**.
- Les données **qualitatives**, qui font référence à un **jugement** ou une **appréciation**.

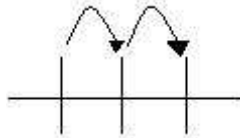
Pour savoir si une grandeur est quantitative ou qualitative, on se pose une question simple : **chaque valeur possible est-il un nombre ?**

- Si **OUI** : la grandeur est **quantitative**.
- Si **NON** : la grandeur est **qualitative**.

1. Données quantitatives

Les données quantitatives sont **mesurables sur une échelle** [elles fournissent un **NOMBRE**] et sont de **2 types** selon l'échelle de mesure :

- Si l'échelle de mesure est à **valeurs discrètes** (entre deux valeurs successives données, il n'existe aucune valeur intermédiaire), la grandeur est **discrète**.



- Si l'échelle de mesure est **continue** (il existe une infinité de valeurs possibles entre deux valeurs données), la grandeur est **continue**.



Pour savoir si une grandeur quantitative est discrète ou continue, on se pose une question simple : **puis-je numéroté les valeurs possibles ?**

- Si **OUI** : la grandeur est **discrète**.
- Si **NON** : la grandeur est **continue**.

2. Données qualitatives

Les données qualitatives ne sont **pas mesurables sur une échelle** et sont de **2 types** :

- Les grandeurs **ordinales** : les valeurs possibles sont **ordonnées**, ou **classées** (préférence).
- Les grandeurs **catégoriques** : dans tous les **autres** cas.

Pour savoir si des grandeurs qualitatives sont ordinales ou catégoriques, on se pose une question simple : **puis-je ordonner les valeurs possibles ?**

- Si **OUI** : les grandeurs sont **ordinales**.
- Si **NON** : les grandeurs sont **catégoriques**.

De nombreuses grandeurs sont possibles :

- **Exemple 1 : les « constantes » biologiques** : glycémie, calcémie, nombre de globules rouges, hémoglobine... Ce sont des nombres à valeurs réelles (avec un instrument de mesure parfait, on pourrait trouver un infinité de chiffres après la virgule pour la glycémie, on ne peut donc pas énumérer toutes les grandeurs possibles) donc ce sont des grandeurs **quantitatives continues**.
- **Exemple 2 : la survenue d'une maladie**. Les « valeurs » possibles sont {malade ; non malade} donc **qualitative à deux classes** donc **catégorique** (on ordonne pas deux possibilités).
- **Exemple 3 : le nombre de métastases**. Les valeurs possibles sont {0 ; 1 ; 2}, **données numériques** et que l'on peut **numéroter** donc le nombre de métastases est une variable **quantitative discrète**.
- **Exemple 4 : la satisfaction du malade**. Les « valeurs possibles » sont {très satisfait ; satisfait ; peu satisfait ...} avec un **ordre entre les valeurs possibles** donc **qualitative ordinale**.

C. Statistiques et statistiques descriptives

La **statistique** intervient quand il est impossible ou inutile d'observer la grandeur d'intérêt sur une **population**, qui est l'**ensemble exhaustif d'individus partageant une(des) caractéristique(s) communes(s)** (par exemple la population française, les étudiants inscrits en PCEM1 à Denis Diderot en octobre 2008, les sujets atteints d'une infection à VIH). On l'observe alors sur un **groupe issu de la population**, que l'on appelle **échantillon**.

On parlera d'« **échantillon représentatif** » uniquement si l'échantillon a été constitué par **tirage au sort** à partir de la population.

Les mesures effectuées sont appelées « **données** ».

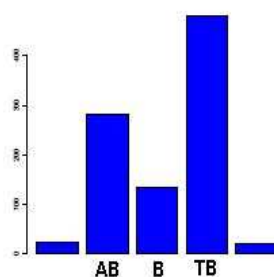
[La statistique correspond à la deuxième partie de l'enseignement des biostatistiques, soit à partir du cours N°6.]

La **statistique descriptive** a pour but de **décrire des données sur un groupe** (échantillon ou population de petite taille). Pour cette description, on fait appel à des **méthodes graphiques ou numériques**, adaptées au type de données.

1. Méthodes graphiques

a. Le diagramme en bâtons

EX: Mentions BAC

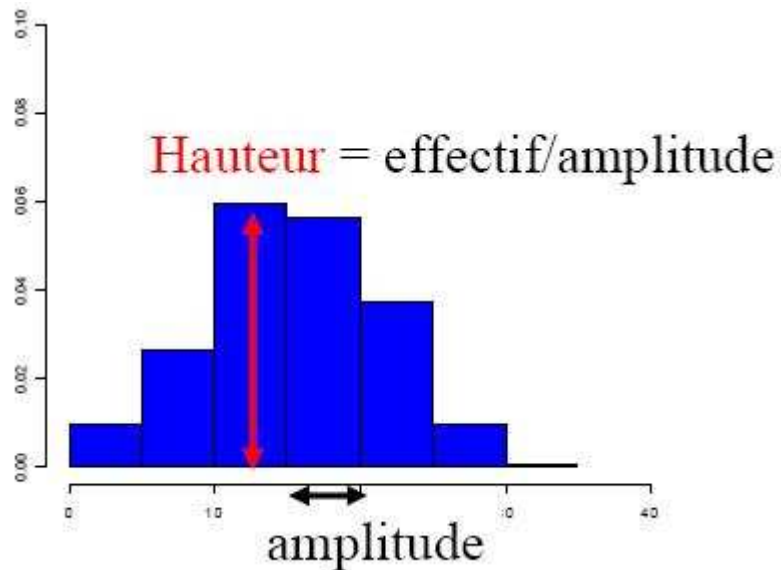


On utilise le **diagramme en bâtons** pour les données **quantitatives discrètes** (note que vous aurez en biostatistiques) ou les données **qualitatives** (mention au bac).

Pour le diagramme en bâtons, on trace en **ordonnées** un bâton dont la **hauteur** est égale à l'**effectif de la catégorie**.

b. Histogramme

Pour les données **quantitatives continues** (ou discrètes mais avec un très grand nombre de valeurs possibles), la représentation en bâtons synthétise très mal la distribution. On trace alors un **histogramme**.



En **abscisse**, on trouve des **classes de valeurs**.

En **ordonnées**, on trouve un rectangle, dont la **hauteur** est égale à l'**effectif divisé par l'amplitude** de la classe.

[L'histogramme est donc en quelque sorte en « deux dimensions », ce qui permet une étude plus poussée des données.]

2. Méthodes numériques

Les méthodes numériques sont réservées aux **données quantitatives**. Elles ont pour but de résumer :

- La **tendance centrale** des données (**moyenne**).
- La **dispersion** des données (l'**étendue** et la **variance**).

a. Moyenne

La **moyenne empirique**, ou **expérimentale**, est donnée par la relation :

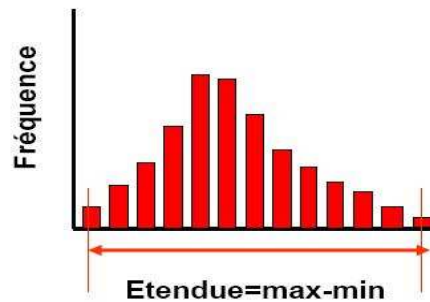
$$m = \bar{x} = \frac{1}{n} \sum x_i$$

Elle s'interprète comme le **centre de gravité** d'un nuage de points.

La moyenne n'est pas suffisante pour étudier des données. Des valeurs peuvent avoir la même moyenne mais ne pas avoir du tout les mêmes caractéristiques.

b. Étendue

L'**étendue** est définie par la **différence entre la valeur la plus grande et la valeur la plus petite**. Elle ne dépend que de 2 observations, elle est donc peu stable [donc on utilise la variance].



c. Variance

La **variance empirique**, ou **expérimentale**, mesure la **dispersion des données autour de sa moyenne**. Elle s'exprime dans l'unité élevée au carré .

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

Sa **racine carrée positive** $\sqrt{s^2}$ (dans la même unité que les données) est appelée **écart-type**, s .

II. Probabilités

A. Intérêts et conditions

1. Intérêts

En ce qui concerne les probabilités, on s'intéresse (toujours) à une **grandeur biologique** qui est **variable** selon les individus.

L'objectif des probabilités est de :

- **Décrire la dispersion**, ou **variabilité**, des données sur l'ensemble de la population (et pas seulement sur un échantillon).
- **Quantifier les « chances » de réalisation des différentes valeurs possibles** de cette grandeur.

Ces deux objectifs sont des intermédiaires pour **modéliser les expériences aléatoires** [voir un peu plus loin], but final des probabilités.

2. Conditions

Pour réaliser des probabilités, la grandeur d'intérêt doit donc être « **variable** », c'est-à-dire :

- Qu'il y a **plusieurs résultats possibles** (dits événements élémentaires).
- Que son résultat est **IMPREVISIBLE**. On ne peut pas connaître la valeur de cette grandeur tant qu'on ne l'a pas mesurée (observée).
- Qu'**on peut répéter la mesure** d'un individu à l'autre.

Ces 3 conditions définissent une **expérience aléatoire**. On dit que ce sont des **Conditions Nécessaires et Suffisantes (CNS)** : cela veut dire qu'il faut que ces 3 conditions soient réunies pour parler d'expérience aléatoire, et en même temps qu'elles suffisent, aucune 4^{ème} condition n'est attendue.

Le résultat d'une expérience aléatoire est un événement aléatoire, donc imprévisible.

[En contre exemple, on peut prendre l'issue de la vie, soit la mort. Même si son arrivée dans le temps est inconnue, c'est tout de même une certitude. La vie n'est donc pas une expérience aléatoire.]

B. Modélisation d'une expérience aléatoire

On a vu qu'un des objectifs des probabilités était de décrire la dispersion des données sur l'ensemble de la population. Ainsi, la **première étape de la modélisation d'une expérience aléatoire** est de **définir l'ensemble des résultats possibles**, ou l'**ensemble des événements élémentaires**, représenté par l'univers **E**. **Chaque événement élémentaire est un point de l'ensemble des résultats possibles.**

On peut définir la notion de **cardinal (E)**, ou **card (E)**, qui correspond au **nombre d'événements élémentaires**.

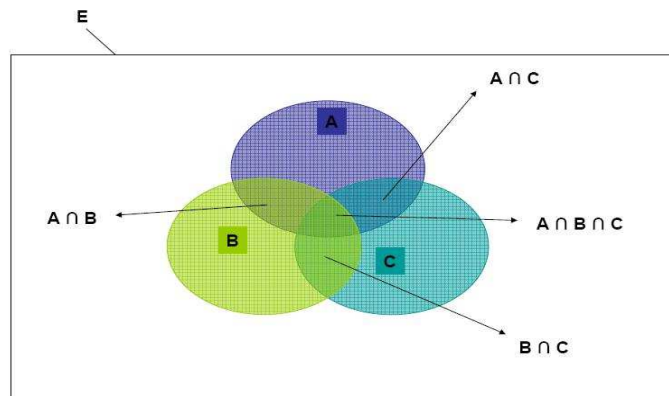
Par exemple, l'expérience aléatoire consistant à traiter un malade avec un médicament et d'observer sa réponse au traitement contient 2 événements élémentaires : $E = \{ \text{succès ; échec} \}$.

Ici, $\text{card}(E) = 2$.

On peut également **définir des événements (quelconques) en réunissant des événements élémentaires**. Un événement (quelconque) est donc une **partie de l'ensemble E**.

On peut construire d'autres événements :

- Par l'**union d'événements** (A OU B noté $A \cup B$)
- Par l'**intersection d'événements** (A ET B noté $A \cap B$)



Un événement qui ne peut se produire est un **événement impossible**. L'univers E est dit **certain** car il se produira forcément. Deux événements qui ne peuvent se produire simultanément sont **incompatibles** ou **exclusifs**.

L'autre objectif des probabilités est de **quantifier les « chances » de réalisation des différentes valeurs possibles** de cette grandeur. Ces deux objectifs sont remplis par l'intermédiaire de la **loi de probabilité de l'expérience**.

Définir la loi de probabilité d'une expérience consiste à définir :

- L'**ensemble de ses événements élémentaires**, soit **E**.
- Une **quantité définissant « la chance » de survenue de chacun** pour chaque résultat possible. On l'appelle la **probabilité de l'événement élémentaire** notée **p(e_i)**.

La **probabilité p_i** d'un événement élémentaire e_i est un **nombre compris entre 0 et 1**.

Pour des expériences à valeurs qualitatives ou discrètes, l'ensemble E est un ensemble **dénombrable** {e₁ ; e₂ ; ...}.

La somme de toutes les probabilités des événements élémentaires sur E doit être égale à 1 [car le résultat de l'expérience est nécessairement dans E !].

Si tous les événements élémentaires (é.é) de l'expérience ont la **même chance de survenue**, on dit qu'ils sont **équiprobables**. Comme leur somme est égale à 1, on a nécessairement :

$$p_i = \frac{1}{(\text{nombre d'éléments de E})} = \frac{1}{[\text{Card}(E)]}$$

Attention, **il ne faut pas confondre imprévisible et équiprobable !**

[Par exemple, passer un concours est une expérience aléatoire car :

- Il y a plusieurs résultats possibles.
- On peut répéter l'expérience (il y a plus de 2 000 inscrits).
- Le résultat d'une expérience est imprévisible (on ne sait pas à l'avance si on va réussir ou échouer), même si les événements ne sont pas équiprobables (en PCEM 1 à Paris 7, vous avez 16,6% de réussir et donc 83,7% d'échouer.)]

La **probabilité d'un événement quelconque A = {e₁ ; e₂ ; ...}** est donnée par :

$$P(A) = \sum_{e_i \in A} p_i = \frac{\text{nombre de cas favorables à A}}{\text{nombre de cas possibles en tout}}$$

C. Axiomes de calcul des probabilités

Les **axiomes de calcul des probabilités** permettent de trouver, connaissant la loi de probabilité d'une expérience, des événements quelconques, en partant de 3 données :

- **P(A) est comprise entre 0 et 1.**
- **P(E) = 1**
- **Si A et B sont incompatibles** ($A \cap B = \text{vide}$) [voir ci-dessous], alors **$P(A \cup B) = P(A) + P(B)$**

Si A et B sont deux événements quelconques, alors :

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

D. Dénombrements

Les dénombrements consistent à trouver, à partir d'une situation, le **nombre de tirages différents possibles**, afin d'établir des probabilités. Il existe trois sortes de tirages :

- Les tirages successifs avec remise.
- Les tirages successifs sans remise.
- Les tirages simultanés.

1. Tirages successifs avec remise

Prenons une urne contenant n jetons numérotés. On prend p jetons, **en remettant avant chaque nouveau tirage le jeton tiré**. Pour le premier tirage, on a n possibilités. Pour le second également, et ainsi de suite. Comme on fait p tirages, on en déduit que le nombre de tirages différents possibles est :

$$n \times n \times n \times \dots \times n = n^p$$

2. Tirages successifs sans remise

Prenons une urne contenant n jetons numérotés. On prend p jetons ($p \leq n$), mais **chaque jeton tiré ne retourne pas dans l'urne**. Pour le 1^{er} jeton, on a n possibilités, pour le 2nd, on a $n-1$ possibilités et ainsi de suite. Le nombre de tirages différents possibles est donc :

$$n \times (n-1) \times (n-2) \times \dots \times (n-p+1) = \frac{n!}{(n-p)!}$$

Où $n! = n \times (n-1) \times (n-2) \times \dots \times 2 \times 1$ ($0! = 1$)

Cas particulier : si l'on fait des tirages sans remise jusqu'à vider l'urne, alors le nombre de résultats différents possibles est $n!$. C'est aussi le nombre de façons de ranger n objets les uns par rapport aux autres, ce qu'on appelle aussi **permutations**.

3. Tirages simultanés

Prenons une urne contenant n jetons numérotés. **On prend p jetons simultanément**, c'est-à-dire **sans ordre ni répétition**. Le nombre de tirages différents possibles est :

$$\binom{n}{p} = \frac{n \times (n-1) \times (n-2) \times \dots \times (n-p+1)}{p!} = \frac{n!}{p!(n-p)!}$$

Propriétés :

$$\binom{n}{0} = \binom{n}{n} = 1 ; \binom{n}{p} = \binom{n}{n-p} ; \binom{n}{1} = \binom{n}{n-1} = n$$

4. Résumé

Comment savoir quel modèle utiliser ?

On se pose deux questions : **les critères peuvent-ils être répétés ? L'ordre des éléments intervient-il ?**

Critères	<u>Les éléments peuvent être répétés</u>	<u>Les éléments sont distincts</u>
<u>On tient compte de l'ordre</u>	Tirages successifs avec remise	Tirages successifs sans remise
<u>On ne tient pas compte de l'ordre</u>		Tirages simultanés

III. Variable aléatoire discrète X

A. Définition

Une **variable aléatoire quantitative** correspond à une expérience aléatoire dont l'ensemble E de ses résultats possibles (événements élémentaires) est composé de **valeurs numériques mesurables sur une échelle**.

Une **variable aléatoire DISCRETE** correspond à une expérience aléatoire dont l'ensemble E de ses résultats possibles (événements élémentaires) est composé de **valeurs numériques mesurables sur une échelle DISCRETE** $E = \{x_1 ; x_2 ; \dots\}$.

Par exemple, le nombre d'enfants : $E = \{0 ; 1 ; 2 ; 3 ; 4 ; \dots\}$

Cela nécessite de **connaître l'échelle de mesure**.

On peut « transformer » toute expérience aléatoire qualitative en variable aléatoire discrète.

Par exemple la réponse à un traitement :

- $E = \{\text{succès} ; \text{échec}\}$: c'est une **expérience qualitative**
- Si on pose $X(\text{succès}) = 1$ et $X(\text{échec}) = 0$, on a $E = \{0 ; 1\}$: c'est une **variable aléatoire discrète**.

Définir la **loi de probabilité d'une expérience** consiste à définir l'ensemble de ses événements élémentaires puis à quantifier, pour chaque résultat possible, une quantité définissant « la chance » de survenue de chacun. On l'appelle la **probabilité de l'événement élémentaire x_i** définie par :

$$P_X(x) = P(X = x) = \sum_{X(e_i)=x} p_i$$

B. Loi de Bernoulli

La **loi de Bernoulli**, notée **B(p)**, est une loi **discrète**, de **paramètre p** ($0 < p < 1$), définie par :

- $E = \{0 ; 1\}$
- $P(X = 1) = p$

D'où $P(X = 0) = 1 - p$

Comme les résultats possibles sont des nombres, on peut synthétiser la loi de probabilité par des **quantités synthétiques numériques**. La **position moyenne** correspond à l'**espérance** et la **dispersion** à la **variance**.

C. Espérance

1. Espérance de X

L'espérance de X est notée $E(X)$. C'est le **barycentre des valeurs de X affectées de leur probabilité**. **Ce n'est pas nécessairement une valeur de E**. On trouve comme synonymes la **moyenne de X** ou la **valeur attendue de X**. Elle est donnée par la relation :

$$E(X) = \sum_{x \in E} x \times P(X = x) = \sum_{x_i \in E} x_i \times p_i$$

2. Espérance d'une loi de Bernoulli

L'espérance d'une loi de Bernoulli est forcément **comprise entre 0 et 1**. Comme $P(X = 1) = p$, on a :

$$E(X) = 0 \times (1 - p) + 1 \times p = p$$

Le **paramètre p** de la loi de Bernoulli est la **probabilité que X = 1**. C'est également l'**espérance de la loi**.

3. Espérance de h(X)

Soit une **fonction linéaire de X**, alors :

$$E(aX + b) = aE(X) + b$$

Soit $h(X)$, une **fonction quelconque de X**, on a :

$$E(h[X]) = \sum_{x \in E} h(x) \times P(X = x)$$

On a aussi plus généralement $E(aX + bY) = aE(X) + bE(Y)$, avec **X et Y deux variables aléatoires définies sur le même univers et a et b des réels**.

D. Variance

1. Variance de X ou Dispersion

La **variance de X**, ou **dispersion**, est l'**espérance d'un carré**. Elle ne peut être **que positive ou nulle**, sachant qu'une variance nulle implique nécessairement que tous les événements élémentaires sont identiques à l'espérance. Elle est donnée par la relation :

$$Var(X) = E[(X - E[X])^2]$$

D'où $Var(X) = E[x^2 + E(X)^2 - 2XE(X)] = E(X^2) + E(X)^2 - 2E(X) \times E(X)$

Donc

$$Var(X) = E(X^2) - [E(X)]^2$$

Avec $E(X^2) = \sum_{x \in E} x^2 \times P(X = x)$

Plus généralement, on a :

$$V(aX) = a^2 V(X)$$

$$V(X+b) = V(X)$$

$$V(aX+b) = a^2 V(X)$$

Si X et Y, deux variables aléatoires, sont indépendantes, on a :

$$V(aX + bY) = a^2 V(X) + b^2 V(Y)$$

Attention !

Ne pas confondre moyenne expérimentale ou empirique avec moyenne ou espérance d'une variable aléatoire.

- La moyenne expérimentale est mesurée sur un échantillon et a pour but de décrire la position des seules valeurs de l'échantillon. Elle est utilisée pour des statistiques

DESCRIPTIVES donc concrètes. Elle est donnée par $m = \bar{x} = \frac{1}{n} \sum x_i$ [Nous étudions un échantillon dont on connaît certaines données, et les exploite pour les caractériser].

- La moyenne ou espérance d'une variable aléatoire indique la position de la distribution de probabilité d'une expérience aléatoire à valeurs numériques. Elle correspond non à quelque chose de concret mais plus à une prévision [on essaie de prévoir les valeurs les plus aptes à tomber]. Elle est donnée par $E(X) = \sum_{x \in E} x \times P(X=x)$

- Idem pour les variances...

2. Variance d'une loi de Bernoulli

Nous sommes dans une loi de Bernoulli, donc l'ensemble des cas possibles est **0** ou **1**. on a :

$$P(X=1) = E(X) = p \quad \text{et} \quad \text{Var}(X) = E(X^2) - [E(X)]^2$$

Où $E(X^2) = 0^2 \times P(X=0) + 1^2 \times P(X=1) = p$

Donc :

$$\text{Var}(X) = p - (p^2) = p \times (1-p)$$

La variance de la loi de Bernoulli est le produit des probabilités que X = 0 et que X = 1.

E. Écart-type de X

L'écart-type de X est la racine carrée positive de la variance, qu'on note $\sqrt{\text{Var}(X)}$. Elle est donc donnée par la relation :

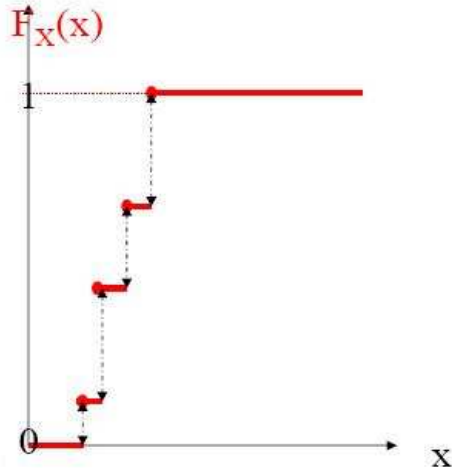
$$\sqrt{p \times (1-p)}$$

F. Fonction de répartition

La fonction de répartition d'une variable aléatoire X, notée **F**, est définie sur \mathbb{R} par :

$$F(x) = p(X \leq x) = \sum_{x_i \leq x} P(x_i)$$

La fonction de répartition d'une variable aléatoire **quantitative discrète** est **monotone**, **croissante**, en **marches d'escalier** et on a $\lim_{x \rightarrow -\infty} F(x) = 0$ et $\lim_{x \rightarrow +\infty} F(x) = 1$



G. Quelques définitions

1. La médiane

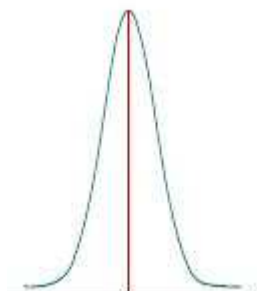
La médiane est définie telle que $F(\text{médiane}) = 0,5$.

2. Le mode

Le **mode** est la valeur de **x** correspondant au **plus grand effectif** (c'est le **x** tel que **y** est le plus grand possible).

3. Symétrie, dissymétrie et position relative de la médiane, de la moyenne et du mode

La distribution de la variable aléatoire peut être **symétrique**. On dit qu'elle est **unimodale**. Dans ce cas, **mode, médiane et moyenne** coïncident.



La distribution de la variable aléatoire peut aussi être **dissymétrique**, à droite ou à gauche.

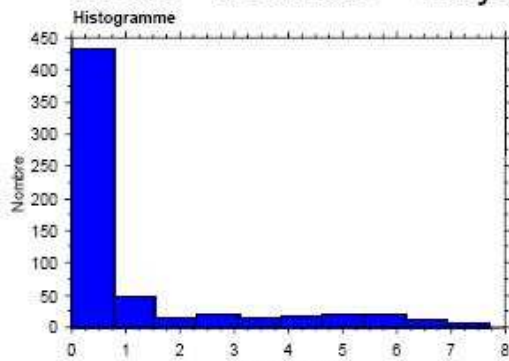
Si elle est **dissymétrique à droite** (donc **mode à gauche** et « queue » à droite), on a, de gauche à droite, d'abord le mode, puis la médiane au milieu et enfin la moyenne à droite.

Si elle est **dissymétrique à gauche** (donc **mode à droite** et « queue » à gauche), on a, toujours de gauche à droite, d'abord la moyenne, puis la médiane au milieu et enfin le mode à droite.

- Si distribution **dissymétrique**

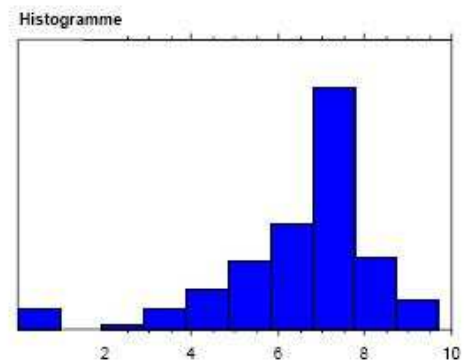
à droite

mode < médiane < moyenne



à gauche

moyenne < médiane < mode



Ce document, ainsi que l'intégralité des cours P1, sont disponibles gratuitement sur <http://coursplbichat-larib.weebly.com/index.html>