UNIVERSITÉ DE RENNES 1

Introduction à la statistique descriptive

Préparation au Capes - Université Rennes 1

But : La statistique descriptive a pour but d'étudier une **population** à partir de **données**. Cette description se fait à travers la présentation des données (la plus synthétique possible), leur représentation graphique et le calcul de résumés numériques.

Hélène Guérin, helene.guerin@univ-rennes1.fr

Partie I - Un peu de vocabulaire

On parle de **recensement** lorsque l'on fait une étude exhaustive d'une **population**. Lorsqu'on n'étudie qu'une sous-population, on parle alors d'**échantillon**.
Les éléments de cette population sont appelés **individus**.

Il convient de définir avec précision les ensembles que l'on étudie et notamment leurs fontières.

Voici une version allégée de la définition d'une agglomération au sens de l'INSEE :

Exemple : Définition d'une agglomération selon l'INSEE

Les limites entre territoires urbain et rural sont redéfinies à l'occasion de chaque recensement. Leur tracé fait intervenir la notion d'agglomération de population définie comme un ensemble d'habitations. Dans cet ensemble, qui doit abriter au moins 2000 habitants, aucune habitation ne doit être séparée de la plus proche de plus de 200 mètres. Les frontières de ces zones coincident dans tous les cas avec les limites communales. En revanche, les limites des autres circonscriptions administratives (cantons, arrondissements, départements) ne sont pas prises en compte lors de leur délimitation. Une même unité urbaine peut s'étendre sur deux départements. Si l'agglomération de population s'étend sur plusieurs communes, l'ensemble de ces communes forme une agglomération urbaine. Si l'agglomération s'étend sur une seule commune, celle-ci est une ville isolée. Toutes ces communes sont considérées comme urbaines. Les autres communes sont classées communes rurales.

À chaque individu de la population sont associés des **caractères**, appelés aussi **variables**.

Exemple : Le personnel d'une entreprise peut être décrit selon divers caratères : âge, sexe, qualification, ancienneté dans l'entreprise, commune de résidence...

Un lot de pièce mécanique peut être décrit suivant le poids, le diamètre, la matière...

Chacun des caractères étudiés peut présenter deux ou plusieurs **modalités**. Les modalités sont les différentes situations où les individus peuvent se trouver à l'égard du caractère considéré. Les modalités d'un même caractère doivent être incompatibles et exhaustives.

Exemple 1 Nomenclature des professions et des catégories socioprofessionnelles

Trois niveaux de regroupements sont proposés :

1. Niveau agrégé (8 postes)

Code	Libellés
1	Agriculteurs exploitants
2	Artisans, commerçants, chefs d'entreprise
3	Cadres et professions intellectuelles supérieures
4	Professions intermédiaires
5	Employés
6	Ouvriers
7	Retraités
8	Autres sans activité professionnelle

2. Niveau de publication courante (24 postes)...extrait

Code	Libellés
10	Agriculteurs exploitants
21	Artisans
22	Commerçants et assimilés
23	Chefs d'entreprise de 10 salariés ou plus
31	Professions libérales
32	Cadres de la fonction publique, professions intellectuelles et artistiques
36	Cadres d'entreprise
41	Professions intermédiaires de l'enseignement, de la santé, de la fonction publique et assimilés
46	Professions intermédiaires administratives et commerciales des entreprises
47	Techniciens
48	Contremaitres, agent de maitrise

On distingue deux types de caractères : les caractères **qualitatifs**, et les caractères **quantitatifs**.

Lorsque le caractère est quantitatif, on parle de **variable statistique**. Une variable statistique est soit discrète soit continue.

On va maintenant s'intéresser à l'étude des caractères quantitatifs.

On s'intéresse à deux études. La première concerne l'âge des étudiants préparant le capes de mathématiques à la fac de Rennes en 2005-2006, l'autre concerne la durée de vie des ampoules d'une grande marque connue. On a relevé dans chacun des cas les données suivantes.

Exemple 1 Âge des étudiants de Capes en 2005-2006 : étude faite sur 53 personnes.

Age	20	21	22	23	24	25	29	31	36	39	40
Effectif	1	7	17	14	7	2	1	1	1	1	1

Exemple 2 On relève la durée de vie de 500 ampoules dites "économiques".

Ampoule	A1	A2	A3	A4	...	A_{500}
Nbre d'heures	1310	874,3	609,2	4106,6	...	2859,7

Partie II - Représentations graphiques

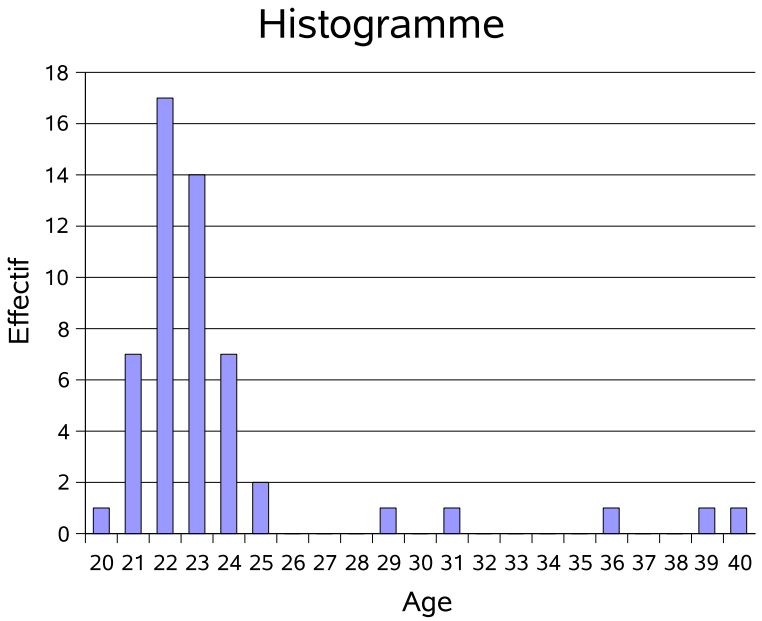
1. Histogrammes et camemberts

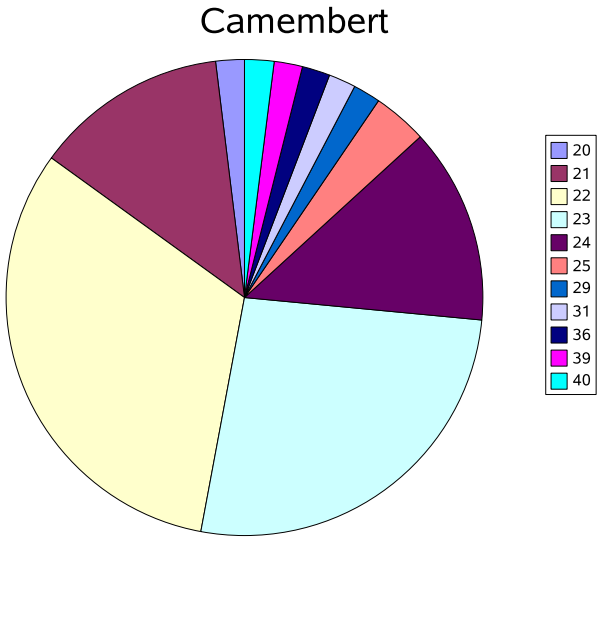
Cas des variables discrètes : on étudie une variable discrète X à p modalités dans une population de taille N.

Modalités	x_1	x_2	$\ldots$	x_p
Effectifs	N_1	N_2	$\ldots$	N_p
Fréquences : $f_i = \frac{N_i}{N}$	f_1	f_2	$\ldots$	f_p

Pour l'histogramme la hauteur des barres est proportionnelle à la fréquence. Pour le camembert, c'est la surface allouée à la modalité qui est proportionnelle à la fréquence.

Exemple 1 : Âge des étudiants de Capes en 2005-2006.





Cas des variables continues : on regroupe les individus par classes. On décompose l'intervalle des valeurs possibles en une partition d'intervalles.

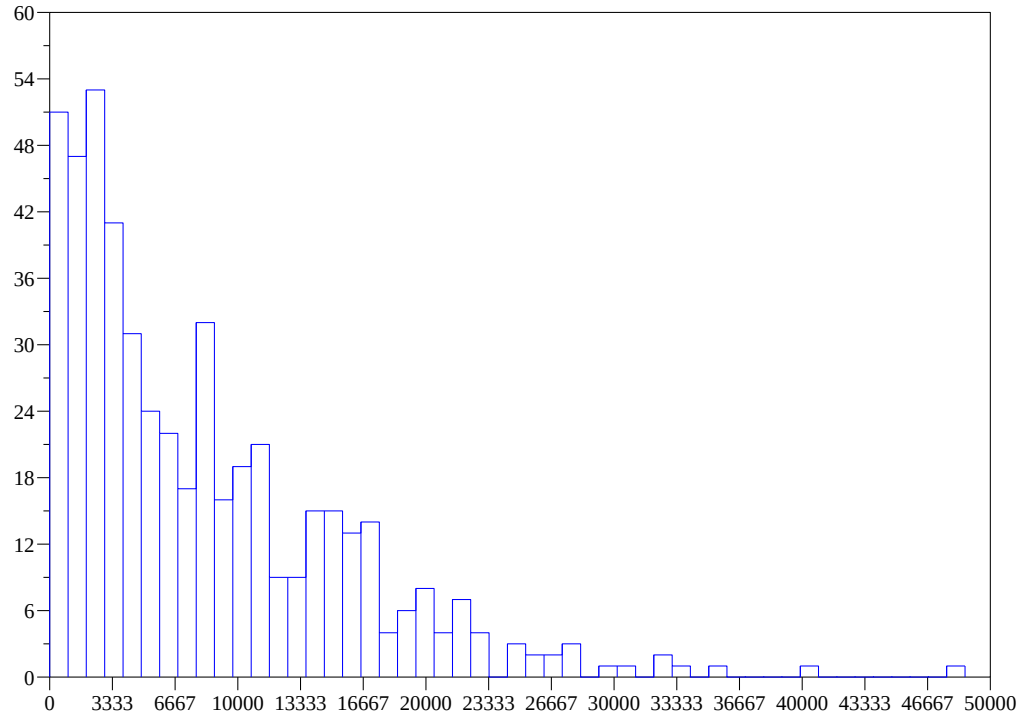
Soit p le nombre d'intervalles qui découpent la plage de variation de la variable étudiée et $\{e_{i-1}, e_i\}$ les bornes de l'intervalle i. On se limite ici au cas où tous les sous intervalles sont de même longueur.

Les données se présentent sous la forme suivante

Classes	e_0 à e_1	e_1 à e_2	...	e_{p-1} à e_p
Effectifs	N_1	N_2	...	N_p
Centres de classe : $c_i = \frac{e_i - e_{i-1}}{2}$	c_1	c_2	...	c_p
Fréquences : $f_i = \frac{N_i}{N}$	f_1	f_2	...	f_p

Exemple 2 : Durée de vie des ampoules

On a découpé l'intervalle des durées observées en 50 sous intervalles de même longueur.



2. La boite à moustache (box-plot)

La boite à moustache permet de résumer les principales caractéristiques d'un tableau de données en un graphique assez simple.

On observe une variable X à p modalités sur une population de taille N. On note $x_1, \ldots, x_p$ les différentes modalités.

- La moyenne arithmétique m (souvent notée $\overline{x}$) :

$$\begin{aligned} m &= \frac{1}{N}\sum_{i=1}^{p} N_i x_i \\ &\simeq \frac{1}{N}\sum_{i=1}^{p} N_i c_i \quad \text{dans le cas d'une variable continue} \end{aligned}$$

Elle représente la valeur moyenne des données.

• Les quantiles : les plus utilisés sont la médiane, les quartiles et les déciles.

La médiane *med* correspond à la valeur de la variable qui partage la population en deux sous populations d'effectifs égaux. *(Il y a autant d'individus dont la valeur de la modalité est supérieure à* med *que d'individus dont la valeur de la modalité est inférieure à* med*.)*

Les quartiles $\{Q_k,\, k = 1, 2, 3\}$ divisent la population en 4 sous populations et les déciles $\{d_k,\, k = 1, \ldots, 9\}$ en 10 sous populations d'effectifs égaux.

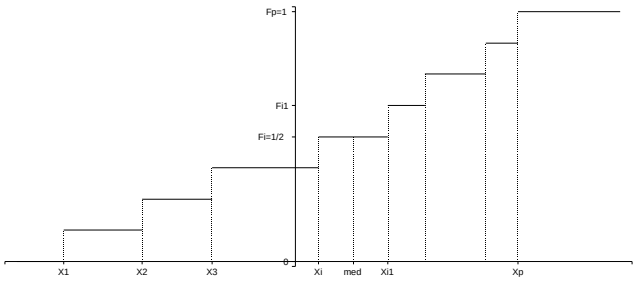
Remarque : la médiane est moins sensible que la moyenne aux valeurs extrêmes.

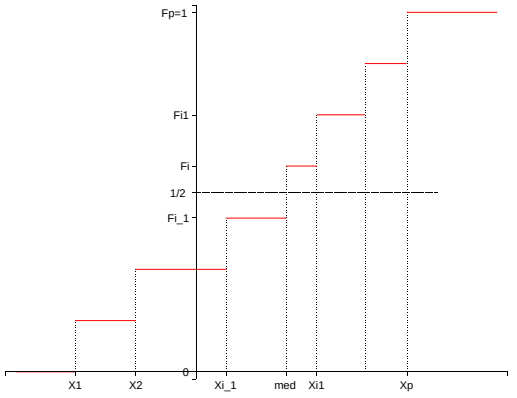
Comment calculer ces quantités ?

$\Rightarrow$ En utilisant les fréquences cumulées : $F_i = f_1 + \ldots + f_i$,
$i = 1, \ldots, p$. (*Notion de fonction de répartition de l'échantillon*).

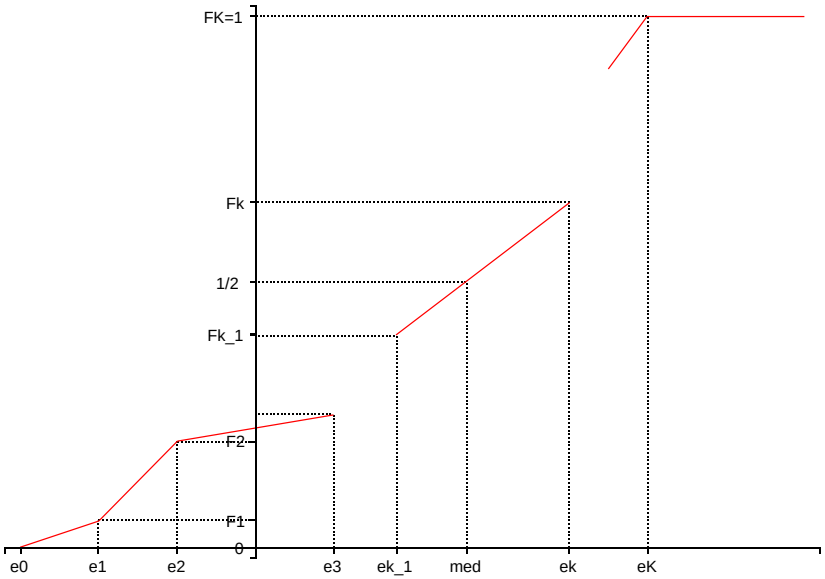
On se limite ici à la médiane. L'idée est la même pour les autres quantiles.

<u>Dans le cas Discret</u> :
S'il existe un i tel que $F_i = 1/2$, alors l'intervalle $[x_i, x_{i+1}[$ est médian. On choisit souvent $med = \frac{x_i + x_{i+1}}{2}$.
S'il existe i tel que $F_{i-1} < 1/2 < F_i$ alors la médiane est x_i.





Dans le cas continu : On approche par interpolation linéaire la fonction de répartition de l'échantillon.

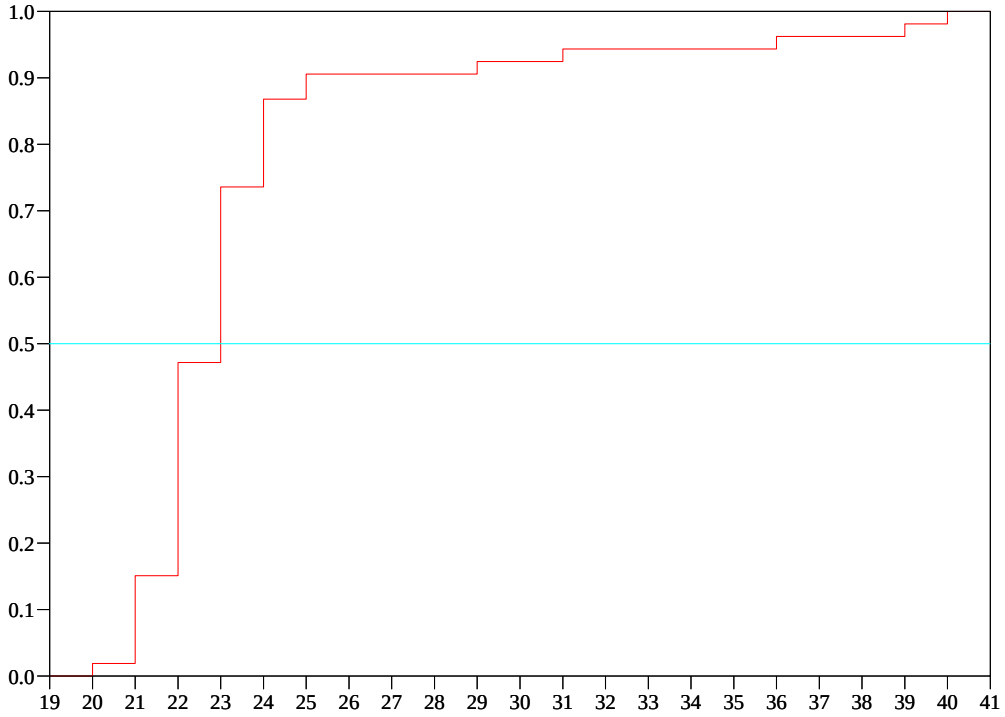


On cherche l'indice i tel que $F_{i-1} < 1/2$ et $F_i > 1/2$. Par interpolation linéaire on obtient :

$$med = e_{i-1} + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(e_i - e_{i-1}).$$

Exemple 1 : Moyenne et médiane de l'âge des étudiants de Capes en 2005-2006

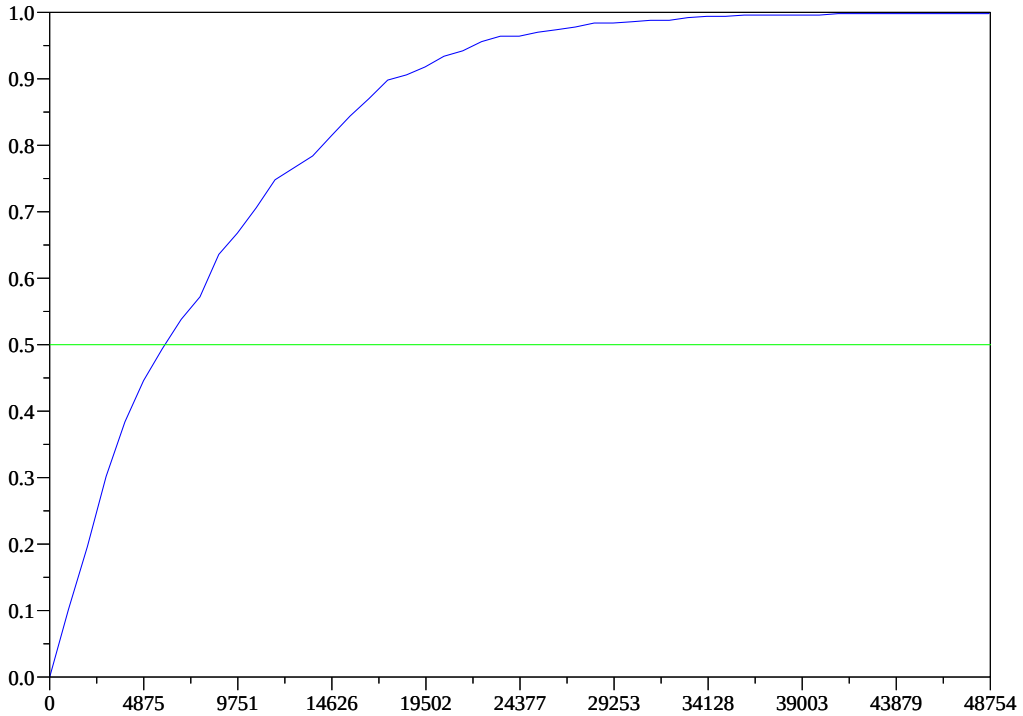
La moyenne est $m = 23.7$.



La médiane est $med = 23$ ans.

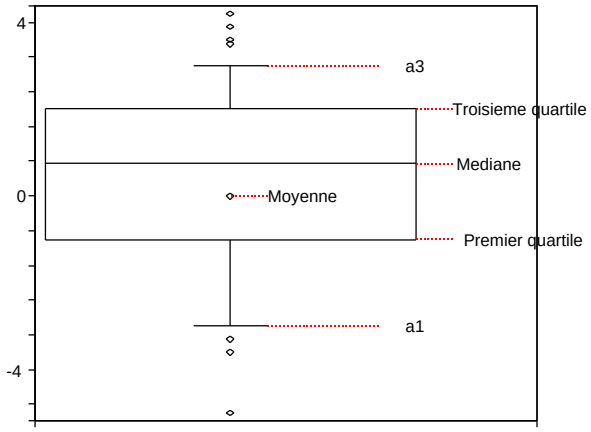
Exemple 2 : Moyenne et médiane de la durée de vie d'une ampoule.

La moyenne est $m = 8144.95$ heures.



La médiane est $med = 5904.45$ heures.

Construction de la boite à moustaches



- a_1 est la plus petite valeur supérieure à $Q_1 - 1.5 \times IQ$
- a_3 est la plus grande valeur inférieure à $Q_3 + 1.5 \times IQ$,

où $IQ = Q_3 - Q_1$ est l'intervalle interquartile.

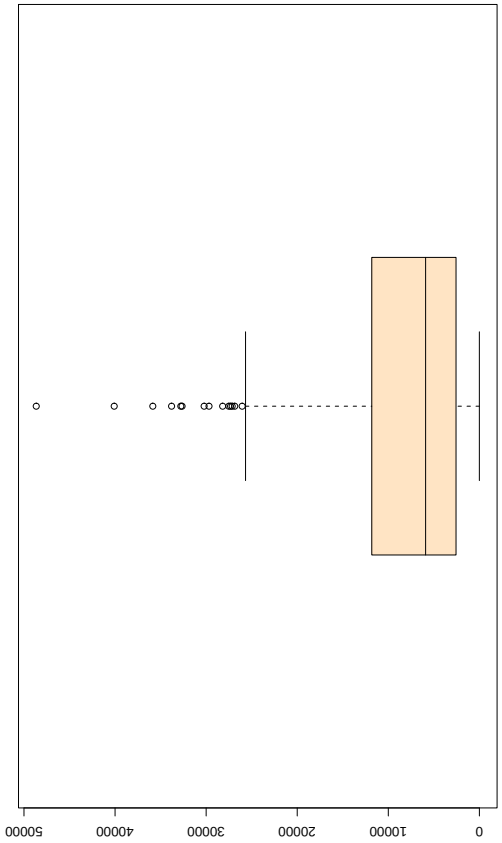
Les valeurs en dehors de ces bornes sont des valeurs "extrêmes" qui sont représentées par des points.

En général, lorsque la moyenne est supérieure à la médiane c'est le signe que la distribution est étalée vers la droite (et inversement).

Exemple 1 : Boite à moustaches de l'âge des élèves de Capes.

Allez-y !

Exemple 2 : Boite à moustaches de la durée de vie d'une ampoule dite 'économique'.



3. Le mode Le mode est défini comme étant la modalité du caractère la plus souvent prise dans la population. On s'aperçoit immédiatement des limites d'un tel indicateur :

1. il n'a de sens que dans le cas d'un faible nombre de modalité,
2. il peut exister plusieurs modes.

4. Les caractéristiques de dispersion

La **variance** : soit m la moyenne, on définit

$$var(X) = \sum_{i=1}^{p} f_i(x_i - m)^2 = \sum_{i=1}^{p} f_i x_i^2 - m^2.$$

L'**écart type** :

$$\sigma_X = \sqrt{var(X)}.$$

Partie III - Que faire en dimension supérieure ?

Lorsque l'on étudie plusieurs caractères simultanément, on souhaite évaluer le lien entre les caractères, leur dépendance.

On va se limiter ici à la dimension 2.

1 - Présentation des données : Les tableaux statistiques

Considérons N individus décrits simultanéments suivant deux caractères X et Y.

Pour $k \in \{1, \ldots, N\}$ l'individu ω_k présente les modalités $X(\omega_k) \in \{x_1, \ldots, x_p\}$ et $Y(\omega_k) \in \{y_1, \ldots, y_q\}$.

On note N_{ij} le nombre d'individus présentant les modalités x_i et y_j.

On a $\sum_{i=1}^{p} \sum_{j=1}^{q} N_{ij} = N$.

Tableau statistique d'une étude simultanée de deux caractères.

Modalités du caractère X	Modalités du caractère Y					Distribution marginale de X
	y_1	$\dots$	y_j	$\dots$	y_q	
x_1	N_{11}	$\dots$	N_{1j}	$\dots$	N_{1q}	$N_{1\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_i	N_{i1}	$\dots$	N_{ij}	$\dots$	N_{iq}	$N_{i\bullet}$
$\vdots$	$\vdots$		$\vdots$		$\vdots$	$\vdots$
x_p	N_{p1}	$\dots$	N_{pj}	$\dots$	N_{pq}	$N_{p\bullet}$
Distribution de Y	$N_{\bullet 1}$	$\dots$	$N_{\bullet j}$	$\dots$	$N_{\bullet q}$	$N_{\bullet\bullet} = N$

On appelle **fréquence du couple** (ou fréquence totale) des modalités x_i et y_j la proportion d'individus qui présentent simultanément les modalités x_i et y_j : $f_{ij} = \frac{N_{ij}}{N}$.

• **Distributions marginales**

Les effectifs $N_{i\bullet}$ définissent la distribution marginale de X. La **fréquence marginale** de la modalité x_i est

$$f_{i\bullet} = \frac{N_{i\bullet}}{N}.$$

De même pour Y, on définit les fréquences marginales $f_{\bullet j} = \frac{N_{\bullet j}}{N}$.

On définit alors $\overline{x}$, $var(X)$, $\overline{y}$ et $var(Y)$.

• **Distributions conditionnelles**

La j$^{\text{ème}}$ colonne du tableau statistique décrit la sous population des individus possédant la modalité y_j suivant le caractère X. La fréquence conditionnelle de la modalité x_i **sachant** y_j (ou **liée à** y_j) est

$$f_i^j = \frac{N_{ij}}{N_{\bullet j}}$$

(*lire "f" "i" sachant "j"*).

De même, la distribution conditionnelle sachant x_i : $f_j^i = \frac{N_{ij}}{N_{i\bullet}}$.

Remarque 2 $f_{ij} = \frac{N_{ij}}{N} = f_{i\bullet} f_j^i = f_{\bullet j} f_i^j$.

2 - Indépendance et dépendance

Le caractère X est **indépendant** du caractère Y si les distributions conditionnelles $(X|y_j)$ sont identiques entres elles : f_i^j ne dépend pas de j et sont alors identiques à la distribution de X.

$\Rightarrow$ les colonnes du tableau stat. sont proportionnelles entre elles.

Exemple 3 Exemple de caractères indépendants

Modalités du caractère X	Modalités du caractère Y			
	y_1	y_3	y_3	y_4
x_1	3	5	2	4
x_2	6	10	4	8
x_3	12	20	8	16

L'indépendance est un cas extrême que l'on rencontre rarement à l'état pur dans la pratique. On peut cependant mesurer l'intensité de la dépendance entre deux caractères X et Y.

• Le Chi-deux

Le Chi-deux permet de comparer le tableau des effectifs relevés à ce qu'il aurait du être si les caractères avaient été indépendants.

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \frac{(N_{ij} - \frac{N_{i\bullet}N_{\bullet j}}{N})^2}{\frac{N_{i\bullet}N_{\bullet j}}{N}} = N\left[\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{N_{ij}^2}{N_{i\bullet}N_{\bullet j}} - 1\right]$$

Propriétés

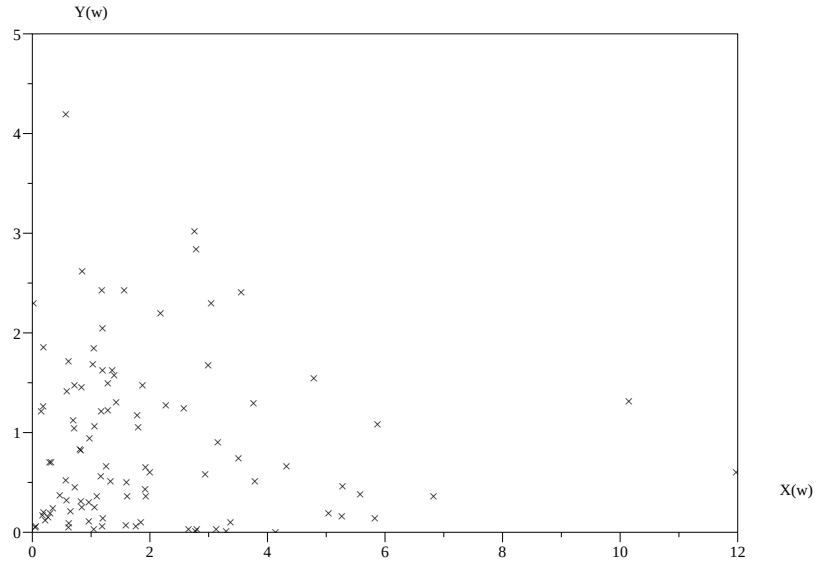
1. Les caractères X et Y sont indépendants ssi $\chi^2 = 0$.
2. $\chi^2 \geq 0$. Il est d'autant plus grand que la liaison entre X et Y est forte.

Le problème du Chi-deux est qu'il dépend de la taille de la population N et des nombres de modalités p et q.

Que peut alors signifier grand dans ce cas ?

• Le nuage de points, diagramme de dispersion (scatter-plot)

On trace sur un graphique l'ensemble des points de coordonnées $(X(\omega_k), Y(\omega_k))$ correspondant à chacun des individus ω_k, $k \in \{1, \ldots, N\}$.



Le choix des échelles est délicat.

• Le coefficient de corrélation

C'est un indice rendant compte numériquement de la manière dont les deux variables considérées varient simultanément.

La **covariance** est définie par

$$cov(X,Y) \quad = \quad \sum_{i=1}^{p}\sum_{j=1}^{q} f_{ij}x_i y_j - \overline{xy}$$

Propriétés

1. la covariance est symétrique : $cov(X,Y) = cov(Y,X)$,
2. $var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$,
3. $cov(X,Y)^2 \leq var(X)var(Y)$.

Le **coefficient de corrélation linéaire** est défini par

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y}.$$

Propriétés

1. $corr(X,Y) = cov(\frac{X-\overline{x}}{\sigma_X}, \frac{Y-\overline{y}}{\sigma_Y})$,
2. Symétrie : $corr(X,Y) = corr(Y,X)$,
3. $corr(X,Y) \in [-1,1]$
4. $|corr(X,Y)| = 1$ ssi il existe une liaison linéaire entre X et Y ($\exists a,b,c \in \mathbb{R}$ tels que $aX + bY + c = 0$),
5. Si X et Y indépendantes alors $corr(X,Y) = 0$.

• Regression linéaire

Quand $|corr(X, Y)|$ proche de 1, on souhaite trouver la fonction de X approchant "le mieux possible" Y. La régression linéaire consiste à chercher des fonctions affines (du type $aX + b$).

On choisit a et b qui minimisent la distance entre le nuage de points et la droite d'équation $y = ax + b$. On obtient

$$\hat{a} = \frac{cov(X, Y)}{var(X)} \quad \text{et} \quad \hat{b} = \overline{y} - \hat{a}\overline{x}.$$

Propriétés

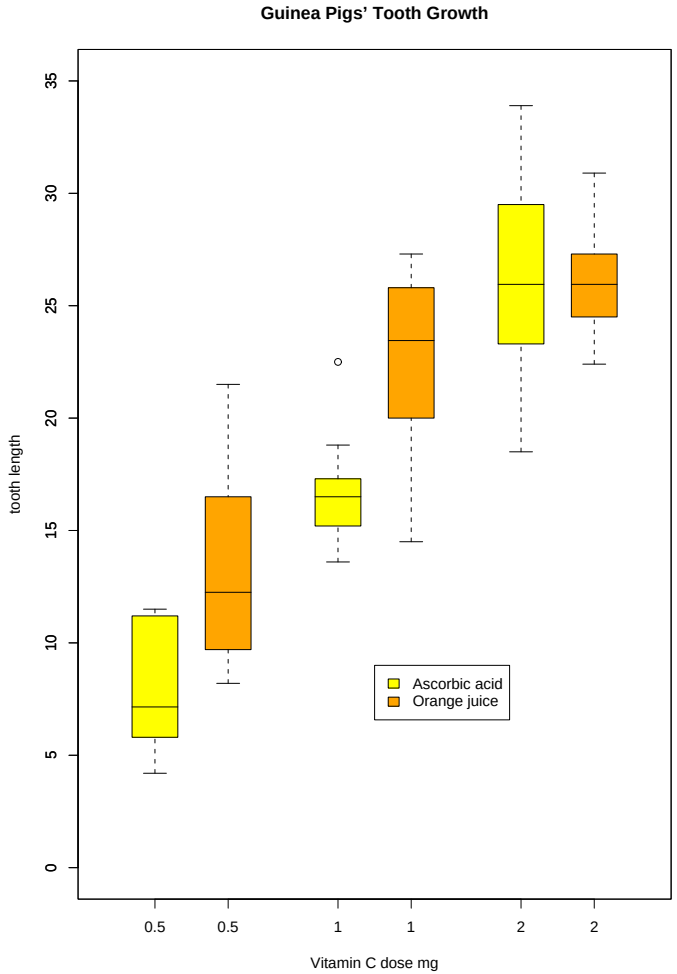
1. La droite d'équation $y = \hat{a}x + \hat{b}$ est appelée **droite de regression de Y sur X**. Elle passe par le barycentre du nuage de points de coordonnées $(\overline{x}, \overline{y})$.
2. Les valeurs $\hat{y}_k = \hat{a}X(\omega_k) + \hat{b}$ sont appelées les **valeurs ajustées**. Elles ont la même moyenne $\overline{y}$ que Y.
3. Les valeurs $\hat{E}_k = Y(\omega_k) - \hat{y}_k$ sont appelées les **résidus**. Ils sont de moyenne nulle et de variance $\frac{1}{N}S(\hat{a}, \hat{b})$.
4. La variable causale $X = (X_1, \ldots, X_N)$ et la variable résiduelle $\hat{E} = (\hat{E}_1, \ldots, \hat{E}_N)$ sont non corrélées : $corr(X, \hat{E}) = 0$.

• Que faire si un caractère est quantitatif et l'autre qualitatif

Supposons que X soit qualitative à p modalités et Y soit quantitative de moyenne $\overline{y}$ et de variance $var(Y)$.

On peut faire des histogrammes parallèles ou des boites à moustaches parallèles où graphique est la représentatation des lois conditionnelles $(Y|X = x_i)$, $i \in \{1, \ldots, p\}$.

Exemple Une étude a été menée pour évaluer l'influence de la vitamine C sur la croissance des dents de 10 cobayes selon la quantité (trois doses ont été administrée : 0.5, 1 et 2 mg) et selon le mode de délivrance (jus d'orange ou ascide ascorbique)



Exemples de boites à moustaches parallèles.

Partie IV - Lien avec les probabilités

On se limite au cas de la dimension 1 (ceci se généralise facilement en dimension ≥ 1).

On a introduit un certain nombre de quantités afin de résumer un tableau de données pour en tirer des règles générales. En allant plus loin, on cherche en fait une *formule mathématique* qui régit le comportement de chaque chose ....Illusoire!!

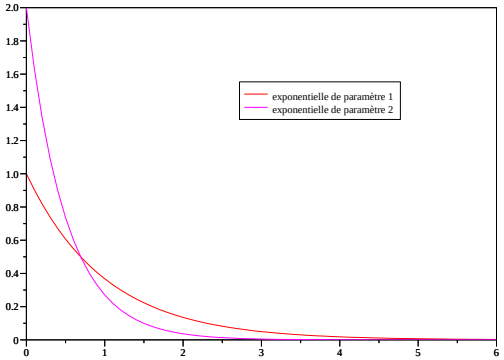
Résumons les choses de façon simple *(voir simpliste)* :

On modélise le comportement d'un objet, d'un individu par la loi de probabilité la plus semblable possible afin de pouvoir anticiper l'avenir ou d'améliorer le présent.

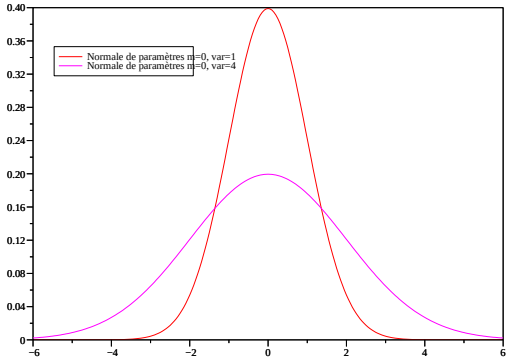
Par exemple, en ce qui concerne l'ampoule éléctrique étudiée, il serait intéressant de connaitre la loi de sa durée de vie afin de pouvoir la comparer à d'autres ampoules et évaluer son efficacité.

La *statistique mathématique* (ou inférentielle) consiste à chercher parmis toutes les lois de probabilités connues la loi dont la courbe s'approche le mieux de l'histogramme des données étudiées. Et ensuite d'ajuster les paramètres de la loi en fonction des principales caractéristiques des données.

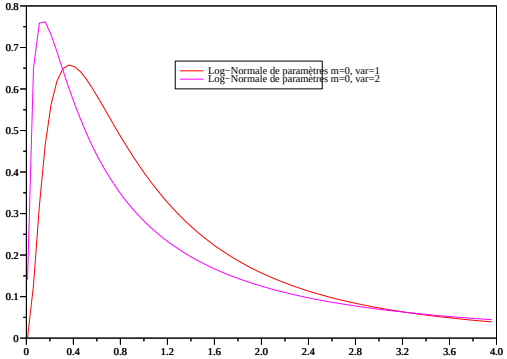
Quelques densités de loi de probabilités :



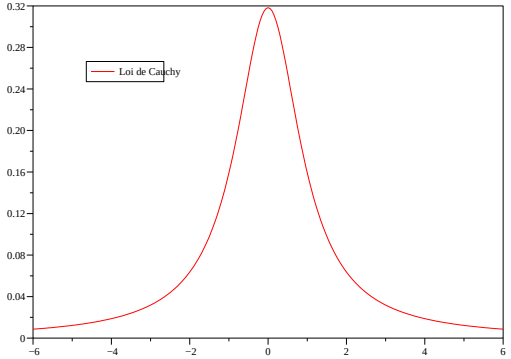
Loi exponentielle



Loi normale

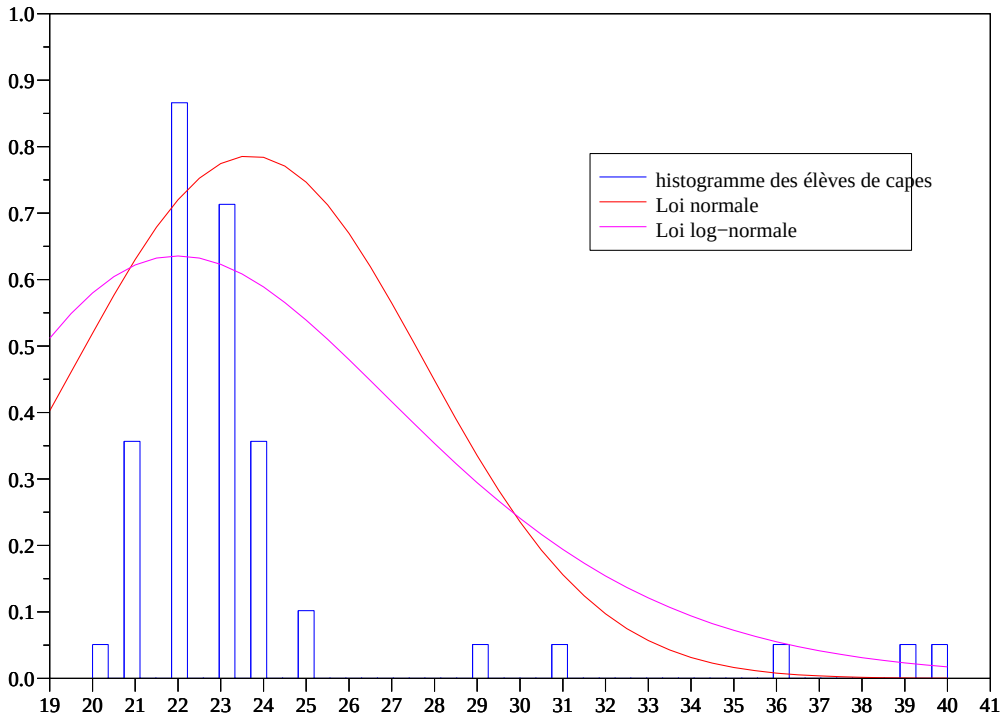


Loi Log-Normale

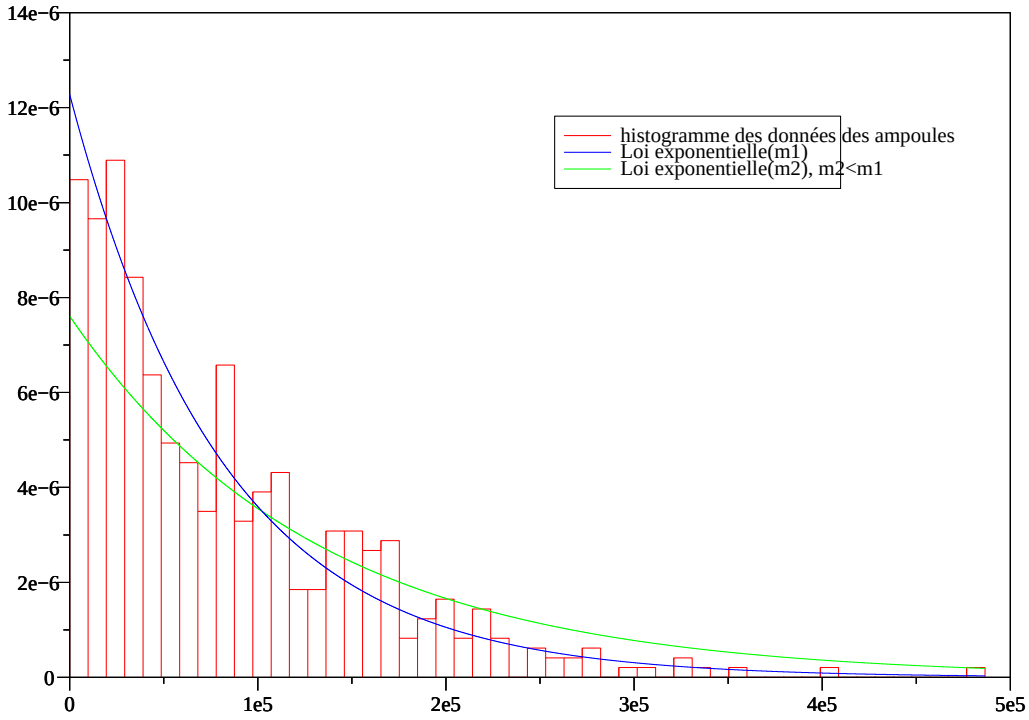


Loi de Cauchy

Exemple 1 : Âge des étudiants de Capes en 2005-2006.



Exemple 2 : Durée de vie d'une ampoule dite 'économique'.



Comment trouver *la bonne loi de probabilité* ?

$\Rightarrow$ Modélisation : Essayer de trouver parmi les lois connues celle dont l'allure s'approche le mieux de l'histogramme.

$\Rightarrow$ Estimation : Trouver des approximations des paramètres (*espérance, variance*) de la loi pour bien *coller* à l'histogramme.

$\Rightarrow$ Test : Validation du choix. Pour cela, on fait une nouvelle étude, un nouveau sondage. On teste si notre choix est conforme aux résultats du nouveau sondage.

Exemples :

En assurance : Le nombre de sinistres par contrat est modélisé par une loi de Poisson. Le cout moyen des accidents est modélisé par une loi log-normale (*assurance auto*) ou par une loi de Pareto (*incendie*).

En économie : Le temps de chômage d'un chômeur est modélisé par une loi exponentielle. Les taux d'intêret ou les taux de change par la loi log-normale.

Souvent on suppose que tout est régit par les lois normales (gaussiennes). Il existe cependant quelques tests simples pour tester si le comportement est similaire à un comportement gaussien.

La droite de Henry : On note ϕ la fonction de répartition de la loi normale centrée réduite.

Soit X un caractère à p modalité, notées $x_1, \ldots, x_p$. Pour chaque i, on calcule y_i tel que $\phi(y_i) = F_i = P(X \leq x_i)$.

Si la variable X est gaussienne, alors les points (x_i, y_i) sont alignés sur la droite d'équation $y = \dfrac{x - \overline{x}}{\sigma}$.