SciencesPo

# STATISTIQUES APPLIQUÉES AUX SCIENCES HUMAINES & SOCIALES
## Niveau introductif

### DESCRIPTIF DU COURS

Très axé sur la pratique de la recherche empirique dans les sciences sociales, ce cours a vocation à donner les bases en statistiques descriptives et inférentielles.

**Outil informatique**

Excel

### MODALITÉS D'ÉVALUATION

Le contrôle continu vaut ⅔ de la moyenne, et l'examen final ⅓. Le contrôle continu comprendra des interrogations de 10 minutes à chaque séance. L'examen final correspondra à un sujet de 2h. L'ensemble des sujets, par groupe, devra être préparé par l'enseignant.

### BIBLIOGRAPHIE


- **Méthodes statistiques en sciences humaines**, Howell, D., Paris, DeBoeck Université
- **Pourcentages et tableaux statistiques**, Novi, M., Paris, PUF
- **Teaching Statistics: A Bag of Tricks**, Andrew Gelman (2002)
- **Statistical Methods for the Social Sciences** (4th Edition), Agresti & Finley
- **Les différents types de variables, leurs representations graphiques et paramètres descriptifs**, Julien Labreuche, INSERM
- **Statistique pour les Sciences Humaines I**, Agnès Lagnoux, université de Toulouse-Mirail
- **Note de cours de Méthodes Quantitatives**, Éric Brunelle et Josée Riverin
- **Principes et Méthodes Statistiques, Notes de cours**, Olivier Gaudoin, Ensimag
- **Les concepts élémentaires des statistiques,** Zarrouk Fayçal, LFEP
- **Cours de simulations et probabilités**, Hélène Guérin, université de Rennes I


## SÉANCE 1 INTRODUCTION À LA RECHERCHE EMPIRIQUE EN SCIENCES SOCIALES

**Notions abordées**

La **statistique** comme outil pour répondre à des questions de recherche empirique dans les sciences sociales
Quelques exemples de question de recherche

**Notions de base**

Individu statistique, population, échantillon, variable, modalité, effectif, fréquence, distribution statistique

**Types de variables**

Nominale, ordinale, d'intervalles et de ratios

### I. VOCABULAIRE STATISTIQUE

La **statistique descriptive** recueille des données, les exploite sous forme graphique (diagrammes…) ou numérique (moyenne, médiane), mais elle se limite à l'analyse d'observations.
La **statistique inférentielle** établit des prévisions en fonction des observations collectées (tests, interpolation…). Elle utilise pour cela des modèles probabilistes.
Les statisticiens distinguent la population et les échantillons : la **population** est l'ensemble des **individus** que l'on cherche à étudier, un **échantillon** est un sous-ensemble de la population. Le nombre d'individus de la population (de l'échantillon) s'appelle la **taille** ou l'**effectif** de la population (de l'échantillon).

Une **variable** ou **caractère statistique** est une caractéristique étudiée pour une population donnée (âge, taille, niveau d'études, opinions politiques…). Une variable permet de distinguer plusieurs **classes** dans un population.
Les **modalités** (ou **valeurs**) du caractère sont les identificateurs de ces classes d'individus.

Les **caractères (ou variables)** sont de deux natures.

- **Quantitatifs** si le caractère est mesurable par un nombre.

**Exemple**
Le nombre d'enfants par ménage ou la taille des individus d'une population.

Une **variable d'intervalle** est une variable quantitative qui se réfère à un zéro relatif. L'échelle de valeurs ne traduit pas un ordre de grandeur entre les valeurs.

**Exemple**
On recense les notes à un examen. Obtenir zéro ne veut pas (forcément) dire que l'étudiant n'a rien fait. L'élève qui obtient 16 n'est pas deux fois meilleur que celui qui obtient 8.

Une **variable de rapport** ou de **ratio** est une variable quantitative qui se réfère à un zéro absolu : le 0 désigne une absence de valeurs. L'échelle de valeurs traduit une situation de proportionnalité entre les valeurs.

**Exemple**
On regarde le nombre de frères et sœurs d'un individu dans une population. La valeur 0 indique bien une absence de frères et sœurs. Un individu ayant 4 frères et sœurs en a bien deux fois plus que celui qui en a 2.

Une variable (ou caractère) quantitative peut être :

✔ **Discrète**
La variable est dénombrable. On peut compter les modalités.
**Exemple** le nombre d'enfants par ménage

✔ **Continue**
La mesure peut être en théorie n'importe quel nombre réel.
**Exemple** la taille des individus d'une population.

- **Qualitatifs** si le caractère n'est pas mesurable par un nombre.

Une variable est **ordinale** si ses modalités sont naturellement ordonnées.
**Exemple** Niveau d'études
Une variable est **nominale** lorsque ses modalités ne sont pas hiérarchisées.
**Exemple** La couleur des yeux.
Les variables qualitatives ordinales se différencient des variables quantitatives discrètes par l'absence d'information sur la distance séparant les nombres. Il ne s'agit que d'une codification sans valeur arithmétique. Dans la littérature scientifique, il est fréquent de trouver les termes de « **variable dichotomique** » ou « variable binaire » qui correspondent à une variable qualitative qui ne peut prendre que deux modalités souvent codées 0 et 1 (exemple ; 1 pour homme et 0 pour femme).

**La fréquence**
La **fréquence** d'une modalité est le quotient du nombre de fois où elle apparaît par l'effectif total.

**Exemple**
On a repéré qui portait des lunettes dans un groupe d'étudiants. Sur 22 étudiants, 6 portent des lunettes.
La fréquence des étudiants portant des lunettes est de

$$\frac{6}{22} \approx 0,27 \approx \frac{27}{100}$$

27% de ces étudiants portent des lunettes.

**Effectifs et fréquences cumulés croissants**
Lorsque la variable est quantitative, on peut calculer les **effectifs cumulés croissants** et les **fréquences cumulées croissantes** correspondantes. Les effectifs cumulés croissants correspondent à l'effectif total de toutes les modalités inférieures à une certaine valeur.
Par exemple on peut regarder le nombre de frères d'un groupe d'étudiants.

| **Nombre de frères et sœurs** | **0** | **1** | **2** | **3** | **4** | **5** |
|---|---|---|---|---|---|---|
| **Effectifs** | 4 | 6 | 8 | 4 | 2 | 1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Effectifs cumulés croissants (ECC)** | 4 | 10 | **18** | 22 | 24 | 25 |
| **Fréquence** | 0,16 | 0,24 | 0,32 | 0,16 | 0,08 | 0,04 |
| **Fréquences cumulées croissantes** | 0,16 | 0,4 | **0,72** | 0,88 | 0,96 | 1 |

Ainsi, 18 étudiants ont deux frères et sœurs ou moins, ce qui réprésente 72% des étudiants.

### Effectifs et fréquences marginaux

On peut aussi croiser **deux caractères** dans une population et calculer les **effectifs marginaux** et les **fréquences marginales**. Les **effectifs marginaux** correspondent aux **effectifs de chaque classe en ne tenant compte que d'un seul des deux caractères** étudiés. Les fréquences marginales sont les fréquences correspondantes.

**Exemple**

On a réparti dans ce tableau la population française en 2018 en fonction de l'âge et du sexe.

| Groupe d'âges | Femmes | Hommes | Total (Effectifs marginaux) |
|---|---|---|---|
| Moins de 15 ans | 5 978 099 | 6 248 352 | **12 226 451** |
| 15-19 ans | 2 038 916 | 2 140 382 | **4 179 298** |
| 20-24 ans | 1 860 041 | 1 903 554 | **3 763 595** |
| 25-29 ans | 1 970 018 | 1 923 405 | **3 893 423** |
| 30-34 ans | 2 082 129 | 1 981 056 | **4 063 185** |
| 35-39 ans | 2 162 277 | 2 075 413 | **4 237 690** |
| 40-44 ans | 2 132 035 | 2 088 366 | **4 220 401** |
| 45-49 ans | 2 304 165 | 2 256 655 | **4 560 820** |
| 50-54 ans | 2 294 952 | 2 219 361 | **4 514 313** |
| 55-59 ans | 2 216 883 | 2 091 464 | **4 308 347** |
| 60-64 ans | 2 129 262 | 1 943 594 | **4 072 856** |
| 65-69 ans | 2 085 309 | 1 867 535 | **3 952 844** |
| 70-74 ans | 1 631 170 | 1 412 809 | **3 043 979** |
| 75 ans ou plus | 3 768 229 | 2 381 207 | **6 149 436** |
| **Total (Effectifs marginaux)** | **34 653 485** | **32 533 153** | **67 186 638** |

| Groupe d'âges | Femmes | Hommes | Total (Fréquences marginales) |
|---|---|---|---|
| **Moins de 15 ans** | 8,9% | 9,3% | **18,2%** |
| **15-19 ans** | 3,0% | 3,2% | **6,2%** |
| **20-24 ans** | 2,8% | 2,8% | **5,6%** |
| **25-29 ans** | 2,9% | 2,9% | **5,8%** |
| **30-34 ans** | 3,1% | 2,9% | **6,0%** |
| **35-39 ans** | 3,2% | 3,1% | **6,3%** |
| **40-44 ans** | 3,2% | 3,1% | **6,3%** |
| **45-49 ans** | 3,4% | 3,4% | **6,8%** |
| **50-54 ans** | 3,4% | 3,3% | **6,7%** |
| **55-59 ans** | 3,3% | 3,1% | **6,4%** |
| **60-64 ans** | 3,2% | 2,9% | **6,1%** |
| **65-69 ans** | 3,1% | 2,8% | **5,9%** |
| **70-74 ans** | 2,4% | 2,1% | **4,5%** |
| **75 ans ou plus** | 5,6% | 3,5% | **9,2%** |
| **Total (Fréquences marginales)** | **51,6%** | **48,4%** | **100,0%** |

Pour une variable considérée,

- Certaines caractéristiques indiquent une modalité qui représente l'ensemble des modalités de la variable, on les appelle les caractéristiques de **position** ou de **tendance centrale** (la moyenne par exemple) ;
- D'autres caractéristiques indiquent la « disparité » des modalités de la variable, on les appelle les caractéristiques de **dispersion** (l'écart-type par exemple).

## SÉANCE 2 SÉRIES STATISTIQUES À UNE VARIABLE

**Notions abordées**
Caractéristiques de position (médiane, moyenne, quantiles)
Caractéristiques de dispersion (variance, écart-type, étendue, écart inter-quantiles)

## II. OUTILS POUR LE TRAITEMENT STATISTIQUE

### 1. Les caractéristiques de tendance centrale ou caractéristique de position

#### a. Le mode

Le **mode** correspond à la valeur (cas discret) ou à la classe (dans le cas continu) qui possède le plus grand effectif. Le mode est la caractéristique de tendance centrale utilisée avec des variables qualitatives nominales là où moyenne et médiane n'ont pas de sens. On pourra utiliser le mode pour trouver quel mot ou quelle lettre apparaît le plus souvent dans un texte.

**Exemple 1**
Le tableau ci-dessous présente le retard accumulé (en minutes) au soir de la quinzième étape du Tour de France par les 150 coureurs qui n'étaient pas Maillot Jaune ce jour-là.

| **Retard** | [0; 30[ | [30; 60[ | [60; 90[ | [90; 120[ | [120; 150[ | [150; 180[ | [180; 210[ | [210; 240[ | [240; 270[ |
|---|---|---|---|---|---|---|---|---|---|
| **Effectif** | 19 | 7 | 13 | 23 | 23 | 24 | 33 | 7 | 1 |

Le mode est ici la classe [180; 210[ (on parle dans ce cas d'**intervalle de classe modale**). Dans ce cas, le centre de la classe (ici 195) peut aussi désigner le mode. La série est **unimodale**. Une seule classe possède un effectif de 33.

**Exemple 2**
Lors d'un devoir on a obtenu les notes suivantes.

| **Notes** | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Effectifs** | 3 | 6 | 5 | 7 | 7 | 10 | 8 | 10 | 3 | 2 | 2 |

Le **mode** est ici la note **5** et la note **7**. La série est **bimodale**. Deux valeurs ont un effectif de 10.

**Remarque**
Contrairement à la médiane, cela n'aurait aucun sens ici de dire que le mode est 6, ce qui correspond à la moyenne arithmétique de 5 et 7.

#### b. La moyenne arithmétique

Soit $X = (x_1; x_2; \ldots; x_N)$ une variable quantitative de taille N. **La moyenne** $E(X)$ **souvent notée** $\bar{x}$ est le quotient de la somme de toutes les modalités par l'effectif total.

$$E(X) = \bar{x} = \frac{somme\ des\ valeurs}{nombre\ de\ valeurs} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

Dans l'exemple 2, la moyenne est :

$$\bar{x} = \frac{0\times3+1\times6+2\times5\ldots+8\times3+9\times2+10\times2}{3+6+\ldots+3+2+2} = \frac{260}{58} \approx 4,5$$

Dans l'exemple 1, le caractère continu impose un regroupement par classe. On va faire l'hypothèse que les valeurs à l'intérieur de cet intervalle sont uniformément réparties. Ainsi, on résumera la classe par une valeur moyenne, nommée **centre de la classe**.

| Retard | 0 min à 30 min | 30 min à 1h | 1 h à 1 h 30 | 1 h 30 à 2 h | 2 h à 2 h 30 | 2 h 30 à 3 h | 3 h à 3 h 30 | 3 h 30 à 4 h | 4 h à 4 h 30 |
|---|---|---|---|---|---|---|---|---|---|
| Effectif | 19 | 7 | 13 | 23 | 23 | 24 | 33 | 7 | 1 |
| Centres | 0,25 | 0,75 | 1,25 | 1,75 | 2,25 | 2,75 | 3,25 | 3,75 | 4,25 |

La moyenne est :

$$\bar{x} \approx \frac{19\times0,25+7\times0,75+\ldots+7\times3,75+1\times4,25}{19+7+\ldots+33+7+1} = \frac{322}{150} \approx 2,15h \approx 2h09min$$

## Quelques propriétés de la moyenne

**Propriété 1** Si $a \in R,\ E(a \cdot X) = a \cdot E(X)$

**Exemple**

Si tous les prix augmentent de 5% ils sont donc multipliés par 1,05 donc la moyenne sera multipliée également par 1,05 : la moyenne augmentera également de 5% !

**Propriété 2** Si $b \in R,\ E(X + b) = E(X) + b$

**Exemple**

Si tous les prix baissent de 1€ alors la moyenne des prix sera également baissée de 1€

## c. La médiane et les quartiles

Ils ne peuvent être calculés que si la variable est **quantitative** et dispose que d'un **nombre fini de modalités**.

Notons $n$ le nombre de valeurs de la série.

**La médiane** est une valeur séparant la série ordonnée en deux séries de même effectif.

- Si $n$ est **impair** alors la médiane est la valeur de la série de rang $\frac{n+1}{2}$.
- Si $n$ est **pair** alors la médiane est la **moyenne** entre les valeurs de rangs $\frac{n}{2}$ et $\frac{n}{2} + 1$.

**Remarque**

La médiane est moins sensible que la moyenne aux valeurs extrêmes et c'est pour cette raison qu'elle est parfois utilisée (longévité, salaire,…) à la place de la moyenne.

**Exemples**

- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 7 ; 8 ; 8 ; 9 ; 10, il y a 9 valeurs donc la médiane est la 5ème valeur : la médiane est donc 7.
- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 6 ; 7 ; 8 ; 8 ; 9 ; 10, il y a 10 valeurs donc la médiane est la moyenne entre la 5ème valeur (6) et la 6ème valeur (7) : la médiane est donc 6,5.

**Remarque**

Si les valeurs ne sont pas ordonnées, il faut les ranger dans l'**ordre croissant**.

## Effectifs cumulés croissants et médiane

On dispose de la série statistique suivante représentant la production de chaussures dans une usine.

| **Pointure** | **40** | **41** | **42** | **43** | **44** | **45** |
|---|---|---|---|---|---|---|
| **Effectif** | 120 | 450 | 240 | 410 | 100 | 80 |

*Comment calcule-t-on la médiane de cette série ?*

On utilise les effectifs cumulés croissants.

| **Pointure** | **40** | **41** | **42** | **43** | **44** | **45** |
|---|---|---|---|---|---|---|
| **Effectif** | 120 | 450 | 240 | 410 | 100 | 80 |
| **Effectifs cumulés croissants** | 120 | 570 | 810 | 1220 | 1320 | 1400 |

1220 est le nombre de chaussures dont la pointure est inférieure ou égale à 43. 1400 est l'effectif total.
Comme il y a eu 1400 chaussures produites, la médiane est une valeur comprise entre la 700ème et la 701ème valeur.
La pointure est 40 de la 1ère valeur jusqu'à la 120ème.
La pointure est 41 de la 121ème valeur jusqu'à la 570ème valeur.
La pointure est 42 de la 571ème valeur jusqu'à la 810ème valeur.
La 700ème et la 701ème valeur sont donc 42 : **la médiane est donc 42**.

**Remarque**
Concrètement, la médiane correspond à la valeur où la première fois l'effectif cumulé atteint la moitié de l'effectif.

**Exemple de regroupement par classe**
On relève la taille en cm des élèves de la classe de 4°1 :
160 – 140 – 165 – 150 – 138 – 158 – 170 – 160 – 155 – 160 – 140 – 150 – 170 – 155 – 155 – 129 – 160 – 163
L'étendue est de $170 - 129 = 41\ cm$.
On regroupe ces tailles en plusieurs classes et dans le tableau suivant :

| Taille | [120 ;130[ | [130 ;140[ | [140 ;150[ | [150 ;160[ | [160 ;170[ | [170 ;180[ |
|---|---|---|---|---|---|---|
| Effectif | 1 | 1 | 2 | 6 | 6 | 2 |

[120 ;130[ est un intervalle qui rassemble tous les élèves dont la taille varie entre 120 cm et 130 cm. Les crochets indiquent que les élèves qui font exactement 120 cm sont dans cette classe et ceux qui font exactement 130 cm ne sont pas dans cette classe.
La fréquence des élèves dont la taille est comprise entre 140 et 150 cm est de $\frac{2}{18} = \frac{1}{9} \approx 0,11 \approx \frac{11}{100}$. 11% des élèves mesurent entre 140 et 150 cm
Avec ce tableau, on peut calculer une moyenne approchée à l'aide du centre des classes et la classe médiane à l'aide des effectifs cumulés croissants (ECC).

| Taille | [120 ;130[ | [130 ;140[ | [140 ;150[ | [150 ;160[ | [160 ;170[ | [170 ;180[ |
|---|---|---|---|---|---|---|
| Centre | $\frac{120+130}{2} = 125$ | 135 | 145 | 155 | 165 | 175 |
| Effectif | 1 | 1 | 2 | 6 | 6 | 2 |
| ECC | 1 | $1 + 1 = 2$ | 4 | 10 | 16 | 18 |

La taille moyenne est de
$\frac{125\times1+135\times1+\ldots+175\times2}{1+1+\ldots+2} = \frac{2820}{18} \approx 157\ cm$
**Remarque**
Ceci est une **valeur approchée de la moyenne**.
$\frac{160+140+\ldots+163}{18} = \frac{2778}{18} \approx 154\ cm$
Elle peut être **légèrement différente** de la valeur réelle.
$18/2 = 9$. La médiane est donc comprise entre la 9ème et la 10ème valeur. La classe médiane est donc [150 ;160[.
Dans cette classe, les valeurs sont 150-150-155-155-155-158.
155 est la 9ème valeur, 158 la 10ème valeur. La médiane est donc de 156,5 cm.

**Le premier quartile ($Q_1$)** est la valeur telle que 1/4 des valeurs de la série lui soient inférieures et telle que 3/4 des valeurs lui soit supérieure.
En pratique, pour déterminer $Q_1$ on calcule $\frac{n}{4}$

- Si $\frac{n}{4}$ n'est pas **entier,** alors $Q_1$ est la valeur de la série ordonnée dont le rang est le plus petit entier supérieur à $\frac{n}{4}$
- Si $\frac{n}{4}$ est **entier,** alors $Q_1$ est la **moyenne** entre les valeurs de rangs $\frac{n}{4}$ et $\frac{n}{4} + 1$

**Exemples**

- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 7 ; 8 ; 8 ; 9 ; 10, il y a $n = 9$ valeurs

$$\frac{n}{4} = 2,25$$

donc le 1er quartile est la 3ème valeur, soit $Q_1 = 4$.

- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 6 ; 7 ; 7 ; 8 ; 8 ; 9 ; 10 ; 11, il y a 12 valeurs

$$\frac{n}{4} = 3$$

donc le 1er quartile est la moyenne entre la 3ème valeur (4) et la 4ème valeur (5), soit $Q_1 = 4,5$.

**Le troisième quartile ($Q_3$)** est la valeur telle que 3/4 des valeurs de la série lui soient inférieures et telle que 1/4 des valeurs lui soit supérieure.

En pratique, pour déterminer $Q_3$, on calcule $\frac{3n}{4}$.

- Si $\frac{3n}{4}$ n'est pas **entier,** alors $Q_3$ est la valeur de la série ordonnée dont le rang est le plus petit entier supérieur à $\frac{3n}{4}$
- Si $\frac{3n}{4}$ est **entier,** alors $Q_3$ est la **moyenne** entre les valeurs de rangs $\frac{3n}{4}$ et $\frac{3n}{4} + 1$

**Exemples**

- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 7 ; 8 ; 8 ; 9 ; 10, il y a $n = 9$ valeurs

$$\frac{3n}{4} = 6,75$$

donc le 3ème quartile est la 7ème valeur, soit $Q_3 = 8$.

- Dans la série ordonnée 0 ; 3 ; 4 ; 5 ; 6 ; 7 ; 7 ; 8 ; 9 ; 10 ; 10 ; 11, il y a 12 valeurs donc

$$\frac{3n}{4} = 9$$

donc le 3ème quartile est la moyenne entre la 9ème valeur (9) et la 10ème valeur (10), soit $Q_3 = 9,5$.

## 2. Les caractéristiques de dispersion

### a. L'étendue

L'**étendue** d'une série statistique est la **différence entre la plus grande valeur et la plus petite valeur** (de la variable) de cette série statistique.

### b. L'écart interquartile

On définit l'**écart interquartile** comme la différence du troisième et du premier quartile.

$$EI = Q_3 - Q_1$$

### c. La variance et écart-type

On définit donc la variance d'une variable quantitative X comme la moyenne des carrés des écarts entre les données de X et leur moyenne.

Soit $X = (x_1; x_2; \ldots; x_N)$ une variable quantitative de taille N.

$$Var(X) = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}\right)^2 = E(X - \bar{x})^2$$

Si on développe $(xi - \bar{x})^2$, on obtient la formule suivante :

$$Var(X) = \left(\frac{1}{N}\sum_{i=1}^{N} x_i^{\,2}\right) - \bar{x}^2 = E\left(X^2\right) - \bar{x}^2$$

L'écart-type σ est la racine carrée de la variance.

$$\sigma(X) = \sqrt{Var(X)}$$

**Remarque**

On utilise l'écart-type plutôt que la **déviation absolue moyenne** donnée par la formule

$$E(|X - \bar{x}|) = \frac{1}{N}\sum_{i=1}^{N}\left|x_i - \bar{x}\right|$$

L'écart-type est proche de la déviation absolue moyenne dans la plupart des cas même s'il est systématiquement supérieur.

**Exemple**

On a indiqué dans le tableau suivant le PIB des pays de la zone européenne en 2016.

Le PIB moyen par habitant est de

$\bar{x} = \frac{34\,500+\ldots+12\,000}{19} = \frac{539\,900}{19} \approx 28\,400€$

L'écart-type est de

$\sigma \approx \sqrt{\frac{(34\,500-28\,400)^2+\ldots+(12\,000-28\,400)^2}{19}}$

$\sigma \approx 17\,300€$

La déviation absolue moyenne est de

$\frac{|34\,500-28\,400|+\ldots+|12\,000-28\,400|}{19} \approx 12\,400€$

| Drapeau | Pays | Date d'entrée | Population (en millions d'habitants, 2016) | PIB (en milliards d'euros, 2016) | PIB/habitant (en euros, 2016) |
|---|---|---|---|---|---|
| | Allemagne | 1999 | 82,2 | 2 758,8 | 34 500 |
| | Autriche | 1999 | 8,7 | 315,1 | 36 100 |
| | Belgique | 1999 | 11,3 | 388,2 | 34 400 |
| | Espagne | 1999 | 46,4 | 1 102,8 | 23 700 |
| | Finlande | 1999 | 5,5 | 189,6 | 34 500 |
| | France | 1999 | 66,8 | 2 122,1 | 31 700 |
| | Irlande | 1999 | 4,7 | 240,7 | 51 400 |
| | Italie | 1999 | 60,7 | 1 568,7 | 25 900 |
| | Luxembourg | 1999 | 0,6 | 48,9 | 83 700 |
| | Pays-Bas | 1999 | 17 | 670,1 | 39 300 |
| | Portugal | 1999 | 10,3 | 174,2 | 16 900 |
| | Grèce | 2001 | 10,9 | 184,5 | 17 100 |
| | Slovénie | 2007 | 2,1 | 38,1 | 18 400 |
| | Chypre | 2008 | 0,8 | 18,2 | 21 300 |
| | Malte | 2008 | 0,4 | 8,7 | 20 000 |
| | Slovaquie | 2009 | 5,4 | 79,1 | 14 500 |
| | Estonie | 2011 | 1,3 | 18,0 | 13 500 |
| | Lettonie | 2014 | 2,0 | 21,6 | 11 000 |
| | Lituanie | 2015 | 2,9 | 34,4 | 12 000 |

Source : Commission européenne

**Propriété de la variance**

**Propriété** Si $a, b \in R$, $Var(a \cdot X + b) = a^2 \cdot Var(X)$

**SÉANCES 3 ET 4 VARIABLES ALÉATOIRES RÉELLES DISCRÈTES**

**Notions abordées**

Définition

Loi de probabilité (Bernoulli, binomiale, Poisson)

## III. VARIABLES ALÉATOIRES RÉELLES

### 1. Définitions

Une **expérience aléatoire** est une expérience dont on ne peut prédire le résultat. L'**ensemble fondamental** (ou **univers**) d'une expérience aléatoire est l'ensemble de tous les résultats possibles de l'expérience. Cet ensemble est en général noté Ω. Il peut être fini, dénombrable, ou infini non dénombrable. Chaque élément ω de Ω est une réalisation de l'expérience, nommée aussi **issue** ou **événement élémentaire**.

**Exemples**

- On lance un dé, on s'intéresse au chiffre obtenu. L'ensemble fondamental est fini : $\Omega = \{1 ; 2 ; 3 ; 4 ; 5 ; 6\}$.
- On jette une pièce autant de fois que nécessaire pour obtenir une fois « face ». L'ensemble fondamental est alors infini dénombrable : $\Omega = \{F; PF; PPF; PPPF; PPPPF \ldots\}$.
- On choisit au hasard un point sur une demi-droite et on regarde la distance entre ce point et l'origine de la demi-droite. L'ensemble fondamental est infini non dénombrable : $\Omega = [0; +\infty[$

Pour chaque expérience, on peut définir des **événements** réalisés par une, plusieurs ou même aucune issue.
Un événement réalisé par **aucune issue** est un **événement impossible**.
Un événement B réalisé par **toutes les issues** est un **événement certain**.

**Exemple**

On lance un dé à 6 faces et on regarde le résultat indiqué sur la face supérieure.
« Tirer un nombre pair » (événement A) est un événement lié à cette expérience. $A = \{2; 4; 6\}$
« Tirer un nombre » (événement B) est un événement **certain**. $B = \Omega$
« Ne tirer aucun nombre » (événement C) est un événement **impossible**. $C = \emptyset$

Deux événements sont **INCOMPATIBLES** s'ils ne peuvent pas se produire en même temps.
Autrement dit, deux événements A et B sont incompatibles si $A \cap B = \emptyset$
Deux événements sont **CONTRAIRES** s'ils sont incompatibles et si leur réunion est l'univers tout entier.

Autrement dit, deux événements A et B sont contraires si $A \cap B = \emptyset$ et si $A \cup B = \Omega$.
On note alors $A = \Omega \backslash B = \overline{B}$

**Remarque**
$A \cap B$ est l'événement qui réalise l'événement A et l'événement B en même temps (autrement dit, l'ensemble qui contient tous les éléments qui sont à la fois dans A et dans B).
$A \cup B$ est l'événement qui réalise l'événement A ou l'événement B ou A et B en même temps (autrement dit, l'ensemble qui contient tous les éléments qui sont soit dans A soit dans B soit dans A et B).
$A \backslash B = A \cap \overline{B}$ est l'événement qui réalise l'événement A mais pas l'événement B (autrement dit, l'ensemble qui contient tous les éléments qui sont dans A mais pas dans B).

Une **probabilité P** sur Ω est une application sur l'ensemble des événements telle que
- $P(\Omega) = 1$
- Pour tout événement A, $0 \leq P(A) \leq 1$
- Pour toute suite d'événements **incompatibles** $(A_i)_{i \in N}$

$$P\left(\bigcup_{i \in N} A_i\right) = \sum_{i \in N} P(A_i)$$

De cette définition, on peut tirer plusieurs propriétés.

**Propriété**
- *Si* $B \subset A$, *alors* $P(A \backslash B) = P(A) - P(B)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(\overline{A}) = 1 - P(A)$
- Si A est un événement impossible, $P(A) = 0$

**Preuve**
- Soit $B \subset A$. Les événements $A \backslash B$ et $B$ sont incompatibles.
  $P(A) = P(A \backslash B \cup B) = P(A \backslash B) + P(B)$
  Donc $P(A \backslash B) = P(A) - P(B)$
- Les événements $A \backslash A \cap B$, $B \backslash A \cap B$ et $A \cap B$ sont incompatibles.
  $P(A \cup B) = P(A \backslash A \cap B) + P(B \backslash A \cap B) + P(A \cap B)$
  $P(A \cup B) = P(A) - P(A \cap B) + P(B) - P(A \cap B) + P(A \cap B)$
  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(\overline{A}) = P(\Omega \backslash A) = P(\Omega) - P(A) = 1 - P(A)$
- Soit A est un événement impossible.
  $P(A) = P(\emptyset) = P(\Omega \backslash \Omega) = P(\Omega) - P(\Omega) = 1 - 1 = 0$

Si **Ω est discret**, la **probabilité d'un événement** est la somme des probabilités des issues qui le compose. La probabilité de chaque issue est souvent induite par l'expérience aléatoire.
Par exemple, la probabilité d'obtenir un nombre impair avec un dé à 6 faces équilibré (non truqué) est de $\frac{3}{6}$.
La probabilité d'obtenir un valet en tirant une carte dans un jeu de 52 cartes est de $\frac{4}{52}$.
Si toutes les issues ont la même chance de se produire, on est dans une **situation d'équiprobabilité**.
Par exemple, si on lance un dé à six faces équilibré (non truqué) et si on regarde le nombre obtenu, il s'agit d'une situation d'équiprobabilité : on a autant de chance de faire 1, 2, 3, 4, 5 ou 6.
Si on regarde la somme des résultats du lancer de 2 dés à six faces, ce n'est pas une situation d'équiprobabilité : on a plus de chances d'obtenir 7 que 2 (il n'y a que 11 issues).
Par contre, si l'on s'intéresse au résultat de chacun des dés en distinguant les 2 dés, il s'agit d'une situation d'équiprobabilité. Il y a 36 issues : (1;1) ; (1;2) ; (1;3)…

Une propriété est aussi souvent utilisée, c'est la **formule des probabilités totales**.

**Formule des probabilités totales**

Si $\{E_1; E_2; ...; E_n\}$ désigne un système complet d'événements (ou partition) sur Ω, c'est-à-dire

$$E_i \cap E_j = \emptyset \text{ si } i \neq j$$

$$E_1 \cup E_2 \cup ... \cup E_n = \Omega$$

Alors

$$P(A) = P(A \cap E_1) + P(A \cap E_2) + ... + P(A \cap E_n)$$

En particulier, $\{B; \overline{B}\}$ est une partition de Ω, donc

$$P(A) = P(A \cap B) + P(A \cap \overline{B})$$

## Probabilité conditionnelle

**Définition**

Soit $A$ et $B$ deux événements tels que $P(A) \neq 0$. On appelle **probabilité conditionnelle de $B$ sachant $A$**, la probabilité que l'événement $B$ se réalise sachant que l'événement $A$ est réalisé. Elle est notée $P_A(B)$ et est définie par

$$P_A(B) = \frac{P(A \cap B)}{P(A)}$$

**Remarque**

$P_A(B)$ se note aussi $P(B/A)$

La probabilité conditionnelle suit les règles et lois de probabilités vues auparavant. En particulier, soient $A$ et $B$ deux événements tels que $P(A) \neq 0$.

- $0 \leq P_A(B) \leq 1$
- $P_A(\overline{B}) = 1 - P_A(B)$
- $P(A \cap B) = P(A) \times P_A(B)$
- **Si $A \cap B = \emptyset$ alors $P_A(B) = P_B(A) = 0$**

De la définition, on déduit aisément la formule de Bayes.

**Formule de Bayes**

$$P_B(A) = \frac{P(A) \times P_A(B)}{P(B)}$$

Combinée avec la formule des probabilités totales, $P(A) = P(A \cap B) + P(A \cap \overline{B})$

On obtient la formule suivante

$$P_B(A) = \frac{P(A) \times P_A(B)}{P(B \cap A) + P(B \cap \overline{A})} = \frac{P(A) \times P_A(B)}{P(A) \times P_A(B) + P(\overline{A}) \times P_{\overline{A}}(B)}$$

**Application aux tests de dépistage**

Vous êtes directeur de cabinet du ministre de la santé. Une maladie est présente dans la population, dans la proportion **d'une personne malade sur 10 000**. Un responsable d'un grand laboratoire pharmaceutique vient vous vanter son nouveau test de dépistage : si une personne est **malade**, **le test est positif à 99%**. Si une personne n'est **pas malade**, **le test est positif à 0,1%**. Ces chiffres ont l'air excellents, vous ne pouvez qu'en convenir. Toutefois, avant d'autoriser la commercialisation de ce test, vous faites appel au statisticien du ministère : ce qui vous intéresse, ce n'est pas vraiment les résultats présentés par le laboratoire, c'est la probabilité qu'une personne soit malade si le test est positif. La formule de Bayes permet de calculer cette probabilité. On note $M$ l'événement « La personne est malade » et $T$ l'événement « Le test est positif ». Le but est de calculer $P_T(M)$. Les données que vous avez en main sont

$$P(M) = 0,0001$$

$$P(\overline{M}) = 0,9999$$
$$P_M(T) = 0,99$$
$$P_{\overline{M}}(T) = 0,001$$

La formule de Bayes donne

$$P_T(M) = \frac{P_M(T)P(M)}{P_M(T)P(M)+P_{\overline{M}}(T)P(\overline{M})} = \frac{10^{-4}\times 0,99}{10^{-4}\times 0,99+0,9999\times 10^{-3}} \simeq 0,09$$

C'est catastrophique ! Seulement **9%** des personnes positives au test sont effectivement malades ! C'est tout le problème des tests de dépistage pour des maladies rares : ils doivent être excessivement performants, sous peine de donner beaucoup trop de « faux-positifs ».

## 2. Lois de probabilité

On considère un ensemble Ω muni d'une probabilité P. Une **variable aléatoire** X est une fonction de l'ensemble fondamental Ω à valeurs dans $R$.
Lorsque la variable X est à valeurs dans un ensemble discret de $R$ (autrement dit, les valeurs de cet ensemble sont isolées dans $R$), on parle de **variable aléatoire discrète**.
On se place dans le cas discret. I est une partie de $N$. X est une variable aléatoire réelle tel que $X(\Omega) = \{x_i\}_{i\in I}$

L'**espérance** d'une variable aléatoire (qui s'apparente à la moyenne statistique) est donnée par la formule

$$E(X) = \sum_{i\in I} x_i \cdot P(X = x_i)$$

**Propriété**
$E(aX + b) = a\cdot E(X) + b$

La **variance** d'une variable aléatoire est donnée par la formule

$$Var(X) = E\left((X - E(X))^2\right) = \sum_{i\in I} P(X = x_i)\cdot\left(x_i - E(X)\right)^2$$

**Propriété**
$Var(aX + b) = a^2\cdot Var(X)$

### Loi de Bernoulli

On réalise une expérience aléatoire qui a 2 résultats possibles : le succès **S**, de probabilité $p$, et l'Echec **E** de probabilité $1 - p$.
Soit $X$ la variable aléatoire définie sur $\Omega = \{S; E\}$ telle que $X(S) = 1$ et $X(E) = 0$.
$X$ suit une loi de **Bernoulli** (savant suisse 1654-1705). On écrit :

$$X \sim B(p)$$

On note aussi
$P(X = 1) = p$ et $P(X = 0) = 1 - p$
$E(X) = p\cdot 1 + (1 - p)\cdot 0 = p$
$Var(X) = E\left(X^2\right) - E^2(X) = p\cdot 1 + (1 - p)\cdot 0 - p^2 = p - p^2 = p(1 - p)$

### Loi binomiale

Si l'on répète $n$ fois de manière indépendante un schéma de Bernoulli, on dira que la variable aléatoire Y qui comptabilise le nombre de succès suit une loi binomiale de paramètres $n$ et $p$. On note

$$Y \sim B(n, p)$$

Y est à valeurs dans $[\![0; n]\!]$ et on a :

$$\forall k\in[\![0; n]\!],\ P(Y = k) = \left(\frac{n}{k}\right)p^k(1 - p)^{n-k}$$

$k$ désigne le nombre de succès et $n - k$ le nombre d'échecs.

### Remarque

$\left(\frac{n}{k}\right)$ désigne le nombre de combinaisons pour choisir $k$ objets parmi $n$ sans tenir compte de l'ordre dans lequel on les choisit.

$$\left(\frac{n}{k}\right) = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\ldots(n-k+1)}{k(k-1)\ldots\times 2\times 1}$$

Par exemple, si l'on choisit deux étudiants dans un groupe de 5, on aura $\left(\frac{5}{2}\right) = \frac{5\times 4}{2\times 1} = 10$ possibilités.

**Rappel** $n! = n(n-1)\ldots\times 2\times 1$

$$E(Y) = \sum_{k=0}^{n} k\cdot P(Y=k) = \sum_{k=0}^{n} k\cdot\left(\frac{n}{k}\right)p^{k}(1-p)^{n-k} = n\cdot p$$

$$Var(Y) = n\cdot p(1-p)$$

## Loi de Poisson

Siméon-Denis Poisson (1781-1840) est un probabiliste, mathématicien et physicien français à qui l'on doit d'importants développements sur la loi des grands nombres, les suites d'épreuves de Bernoulli mais aussi sur les applications des probabilités dans le domaine du droit.

La variable Z suit une loi de Poisson de paramètre λ et on note $Z \sim P(\lambda)$ si :

$$\forall k \in N,\ P(Z=k) = \frac{\lambda^{k}}{k!}e^{-\lambda}$$

**Modélisation**

Lorsqu'un événement a une faible probabilité $p$ d'apparition lors d'une épreuve de Bernoulli et si l'on répète un grand nombre de fois cette épreuve $(n)$ le nombre total de réalisations de l'événement considéré suit à peu près une loi de Poisson de paramètre $\lambda = np$ (dans la pratique $n > 50$ et $p < 0,1$).

$$E(Z) = Var(Z) = \lambda$$