POLYTECH GRENOBLE

Statistiques E2i5

Nathalie GUYADER
nathalie.guyader@gipsa-lab.fr

Année 2017-2018

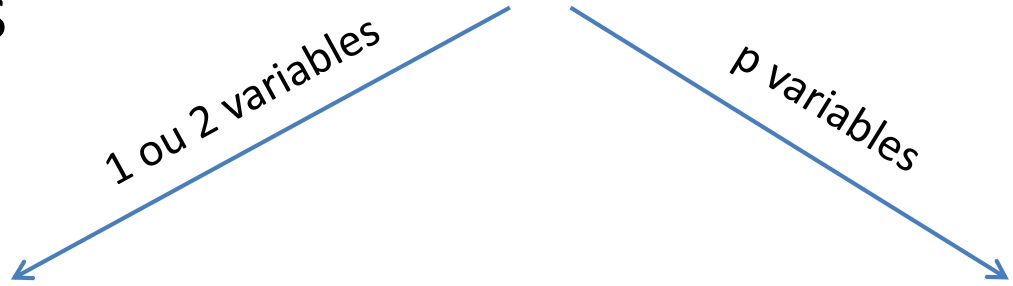
POLYTECH

Classiquement on distingue :

- Les statistiques descriptives ou exploratoires
- Les statistiques inférentielles ou décisionnelles

Statistiques descriptives

- **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



Cours de Statistiques (chap. 1)

Cours de Traitement de données

Statistiques inférentielles

- **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

Tout cela sera abordé
dans ce cours

Planning des séances

- 1 séance d'introduction : BE1

- Statistiques :

5 séances de 4h

- Traitement de données :

6 séances de 4h

Important

Statistiques:

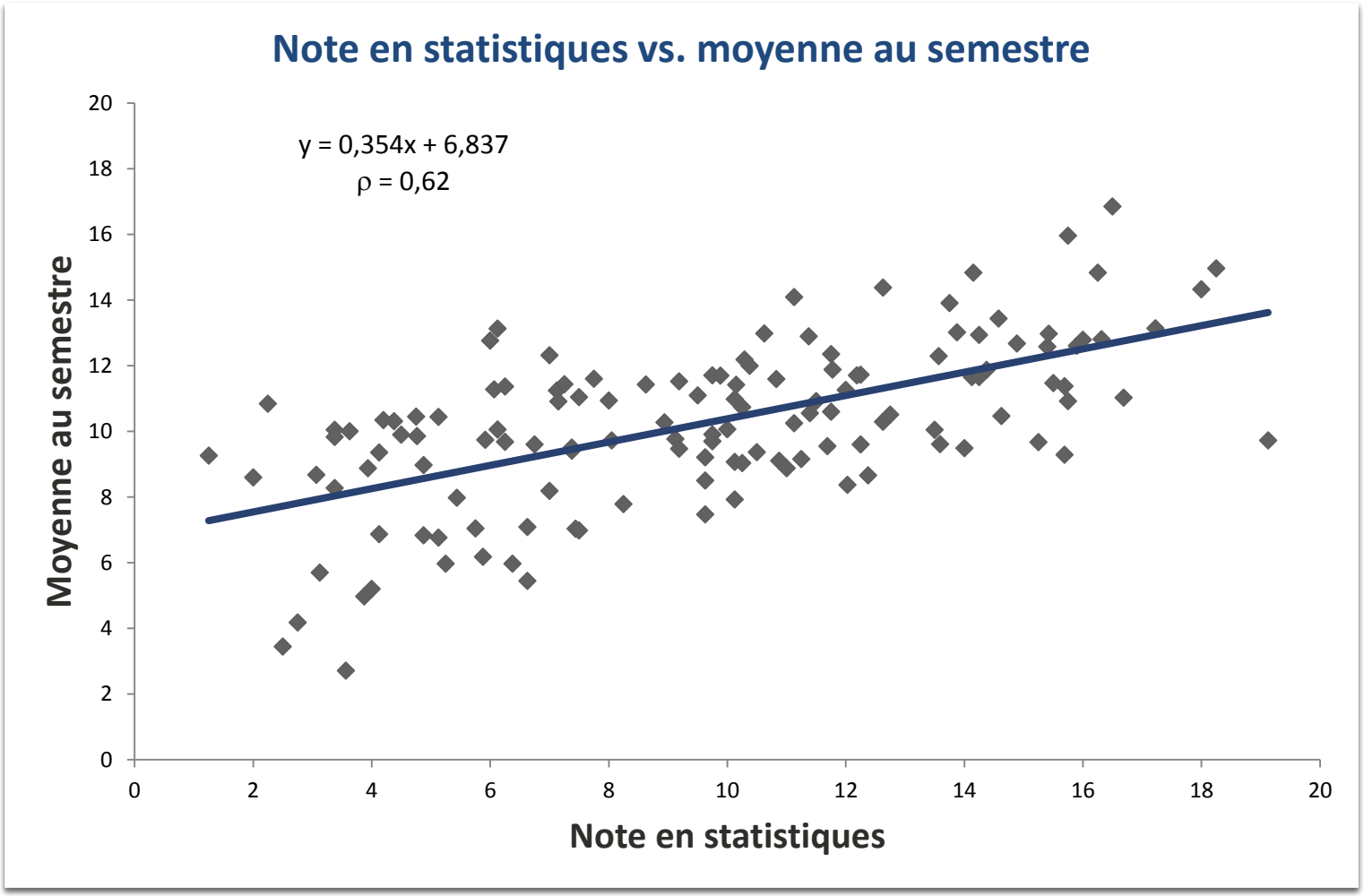
- Travailler au jour le jour! Le partiel est directement à la fin de séances
- Partiel sur machine, conforme aux exercices qui seront faits en séance

Traitement de données:

- sous forme de TPs sur machine

N'hésitez pas à poser des questions et à faire part de vos remarques/critiques

Important


Note en statistiques vs. moyenne au semestre
y = 0,354x + 6,837
ρ = 0,62
Moyenne au semestre
Note en statistiques
0
2
4
6
8
10
12
14
16
18
20

Bibliographie

Pré-requis:

Cours de probabilités

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).

Sites internet:

http://www.univ-tours.fr/ash/psycho

Autres:

Cours d'Alan Chauvin, MdC à l'UGA

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 :

- la fiabilité des matériels
- le contrôle de qualité
- l'analyse des résultats de mesure et leur planification
- 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 ».

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

Régulièrement les étudiants (E2i, IESE, et TIS) ont besoin des statistiques:

- soit directement dans leur travail
- 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 ».
Ainsi, en contrôle de fabrication, on prélèvera un ensemble de pièces dans une production homogène et on mesurera leur poids, leur diamètre… en tests de composants, on relèvera les temps d'exécution des tests, de parties d'algorithmes…

La notion fondamentale en statistique est celle d'ensemble d'objets équivalents (**population**[1]). Ces objets sont des **individus**. La statistique traite des propriétés des populations plus que de celles d'individus.
Généralement, la population à étudier est trop vaste pour pouvoir être observée exhaustivement : c'est évidemment le cas d'une population infinie mais c'est aussi la cas lorsque les observations sont coûteuses (par exemple le contrôle destructif).

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

[1]: Ce terme est hérité des premières applications de la statistique qui concernait la démographie.

Quelques définitions (extraites du Saporta)

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 traite des propriétés de certaines structures modélisant des phénomènes dans lequel le « hasard » intervient. Cette théorie permet de modéliser efficacement certains phénomènes aléatoires et d'en faire l'étude théorique. On peut alors se poser la question : 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**.

Plan du cours de Statistiques

- Rappels
- Statistique descriptive ‘simple’ (1 ou 2 variables)
- Théorie de l’échantillonnage et estimation
- Statistiques inférentielles et tests d’hypothèse

Chapitre 1

Rappels

Rappels - Probabilités

Expérience aléatoire : expérience faisant 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.

Univers Ω : l'ensemble des issues possibles d'une expérience aléatoire

- *Ω* est fini: on lance un dé; le résultat est le n° qui sort
- *Ω* est infini dénombrable: on lance une pièce de monnaie jusqu'à obtenir face; le résultat est le numéro du lancer qui finit l'expérience

Issue ou éventualité ω : un résultat

Evènement A : un sous-ensemble de Ω

Rappels - Probabilités

- Ω est l'évènement certain et { } est l'évènement impossible
- Soient une épreuve E d'univers Ω et A une tribu d'évènements de Ω.

On appelle probabilité toute application telle que:

$$\forall A \in \mathrm{A}, \mathrm{P}(\mathrm{A}) \in [0;1] \quad et \quad P(\Omega) = 1$$

Propriétés:

$$P(\{\ \}) = 0$$

$$P(\overline{A}) = 1 - P(A)$$

$$P(B \setminus A) = P(B) - P(A \cap B)$$

Attention ici B privé de A!

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

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

Rappels - Probabilités

Définition de la probabilité :

$$P(A) = \frac{card(A)}{card(\Omega)}$$

Probabilités conditionnelles:

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

Théorème de Bayes pour 2 évènements A et B quelconques:

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

A et B sont indépendants $\overset{\text{déf}}{\Leftrightarrow} P(A \cap B) = P(A)P(B)$

Rappels - Probabilités

Exercice 1: Probabilités conditionnelles

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?

Rappels - Probabilités

Exercice 2: Loi de probabilités conjointes de 2 variables

P(X,Y)	X=1	X=0
Y=1	0,15	0,05
Y=0	0,15	0,65

X et Y sont – elles indépendantes ?

Rappels – Variables aléatoires

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 croit strictement	
$X(\Omega)$ = ensemble des x en lesquels le saut $P(X=x)$ de F est non nul $=\{x \in R / P(X=x) \neq 0\}$ $X(\Omega) \overset{not}{=} \{x_k\}$ est fini ou dénombrable	$X(\Omega)$ = 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 $X(\Omega) = \{x_k\}$ et les $P(X=x_k)$	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, $P(X \in I) = \int_I f(x)dx$
$P(X \in]-\infty,+\infty[) = 1 = \sum_{x_k \in R} P(X=x_k)$	$P(X \in]-\infty,+\infty[) = 1 = \int_{-\infty}^{+\infty} f(x)dx$
$E(X) = \sum_{x_k \in R} x_k P(X = x_k)$	$E(X) = \int_{-\infty}^{+\infty} x f(x)dx$
$E(\varphi(X)) = \sum_{x_k \in R} \varphi(x_k) P(X = x_k)$	$E(\varphi(X)) = \int_{-\infty}^{+\infty} \varphi(x) f(x)dx$
$V(X) = E\left[(X - E(X))^2\right] = E(X^2) - E^2(X)$	

Extrait du cours de Probabilités de 1ère année – F. Blanchet

Rappels – Espérance et Variance

$$E(a) = a$$

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

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

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

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

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

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

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

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

Rappels – Covariance

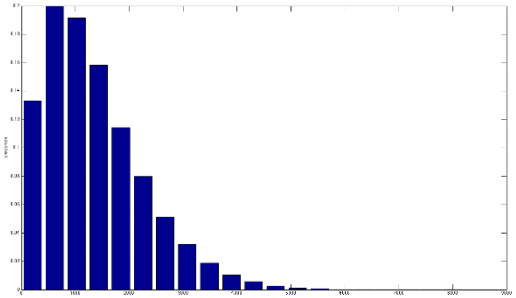
$$\mathrm{cov}(X,Y) = E(XY) - E(X)E(Y)$$

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

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

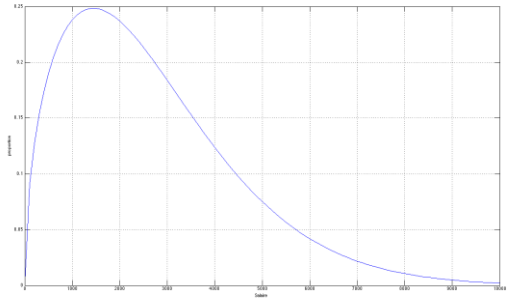
Lois de probabilité

- Distribution



  - La VA peut prendre un nombre fini de valeurs
  - La somme des aires est égale à 1
  - Les valeurs prises par la VA sont en abscisse
  - L'ordonnée représente la densité (ou la fréquence relative associée)

- Densité



  - La VA peut prendre un nombre infini de valeurs
  - L'aire sous la courbe est égale à 1
  - L'ensemble des valeurs prises par la VA sont en abscisse
  - L'ordonnée représente la densité

Rappel: dans le cas continu, la probabilité que la VA prenne une valeur particulière est nulle.

Principales lois discrètes

Loi uniforme : U(n)

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

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

Moments:

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

Exemple : loi de probabilité associée à l'expérience aléatoire consistant à jeter un dé

Principales lois discrètes

Loi de Bernouilli : 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

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

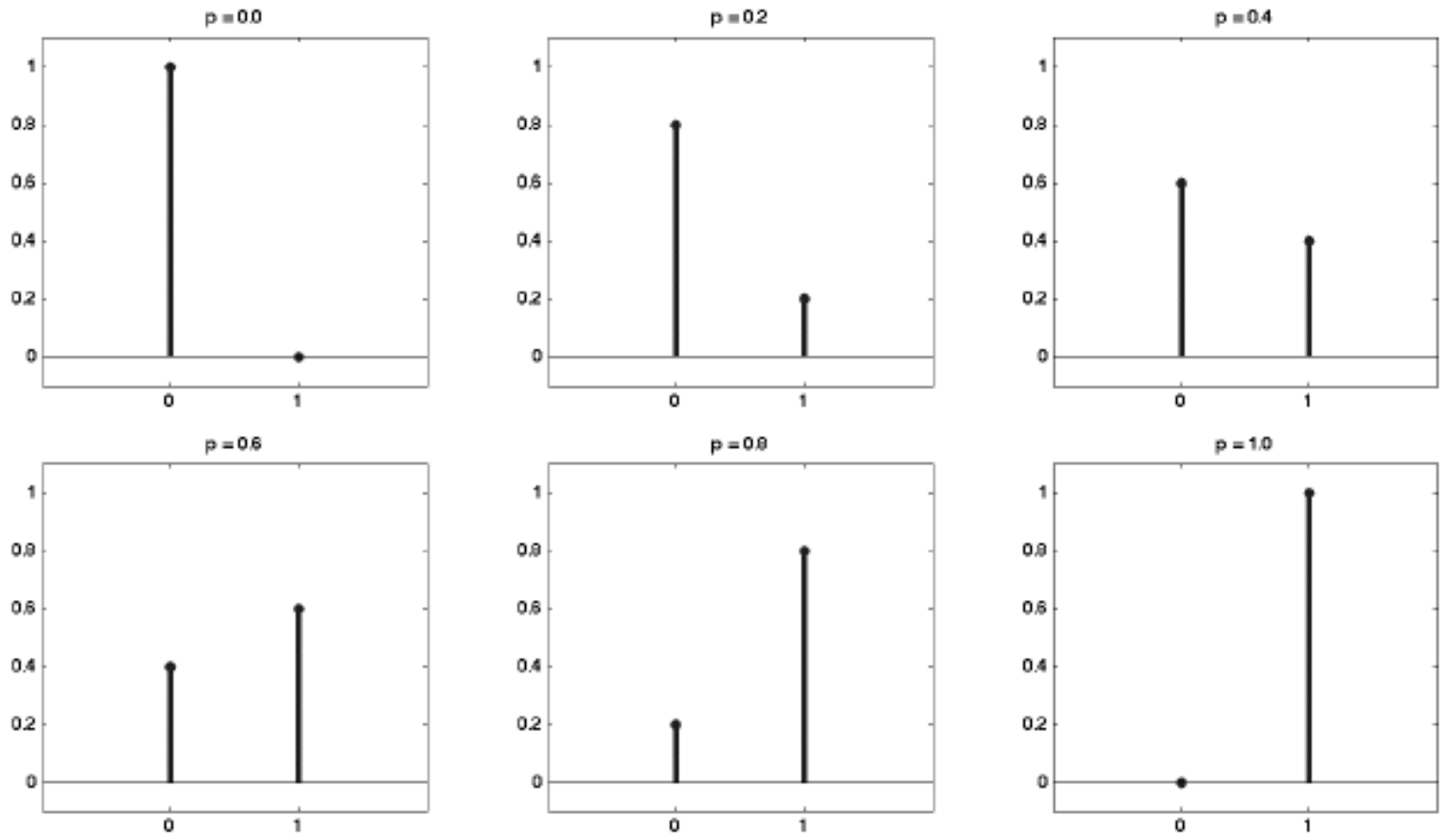
Moments:

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

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

Principales lois discrètes

Loi de Bernouilli : 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

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

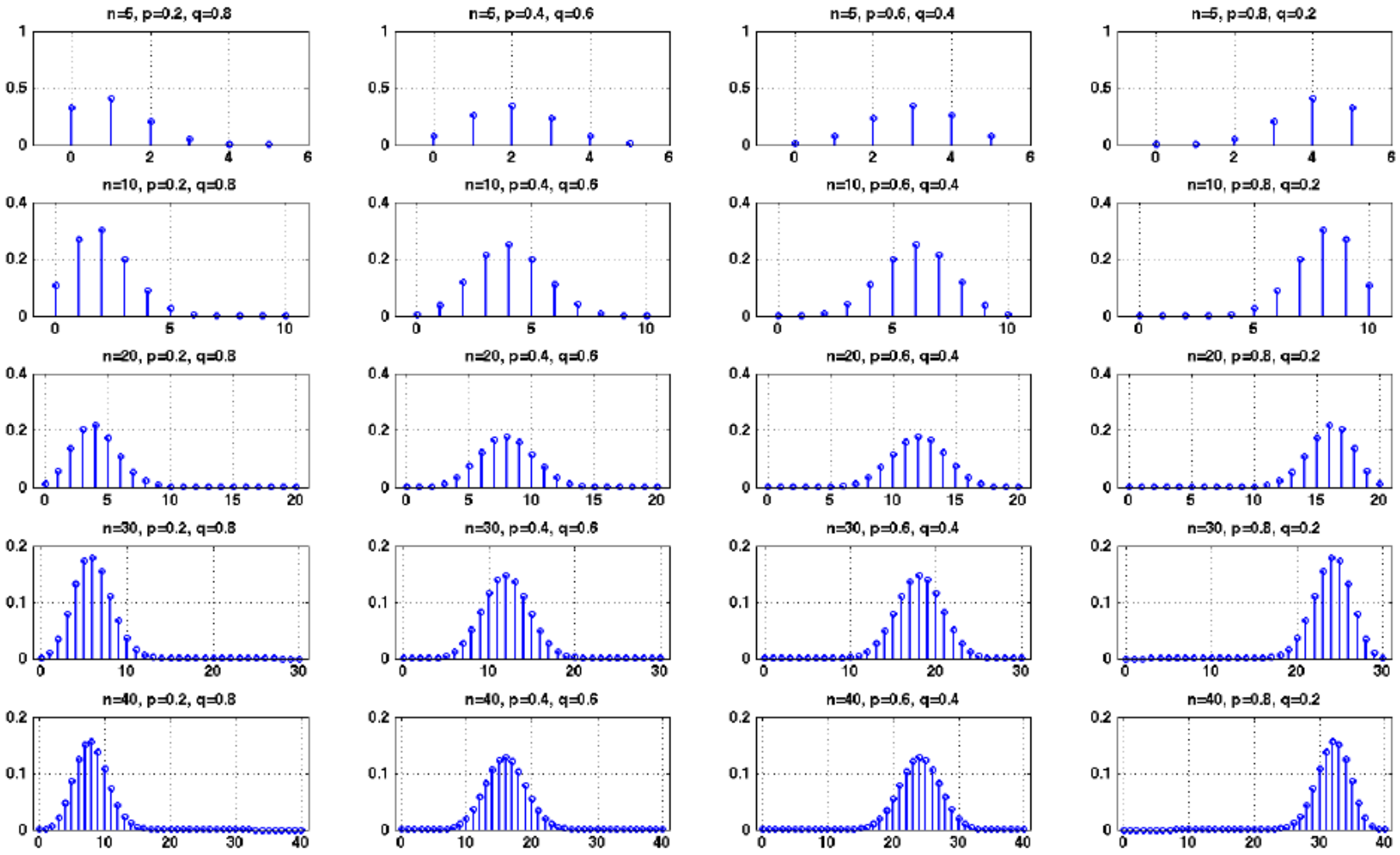
Moments:

$$E(X) = np \qquad V(X) = np(1-p)$$

Exemple : nombre de réalisations de piles après n lancés

Principales lois discrètes

Loi Binomiale: **B**(n,p)



Exercice 3 :

Loi de probabilité de la variable : nombre de garçons dans une famille de 7 enfants:

Principales lois discrètes

Loi de Poisson: P(λ)

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

Moments:

$$E(X) = \lambda \qquad V(X) = \lambda$$

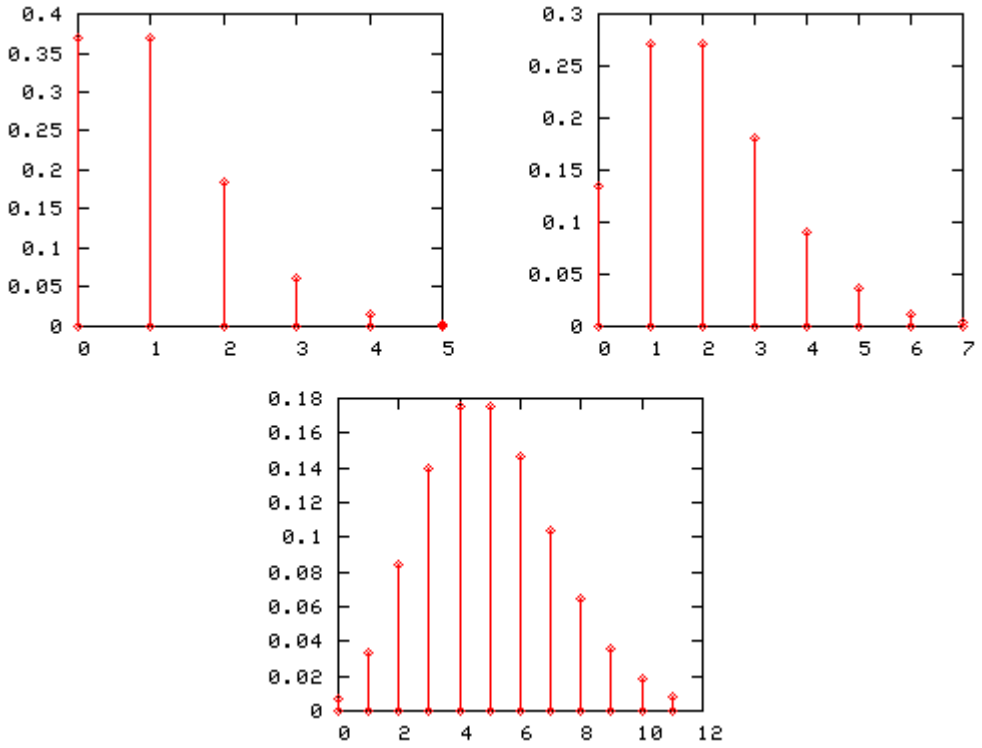
Remarque : si λ est grand alors la loi de Poisson tend vers une loi normale

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

Principales lois discrètes

Loi de Poisson: **P(λ)**

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 paramètre 1, 2 et 5.



Principales lois continues

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

Densité de probabilité:

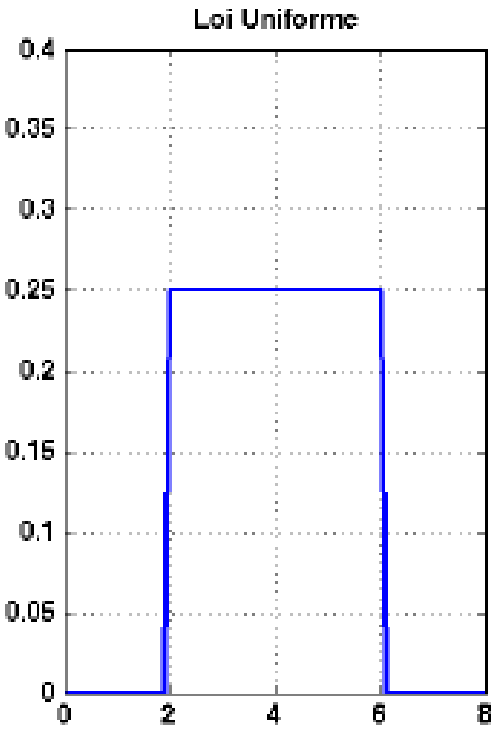
$$f(x) = \begin{cases} \frac{1}{a} \; sur \; [0,a] \\ 0 \; ailleurs \end{cases}$$

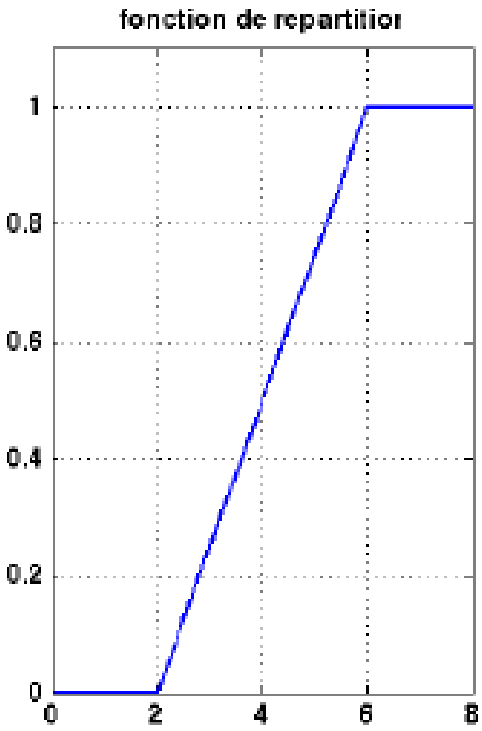
Moments:

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

Principales lois continues

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





Principales lois continues

Loi exponentielle: E(λ)

Densité de probabilité

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

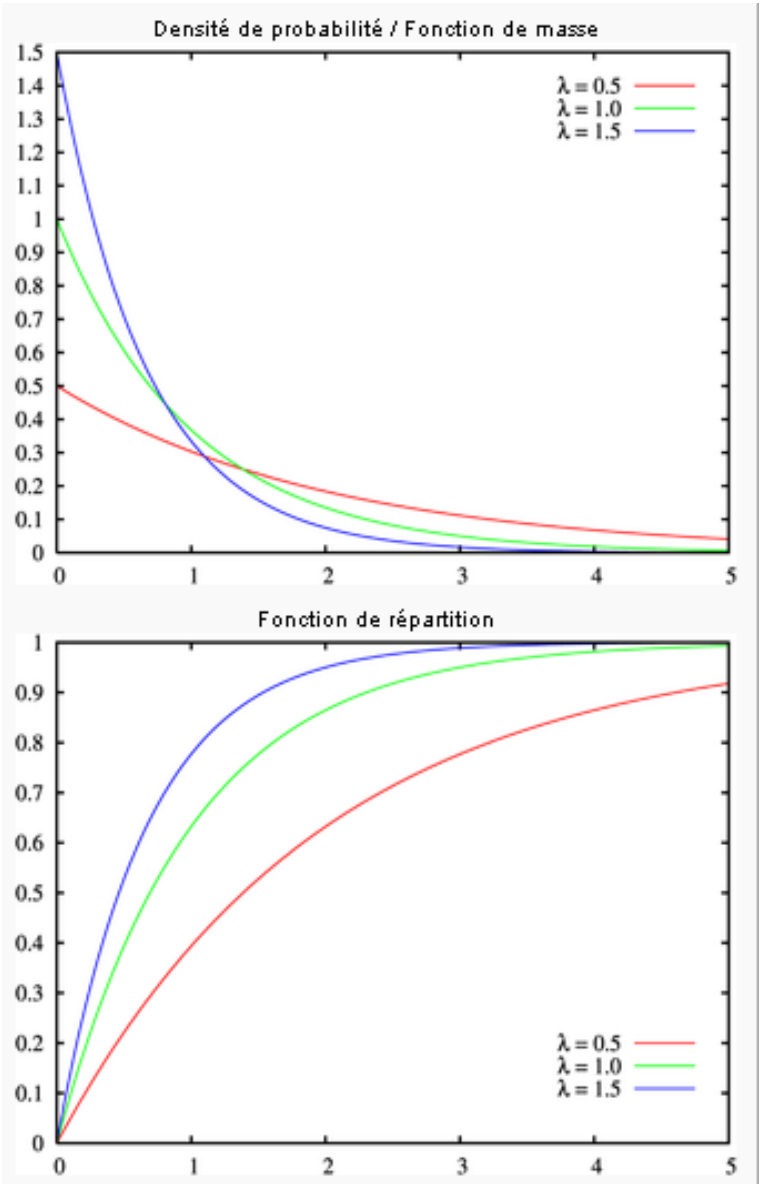
Moments:

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

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

Principales lois continues

Loi exponentielle: E(λ)



Principales lois continues

Loi normale : N(μ, σ^2)

Densité de probabilité

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

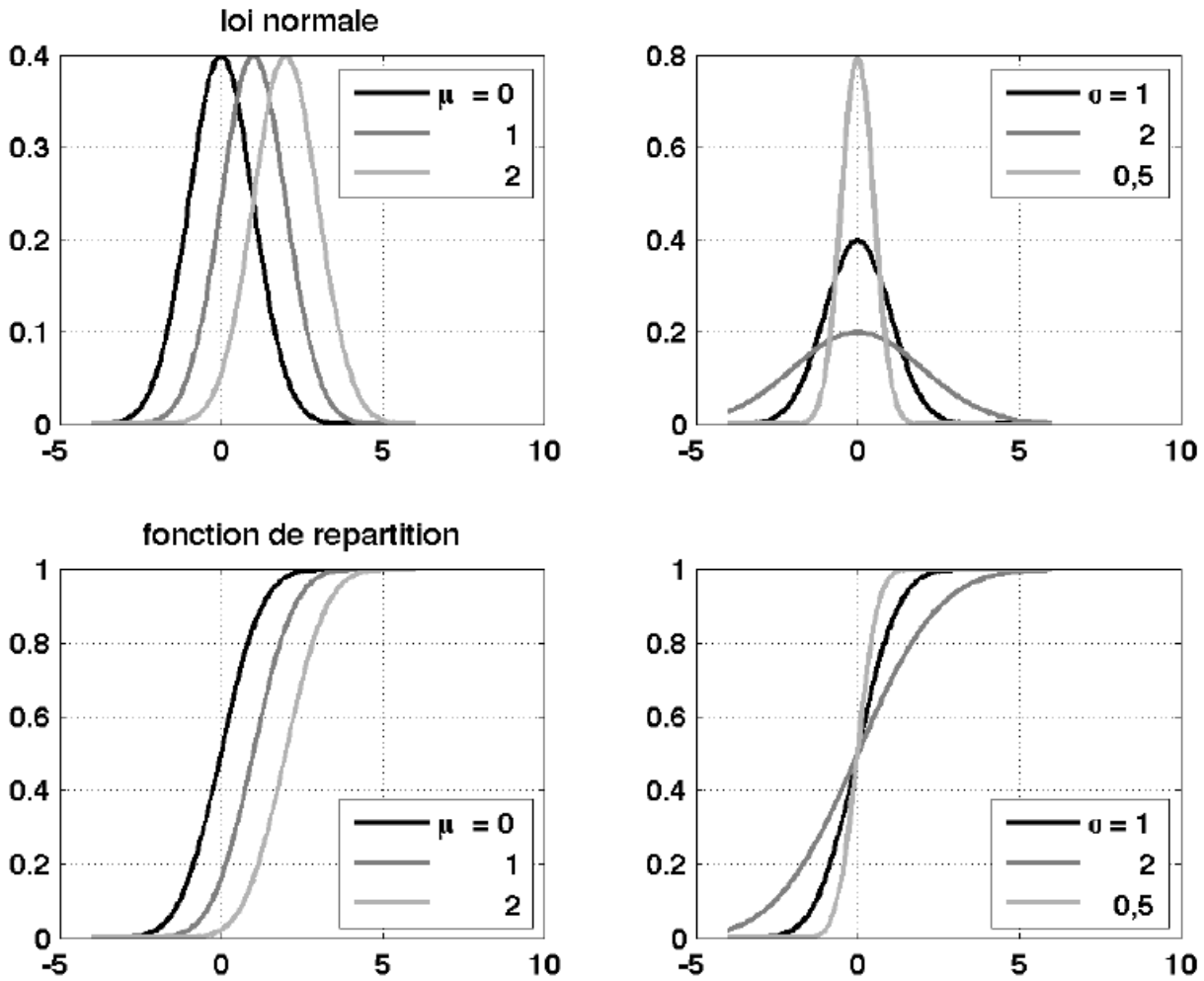
Moments:

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

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

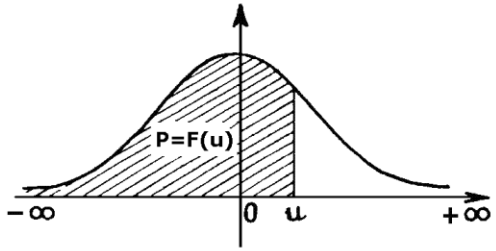
Principales lois continues

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



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

Chapitre 2

Statistiques descriptives

Introduction

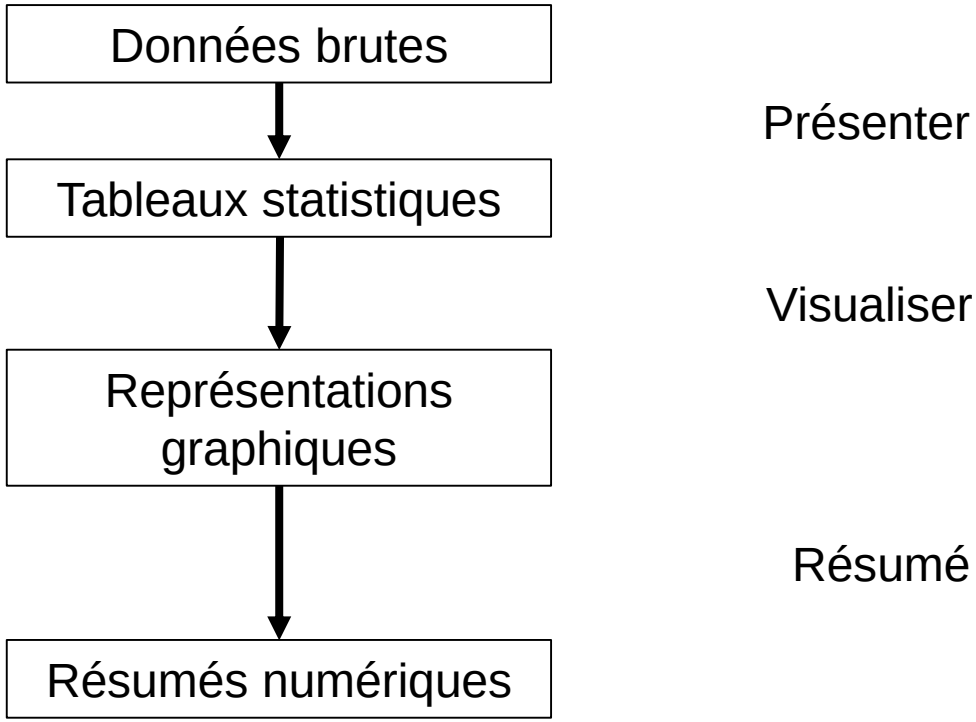
On étudie un ensemble d'objets équivalents sur lesquels on observe des caractéristiques appelées « **variables** ». Nous nous arrêterons ici à l'étude de 1 ou 2 variables. Le cas p variables sera traité plus tard dans le cours de Traitement de Données. Chaque individu d'une population est décrit par 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 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 de classe, de fréquences absolues et relatives

Dépouillement des données

On doit avant toute chose **se fixer le nombre de classes : K**.
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 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 39,1 39,1 39,2 39,4 39,4 39,4 39,4 39,4 39,5 39,5 39,5 39,6 39,7 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 donner l'amplitude de chaque classe:

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 ≤ X < 37,5				
37,5 ≤ X < 38,5				
38,5 ≤ X < 39,5				
39,5 ≤ X < 40,5				
40,5 ≤ X < 41,5				
41,5 ≤ X < 42,5				
42,5 ≤ X < 43,5				

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

Représentations graphiques

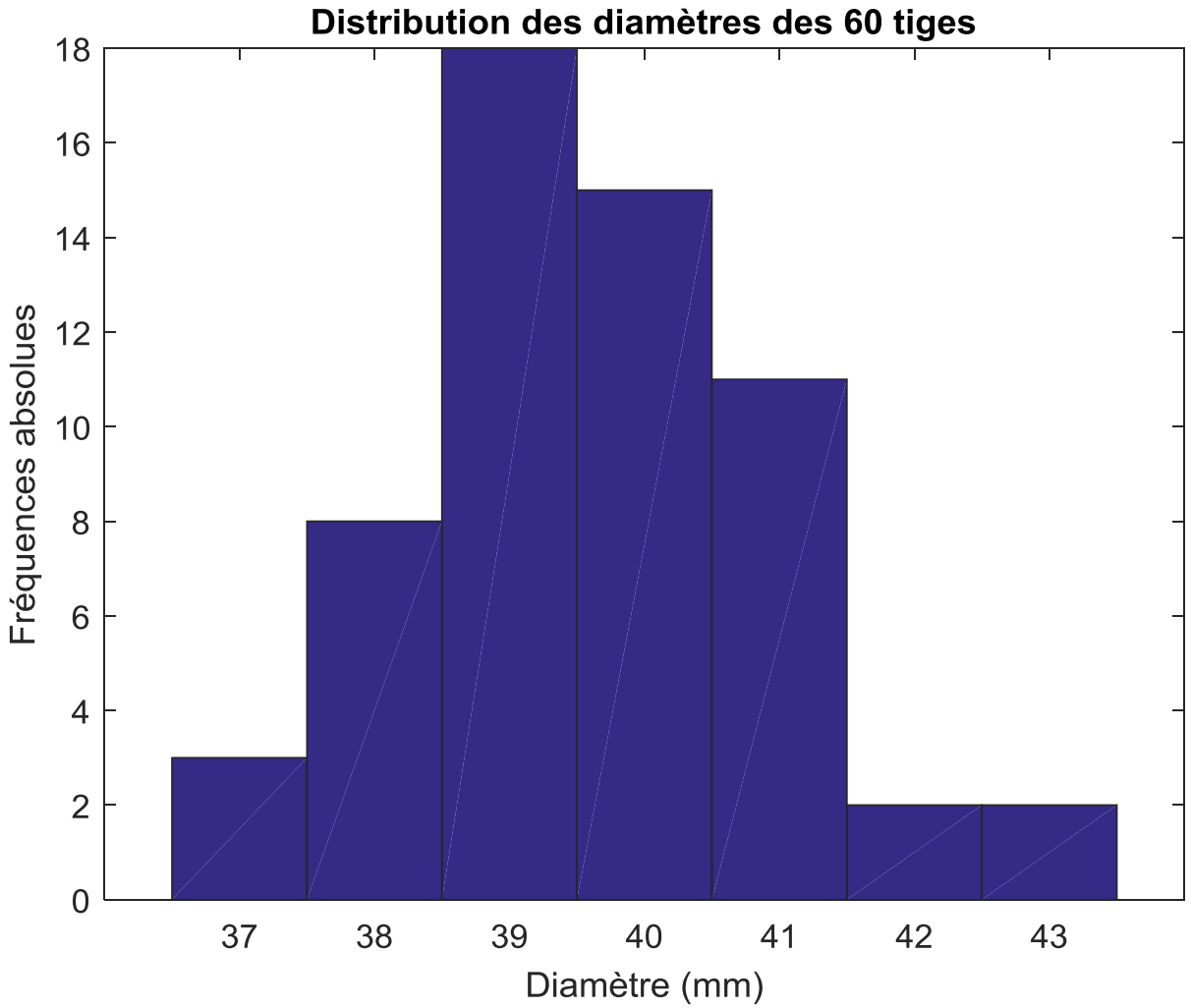
Lorsque la variable est qualitative on utilisera généralement une représentation en diagrammes de bâtons.

Lorsque la variable quantitative est discrète (VAD):
Diagrammes en bâtons : en abscisses les valeurs de la VAD et en 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 : 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.

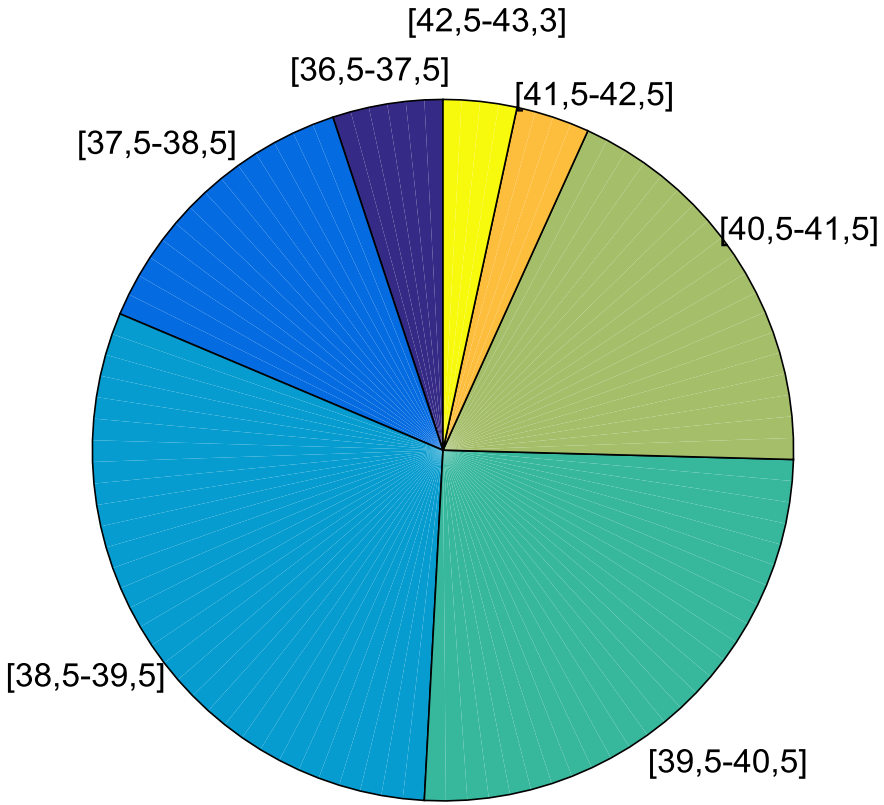
Représentations graphiques

On reprend les premières données; vous pouvez les récupérer en chargeant le fichier data1.mat



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

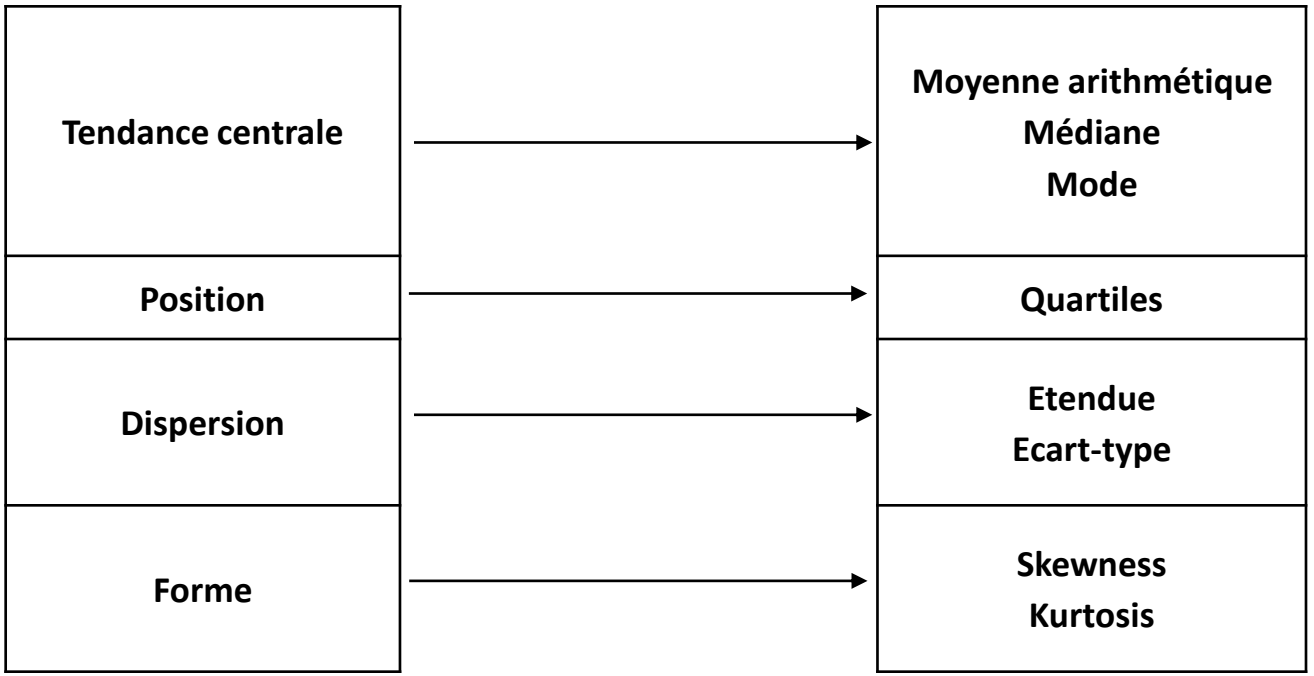
Nous avons vu que nous pouvons représenter une série numérique à l'aide de tableaux et de graphiques. Il est également intéressant de pouvoir décrire la série numérique à l'aide de valeurs caractéristiques. On distingue 3 types de caractéristiques :

- *Les caractéristiques (ou mesures) de tendance centrale* : elles permettent d'obtenir une idée de l'ordre de grandeur des valeurs constituants la série et indique 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 s'étalent autour de la valeur centrale.

- *Les caractéristiques (ou mesure) de forme* : elles donnent une idée de la symétrie et de l'aplatissement d'une distribution. Ces dernières sont moins souvent utilisées.

Résumé par des valeurs



Tendance centrale

La moyenne arithmétique est généralement le premier indicateur de tendance centrale 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 est la valeur *m* telle que le nombre de valeurs de la série supérieures ou égales à *m* est égal au nombre de valeurs inférieures ou égales à *m*.

Le mode ou valeur dominante désigne la valeur la plus représentée de la série. Une répartition peut-être unimodale ou plurimodale (bimodale, trimodale…), si 2 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 mois 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 mois 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 correspond à la valeur maximale moins la valeur minimale de la variable (de la série de valeurs).

L'intervalle interquartile correspond à la différence entre Q_3 et Q_1.

La variance de l'échantillon, 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 de 2 variables

Dans les rappels, nous avons vu la table conjointe de 2 VAs (tableau de contingence ou des probabilités jointes); également la loi marginale et la loi conditionnelle… Régulièrement, il est nécessaire de visualiser conjointement 2 VAs.

Covariance:

La covariance permet d'évaluer le sens de variation de 2 VAs. La covariance permet également de qualifier l'indépendance de ces variables.

Si 2 VAs sont indépendantes alors leur covariance est nulle, mais attention la réciproque est fausse

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

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

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

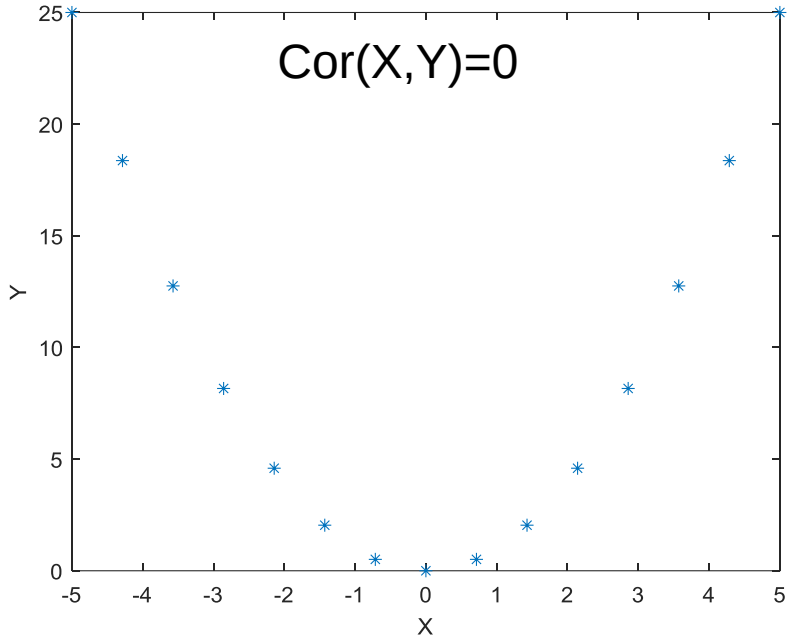
$$\text{cov}(X,X) = V(X)$$

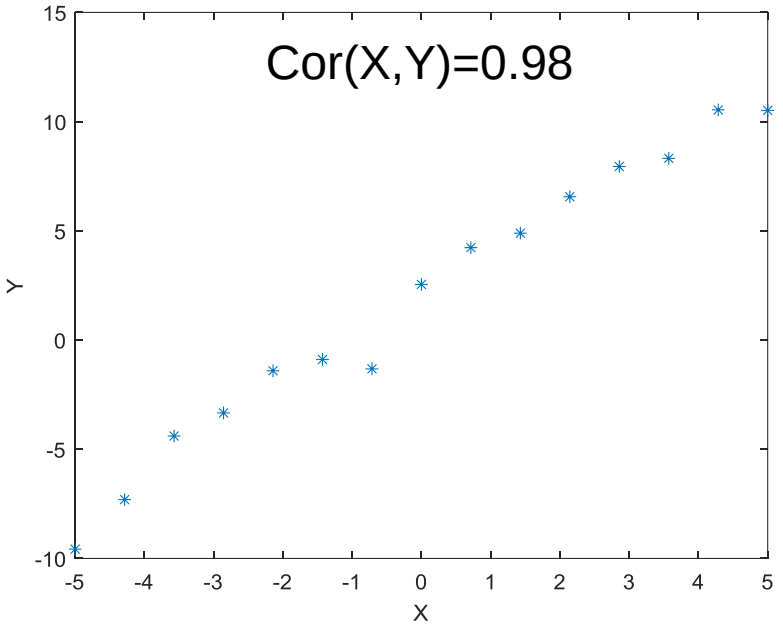
Corrélation

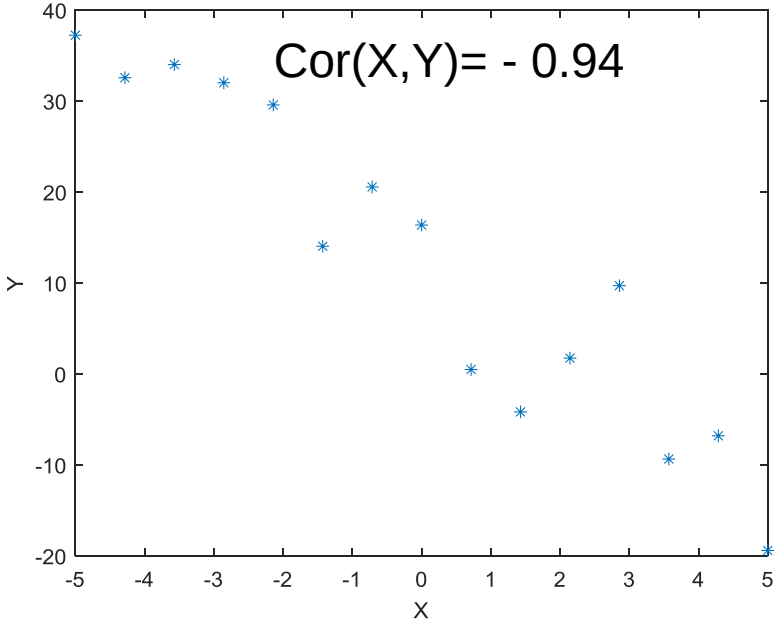
La corrélation entre 2 VA permet de mesurer/quantifier la relation linéaire qui existe entre 2 VAs. La corrélation est obtenue par le calcul du **coefficient de corrélation linéaire.** Ce coefficient est égal au rapport de la covariance et du produit non nul de leurs écarts types.

Le coefficient de corrélation est compris entre -1 et 1

Son signe indique si les 2 variables évoluent ou non dans le même sens


Cor(X,Y)=0
Y
X
25
20
15
10
5
0
-5
-4
-3
-2
-1
0
1
2
3
4
5


Cor(X,Y)=0.98
Y
X
15
10
5
0
-5
-10
-5
-4
-3
-2
-1
0
1
2
3
4
5


Cor(X,Y)= - 0.94
Y
X
40
30
20
10
0
-10
-20
-5
-4
-3
-2
-1
0
1
2
3
4
5

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 VAs ne peut être privilégiée : la liaison entre *X* et *Y* s'apprécie alors de façon symétrique par la mesure de la **corrélation**.