


Statistiques

Nathalie GUYADER
nathalie.guyader@gipsa-lab.fr

Année 2021-2022



Planning des séances

- 7 séances de 4h
- Un examen « machine » de 4h

Le cours vous sera envoyé à la fin de la séance (sauf demande contraire) en version numérique.

Normalement j'ajouterai au fur et à mesure les réponses à vos questions.

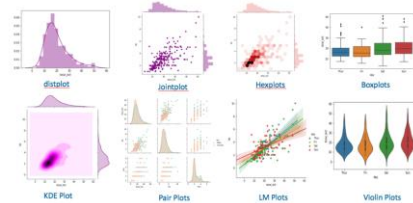
Nous travaillerons très régulièrement sur PC avec le logiciel PYTHON



Classiquement on distingue :

- Les statistiques descriptives ou exploratoires:
 - **Décrire** une variable, un lien entre des variables, un tableau de chiffres
 - **Visualiser** un ensemble de données grâce à des représentations adaptées
 - **Résumer** un ensemble de données par des indices
- Les statistiques inférentielles ou décisionnelles:
 - **Prévoir** un résultat à partir d'un échantillon
 - **Estimer** des paramètres auxquels on n'a pas accès
 - **Généraliser** un résultat observé sur un échantillon à toute la population
 - **Réfuter** une hypothèse grâce à l'utilisation de critères fiables et contrôlables

Seaborn Plots



Dans le cadre de ce cours nous utiliserons Python, NumPy, SciPy, Pandas, Matplotlib, Seaborn

Bibliographie

Pré requis:

Cours de probabilités et plus généralement le cours de Maths d'année 3.

Ouvrages:

- Probabilités, analyse des données et statistique de G. Saporta aux éditions Technip.
- Howell, D. C. (1998). Méthodes statistique en sciences humaines. Ed. De Boeck Université.
- Introduction à l'inférence statistique: Méthodes d'échantillonnage, estimation, tests d'hypothèses, corrélation linéaire, droite de régression et test du khi-deux avec applications diverses de Gérald Baillargeon. Editeur : Smg (5 novembre 1999).

Autres:

Cours d'Alan Chauvin, MCF à l'UGA (Licence Psychologie)

Cours de Guillaume Laget, MCF à l'UGA (IUT Mesures Physiques)

Introduction générale

Les méthodes statistiques sont aujourd'hui utilisées dans presque tous les secteurs de l'activité humaine et font partie des connaissances de base de l'ingénieur, du gestionnaire, de l'économiste, du psychologue...

Parmi les innombrables applications, citons dans le domaine industriel :

- l'analyse des résultats de mesure et leur planification
- la fiabilité du matériel
- le contrôle de qualité
- la prévision

et dans le domaine de l'économie et des sciences de l'homme

- les modèles économétriques
- les sondages et les enquêtes d'opinion

Introduction générale

Selon la définition de l'encyclopédie Universalis :

« Le mot statistique désigne à la fois un ensemble de données issues d'observations et l'activité qui consiste dans leur recueil, leur traitement et leur interprétation ».

Durant ce cours nous ne parlerons pas du recueil des données mais c'est une partie non négligeable dans la pratique!

Régulièrement les étudiants ingénieurs lors des stages ont besoin des Statistiques:

- soit directement durant leur stage/alternance
- soit pour rendre des rapports

Et même les étudiants qui ne s'en servent pas directement, les statistiques sont une des connaissances de base de l'ingénieur et il est donc important de comprendre les différents objectifs de cette discipline.

Quelques définitions (extraites du Saporta)

Faire de la statistique suppose que l'on étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « variables ».

Exemples:

- En contrôle de fabrication on prélèvera un ensemble de pièces dans une production homogène et on mesurera leur poids, leur diamètre
- En contrôle de composants électroniques, par exemple des résistances, on prélèvera quelques résistances et on s'assurera que la résistance ohmique (mesure en hors circuit) est proche de la valeur attendue, à sa tolérance près...
- En tests de circuits on relèvera les temps d'exécution des tests, etc.

La notion fondamentale en statistique est celle d'ensemble d'objets équivalents (**population**). Ce terme est hérité des premières applications de la statistique qui concernait la démographie.

Les objets sont des **individus**. La statistique traite des propriétés des populations plus que de celles d'individus.

Quelques définitions (extraites du Saporta)

Généralement, la population à étudier est trop vaste pour pouvoir être observée exhaustivement : c'est évidemment le cas d'une population infinie (toutes les pièces qui sortent d'une chaîne de fabrication dans des conditions déterminées) mais c'est aussi le cas lorsque les observations sont coûteuses (par exemple le contrôle destructif).

L'étude de tous les individus d'une population s'appelle un **recensement**. Lorsque l'on n'observe qu'une partie de la population on parle de sondage, la partie étudiée s'appelle alors l'**échantillon**.

Le concept clé en statistique est **la variabilité** qui signifie que des individus en apparence semblables peuvent prendre des valeurs différentes : ainsi un processus de fabrication ne fournit jamais des caractéristiques parfaitement constantes.

L'analyse statistique est pour l'essentiel une **étude de la variabilité** : on peut en tenir compte pour prévoir de façon probabiliste le comportement d'individus (non encore observés).

Statistiques et Probabilités

(extraits du Saporta)

La théorie des probabilités est une branche des mathématiques qui traite des propriétés de certaines structures modélisant des phénomènes où le « hasard » intervient. Cette théorie permet de modéliser efficacement certains phénomènes aléatoires et d'en faire l'étude théorique. Quels sont ses liens avec la statistique qui repose plutôt sur l'observation de phénomènes concrets ?

Les données observées sont souvent imprécises avec une erreur. Le modèle probabiliste permet alors de représenter comme des variables aléatoires (VA) les déviations entre les vraies valeurs et les valeurs observées.

On constate souvent que la répartition statistique d'une variable au sein d'une population est voisine de modèles mathématiques (loi de probabilité).

Enfin, les échantillons d'individus sont tirés la plupart du temps au hasard dans la population, ceci pour assurer mathématiquement leur représentativité : si le tirage est fait de manière équiprobable chaque individu de la population a une probabilité constante et bien définie d'appartenir à l'échantillon. Les caractéristiques deviennent, grâce à ce tirage au sort, des VA et le calcul des probabilités permet d'étudier leur répartition.



Chapitre 1

Rappels - Probabilités

Ce chapitre est très largement inspiré du cours de Guillaume Laget (IUT Mesures Physiques)



Combinatoire

Avant de revoir les probabilités il est important de revoir le dénombrement; le dénombrement permet de compter le nombre d'éléments des ensembles finis qui seront les événements étudiés. On doit connaître: le cardinal, la factorielle, les listes, les arrangements et les combinaisons.

1 - Cardinal

Le nombre d'éléments distincts dans un ensemble E est appelé cardinal de E , et on le note $\text{card}(E)$

Exemples:

Si E est l'ensemble vide, noté \emptyset , son cardinal est $\text{card}(E) = 0$

Si $E = \{1,2,3,4,5,6\}$ son cardinal est $\text{card}(E) = 6$

Si E est l'ensemble des entiers alors $\text{card}(E)$ est infini

Combinatoire

2 - Factorielle

La factorielle d'un entier naturel n correspond au nombre de façons de classer les n éléments; on la note $n!$

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

Autrement dit $n!$ est le nombre de permutations d'un ensemble à n éléments. (Rq: la permutation correspond à la disposition ordonnée de tous les éléments d'un ensemble)

3 - Listes

Un problème classique en dénombrement est celui du tirage successif avec remise: typiquement on dispose d'une urne contenant 5 jetons, numérotés de 1 à 5, et on tire 3 fois de suite un jeton dont on note le numéro avant de la remettre dans l'urne. On a alors $5 \times 5 \times 5 = 5^3$ triplets de résultats possibles.

Ainsi le nombre de manières de fabriquer une liste (ordonnée) de n -éléments tous compris entre 1 et p est p^n .

Combinatoire

4 - Arrangement

L'arrangement, défini pour tout entier naturel n et tout entier naturel k inférieur ou égal à n , est le nombre de parties ordonnées de k éléments dans un ensemble à n éléments. Il est noté: A_n^k .

Lorsque l'on choisit k objets parmi n objets et que l'ordre dans lequel les objets sont sélectionnés a une importance, on peut les représenter par un k -uplet d'éléments distincts et on en constitue une liste *ordonnée* sans répétition possible, c'est-à-dire dans laquelle l'ordre des éléments est pris en compte (si l'on permute deux éléments de la liste, on a une liste différente, et un élément ne peut être présent qu'une seule fois). Une telle liste ordonnée est un arrangement:

$$A_n^k = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!} \quad n \leq k$$

Rq: alors que la permutation correspond à la disposition ordonnée de tous les éléments d'un ensemble, l'arrangement correspond à une disposition ordonnée d'un certain nombre d'éléments.

Combinatoire

5 – Combinaison

Lorsque l'on choisit k objets parmi n objets discernables (numérotés de 1 à n) et que l'ordre dans lequel les objets sont placés (ou énumérés) n'a pas d'importance, on peut les représenter par un ensemble à k éléments:

$$C_n^k = \binom{n}{k} = \frac{n \times (n-1) \times \cdots \times (n-k+1)}{k \times (k-1) \times \cdots \times 1} = \frac{n!}{k! (n-k)!} \quad n \leq k$$

La combinaison d'un ensemble d'éléments est une disposition non ordonnée d'un certain nombre d'éléments de cet ensemble.

$\binom{n}{k}$ se prononce « k parmi n ». Les $\binom{n}{k}$ sont les coefficients binômiaux.

Rq: Déterminer un ensemble à k éléments parmi n revient exactement à déterminer son complémentaire soit $n - k$ éléments parmi n :

$$\binom{n}{k} = \binom{n}{n-k}$$

Combinatoire

Triangle de Pascal:

	col.0	col.1	col.2	col.3	col.4	col.5	col.6	col.7	col.8	col.9	...
lig.0	1										
lig.1	1	1									
lig.2	1	2	1								
lig.3	1	3	3	1							
lig.4	1	4	6	4	1						
lig.5	1	5	10	10	5	1					
lig.6	1	6	15	20	15	6	1				
lig.7	1	7	21	35	35	21	7	1			
lig.8	1	8	28	56	70	56	28	7	1		
lig.9	1	9	36	84	126	126	84	36	9	1	
...					...						

Combinatoire

Dans le tableau (triangle de Pascal) chaque nombre à partir de la ligne 1 est la somme des deux nombres de la ligne du dessus, celui sur la même colonne et celui sur la colonne de gauche. Ce tableau contient à la ligne n et à la colonne k la valeur de $\binom{n}{k}$.

Rappel:

On utilise les $\binom{n}{k}$ et le triangle de Pascal pour le développement des expressions algébriques:

$$(a + b)^n = \binom{n}{0} a^n b^0 + \binom{n}{1} a^{n-1} b^1 + \dots + \binom{n}{n-1} a^1 b^{n-1} + \binom{n}{n} a^0 b^n$$

En particulier

$$(a + b)^2 = a^2 + 2ab + b^2 \text{ et } (a + b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

Probabilités

1 – Expérience aléatoire

On appelle expérience aléatoire une expérience qui fait intervenir le hasard. On connaît l'ensemble des issues (ou résultats) possibles sans savoir laquelle de celles-ci se réalisera. Il est possible de répéter un certain nombre de fois cette expérience dans des conditions identiques.

L'ensemble, souvent noté Ω , de toutes les issues possibles est appelé univers ou espace d'échantillonnage de l'expérience.

Exemples:

- On jette un dé à 6 faces, il y a 6 issues possibles: $\Omega = \{1,2,3,4,5,6\}$
- Un fabricant contrôle des composants électroniques en sortie de chaîne de fabrication: il y a 2 issues possibles, ou bien le composant est conforme aux spécifications et il sera vendu ou bien il n'est pas conforme et il sera jeté: $\Omega = \{\text{conforme}, \text{non conforme}\}$
- On choisit un point dans le plan: $\Omega = \mathbb{R}^2$, et l'univers est ici infini

Probabilités

2 – Évènement

Un sous-ensemble, ou partie de Ω est appelé évènement. On note $\mathcal{P}(\Omega)$ l'ensemble des parties de Ω .

Ω est appelé l'évènement certain

\emptyset est appelé l'évènement impossible

L'ensemble qui ne contient qu'une seule issue est un évènement élémentaire (noté par exemple ω)

Exemples:

- Dans l'expérience du dé « on obtient 1 » est un évènement élémentaire, « on obtient un nombre impair » ou « on obtient un nombre inférieur ou égal à 4 » sont deux évènements (non élémentaires).
- Soit $\Omega = \{\omega_1, \omega_2, \omega_3\}$ alors $\mathcal{P}(\Omega) = \{\emptyset, \{\omega_1\}, \{\omega_2\}, \{\omega_3\}, \{\omega_1, \omega_2\}, \{\omega_1, \omega_3\}, \{\omega_2, \omega_3\}, \{\omega_1, \omega_2, \omega_3\}\}$ et $\text{card}(\mathcal{P}(\Omega)) = 8$

Probabilités

3 – Opérations sur les évènements

A et B sont 2 évènements. Alors:

L'évènement contraire de A est son complémentaire dans Ω , noté \bar{A} et se comprend « A n'est pas réalisé ».

La réunion de A et B est noté $A \cup B$ et se comprend « A ou B (ou les deux) sont réalisés ».

L'intersection de A et B est noté $A \cap B$ et se comprend « A et B sont réalisés simultanément ».

Exemple:

- Dans l'expérience du dé si $A = \{1,3,5\}$ = « on obtient un nombre impair » et $B = \{1,2,3,4\}$ = « on obtient un nombre inférieur ou égal à 4 », alors $\bar{A} = \{2,4,6\}$ = « on obtient un nombre pair », $A \cup B = \{1,2,3,4,5\}$ = « on obtient un nombre inférieur ou égal à 5 », $A \cap B = \{1,3\}$ = « on obtient un nombre impair inférieur ou égal à 4 ».

Probabilités

Deux évènements sont incompatibles s'ils ne peuvent se produire simultanément, i.e. si leur intersection $A \cap B$ est vide.

Bien sûr un évènement et son contraire (complémentaire) sont toujours incompatibles.

4 – Loi de probabilité

On peut associer à une expérience aléatoire et à son univers une probabilité qui permet de quantifier le fait qu'un évènement est « probable » ou « peu probable ».

Une probabilité est une application p de $\mathcal{P}(\Omega)$ dans $[0,1]$ telle que $p(\Omega) = 1$ et telle que si A et B sont 2 évènements incompatibles $p(A \cup B) = p(A) + p(B)$

Probabilités

On a donc les propriétés suivantes:

$$0 \leq p(A) \leq 1 \text{ pour tout évènement } A$$

$$p(\emptyset) = 0, p(\Omega) = 1$$

$$p(\bar{A}) = 1 - p(A) \text{ pour tout évènement } A$$

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

Rq: pour chaque univers, on peut imaginer plusieurs lois de probabilités. Dans l'expérience de fabrication de composants électroniques on peut imaginer qu'une chaîne fonctionnant bien ait une probabilité $p(\text{"conforme"}) = 0,95$ et $p(\text{"non conforme"}) = 0,05$ alors qu'une chaîne de moins bonne qualité ait des probabilités associées $p(\text{"conforme"}) = 0,75$ et $p(\text{"non conforme"}) = 0,25$

Il est donc important de définir non seulement l'univers Ω mais aussi la loi de probabilité p dont on le munit. En toute rigueur on parle d'un espace probabilisé (Ω, p)

Probabilités

5 – Cas particulier des univers finis

Pour étudier un phénomène à l'aide des probabilités, on a besoin de connaître la loi de probabilité p , qui est une application de $\mathcal{P}(\Omega)$ dans $[0,1]$, donc a priori on a besoin de connaître sa valeur sur chaque sous-ensemble de Ω . Mais en fait quand Ω est fini, la connaissance de p sur chaque évènement élémentaire suffit: si $A \subset \Omega$ est un évènement quelconque, A est fini et on peut écrire $A = \{a_1, a_2, \dots, a_k\}$ donc $p(A) = p(\{a_1\}) + p(\{a_2\}) + \dots + p(\{a_k\})$

Un cas particulier mais fondamental est le cas de l'équiprobabilité: sur un univers fini, on dit que la loi est équiprobable si tous les évènements élémentaires ont la même probabilité. Dans ce cas, la probabilité de chaque évènement élémentaire est simplement $1/\text{card}(\Omega)$ et la probabilité d'un évènement A est:

$$p(A) = \frac{\text{card}(A)}{\text{card}(\Omega)}$$

Probabilités

6 – Cas particulier des probabilités infinies discrètes

Ω infini est dit discret si on peut énumérer ses éléments, i.e. si on peut écrire $\Omega = \{\omega_1, \omega_2, \dots\}$. Typiquement, cela correspond à des expériences dont le résultat est un entier naturel. Comme dans le cas précédent, on obtient la probabilité d'un évènement quelconque comme somme (éventuellement infinie) des évènements élémentaires qui le composent.

Rq: on a jamais équiprobabilité sur un ensemble infini discret.

7 – Cas particulier des probabilités continues

Pour étudier les probabilités sur des univers continus infinis (par exemple: choix d'un nombre au hasard dans $[0,1]$; durée de vie d'une voiture dans $[0, +\infty[$, ...) on va comme dans le cas fini partir d'« évènements de base » qui permettent de reconstituer tous les évènements et donc toutes les probabilités. Mais ici le problème est un peu plus délicat. En effet, en général, avec un univers continu la probabilité de chaque évènement élémentaire est nulle...

Probabilités

Pour ce qui suit on prend pour Ω un intervalle de \mathbb{R} (par exemple $[0,1]$ ou $[0, +\infty[$, ou \mathbb{R} ...). Ces évènements de base vont ici être les segments $[a, b]$. Dans la plupart des cas, les évènements qui nous intéressent pourront être décrits comme réunion, intersection, complémentaires, ... de segments et on pourra donc déduire ainsi leur probabilité de celles de ces segments grâce aux règles de calcul des probabilités.

Donc dans le cas des probabilités continues on associe à chaque probabilité **une densité de probabilité** qui est une fonction intégrable et positive, telle que $\int_{\Omega} f = 1$. Et la probabilité p est caractérisée par le fait que pour tout évènement A , $p(A) = \int_A f$. Et en particulier, $p([a, b]) = \int_a^b f$ pour tout segment $[a, b]$.

Exemple:

- Le cas le plus simple est celui de la probabilité uniforme sur $[0,1]$, qui correspond à l'expérience « on choisit au hasard un nombre compris entre 0 et 1, sans privilégier aucune valeur »

Probabilités

Alors la densité correspondante est $f = 1$, et la probabilité d'obtenir un nombre entre a et b (pour $0 \leq a \leq b \leq 1$) est égale à $p([a, b]) = \int_a^b 1 = b - a$.

Ainsi avec $a = 0$ et $b = 1$, la probabilité est 1: le choix d'un nombre entre 0 et 1 donne à coup sûr un nombre entre 0 et 1!

Si $a = 0,25$ et $b = 0,75$, on a une chance sur deux que le nombre choisi soit dans l'intervalle $[a; b]$ de longueur $1/2$.

Probabilités

8 – Probabilités conditionnelles

Soit (Ω, p) un espace probabilisé, et A un évènement de probabilité non nulle. On appelle « probabilité que B soit réalisé sachant que A l'est » ou plus simplement « probabilité de B sachant A », la quantité:

$$p(B|A) = \frac{p(A \cap B)}{p(A)}$$

Exemple:

On lance 2 dés équilibrés. Quelle est la probabilité que la somme des résultats soit strictement supérieure à 10 sachant que l'un des dés a donné 6.

« somme > 10 » = $\{(6,6); (6,5); (5,6)\}$ et « l'un des dés donne 6 » est de cardinal 11; l'intersection est de cardinal 3

Donc la probabilité est $3/11$

Probabilités

Connaissant $p(A \setminus B)$ on aimerait parfois connaître $p(B \setminus A)$. C'est souvent possible en écrivant de 2 manières différentes $p(B \cap A)$ à l'aide des définitions de $p(A \setminus B)$ et $p(B \setminus A)$:

$$p(A \cap B) = p(A)p(B \setminus A) = p(B)p(A \setminus B)$$

Exemple:

40 des 55 étudiants de IESE3 ont eu la moyenne au TC Maths, et 22 des 35 étudiants TIS3. Quelle est la probabilité qu'un étudiant ayant eu la moyenne soit en TIS?

On a $p(M/IESE3) = 40/55$ et $p(M/TIS3) = 22/35$

De plus $p(M) = 62/90$ et $p(TIS3 \cap M) = 22/90$

$$\text{Donc } p(TIS3 \setminus M) = \frac{p(TIS3 \cap M)}{p(M)} = \frac{22/90}{62/90} = \frac{22}{62}$$

Plus directement on écrit:

$$p(B \setminus A) = \frac{p(A \setminus B)p(B)}{p(A)}$$

Probabilités

Dans les cas plus compliqués on peut avoir besoin de la **formule de Bayes**.

Considérons donc les événements incompatibles A_1, A_2, \dots, A_n et un événement B qui ne peut se produire que si l'un des A_i se produit, les $p(B \setminus A_i)$ étant connus. On cherche la probabilité pour que B s'étant produit A_k en soit la cause.

On a : $p(B) = p(A_1 \cap B) + p(A_2 \cap B) + \dots + p(A_n \cap B)$

Et comme $p(A_k \cap B) = p(A_k) p(B \setminus A_k)$, on obtient donc la **formule des probabilités totales**:

$$p(B) = p(B \setminus A_1)p(A_1) + p(B \setminus A_2)p(A_2) + \dots + p(B \setminus A_n)p(A_n)$$

Alors en écrivant $p(A_k \setminus B) = p(A_k \cap B)/p(B) = p(B \setminus A_k) p(A_k) / p(B)$, et en remplaçant $p(B)$ par la formule précédente on obtient la **formule de Bayes**:

$$p(A_k \setminus B) = \frac{p(A_k)p(B \setminus A_k)}{\sum_{i=1}^n p(A_i)p(B \setminus A_i)}$$

Probabilités

Exemple:

Un test de dépistage d'une maladie rare touchant une personne sur 10000 semble efficace: il détecte 99% des personnes infectées, avec seulement 0,5% de « faux positifs ». Quelle est la probabilité qu'une personne dont le test est positif soit effectivement malade?

Soit M l'évènement « personne malade », soit P l'évènement « test positif » alors:

$$p(M|P) = \frac{p(P|M)p(M)}{p(P|M)p(M) + p(P|\bar{M})p(\bar{M})} \simeq 1,94\%$$

9 – Evènements indépendants

On dit que 2 évènements sont indépendants quand l'un des 2 est de probabilité nulle ou bien quand les deux sont de probabilité non nulle, si le fait de savoir que l'un est réalisé n'influe pas sur la probabilité que l'autre le soit. Autrement dit 2 évènements de probabilité non nulle sont indépendants quand $p(B|A) = p(B)$ (ou de manière équivalente $p(A|B) = p(A)$).

Probabilités

Ainsi 2 évènements sont indépendants si et seulement si:

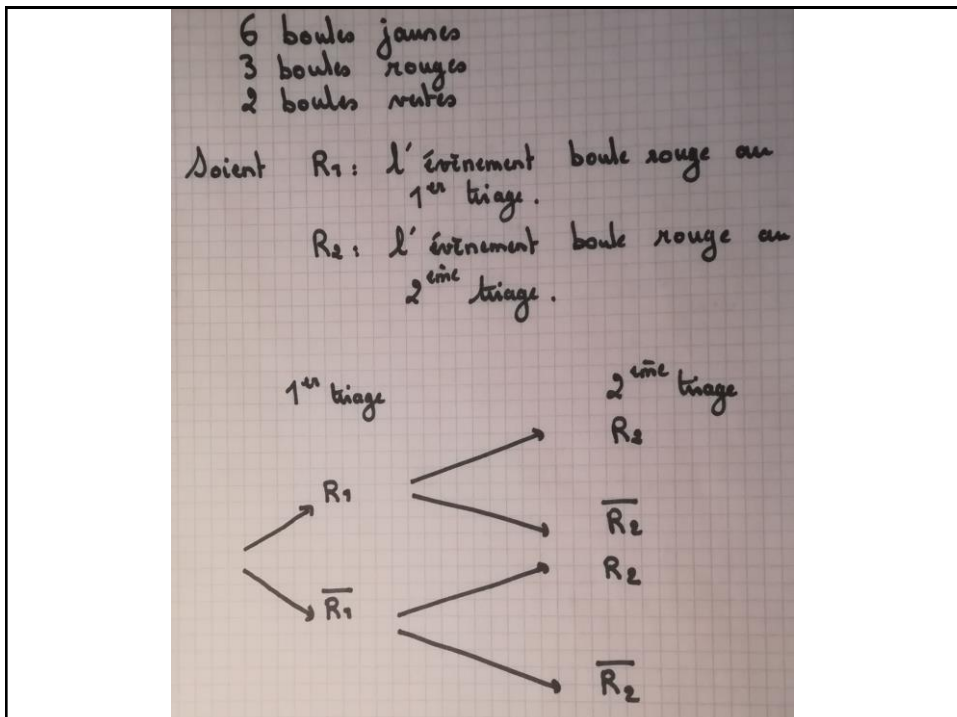
$$p(A \cap B) = p(A)p(B)$$

Rq: ne pas confondre les 2 notions d'évènements indépendants et d'évènements incompatibles!

2 évènements incompatibles ne sont jamais indépendants (sauf si les 2 sont de probabilité nulle). En effet, si A et B sont incompatibles et que l'on sait que A est réalisé, justement, B ne peut pas se produire.... Il n'y a donc pas de dépendance.

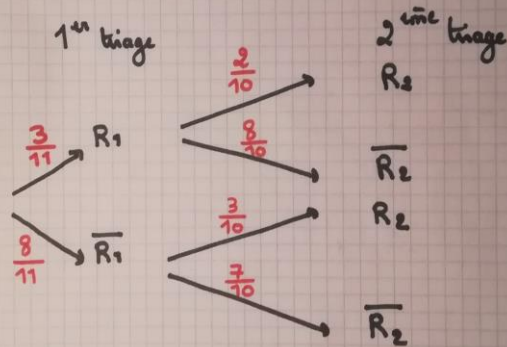
Exercices

Une urne contient 6 boules jaunes, 3 boules rouges et 2 boules vertes. Le mode de tirage des boules dans l'urne est successif sans remise. On tire 2 boules; quelle est la probabilité d'obtenir une boule rouge au deuxième tirage?



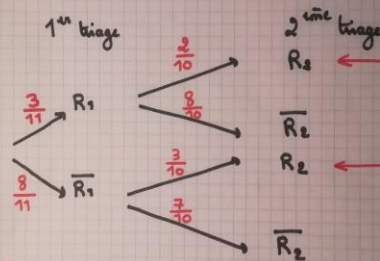
6 boules jaunes
3 boules rouges
2 boules noires

Soient R_1 : l'événement boule rouge au 1^{er} tirage.
 R_2 : l'événement boule rouge au 2^{ème} tirage.



6 boules jaunes
3 boules rouges
2 boules noires

Soient R_1 : l'événement boule rouge au 1^{er} tirage.
 R_2 : l'événement boule rouge au 2^{ème} tirage.



D'où

$$\begin{aligned} P(R_2) &= \frac{3}{11} \times \frac{2}{10} + \frac{8}{11} \times \frac{3}{10} \\ &= \frac{6}{110} + \frac{24}{110} \\ &= \frac{30}{110} \\ &= \frac{3}{11} \end{aligned}$$

On peut aussi directement écrire :

$$\begin{aligned}
 P(R_2) &= P(R_2 \cap R_1) + P(R_2 \cap \bar{R}_1) \\
 &= P(R_1) P(R_2 | R_1) + P(\bar{R}_1) P(R_2 | \bar{R}_1) \\
 &= \frac{3}{11} \times \frac{2}{10} + \frac{8}{11} \times \frac{3}{10} \\
 P(R_2) &= \frac{3}{11}
 \end{aligned}$$

Pour ceux qui veulent revoir les probabilités

- Probabilité conditionnelle
<https://youtu.be/BbD9Dpk3E5Y>
- Probabilité totale
<https://youtu.be/izEY7ggfykM>
- Arbre pondéré et probabilité conditionnelle
<https://youtu.be/YnM9-9K1fWY>
- Evènements indépendants
<https://youtu.be/PupQYTVqJrg>

Chapitre 2

Rappels – Variables Aléatoires (VA)

Variable aléatoire (VA)

Etant donné un univers Ω , une variable aléatoire réelle (VAR) X est une application de Ω dans \mathbb{R}

$$X: \omega \in \Omega \rightarrow X(\omega) \in \mathbb{R}$$

1 - Loi de probabilité

Soit Ω un univers muni d'une probabilité p , et soit X une VAR. On appelle loi de probabilité de X , notée p_X , l'application qui à toute partie A de \mathbb{R} associe

$$p_X(A) = p(\{\omega \in \Omega: X(\omega) \in A\})$$

Variable aléatoire (VA)

2 - Fonction de répartition:

La fonction de répartition de la VAR X est définie par:

$$F_X(x) = p(X \leq x), x \in \mathbb{R}$$

• Propriétés de la fonction de répartition:

$$1/ 0 \leq F_X \leq 1$$

2/ F_X tend vers 0 en $-\infty$ et vers 1 en $+\infty$

3/ F_X est croissante

4/ F_X est continue à droite

Proposition:

$$p(a < X \leq b) = F_X(b) - F_X(a), \forall a < b$$

<https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-l-inf-probas.pdf>

Variable aléatoire discrète

Pour une variable aléatoire discrète X qui prend les valeurs $\{x_1, x_2, x_3, \dots, x_n\}$ avec les probabilités $\{p_1, p_2, p_3, \dots, p_n\}$ (où $p_i = p(X = x_i)$), on définit les 3 grandeurs suivantes:

L'espérance de X :

$$E(X) = \sum_{i=1}^n x_i p_i$$

La variance de X :

$$V(X) = \sum_{i=1}^n p_i (x_i - E(X))^2$$

Rq: les sommes ci-dessus peuvent être infinies

L'écart-type de X:

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

Variable aléatoire continue

Pour une variable aléatoire continue X , on définit les 3 grandeurs suivantes par analogie avec le cas discret:

L'espérance de X :

$$E(X) = \int_I x f_X(x) dx$$

La variance de X :

$$V(X) = \int_I f_X(x) (x - E(X))^2 dx$$

L'écart-type de X :

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

Variable aléatoire

Plus précisément les probabilités élémentaires $p_i = p(X = x_i)$ sont remplacés par la fonction de densité $f_X(x)$, que l'on peut interpréter en disant que $f_X(x)dx$ est la probabilité de l'évènement infinitésimal $p(x \leq X \leq x + dx)$.

Dans le cas discret, on obtient la probabilité de l'évènement A par la formule $p(A) = \sum_{a \in A} p(X = a)$ (la probabilité d'un évènement est la somme des probabilités des évènements élémentaires qui le composent). Dans le cas continu, la somme est remplacée par une intégrale, et on a la formule $p(A) = \int_A f_X(x) dx$.

Variable aléatoire (VA)

VAD X	VAC X
La fonction de répartition F est une fonction en escaliers sur R	La fonction de répartition F est continue sur R
ensemble $X(\Omega)$ des valeurs prises par X = ensemble des x en lesquels F croît strictement	
$X(\Omega)$ = ensemble des x en lesquels le saut $P(X=x)$ de F est non nul = $\{x \in R / P(X=x) \neq 0\}$ $X(\Omega) = \{x_k\}$ est fini ou dénombrable connaître la loi de X \Leftrightarrow connaître $X(\Omega) = \{x_k\}$ et les $P(X=x_k)$	$X(\Omega)$ = ensemble des x en lesquels la dérivée f de F est non nulle = $\{x \in R / f(x) \neq 0\}$ $X(\Omega)$ est un intervalle ouvert ou une réunion d'intervalles ouverts connaître la loi de X \Leftrightarrow connaître $X(\Omega) = \{x \in R / f(x) \neq 0\}$ et l'expression de f(x) sur $X(\Omega)$
$\forall I$ un intervalle de R, $P(X \in I) = \sum_{x_k \in I} P(X=x_k)$	$\forall I$ un intervalle de R, $P(X \in I) = \int_I f(x) dx$
$P(X \in]-\infty, +\infty]) = 1 = \sum_{x_k \in R} P(X=x_k)$	$P(X \in]-\infty, +\infty]) = 1 = \int_{-\infty}^{+\infty} f(x) dx$
$E(X) = \sum_{x_k \in R} x_k P(X=x_k)$	$E(X) = \int_{-\infty}^{+\infty} x f(x) dx$
$E(\varphi(X)) = \sum_{x_k \in R} \varphi(x_k) P(X=x_k)$	$E(\varphi(X)) = \int_{-\infty}^{+\infty} \varphi(x) f(x) dx$
$V(X) = E[(X - E(X))^2] = E(X^2) - E^2(X)$	
Extrait du cours de Probabilités d'année 3 – F. Blanchet	

Espérance et Variance

$$E(a) = a$$

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

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

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

$$V(X) = E(X^2) - E^2(X)$$

$$V(X + a) = V(X)$$

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

$$V(X + Y) = V(X) + V(Y) + 2 \text{cov}(X, Y)$$

$$V(X + Y) = V(X) + V(Y) \text{ si } X \text{ et } Y \text{ sont indépendantes}$$

Covariance

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

$$\text{cov}(X, Y) = \text{cov}(Y, X)$$

$$\text{cov}(X, X) = V(X)$$

Couple de variables aléatoires

Il est courant d'étudier plusieurs variables définies sur une même population, et les liens entre ces variables. Par exemple l'âge, la taille, le poids d'un groupe de personnes, la puissance consommée d'un composant et sa durée de vie etc.

Nous allons ici introduire la notion de couple de variables aléatoires

On considère 2 variables X et Y définies sur un même univers Ω , (X, Y) est une nouvelle variable aléatoire à 2 dimensions : c'est une application de $\mathcal{P}(\Omega)$ dans \mathbb{R}^2 .

1 – Loi d'un couple de variables aléatoires discrètes

Si Ω est discret

Moments des variables aléatoires

Moment centré d'ordre k: $\mu_k = E([X - E(X)]^k)$

Moment centré d'ordre 1: $\mu_1 = 0$

Moment centré d'ordre 2: $\mu_2 = V(X)$

Moment non centré d'ordre k: $m_k = E(X^k)$

Remarques:

- La variance est le moment centré d'ordre 2 et l'espérance est le moment non centré d'ordre 1.
- Le moment d'ordre 3 donne une indication sur l'asymétrie de la distribution d'une VA (skewness).
- Le moment d'ordre 4 donne une indication sur l'aplatissement de la distribution d'une VA (kurtosis).

Principales lois discrètes

Loi uniforme : $U(n)$

Loi d'une VA X prenant les valeurs 1, 2, ..., n avec la même probabilité

$$P(X = x) = \frac{1}{n} \quad \forall x \in \{1, 2, \dots, n\}$$

Moments:

$$E(X) = \frac{n+1}{2} \quad V(X) = \frac{n^2 - 1}{12}$$

Exemple : la loi de probabilité associée à l'expérience aléatoire consistant à jeter un dé à 6 faces non pipé est défini par la tableau suivant:

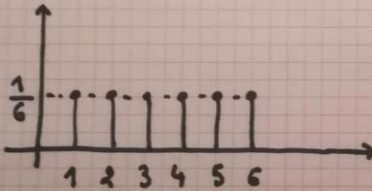
Principales lois discrètes

Loi uniforme : $U(n)$

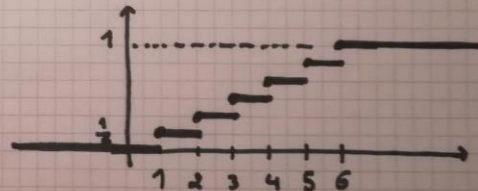
X prend les valeurs $\{1, 2, 3, 4, 5, 6\}$

$$\forall x \in \{1, 2, 3, 4, 5, 6\} \quad P(X=x) = \frac{1}{6}$$

Loi de probabilité :



Fonction de répartition :



Principales lois discrètes

Loi de Bernoulli : $B(1, p)$

Loi d'une VA X ne pouvant prendre que 2 valeurs (par exemple 0 et 1) avec les probabilités associées p et q

$$P(X=1) = p \quad \text{et} \quad P(X=0) = q = 1 - p$$

Moments:

$$E(X) = p \quad V(X) = p(1 - p)$$

Exemple : la loi de probabilité associée à l'expérience aléatoire consistant à jeter une pièce de monnaie non truquée

Principales lois discrètes

Loi de Bernoulli : $B(1, p)$

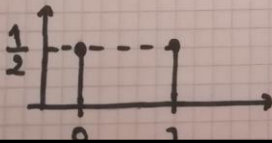
$X = 0$ si pile

$X = 1$ si face

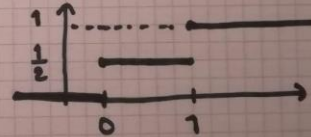
X prend les valeurs $\{0, 1\}$

$$\forall x \in \{0, 1\} \quad P(X=x) = \frac{1}{2}$$

Loi de probabilité :

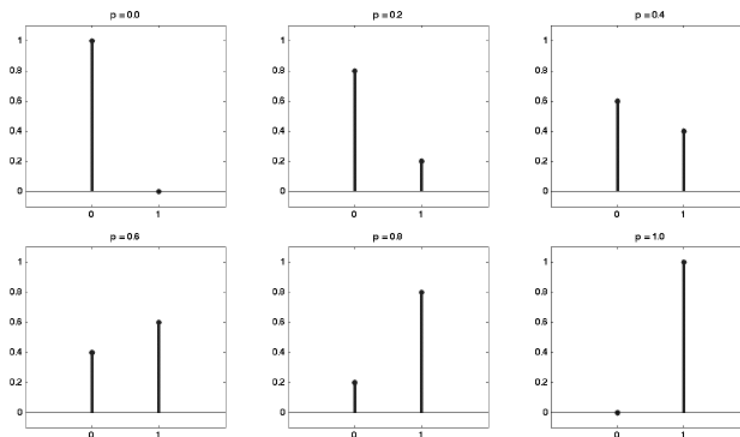


Fonction de répartition :



Principales lois discrètes

Loi de Bernoulli : $B(1, p)$



Principales lois discrètes

Loi Binomiale: $B(n, p)$

Répétition de l'expérience de Bernouilli n fois et X est le nombre de fois où la variable de Bernouilli prend la valeur 1 (ou encore X est la somme des résultats de l'expérience). On connaît alors la probabilité d'obtenir k valeurs 1:

$$P(X = k) = C_n^k p^k (1-p)^{n-k}$$

Moments:

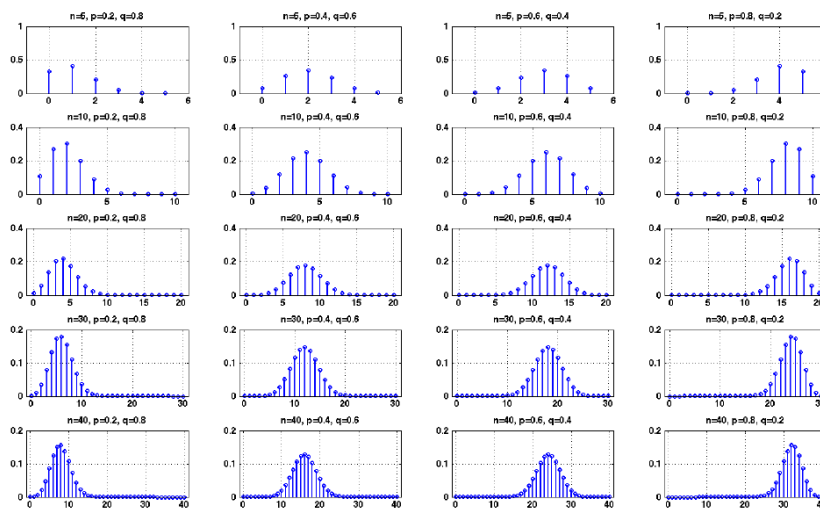
$$E(X) = np$$

$$V(X) = np(1-p)$$

Exemple : nombre de réalisations de piles après n lancers

Principales lois discrètes

Loi Binomiale: $B(n, p)$



Exercice 3 : déterminer la loi de probabilité de la variable : nombre de garçons dans une famille de 7 enfants:

$$\begin{cases} X = 1 \text{ si garçon} \\ X = 0 \text{ si fille} \\ P(X=1) = \frac{1}{2} \end{cases}$$

Rappel:

$$C_m^k = \frac{m!}{k!(m-k)!}$$

$X \sim \mathcal{B}(1; \frac{1}{2})$: Loi de Bernoulli

Y : nbre de garçons dans une famille de 7 enfants

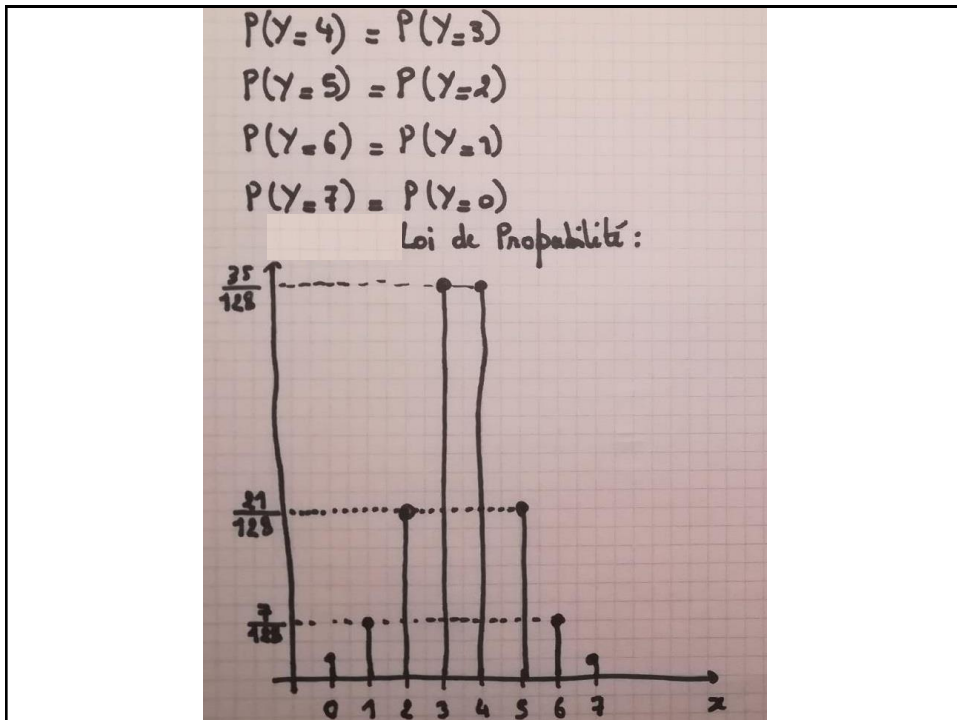
Y peut prendre les valeurs : $\{0, 1, 2, 3, 4, 5, 6, 7\}$

$$P(Y=k) = C_7^k \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{7-k} = C_7^k \frac{1}{2^7} = C_7^k \frac{1}{128}$$

$\forall k \in \llbracket 0, 7 \rrbracket$

D'où : $P(Y=0) = \frac{1}{128}$

$$P(Y=2) = \frac{7!}{5!2!} \frac{1}{128} = \frac{21}{128}$$



Principales lois discrètes

Loi de Poisson: $P(\lambda)$

$$P(X = x) = e^{-\lambda} \frac{\lambda^x}{x!} \forall x \in \mathbb{N}$$

Moments:

$$E(X) = \lambda$$

$$V(X) = \lambda$$

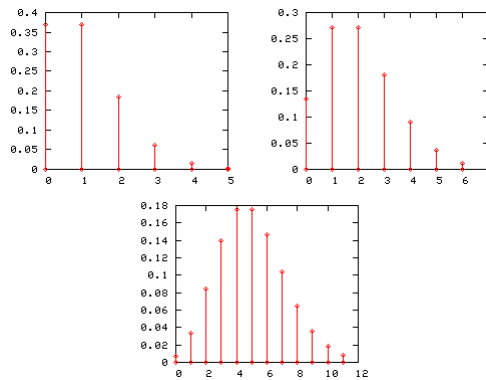
Remarque : si λ est grand alors la loi de Poisson tend vers une loi normale

Exemple : nombre d'appels téléphoniques pendant un intervalle de temps

Principales lois discrètes

Loi de Poisson: $P(\lambda)$

Comme toute loi de probabilité discrète, une loi de Poisson peut être représentée par un diagramme en bâtons. Ci-dessous sont représentés les diagrammes en bâtons des lois de Poisson de paramètres 1, 2 et 5.



Principales lois continues

Loi uniforme : $U_{[0,a]}$ sur $[0,a]$

Densité de probabilité

$$f(x) = \begin{cases} \frac{1}{a} & \text{sur } [0, a] \\ 0 & \text{ailleurs} \end{cases}$$

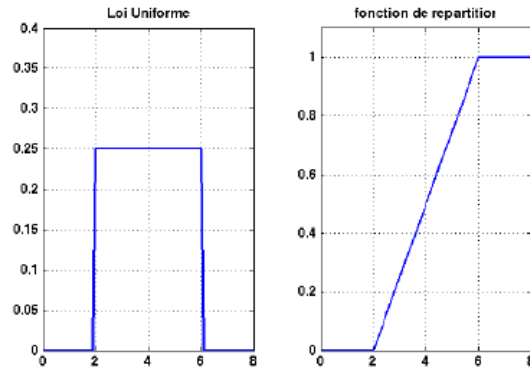
Moments:

$$E(X) = \frac{a}{2} \qquad V(X) = \frac{a^2}{12}$$

Remarque : la somme de 2 lois uniformes n'est pas uniforme!

Principales lois continues

Loi uniforme : $U_{[2,6]}$



Principales lois continues

Loi exponentielle: $E(\lambda)$

Densité de probabilité

$$f(x) = \lambda e^{-\lambda x} \text{ si } x > 0$$

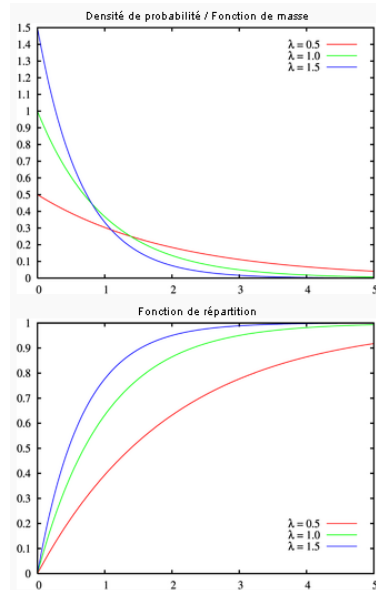
Moments:

$$E(X) = \frac{1}{\lambda} \qquad V(X) = \frac{1}{\lambda^2}$$

Exemple : Durée de vie d'un phénomène ou dans votre domaine d'un composant électrique

Principales lois continues

Loi exponentielle: $E(\lambda)$



Principales lois continues

Loi normale : $N(\mu, \sigma^2)$

Densité de probabilité

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \forall x \in \mathbb{R}$$

Moments:

$$E(X) = \mu \qquad V(X) = \sigma^2$$

Exemples : variation du diamètre d'une pièce, répartition des erreurs de mesure autour de la « valeur vraie »

Principales lois continues

Loi normale : $N(\mu, \sigma^2)$

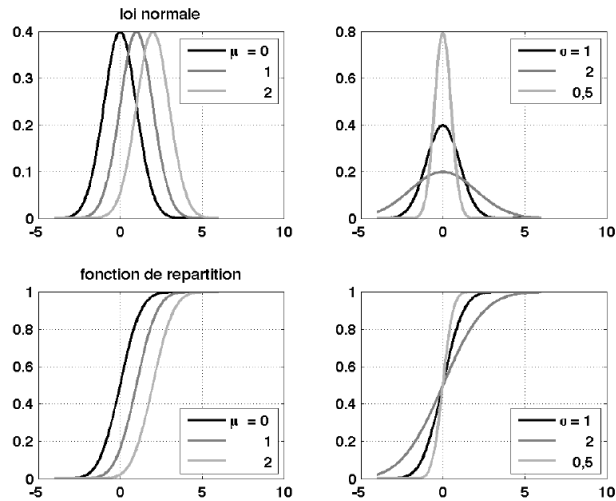
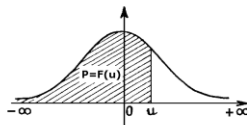


Table de la loi normale centrée réduite

	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0,1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0,2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0,3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0,4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0,5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0,6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0,7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0,8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0,9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1,1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1,2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1,3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1,4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1,5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1,6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1,7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1,8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1,9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2,1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2,2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2,3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2,4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2,5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2,6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2,7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2,8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2,9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

FONCTION DE RÉPARTITION DE LA LOI NORMALE RÉDUITE
(Probabilité de trouver une valeur inférieure à u)



u	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7290	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,0	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,0	0,9772	0,9779	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

Table pour les grandes valeurs de u

u	3,0	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4,0	4,5
$F(u)$	0,99865	0,99904	0,99931	0,99952	0,99966	0,99976	0,999841	0,999928	0,999968	0,999997

Exercice: on suppose que X suit la loi normale centrée réduite

$$\begin{aligned} P(X \leq 0) &= \\ P(X \leq 0,8) &= \\ P(X \leq 1,64) &= \\ P(X \leq 1,96) &= \\ P(X \leq 3) &= \end{aligned}$$

$$\begin{aligned} P(X > 1,64) &= \\ P(X > 2,1) &= \end{aligned}$$

$$\begin{aligned} P(X \leq -2) &= \\ P(-1,96 \leq X \leq 1,96) &= \end{aligned}$$

Trouver le z tel que :
 $P(X \leq z) = 0,95$
 $P(-z < X < z) = 0,95$

Enfin X suit une loi normale centrée en 170 et d'écart-type 40. Calculer les probabilités suivantes :
 $P(X \leq 160) =$

$$P(X \leq 0) = 0,5$$

$$P(X \leq 0,8) = 0,7881$$

$$P(X \leq 1,64) = 0,9495$$

$$P(X \leq 1,96) = 0,9750$$

$$P(X \leq 3) = 0,9986$$

$$P(X > 1,64) = 1 - P(X \leq 1,64) = 1 - 0,9495 = 0,0505$$

$$P(X > 2,1) = 1 - P(X \leq 2,1) = 1 - 0,9821 = 0,0179$$

$$P(X \leq -2) = P(X \geq 2) = 1 - P(X \leq 2) = 1 - 0,9772 = 0,0228$$

$$\begin{aligned} P(-1,96 \leq X \leq 1,96) &= P(X \leq 1,96) - P(X < -1,96) \\ &= 0,9750 - (1 - P(X < 1,96)) \\ &= 0,9750 - 1 + 0,9750 = 0,95 \end{aligned}$$

Trouver z tel que : (rq. il faut lire la table dans le "sens inverse")

$$P(X \leq z) = 0,95 \Rightarrow z = 1,645$$

$$P(-z \leq X \leq z) = 0,95 \Rightarrow P(X \leq z) - 1 + P(X \leq z) = 0,95$$

$$\Leftrightarrow 2P(X \leq z) = 1,95$$

$$\Leftrightarrow P(X \leq z) = \frac{1,95}{2} = 0,975$$

$$\Rightarrow z = 1,96$$

$$X \hookrightarrow \mathcal{N}(\mu=170; \sigma^2=40^2)$$

on ne dispose pas des valeurs de cette loi

\Rightarrow On se ramène à la loi $\mathcal{N}(0,1)$

$$X \hookrightarrow \mathcal{N}(170; 40^2)$$

on centre et on réduit la variable

$$\Rightarrow \underbrace{\frac{X-170}{\sqrt{40^2}}}_{X^*} \hookrightarrow \mathcal{N}(0,1)$$

$$\text{D'où } P(X \leq 160) = P\left(\frac{X-170}{40} \leq \frac{160-170}{40}\right)$$

$$= P(X^* \leq -0,25)$$

$$= P(X^* > 0,25)$$

$$= 1 - P(X^* < 0,25)$$

$$= 1 - 0,5987$$

$$= 0,4013$$

Pour ceux qui veulent revoir la loi normale

- Probabilité normale centrée réduite

<https://youtu.be/8wjwbCxM7G0>

- Probabilité et loi normale

<https://youtu.be/5JTPEEf4wbl>

- Loi normale

<https://youtu.be/SfVuKV4TrGI>



Chapitre 3

Statistiques descriptives



The banner at the top of the slide features the Polytech Grenoble logo on the left and a photograph of a snow-capped mountain range on the right. The main title 'Chapitre 3' is centered in a large, bold, black font, with the subtitle 'Statistiques descriptives' centered below it in a smaller black font. At the bottom right, there are two logos: 'Grenoble INP UGA' and 'Polytech Grenoble'.

Introduction

On va dans ce chapitre étudier un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « **variables** ». Nous nous arrêterons ici à l'étude de 1 ou 2 variables. Le cas p variables sera traité aux prochains chapitres.

Chaque individu d'une population est décrit par un ensemble de caractéristiques appelées **variables** ou caractères.

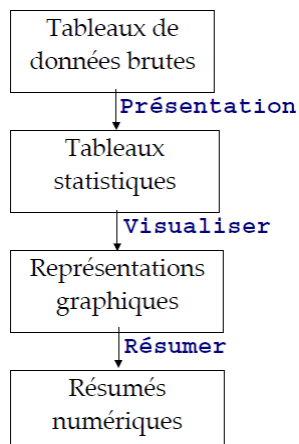
- **Variables quantitatives ou numériques** : par exemple taille, poids, volume s'exprimant par des nombres réels sur lesquels les opérations arithmétiques courantes ont un sens. Elles sont *discrètes* (nombre fini ou dénombrable de valeurs) comme le nombre de défauts d'une pièce ou *continues* si toutes les valeurs d'un intervalle de \mathbb{R} sont acceptables.

- **Variables qualitatives** s'exprimant par l'appartenance à une catégorie ou une modalité d'un ensemble fini. Elles sont purement *nominales* : par exemple, la catégorie socioprofessionnelle d'un actif, ou *ordinales* lorsque l'ensemble des catégories est muni d'un ordre (Exemple: « résistant », « moyennement résistant » et « très résistant »).

Ce chapitre est en grande partie extrait de Introduction à l'inférence statistique de Gérald Baillargeon.
Editeur : Smg (5 novembre 1999)

Introduction

Nous passons ici en revue différentes solutions pour décrire un échantillon de valeurs.



Dépouillement des données

Série numérique	Données rangées (ordre croissant)
78,9 83,4 90,0 88,2 89,3 60,8	56,5 60,3 60,8 65,0 67,4 70,2
75,0 88,0 92,3 73,1 73,7 76,3	71,6 73,1 73,7 74,2 75,0 76,3
60,3 67,4 84,2 70,2 94,6 97,8	77,0 77,2 78,5 78,9 80,0 83,4
92,1 80,0 77,0 77,2 74,2 84,5	84,2 84,2 84,5 88,0 88,2 89,3
93,7 78,5 65,0 56,5 71,6 84,2	90,0 92,1 92,3 93,7 94,6 97,8

Classes	Fréquences absolues
$55 \leq X < 65$	3
$65 \leq X < 75$	7
$75 \leq X < 85$	11
$85 \leq X < 95$	8
$95 \leq X < 105$	1
	Total : 30

Notions de limites de classes, d'amplitude classe, de fréquences absolues et relatives

Dépouillement des données

On doit avant toute chose **se fixer le nombre de classes : K**. Mais comment faire en pratique ?

Ce choix est évidemment fonction du nombre de données à dépouiller mais également de l'étalement de ces données. Le but étant bien entendu de conserver à la distribution sa forme générale.

On peut utiliser :

Formule empirique: $K = \sqrt{n}$

Critère de Brooks-Carruthers: $K < 5 \log_{10}(n)$

Critère de Huntsberger-Sturges: $K = 1 + 10 \log_{10}(n) / 3$

Dans la pratique l'utilisation de logiciels de statistiques permet simplement de tester différentes valeurs de K...

Distributions de fréquences (cas continu)

Exercice:

Dans un atelier mécanique on a vérifié le diamètre de tiges tournées sur un tour automatique. Le diamètre peut fluctuer selon le réglage du tour. Le diamètre devrait normalement se situer entre 36 et 44 mm. 60 tiges ont été mesurées avec un micromètre de précision et les résultats sont présentés dans l'ordonné croissant dans le tableau suivant :

Série numérique																																																								
37	37	37,2	37,6	37,6	37,8	37,8	37,9	38,4	38,4	38,5	38,5	38,6	38,7	38,8	38,9	39	39	39,1	39,1	39,2	39,4	39,4	39,4	39,4	39,5	39,5	39,5	39,6	39,7	39,7	39,9	39,9	39,9	40	40	40	40	40,1	40,3	40,4	40,4	40,6	40,6	40,7	40,8	40,9	40,9	41,2	41,2	41,3	41,5	41,5	42,1	42,2	42,6	43,1

Distributions de fréquences (cas continu)

1/ Estimation du nombre de classes par la formule de Sturges:

2/ Donner l'étendue des données puis en fonction du nombre de classes choisis ci-dessus donner l'amplitude de chaque classe:

L'étendue = valeur max – valeur min

3/ Donner le tableau des classes et des fréquences absolues et relatives associées:

4/ Donner également le tableau des fréquences cumulées:

Fréquences cumulées

Classes	Fréquences absolues	Fréquences absolues cumulées	Fréquences relatives	Fréquences relatives cumulées
$36,5 \leq X < 37,5$	3	3	$3 / 60 = 0,05$	0,05
$37,5 \leq X < 38,5$	7	10	0,12	0,17
$38,5 \leq X < 39,5$	17	27	0,28	0,45
$39,5 \leq X < 40,5$	18	45	0,3	0,75
$40,5 \leq X < 41,5$	9	54	0,15	0,90
$41,5 \leq X < 42,5$	4	58	0,07	0,97
$42,5 \leq X < 43,5$	2	60	0,03	1

Les courbes des fréquences cumulées croissantes (ou décroissantes) permettent de faire correspondre à une valeur d'une série le nombre d'observations qui lui sont inférieures (ou supérieures). Ces courbes permettent de répondre très simplement à des questions du genre : combien de salariés ont un salaire inférieur ou égal à 1300 euros ? Combien de salariés ont 20 ans et plus ? etc.

Distributions de fréquences (cas discret)

Exemple :

A la sortie d'une chaîne d'assemblage on a prélevé 20 échantillons successifs comportant chacun 10 pièces. Un contrôle visuel a été effectué sur chacune des pièces et on a noté le nombre de pièces présentant un défaut. Les résultats sont présentés dans le tableau ci-contre :

Nombre de pièces présentant un défaut																			
0	1	0	2	0	0	1	2	0	0	1	0	1	3	0	1	2	1	0	0

Nombre de pièces défectueuses	Nombres d'échantillons ou fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
0	10	0,5	0,5
1	6	0,3	0,8
2	3	0,15	0,95
3	1	0,05	1

Distributions de fréquences (cas discret)

Exemple :

A la sortie d'une chaîne d'assemblage on a prélevé 20 échantillons successifs comportant chacun 10 pièces. Un contrôle visuel a été effectué sur chacune des pièces et on a noté le nombre de pièces présentant un défaut. Les résultats sont présentés dans le tableau ci-contre :

Nombre de pièces défectueuses	Nombres d'échantillons ou fréquences absolues	Fréquences relatives	Fréquences relatives cumulées
0	10	0,5	0,5
1	6	0,3	0,8
2	3	0,15	0,95
3	1	0,05	1

Représentations graphiques

Lorsque la variable est qualitative on utilisera généralement une représentation en diagrammes en bâtons.

Lorsque la variable quantitative est discrète (VAD):

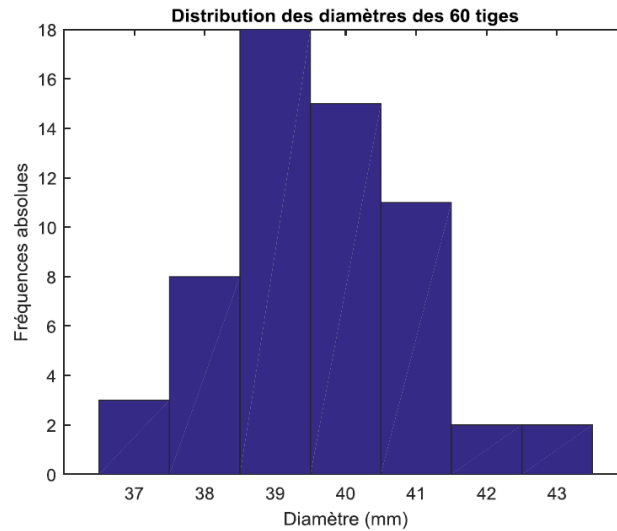
Diagrammes en bâtons : le diagramme en bâtons est constitué en portant en abscisses les valeurs de la variable discrète et en traçant parallèlement à l'axe des ordonnées un bâton de longueur proportionnelle à la fréquence (absolue ou relative) de chaque valeur de la variable.

Lorsque la variable quantitative est continue (VAC):

Histogramme : l'histogramme est constitué de rectangles juxtaposés dont chacune des bases est égale à l'intervalle de chaque classe et dont la hauteur est telle que la surface soit proportionnelle à la fréquence (absolue ou relative) de la classe correspondante.

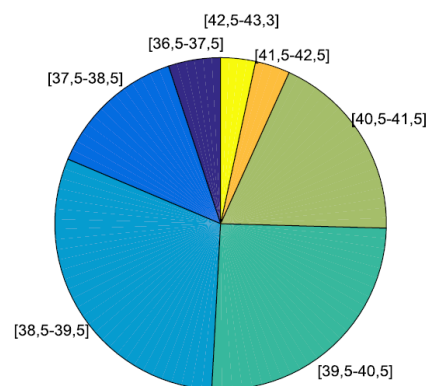
On peut à partir de l'histogramme tracer le polygone de fréquences ; il permet de représenter l'histogramme sous la forme d'une courbe qui va joindre les milieux des sommets des rectangles de l'histogramme.

Représentations graphiques



Représentations graphiques

Un autre type de représentation des fréquences est le diagramme à secteurs circulaires. Il consiste en un cercle dont l'aire est décomposée en secteurs ; la taille du secteur (l'angle au centre de chaque secteur) est proportionnelle à la fréquence absolue ou relative.



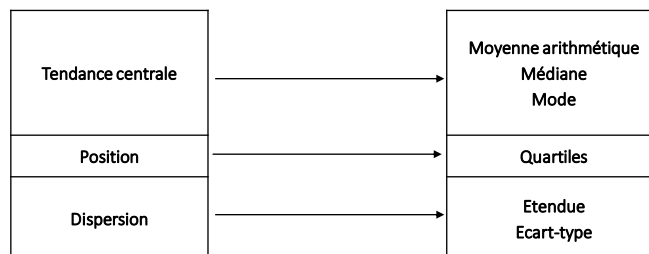
Résumé par des valeurs

Nous avons vu que nous pouvons représenter une série numérique à l'aide de tableaux et de graphiques. Il est également intéressant de pouvoir décrire la série numérique à l'aide de valeurs caractéristiques.

On distingue 3 types de caractéristiques :

- *Les caractéristiques (ou mesures) de tendance centrale* ; elles permettent d'obtenir une idée de l'ordre de grandeur des valeurs constituant la série et indiquent également la position où semblent se rassembler les valeurs de la série.
- *Les caractéristiques (ou mesures) de dispersion* ; elles quantifient les fluctuations de valeurs observées autour de la valeur centrale. Elles permettent d'apprécier l'étalement de la série de valeurs c'est-à-dire si les valeurs s'écartent les unes des autres ou d'étalement de la valeur centrale.
- *Les caractéristiques de forme* ; elles donnent une idée de la symétrie et de l'aplatissement d'une distribution. Toutefois ces dernières sont moins souvent utilisées.

Résumé par des valeurs



Tendance centrale

La moyenne arithmétique :

La caractérisation de la tendance centrale est généralement le premier indicateur qui est regardé. Il permet de connaître la valeur autour de laquelle se répartissent les valeurs de la série. La moyenne permet de résumer par un seul nombre l'ensemble des données.

La médiane :

La **médiane** est la valeur m telle que le nombre de valeurs de l'ensemble supérieures ou égales à m est égal au nombre de valeurs inférieures ou égales à m .

Le mode :

Le mode ou valeur dominante désigne la valeur la plus représentée de la série. Une répartition peut être unimodale ou plurimodale (bimodale, trimodale...), si deux ou plusieurs valeurs de la variable considérée émergent également.

Tendance de position

Les quartiles :

On appelle premier quartile tout réel Q_1 tel que :

- Au moins 25% des termes de la série ont une valeur inférieure ou égale à Q_1
- Et au moins 75% des termes de la série ont une valeur supérieure ou égale à Q_1

On appelle troisième quartile tout réel Q_3 tel que :

- Au moins 75% des termes de la série ont une valeur inférieure ou égale à Q_3
- Et au moins 25% des termes de la série ont une valeur supérieure ou égale à Q_3

Remarques :

- Le deuxième quartile correspond à la médiane
- Les 3 quartiles partagent l'ensemble des valeurs en 4 sous-ensembles

On constate donc que la détermination des quartiles est différente suivant que l'effectif total n est multiple ou non de 4 :

- Si l'effectif total n'est pas un multiple de 4, pas de difficulté les quartiles sont les termes de rang immédiatement supérieur à $n/4$ et $3n/4$
- Lorsque l'effectif est un multiple de 4 alors l'usage veut que l'on choisisse pour les quartiles les termes de rang $n/4$ et $3n/4$

Tendance de dispersion

L'étendue : écart entre les valeurs min et max

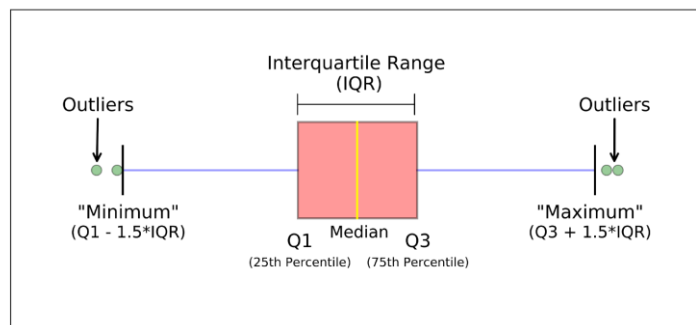
L'intervalle interquartile : écart entre le 1^{er} et le 3^{ème} quartile

La variance de l'échantillon :

La variance, et par conséquent l'écart-type, nous permet de caractériser de quelle façon les valeurs observées se répartissent autour de la moyenne. Elle tient compte de toutes les valeurs.

Représentation des variables

- On peut aussi visualiser les valeurs prises par la variable aléatoire sous la forme d'un « boxplot » ou « boîte à moustaches » qui résume quelques indicateurs de position du caractère étudié (médiane, quartiles, minimum, maximum).



<https://le-datascientist.fr/3-concepts-statistiques>

Axe avec les valeurs prises par les différents indicateurs

Représentation de 2 variables

- Régulièrement il est nécessaire de visualiser conjointement 2 VA
- La covariance permet d'évaluer le sens de variation de 2 VA
- La covariance permet également de qualifier l'indépendance de ces variables
- Si 2 VA sont indépendantes alors leur covariance est nulle, mais la réciproque est fausse
- Dans le cas de 2 variables quantitatives il sera classique de représenter les individus comme un nuage de points dans un espace à 2 dimensions (chaque dimension correspondant à une variable)
- Nous verrons en TP toutes les représentations de 2 variables en fonction de leur type.

Rappels - Covariance

$$\text{cov}(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$\text{cov}(X, Y) = E(XY) - E(X)E(Y)$$

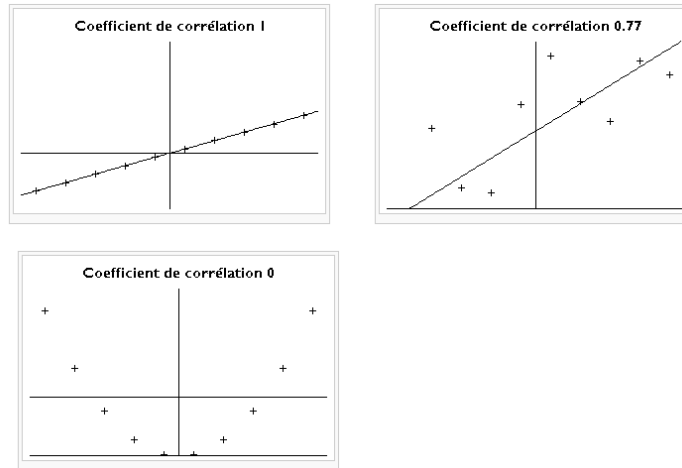
$$\text{cov}(X, Y) = \text{cov}(Y, X)$$

$$\text{cov}(X, X) = V(X)$$

Corrélation

- La corrélation entre 2 VA permet de mesurer la relation linéaire qui existe entre les 2 VA. La corrélation est obtenue par le calcul du **coefficient de corrélation linéaire**
- Ce coefficient est égal au rapport de la covariance et du produit non nul de leurs écarts types. Le coefficient de corrélation est compris entre -1 et 1

Corrélation



Extrait de http://fr.wikipedia.org/wiki/Corrélation_statistiques

Corrélation / Régression linéaire

- Dans certains cas, nous pouvons nous poser la question suivante: la connaissance d'une modalité de la variable X apporte-t-elle une information supplémentaire sur les modalités de la variable Y ? La réponse à cette question est du domaine de la **régression** : dans un tel cas, on dit que X est la variable explicative et Y la variable expliquée.
- Dans d'autres cas, aucune des 2 VA ne peut être privilégiée : la liaison entre X et Y s'apprécie alors de façon symétrique par la mesure de la **corrélation**.