Méthodes statistiques – MST 2

ENSIIE

1ère année

2015/16

Evry : Nicolas Brunel

Strasbourg : Emmanuel Périnel, Nancy Rebout

Sommaire

0. Introduction
1. Notions de base – Statistique descriptive
2. Bases de l'échantillonnage
3. L'estimation
4. L'estimation par intervalle de confiance
5. Les tests statistiques

0. Introduction

Plan

- Qu'est-ce que la statistique ?
- Domaines d'applications
- La démarche statistique
- Quelques exemples
- L'objet de ce cours

Qu'est-ce que la statistique ?

Définition(s)

« Ensemble de données d'observations (une statistique) et activité (la statistique) qui consiste en leur **recueil**, leur **traitement** et leur **interprétation** »

« Compter, dénombrer, **résumer**, synthétiser des **données** afin de mieux comprendre des phénomènes, les expliquer, les modéliser, les prévoir »

- **À l'origine** : ensemble d'informations concernant la population et l'économie
- **Aujourd'hui** : branche des mathématiques appliquées à la frontière de disciplines scientifiques (mathématique et informatique) : théorie des probabilités, algèbre, théorie des graphes, algorithmique, machine learning, datamining (= fouille de données), big data, etc.

Domaines d'applications

Ils sont extrêmement **variés** !

- médecine
- démographie
- agriculture
- économie
- sociologie
- psychologie
- physique
- contrôle de qualité
- fiabilité
- enquête / sondage
- génomique
- écologie
- astronomie
- analyse sensorielle
- sport
- météorologie
- musique
- analyse de textes,
- web mining, etc.

La démarche statistique

Elles sont liées aux différentes phases du travail d'un statisticien

- **Recueil des données**
  Plan d'expérience, plan de sondage
- **Statistique descriptive et exploratoire**
  Préparation des données, représentations graphiques, indicateurs numériques, analyse bivariée et multivariée (liaisons entre variables)
- **Statistique inférentielle**
  Raisonnement à partir d'un échantillon, estimation d'une grandeur, qualité de l'estimation, test d'une hypothèse
- **Modélisation et prévision statistique**
  Expliquer / prévoir un phénomène à l'aide de modèles mathématiques

Quelques exemples

1. Construire un plan de sondage dans une enquête marketing

But : évaluer l'appréciation d'un nouveau produit par des consommateurs

- Quels sont les clients potentiels d'un produit (population étudiée) ?
- Quelle technique de sondage (quotas ? aléatoire ? stratifié ? par grappe ?)
- Comment interroger (téléphone ? internet ? auto administré ? En face à face ?)
- Combien de personnes doit-on / peut-on interroger (taille d'échantillon) ?
- Quelle est la précision attendue sur les résultats obtenus selon la taille ?
- Évaluer la représentativité de l'échantillon (faut-il redresser l'échantillon ?)

2. Construire un plan d'expérience en agriculture

But : Comparer le rendement de variétés de blé

- Quelle est la variable réponse ? *Rendement de chaque variété*
- Quels sont les facteurs contrôlés ? *Les variétés, les doses de fertilisant*
- Quels sont les facteurs aléatoires ? *Hétérogénéité du sol, météo, etc.*
- Quel type de dispositif expérimental ? *Dispositif en blocs, randomisation totale, plans en carrés latins, split plot, criss-cross, α - plans, etc.*
- Combien de variétés peut-on étudier au maximum ?
- Comment maîtriser les effets de bordure ou de voisinage ?
- etc.

3. Analyser le lien de dépendance entre deux caractères

But : étude de la pérennité d'une union selon le type d'habitat

D'après Balakrishnan, 1986

Échantillon de 3864 couples, situation après 5 ans

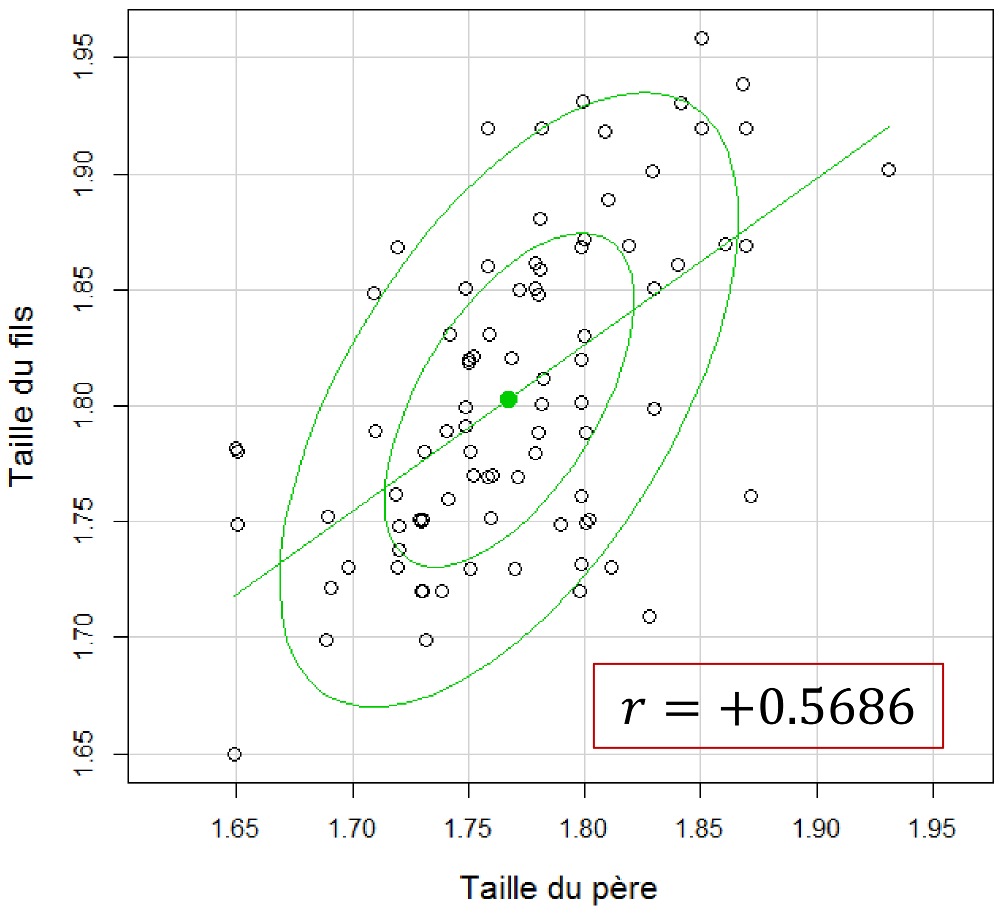
Situation / Habitat	Unis	Séparés	Total
rural	287	18	305
Petite ville	1124	89	1213
Grande ville	2081	265	2346
Total	3492	372	3864

Test d'indépendance (Khi²) :	
Khi² (Valeur observée)	19,685
Khi² (Valeur critique)	5,991
DDL	2
p-value	< 0,0001
alpha	0,05

L'hypothèse d'indépendance est rejetée…

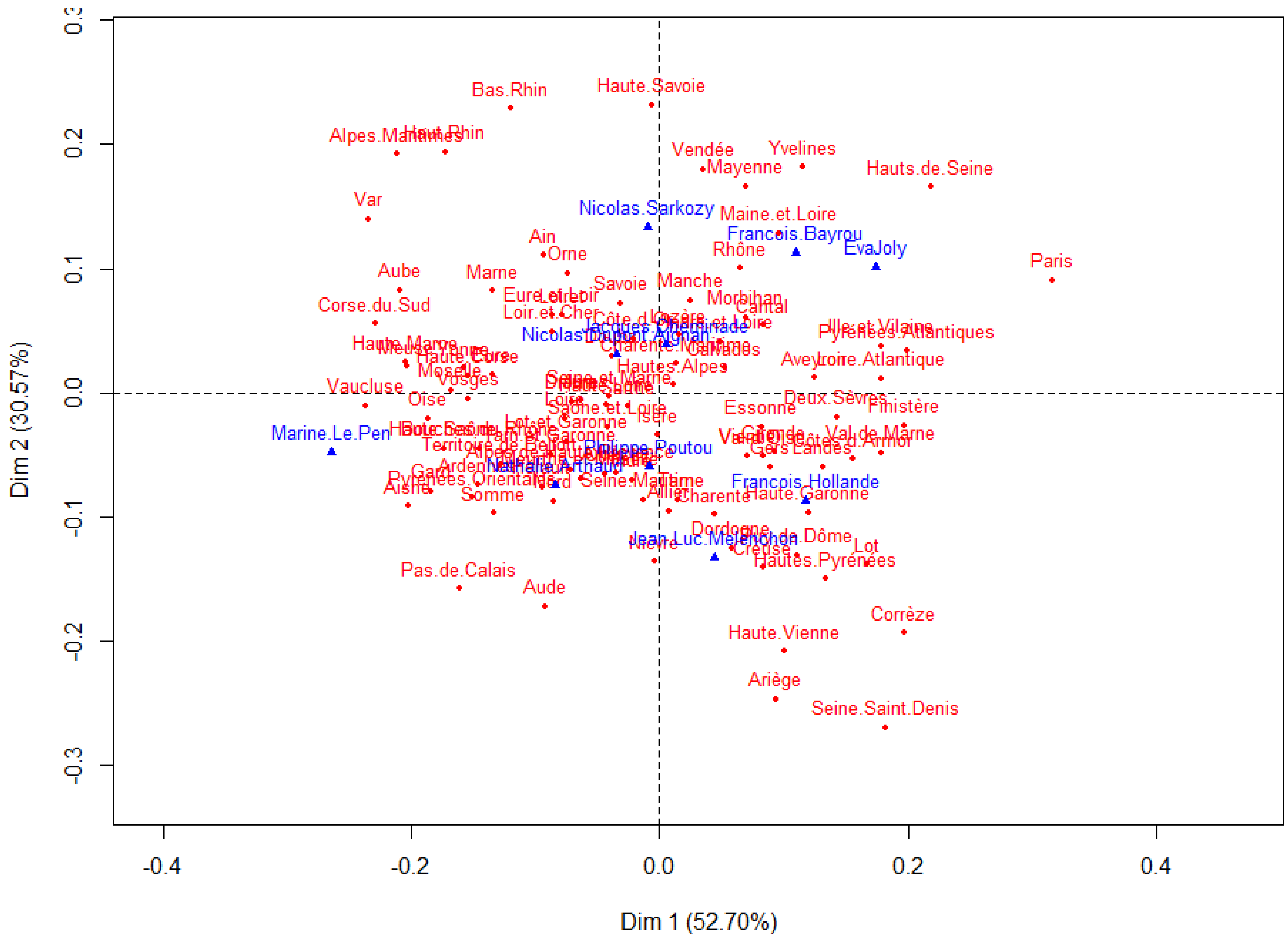
4. Analyse d'une corrélation linéaire

But : Dans quelle mesure la taille d'un enfant est-elle liée à celle de ses parents ?



5. Analyse exploratoire multidimensionnelle (AFC)

But :
Comparer les résultats de la présidentielle 2012 selon les départements



6. Modélisation par analyse discriminante

But : Prévoir l'appartenance d'un individu à un groupe prédéfini

D'après Saporta (Analyse discriminante, support de cours du CNAM)

- **Exemple 1.** Solvabilité d'emprunteurs auprès de banques – Crédit scoring

- Deux groupes : Client à risque (contentieux) ou non
- Variables explicatives : taux d'endettement, revenu disponible du ménage, statut matrimonial, propriétaire/locataire, profession, ancienneté emploi, âge, nombre d'enfants, etc.

- **Exemple 2.** Risque en assurance automobile

- Deux groupes : « Bon ou mauvais conducteur »
- Variables explicatives : CSP, sexe, tranche d'âge, catégorie de véhicule, etc.

0. Introduction

Exemple 1. Solvabilité d'emprunteurs

Principe : établir un score, fonction des variables explicatives, traduisant le niveau de solvabilité du client

$$\boxed{score = fonction\ (X_1, X_2, \dots, X_p)}$$

Exemple

Variable	Valeur – catégorie	Score
Ratio d'endettement	15%	+16
Revenu disponible par personne	2300 F	+12
Situation dans le logement	locataire	0
Etat matrimonial, nombre d'enfants	Marié sans enfant	+10
Ancienneté dans l'emploi	6 ans	+22

$$score = +60$$

0. Introduction

Répartition du nombre de contentieux par tranche de score

Tranche de score	Nbre de deman.	Nbre de conten.	Taux de conten.
90-100	1000	10	1 %
80-90	1500	35	2,3 %
70-80	1500	55	3,6 %
60-70	2000	80	4 %
50-60	2000	100	5 %
40-50	2000	140	7 %
30-40	2000	180	9 %
20-30	1000	110	11 %
10-20	1000	130	13 %
0-10	1000	160	16 %

7. Web mining – Text mining

Web mining

Ensemble des techniques qui visent à explorer, traiter et analyser les grandes masses d'informations consécutives à une activité Internet
But = valoriser un site ; personnaliser un contenu selon le profil de l'utilisateur

Type de données traitées : contenu d'une page (textes, graphiques), sa structure, son usage (adresses IP, date, temps des requêtes), profil de l'utilisateur

Text mining

Extraction de connaissances dans les textes ; spécialisation de la fouille de données (*data mining*) ; fait partie du domaine de l'intelligence artificielle
Applications : indexation de textes, détection d'anomalies, anti spam, mesure de ressemblance / coïncidence entre textes, etc.

L'objet de ce cours

- Notions de base, statistique **descriptive**
- Les fondements de la statistique **inférentielle**

Un cours de statistique inférentielle / statistique mathématique

- **Échantillonnage** : déduire des renseignements sur un échantillon à partir de la connaissance de la population
- **Estimation** : déduire des renseignements sur une population à partir de la connaissance d'un échantillon
- Mettre en place un **test d'hypothèse**

1. Notions de base

Plan

- Exemples de jeux de données
- Population, échantillon, individus
- Variables statistiques
- Le tableau *individus* x *variables*
- Représentations graphiques usuelles
- Principaux indicateurs numériques

1.1 - Exemples de jeux de données

Exemple 1. Chevesne

Longueur (en mm) et poids (en grammes) de 20 chevesnes

longueur	poids
105	11
155	36
159	41
165	43
170	46
173	56
181	66
187	70
191	76
195	76
202	101
211	102
220	114
221	118
225	125
232	136
238	139
248	158
252	159
301	274

Exemple 2. Prison

Enquête par questionnaire auprès de 1500 adultes français
sur le thème de la prison

Ident	âge	sexe	âge en classes	diplôme	PCS	orientation politique	Superficie cellule	détenus par cellule	WC cloisonnés	préservatifs en prison	sexe lors des visites	travail souhaitable	conditions de détention	respect des droits de l'homme	Poids
1	44	homme	âge (36,50)	DPS	ouvrier	centre droit	10	2	non	oui	dépend	oui	a.bonnes	plutôt	2,16
2	24	femme	âge (18,25)	DPS	employé	autre	6	2	non	oui	non	oui	a.mauvaises	pasdutout	0,39
3	37	femme	âge (36,50)	sans diplôme	commerçant ,artisan	centre	6	2	non	non	non	oui	mauvaises	pasdutout	0,66
4	20	homme	âge (18,25)	BAC	inactif	droite	10	3	non	oui	oui	oui	bonnes	plutôtpas	0,66
5	75	femme	âge 66 et +	>BAC+2	profintermé diaires	centre gauche	6	4	non	non	?	oui	mauvaises	pasdutout	0,49
6	45	homme	âge (36,50)	DPS	commerçant ,artisan	centre gauche	5	4	non	oui	oui	oui	a.mauvaises	plutôtpas	0,19
7	19	femme	âge (18,25)	BAC+2	inactif	gauche	10	4	non	non	non	oui	mauvaises	pasdutout	0,32
8	46	femme	âge (36,50)	BAC	profintermé diaires	centre gauche	10	4	?	oui	non	oui	bonnes	plutôt	0,28
9	30	homme	âge (26,35)	>BAC+2	cadre, profint.sup.	autre	6	4	non	oui	non	oui	mauvaises	pasdutout	0,11
...															
1497	25	homme	âge (18,25)	>BAC+2	inactif	centre gauche	15	6	non	?	non	oui	a.mauvaises	plutôtpas	0,72
1498	58	homme	âge (51,65)	CEP	commerçant ,artisan	centre droit	7	8	non	oui	oui	oui	a.bonnes	plutôt	0,6
1499	25	homme	âge (18,25)	BAC	employé	centre gauche	5	6	non	?	oui	oui	a.mauvaises	plutôtpas	0,34
1500	34	femme	âge (26,35)	BAC	commerçant ,artisan	centre	10	4	oui	oui	non	oui	a.mauvaises	plutôtpas	0,6

Exemple 3. Poussins

Etude de l'effet de trois traitements sur le poids de 24 poussins mâle ou femelle

	Traitement 1	Traitement 2	Traitement 3
Mâles	25 30 26 33	21 26 22 27	23 28 24 29
Femelles	15 20 18 21	16 18 17 20	15 19 17 22

Exemple 4. Yaourts

Etude de la viscosité (en mPa.s) de quatre yaourts à deux dates différentes

produit	semaine	répétition	viscosité Brookfield (mPa.s)
F1	J+7	1	39400
F1	J+7	2	42000
F2	J+7	1	49200
F2	J+7	2	51000
F3	J+7	1	59400
F3	J+7	2	60800
F4	J+7	1	80000
F4	J+7	2	79600
F1	J+15	1	38800
F1	J+15	2	42400
F2	J+15	1	57800
F2	J+15	2	52600
F3	J+15	1	78200
F3	J+15	2	78000
F4	J+15	1	101400
F4	J+15	2	101600

1.2 - Population, échantillon, individus, variables

Population et individu statistique

- Très souvent : notion **démographique**, **biologique** – écologique
- **En statistique** : ensemble des objets ou individus statistiques étudiés
- **Individu** statistique = **unité** statistique : très diverses !

Population *versus* échantillon

- **échantillon** : fraction, sous-ensemble de la population étudiée
- **Pourquoi** n'étudier qu'une fraction de la population ?
- Recensement *versus* sondage

Description et inférence statistique

Description

- Résumer, synthétiser les résultats l'information contenue dans des données à l'aide de **graphiques** ou d'**indicateurs numériques**
- Les résultats obtenus se limitent aux individus observés

Inférence

- **Extrapoler, généraliser** les résultats observés sur un échantillon à la population dans sa globalité
- Qualité essentielle attendue pour un échantillon : sa **représentativité**
- Un exemple typique : l'enquête par **sondage**
- Contexte **d'incertitude**. Importance de la théorie des **probabilités**
- Notions importantes : tests d'hypothèses, risque, confiance, prévision, estimation, probabilité critique, etc.

1.3 - Variables statistiques

- **Caractéristiques** utilisées pour décrire les individus de la population étudiée
  Informations recueillies sur les unités statistiques
- Terminologie différente selon les domaines d'application :
  Variables, attributs, paramètres, descripteurs, facteurs, caractères, etc.

Deux grands types de variables

Variables quantitatives

(taille, salaire, température, etc.)

Variables qualitatives

(sexe, profession, habitat, etc.)

Variables **quantitatives**

Les valeurs prises sont des grandeurs mesurables, numériques

Elles se subdivisent en 2 types :

- **Continues** : observables sur un intervalle continu
  Nombre de valeurs possibles *a priori* infini
- **Discrètes** : prennent un nombre fini de valeurs
  (en général entières et peu nombreuses)

Variables **qualitatives**

- Les valeurs prises par la variable sont des « qualités », non numériques, appelées **modalités** ou **catégories**
- Les modalités définissent des **sous populations** dans la population étudiée (hommes/femmes, rural/urbain, etc.)
- Variables **ordinales** (modalités ordonnées) ou **nominales** (aucun ordre)

REMARQUES

- Comment **distinguer qualitative / quantitative** ?
  Opération arithmétique (min, max, somme, moyenne, etc.) possible?
- **Discret – Continu** : distinction parfois **arbitraire**...
- Une variable continue : presque autant de valeurs différentes que d'individus
- Souvent, seule la distinction *quanti / quali* est importante

Variable qualitative et sous populations associées

produit	semaine	répétition	viscosité
F1	J+7	1	39400
F1	J+7	2	42000
F2	J+7	1	49200
F2	J+7	2	51000
F3	J+7	1	59400
F3	J+7	2	60800
F4	J+7	1	80000
F4	J+7	2	79600
F1	J+15	1	38800
F1	J+15	2	42400
F2	J+15	1	57800
F2	J+15	2	52600
F3	J+15	1	78200
F3	J+15	2	78000
F4	J+15	1	101400
F4	J+15	2	101600

produit	semaine	répétition	viscosité
F1	J+7	1	39400
F1	J+7	2	42000
F1	J+15	1	38800
F1	J+15	2	42400
F2	J+7	1	49200
F2	J+7	2	51000
F2	J+15	1	57800
F2	J+15	2	52600
F3	J+7	1	59400
F3	J+7	2	60800
F3	J+15	1	78200
F3	J+15	2	78000
F4	J+7	1	80000
F4	J+7	2	79600
F4	J+15	1	101400
F4	J+15	2	101600

1.4 - Le tableau *individus* x *variables*

Principal format de données sous lequel les informations sont saisies afin d'être traitées par un logiciel statistique

variables

		x_1	$\cdots$	x_j	$\cdots$	x_p
	1	x_{11}		x_{1j}		x_{1p}
	$\vdots$			$\vdots$		
individus	i	x_{i1}	$\cdots$	x_{ij}	$\cdots$	x_{ip}
	$\vdots$			$\vdots$		
	n	x_{n1}		x_{nj}		x_{np}

1. Notions de base – Le tableau *individus* x *variables*

Identificateur des individus

- Identificateur = nom des individus
- Obligatoirement tous différents !
- Appelé aussi « *libellé des observations* », « *nom des cas* » (logiciel R – Package Rcmdr)

produit	semaine	répétition	viscosité Brookfield (mPa.s)
F1	J+7	1	39400
F1	J+7	2	42000
F2	J+7	1	49200
F2	J+7	2	51000
F3	J+7	1	59400
F3	J+7	2	60800
F4	J+7	1	80000
F4	J+7	2	79600
F1	J+15	1	38800
F1	J+15	2	42400
F2	J+15	1	57800
F2	J+15	2	52600
F3	J+15	1	78200
F3	J+15	2	78000
F4	J+15	1	101400
F4	J+15	2	101600

Individus anonymes

Athlète	100m	Longueur	Poids	Hauteur	400m
Sebrle	10,85	7,84	16,36	2,12	48,36
Clay	10,44	7,96	15,23	2,06	49,19
Karpov	10,5	7,81	15,93	2,09	46,81
Macey	10,89	7,47	15,73	2,15	48,97
Warners	10,62	7,74	14,48	1,97	47,97
Zsivoczky	10,91	7,14	15,31	2,12	49,4
Hernu	10,97	7,19	14,65	2,03	48,73
Nool	10,8	7,53	14,26	1,88	48,81
Bernard	10,69	7,48	14,8	2,12	49,13
Schwarzl	10,98	7,49	14,01	1,94	49,76
Pogorelov	10,95	7,31	15,1	2,06	50,79
Schoenbeck	10,9	7,3	14,77	1,88	50,3
Barras	11,14	6,99	14,91	1,94	49,41
Smith	10,85	6,81	15,24	1,91	49,27

Nom des observations

1. Notions de base – Le tableau *individus* x *variables*

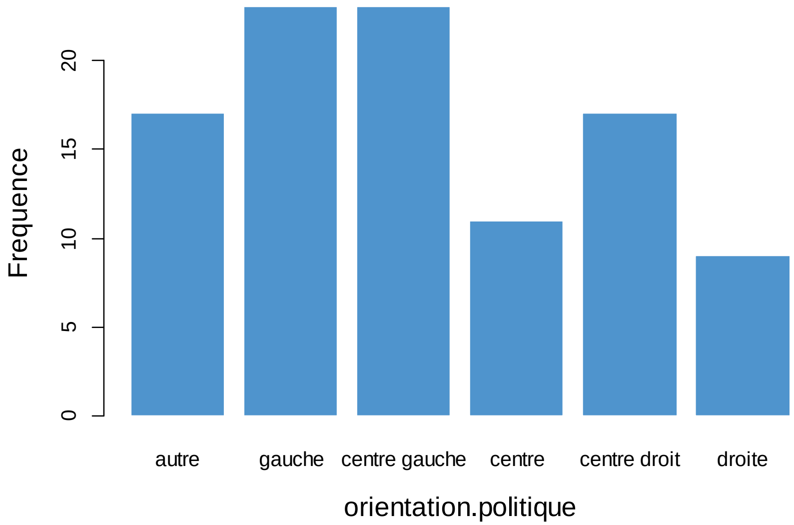
Quel format pour un tableau individus x variables ?

	Traitement 1	Traitement 2	Traitement 3
Mâles	25 30 26 33	21 26 22 27	23 28 24 29
Femelles	15 20 18 21	16 18 17 20	15 19 17 22

1.5 - Représentations graphiques usuelles

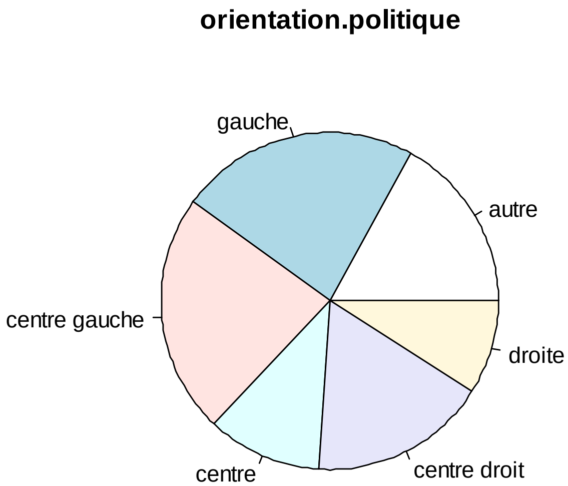
Pour une variable qualitative

Diagramme en bâton




```
barplot(table(genepi$orientation.politique),
xlab="orientation.politique", cex.lab=1.3,
  ylab="Frequence", border="white", col="steelblue3")
```

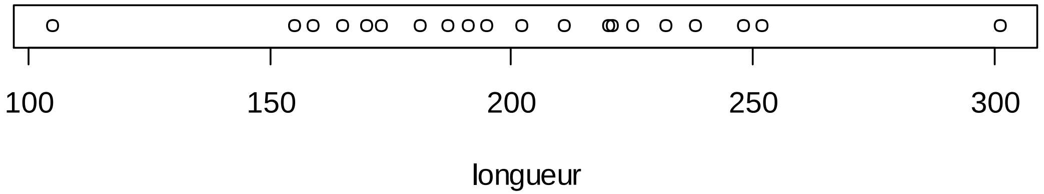
Diagramme en secteurs - Camembert




```
pie(table(genepi$orientation.politique),
    main="orientation.politique")
```

Pour une variable quantitative – *Représentation des valeurs individuelles*

Exemple : longueur d'un chevesne

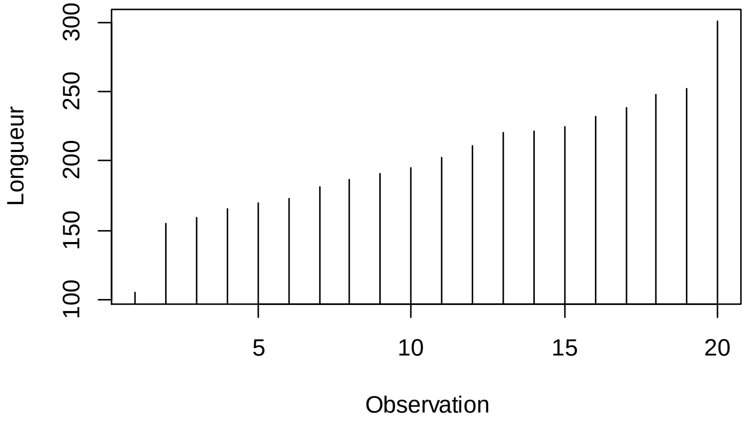


Représentation axiale

Nuage de points

Graphique « en bande »

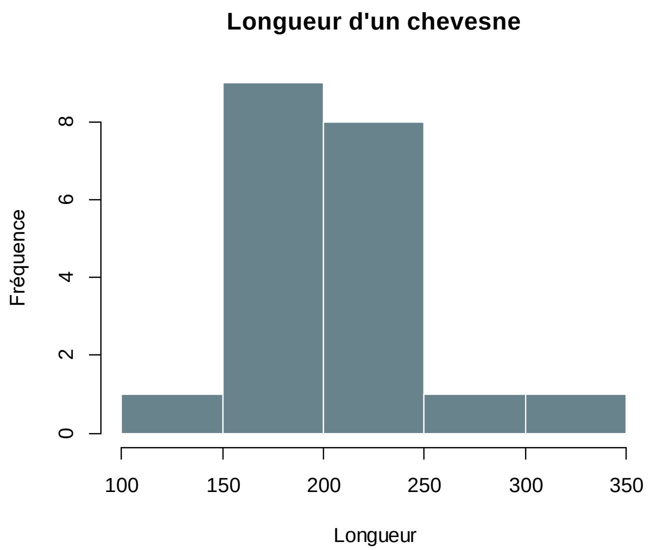
```
stripchart(chevesne$longueur,
pch=1, xlab="longueur")
```



Graphique indexé

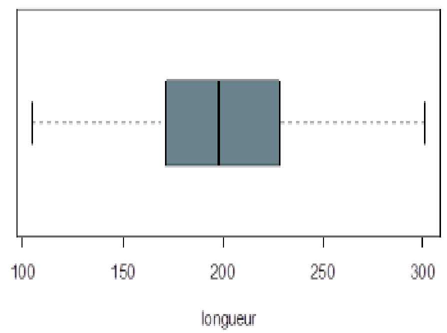
```
plot(chevesne$longueur, type="h", ylab="Longueur",
     xlab="Observation")
```

Pour une variable quantitative – *Représentation synthétique*



Histogramme

```
hist(chevesne$longueur, nclass=6,
col="lightblue4", border="white", main =
"Longueur d'un chevesne", ylab="Fréquence",
xlab="Longueur")
```

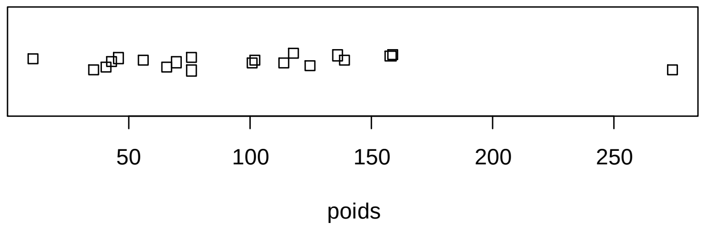


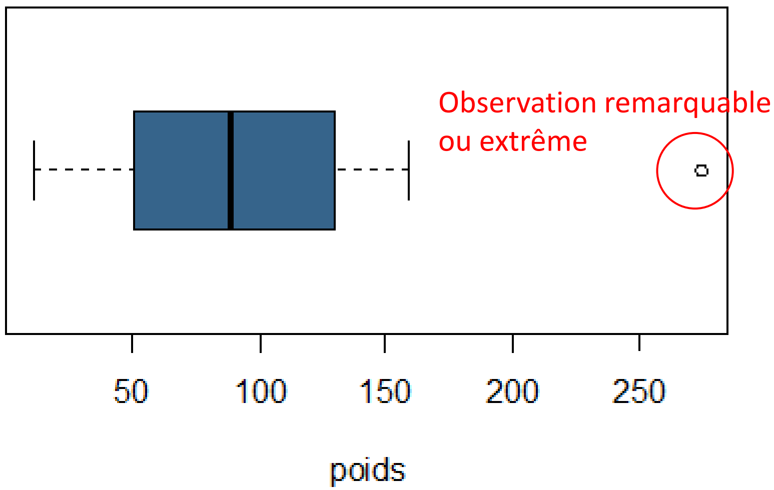
Boîte à moustaches - *boxplot*

```
boxplot(chevesne$longueur, xlab="longueur",
col="lightblue4", horizontal=T)
```

1. Notions de base – Représentations graphiques

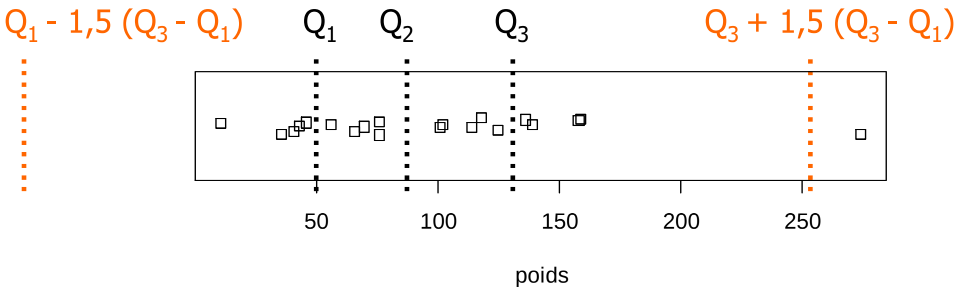
Pour la variable *Poids*

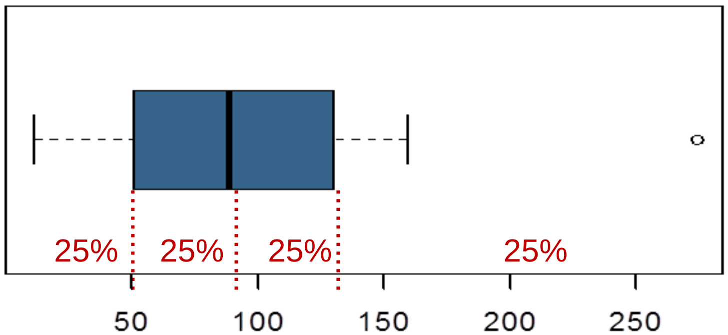




Construction d'un box plot

Représentation basée sur les *quartiles*



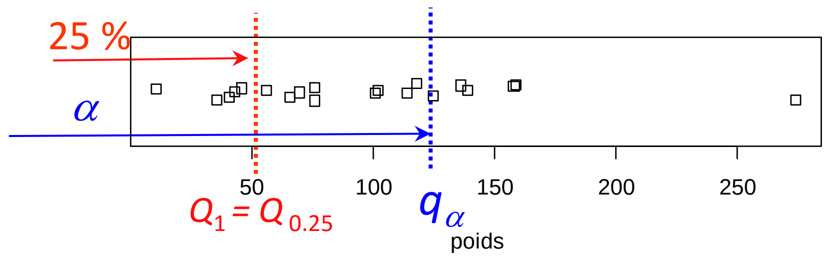


Les observations qui s'écartent du bord de la boîte de plus d'une fois et demi la longueur de la boîte sont considérées comme « remarquables », « extrêmes »

Quantile d'ordre α

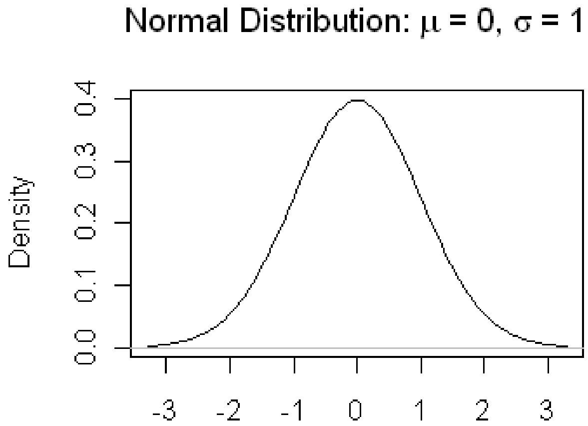
C'est la valeur d'une variable, notée q_α, associée à une **fréquence cumulée = α**

Q1 = $Q_{0.25}$ = quantile d'ordre 25 %
Q2 = $Q_{0.50}$ = quantile d'ordre 50 %
Q3 = $Q_{0.75}$ = quantile d'ordre 75 %



« Le pourcentage de valeurs inférieures à q_α est égal à α »

Illustration pour la loi Normale *N* (0,1)



$q_{0.975}$ = ?

$q_{0.95}$ = ?

Indicateurs numériques

Pour des variables quantitatives

- Indicateurs de tendance centrale
- Indicateurs de dispersion

Pour des variables qualitatives

- Tableaux d'effectifs ou de fréquences
- Mode

1.6 – Indicateurs de tendance centrale

Pour des variables quantitatives

Illustration

- Population : 25 poudres de lait
- Variable MAT/MST
  Teneur en protéine / Matière sèche

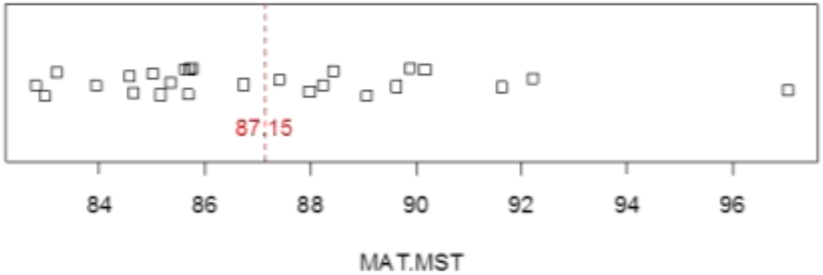
N° Poudre lait	MAT/MST
17	82,79
22	82,96
14	83,17
21	83,92
11	84,57
20	84,65
25	85,02
19	85,14
13	85,34
12	85,62
16	85,68
24	85,7
23	85,77
15	86,73
9	87,4
8	87,97
10	88,24
1	88,44
7	89,06
6	89,63
3	89,88
2	90,17
4	91,64
5	92,21
18	97,06

La moyenne arithmétique

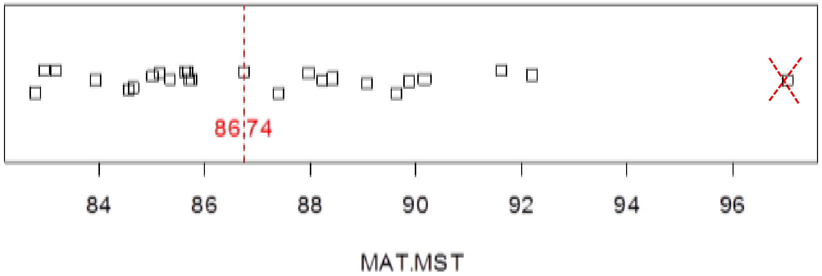
Une série statistique de *n* valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- Indicateur « universel »
- Sensibilité aux valeurs extrêmes

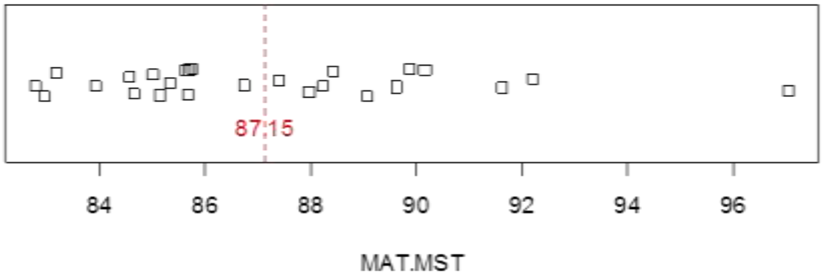


Moyenne (avec) = 87,15

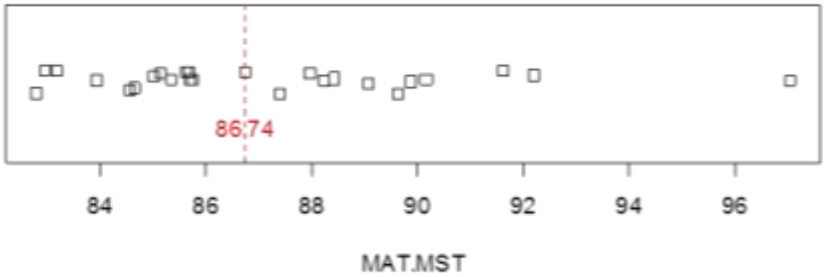


Moyenne (sans) = 86,74

1. Notions de base – Indicateurs de tendance centrale



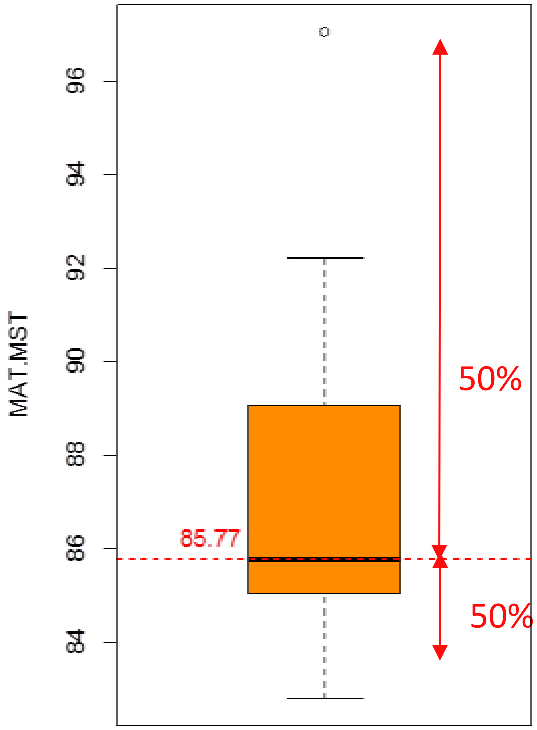

```
stripchart(MAT.MST, method="jitter", xlab="MAT.MST")
text(mean(MAT.MST),0.7, round(mean(MAT.MST),2),col="red")
abline(v=mean(MAT.MST),col="red", lty=2))
```




```
stripchart(MAT.MST, method="jitter", xlab="MAT.MST")
text(mean(MAT.MST[1:24]),0.7, round(mean(MAT.MST[1:24]),2),col="red")
abline(v=mean(MAT.MST[1:24]),col="red", lty=2)
```

La médiane

Une série statistique de *n* valeurs rangées $x_{(1)} < x_{(2)} < \ldots < x_{(i)} < \ldots < x_{(n)}$



$$\text{Me} = x_{\left(\frac{n+1}{2}\right)} \qquad n \text{ impair}$$

$$\text{Me} = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} \qquad n \text{ pair}$$

La médiane partage la série en deux parties égales

- Indicateur « robuste »
- peu sensible aux valeurs extrêmes

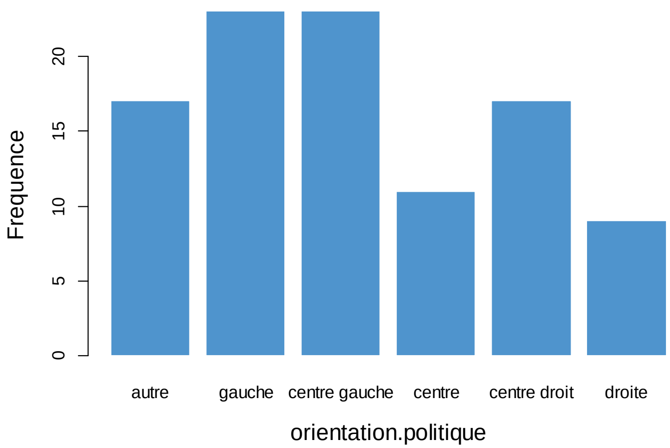
```
Médiane (avec) = 85,77
Médiane (sans) = 85,74
```

Pour des variables <u>qualitatives</u>

Tableau de fréquences

Tableau associant à chaque modalité d'une variable qualitative, sa fréquence (ou son effectif) observé dans l'échantillon

autre	gauche	centre gauche	centre	centre droit	droite
17	23	23	11	17	9



Le Mode

C'est la valeur de la variable **la plus fréquente** (ou d'effectif maximum)
Déterminé en général pour des *variables qualitatives*

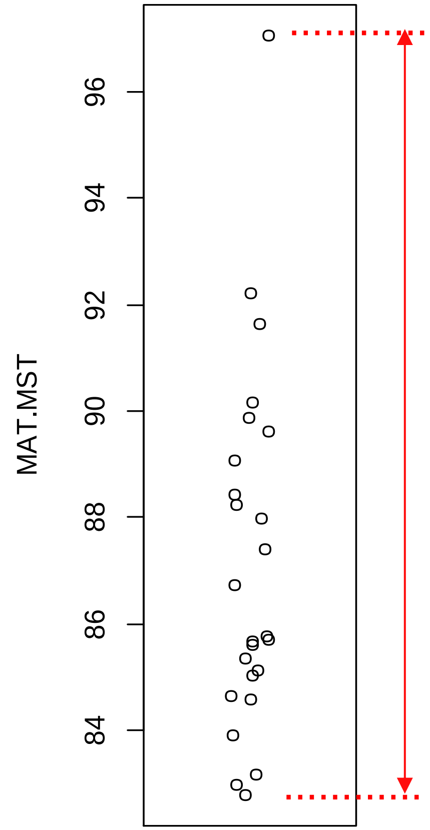
1.7 – Indicateurs de dispersion

L'étendue

Une série statistique de *n* valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$

$$\text{Etendue} = \text{Amplitude} = (x_{max} - x_{min})$$

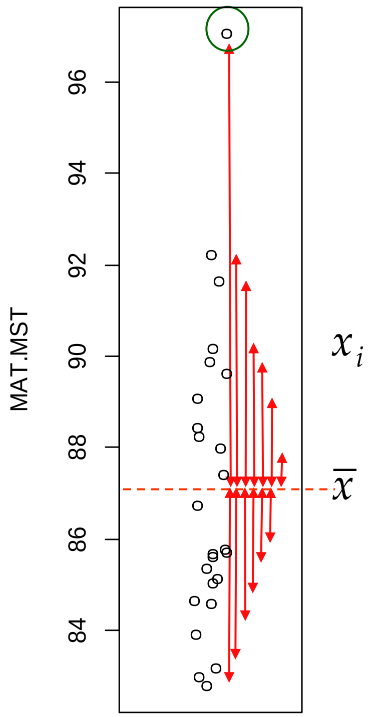
- Le plus simple et le plus intuitif
- Très sensible aux valeurs extrêmes !
- Seules deux valeurs de la série participent au calcul




```
Étendue = 97,06 - 82,79 = 14,27
```

L'écart absolu moyen : EAM

Une série statistique de *n* valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$



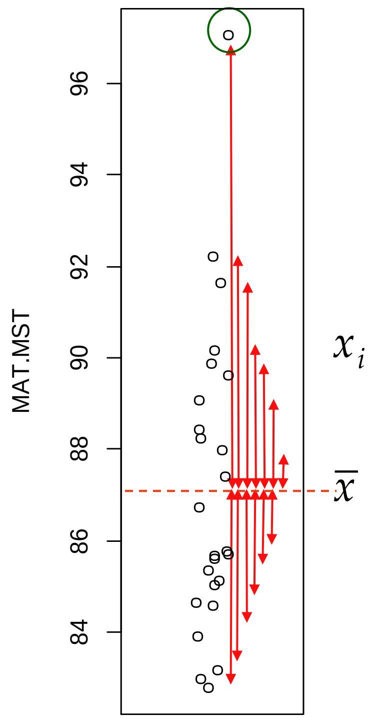
$$\text{EAM} = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

« Écart moyen à la moyenne »
« Dispersion moyenne autour de la moyenne »

EAM = 2,64 (avec ◯)
EAM = 2,27 (sans ◯)

La variance : s^2

Une série statistique de n valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$



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

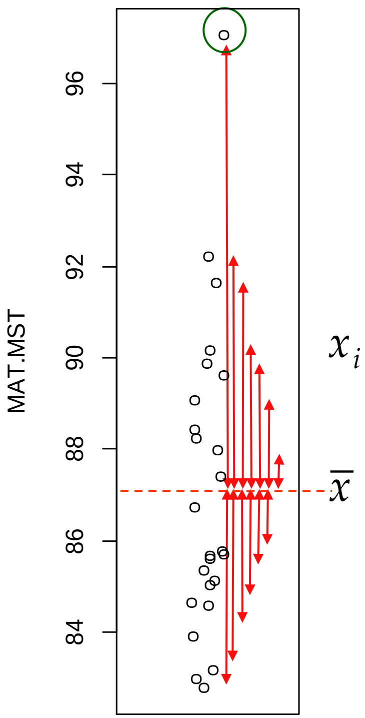
« *Moyenne des écarts à la moyenne au carré* »

- Pas d'interprétation concrète (unité de mesure au carré)
- Importance fondamentale en statistique (nombreuses propriétés mathématiques)

```
s2 = 10,81   (avec ○ )
s2 = 7,00    (sans ○ )
```

L'écart-type : *s*

Une série statistique de *n* valeurs $(x_1, x_2, \dots, x_i, \dots, x_n)$



$$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)^2}$$

« *Écart moyen à la moyenne* » (par abus…)

- Souvent proche de l'EAM (mais supérieur)
- Plus sensible aux valeurs extrêmes
- Interprétable dans l'unité de mesure étudiée

```
S  = 3,29   (avec ○ )
S  = 2,65   (sans ○ )
```

Le coefficient de variation : CV

Une série statistique de *n* valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$

$$CV = \frac{s}{\overline{x}}$$

« *écart-type normalisé / standardisé* »
« *écart-type en pourcentage de la moyenne* »

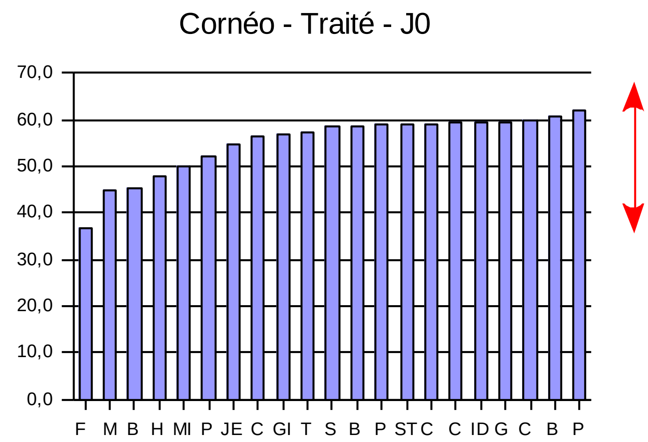
Quel intérêt ?

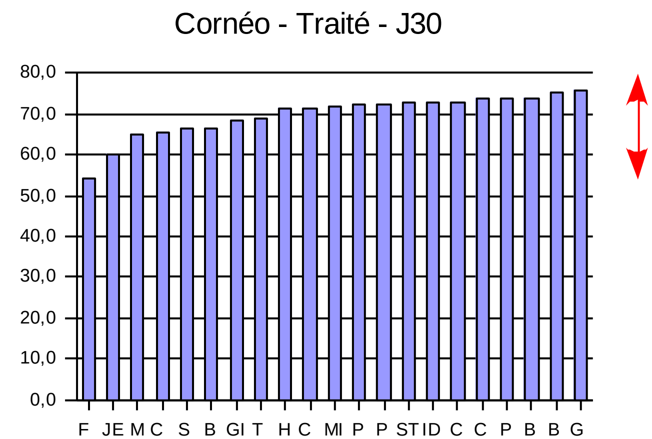
- **Comparer des dispersions** entre elles
- Le CV permet de comparer la *dispersion de variables* ayant :
  - *des unités de mesure différentes*
  - *des moyennes différentes*

1. Notions de base – Indicateurs de dispersion

Exemple. Mesure cornéométrique à deux dates

	J0	J30	
JEG	54,7	60,1	
PUJ	62,1	72,5	
CHA	56,3	73,0	
STR	58,9	72,7	
BER	45,3	66,6	
GIR	56,9	68,6	
POM	52,2	72,4	
CLA	59,8	65,6	
IDR	59,5	72,8	
MEN	45,0	64,9	
CAT	59,0	71,5	
SER	58,4	66,4	
BERG	58,5	75,2	
MIN	50,1	71,9	
BUD	60,6	73,7	
TRE	57,1	68,9	
FEI	36,6	54,2	
CEY	59,3	73,6	
GUI	59,5	75,9	
POU	58,9	73,7	
HAB	48,1	71,2	
écart-type	6,38	5,19	-18,70%
moyenne	55,09	69,78	
CV	0,12	0,07	-35,80%





Le cas de données « groupées »

Exemple

Examen anticipé d'Analyse des Données
16 Décembre 2015
- Documents interdits - Calculette autorisée.

Exercice 1 Une entreprise est constituée de deux usines, appelées A et B. Le tableau suivant récapitule les salaires en euros par catégorie de personnel et par usine :

Usine A	Salaires	Effectifs	Usine B	Salaires	Effectifs
Ouvriers	700	200	Ouvriers	900	60
Employés	1400	20	Employés	1600	40
Cadres	5300	10	Cadres	7300	20

1. Calculer la moyenne des salaires dans chacune des usines, dans l'entreprise. Vérifier que la moyenne des salaires dans l'entreprise est la moyenne des salaires moyens de chaque usine.
2. Calculer la moyenne des salaires des ouvriers, puis des employés et enfin des cadres dans l'entreprise.
3. Calculer la variance des salaires dans chacune des usines et dans l'entreprise.
4. Vérifier que la variance des salaires dans l'entreprise est égale à la moyenne des variances des usines augmentée de la variance des moyennes calculées dans chaque usine. Quelle est la propriété du cours illustrée ici ?

Les données centrées réduites

Intérêt ?

- Évaluer le caractère **remarquable** / **extrême** d'une valeur dans une série statistique
- Comparer des données exprimées dans des **unités de mesure différentes**

Données initiales

Une série statistique de *n* valeurs $(x_1, x_2, \ldots, x_i, \ldots, x_n)$

Données centrées réduites

centrage

$$\frac{(x_1 - \overline{x})}{s_x}, \ldots, \frac{\boxed{(x_i - \overline{x})}}{\boxed{s_x}}, \ldots, \frac{(x_n - \overline{x})}{s_x}$$

réduction

1. Notions de base – Données centrées réduites

Exemple. Données météorologique à Strasbourg, le 8 février

Année	minimale (°C)	maximale (°C)	ensoleillement (heures)	précipitations (mm)
2006	2,20	5,30	0,20	1,60
2007	2,10	10,70	2,20	3,00
2008	-2,20	9,40	9,20	0,00
2009	1,30	3,60	0,10	0,00
2010	-1,20	2,40	1,80	0,00
2011	0,40	8,50	0,70	0,20
2012	-9,20	-2,80	7,20	0,00
2013	-0,80	3,60	0,70	0,40
2014	0,60	8,00	0,00	2,20
2015	-4,90	4,10	0,40	0,40
moyenne	-1,17	5,28	2,25	0,78
écart type	3,37	3,80	3,08	1,03

Question : Quelle est l'observation la plus « remarquable », la plus « extrême » ?

1. Notions de base – Données centrées réduites

- Une précipitation de 3 mm se situe à 2,15 écart-type au dessus de sa moyenne
Données « brutes » : 3,00 = 0,78 + 2,15 x 1,03
Données C-R : 2,15 = 0 + 2,15 x 1
- La température de -9,2°C se situe à 2,39 écart type sous la moyenne

Année	minimale (°C)	maximale (°C)	ensoleillement (heures)	précipitations (mm)
2006	1,00	0,01	-0,66	0,79
2007	0,97	1,42	-0,02	2,15
2008	-0,31	1,08	2,25	-0,75
2009	0,73	-0,44	-0,70	-0,75
2010	-0,01	-0,76	-0,15	-0,75
2011	0,47	0,85	-0,50	-0,56
2012	-2,39	-2,12	1,60	-0,75
2013	0,11	-0,44	-0,50	-0,37
2014	0,53	0,71	-0,73	1,37
2015	-1,11	-0,31	-0,60	-0,37
moyenne	0,00	0,00	0,00	0,00
écart type	1,00	1,00	1,00	1,00

Une autre interprétation des données C-R

$$\frac{\boxed{(x_i - \bar{x})}}{\boxed{s_x}}$$

Écart de l'observation (*i*) à sa moyenne

Écart moyen à la moyenne

Propriétés des données centrées - réduites

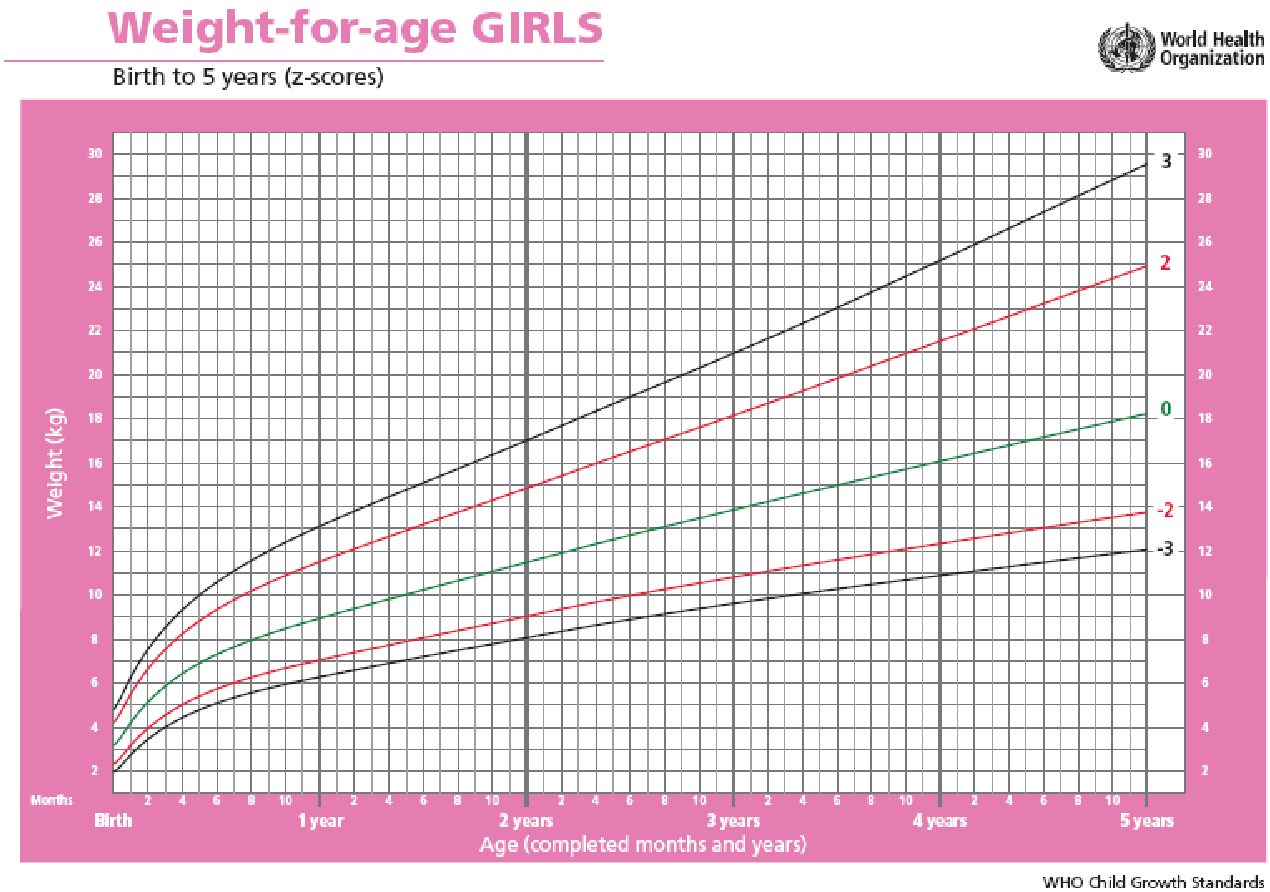
Les données centrées réduites :

- sont :
  - *de moyenne* = 0 (conséquence du centrage)
  - *d'écart-type* = 1 (conséquence de la réduction)
- s'expriment en **nombre d'écart-type**
- sont **indépendantes de l'unité de mesure** de la variable

  On parle également de données « normalisées » ou « standardisées »

1. Notions de base – Données centrées réduites

Evolution du poids des filles entre 0 et 5 ans



1.8 – Indicateurs de forme

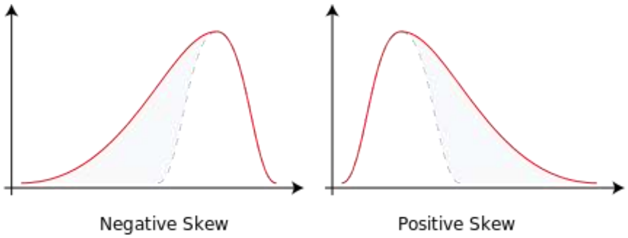
Coefficient d'asymétrie ou *skewness*

- C'est le moment centré d'ordre 3, normalisé

$$\gamma_1 = \frac{E\left[\left(X - E(X)\right)^3\right]}{\sigma^3}$$

- Référence pour une loi normale : $\gamma_1 = 0$

$\gamma_1 > 0$: asymétrie à droite
$\gamma_1 < 0$: asymétrie à gauche



source : Wikipedia

- Estimateur **sans biais**

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s'_x}\right)^3$$

Coefficient d'aplatissement ou *kurtosis*

- C'est le moment centré d'ordre 4, normalisé $$\gamma_2 = \frac{E\left[\left(X - E(X)\right)^4\right]}{\sigma^4}$$

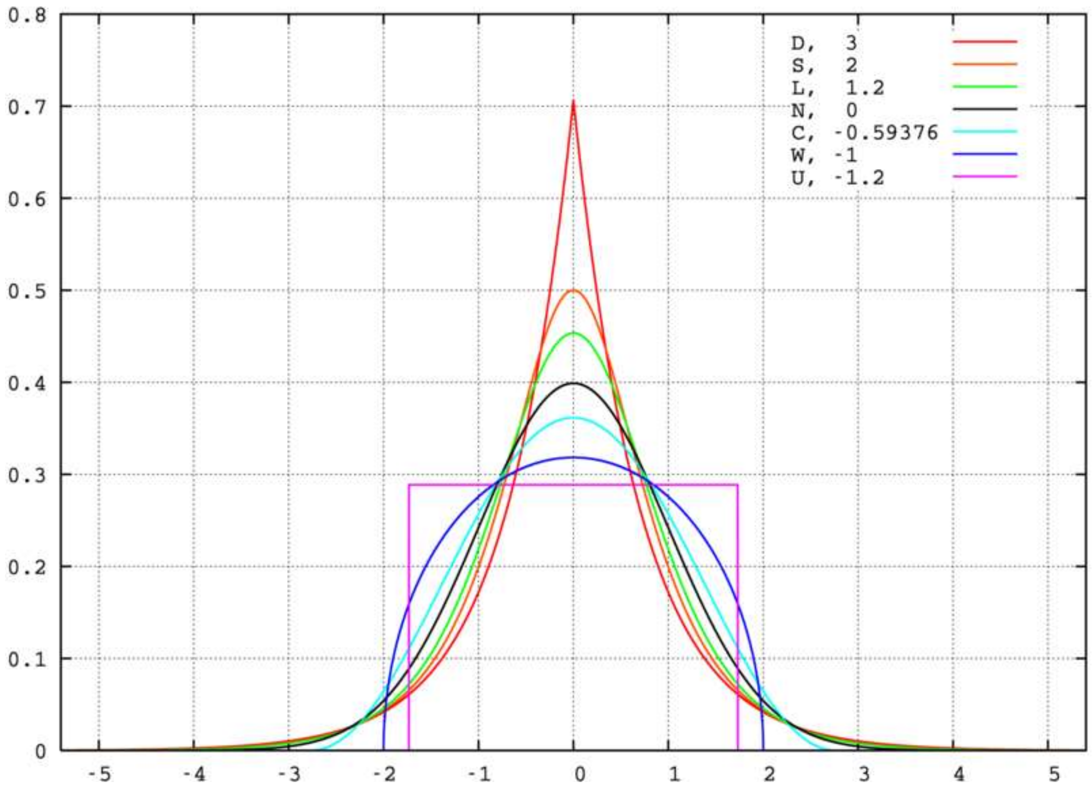
- Référence pour une loi normale : $\gamma_2 = 3$

  $\gamma_2 > 3$: queues de distribution plus « épaisses », plus pointue en sa moyenne
  $\gamma_2 < 3$: queues de distribution plus « fines »

- Le coefficient est souvent exprimé en termes d'écart à la valeur 3 (*excess kurtosis*)

- Estimateur sans biais $$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s'_x}\right)^4 - 3\frac{(n-1)^2}{(n-2)(n-3)}$$

Illustration : valeur de : $(\gamma_2 - 3)$ pour quelques distributions



source : Wikipedia

- Distribution *leptokurtique*
  $$\gamma_2 - 3 > 0$$
- Distribution *platikurtique*
  $$\gamma_2 - 3 < 0$$
- Loi normale = *mésokurtique*

1.9 – Fonction de répartition empirique

- Soit $\{x_1, x_2, \dots, x_n\}$ un échantillon de taille *n*
- On définit la fonction de répartition empirique pour tout $x \in \mathrm{R}$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$$

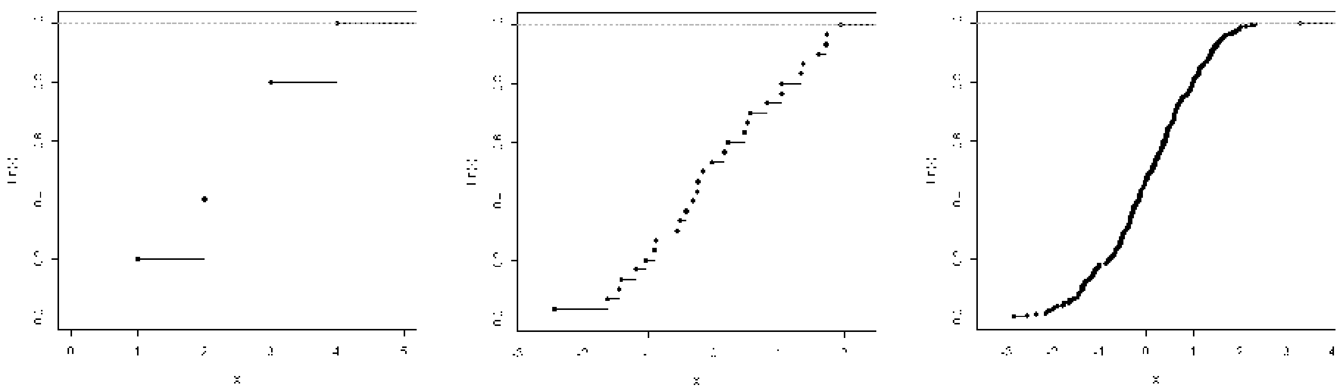


Illustration (logiciel R)

```
> x=c(1,2,3,3,4)
> plot(ecdf(x))
```

```
> x=rnorm(30)
> plot(ecdf(x))
```

```
> x=rnorm(300)
> plot(ecdf(x))
```

Inégalité Dvoretzky – Kiefer - Wolfowitz

- Soit F_X la fonction de répartition de la variable *X* de la population
- Soit $\hat{F}_n$ la fonction de répartition empirique pour un échantillon de taille *n*

$$\forall \varepsilon > 0, \qquad P\left(\sup_{x \in R} \left| F_X(x) - \hat{F}_n(x) \right| > \varepsilon\right) \leq 2e^{\left(-2n\varepsilon^2\right)}$$

1.10 – La loi Normale ou de Laplace-Gauss

Pourquoi est-elle si importante ?

1. Loi de probabilité très utilisée pour **modéliser des phénomènes** naturels (phénomènes résultant de l'addition de multiples aléas indépendants)
2. Loi des **erreurs**
3. Loi d'une **moyenne d'échantillon** ; loi **limite** de certaines distributions
4. À l'origine de la définition des nombreuses **autres lois** (ex. : *Student*, *Fisher*, *Khi2*)
5. La normalité des données est une **hypothèse nécessaire** pour la réalisation de nombreuses analyses statistiques

1. Notions de base – La loi normale

Carl Friedrich Gauss
(1777 - 1855)

Pierre Simon de Laplace
(1749 - 1827)

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY ♦ IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING ♦
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE BASIC DATA OBTAINED BY OBSERVATION AND EXPERIMENT

Citation de
William Youden
(*source : Wikipedia*)

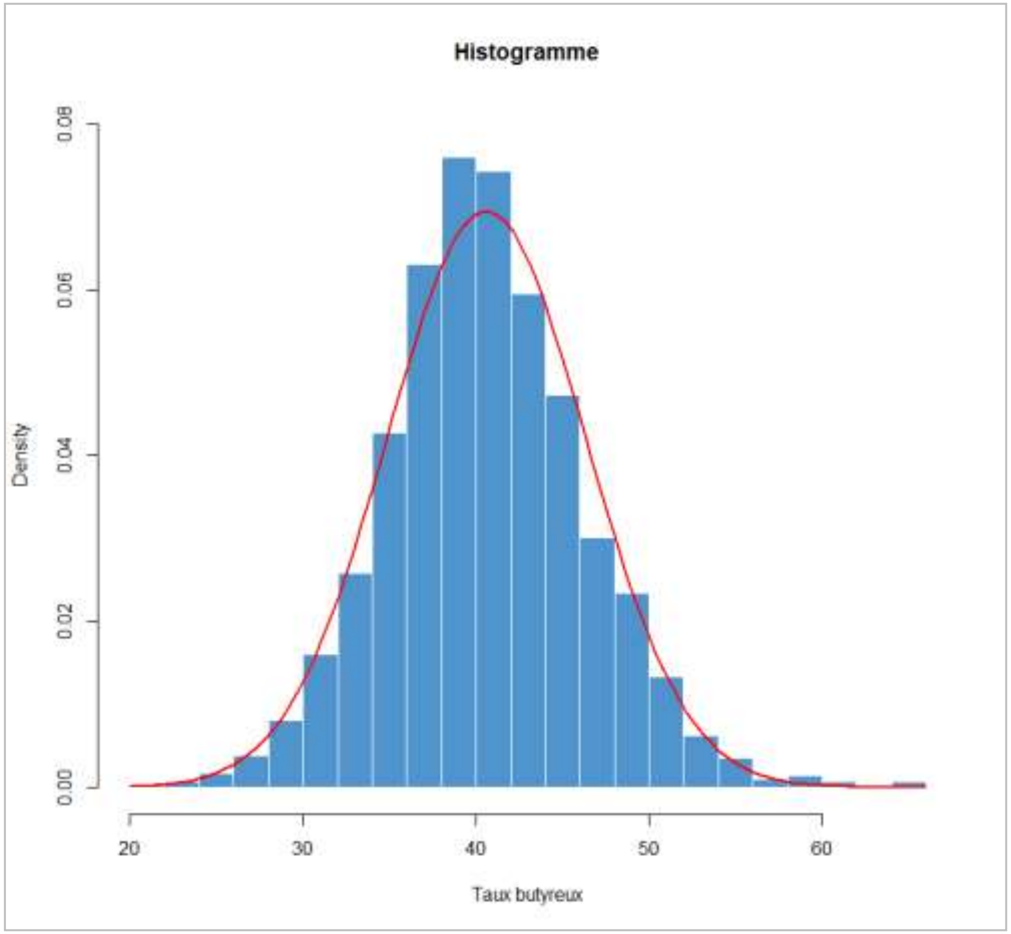
« La loi normale des erreurs se distingue dans l'expérience de l'humanité comme une des plus larges généralisations de la philosophie naturelle ♦ Elle sert de guide dans la recherche en sciences physique et sociale, en médecine, en agriculture et en ingénierie ♦ C'est un outil indispensable pour l'analyse et l'interprétation des données de base obtenues par l'observation et l'expérience. »

La loi de LG permet de modéliser un grand nombre de phénomènes

Exemple 1.

Taux butyreux du lait de 1428 vaches montbéliardes

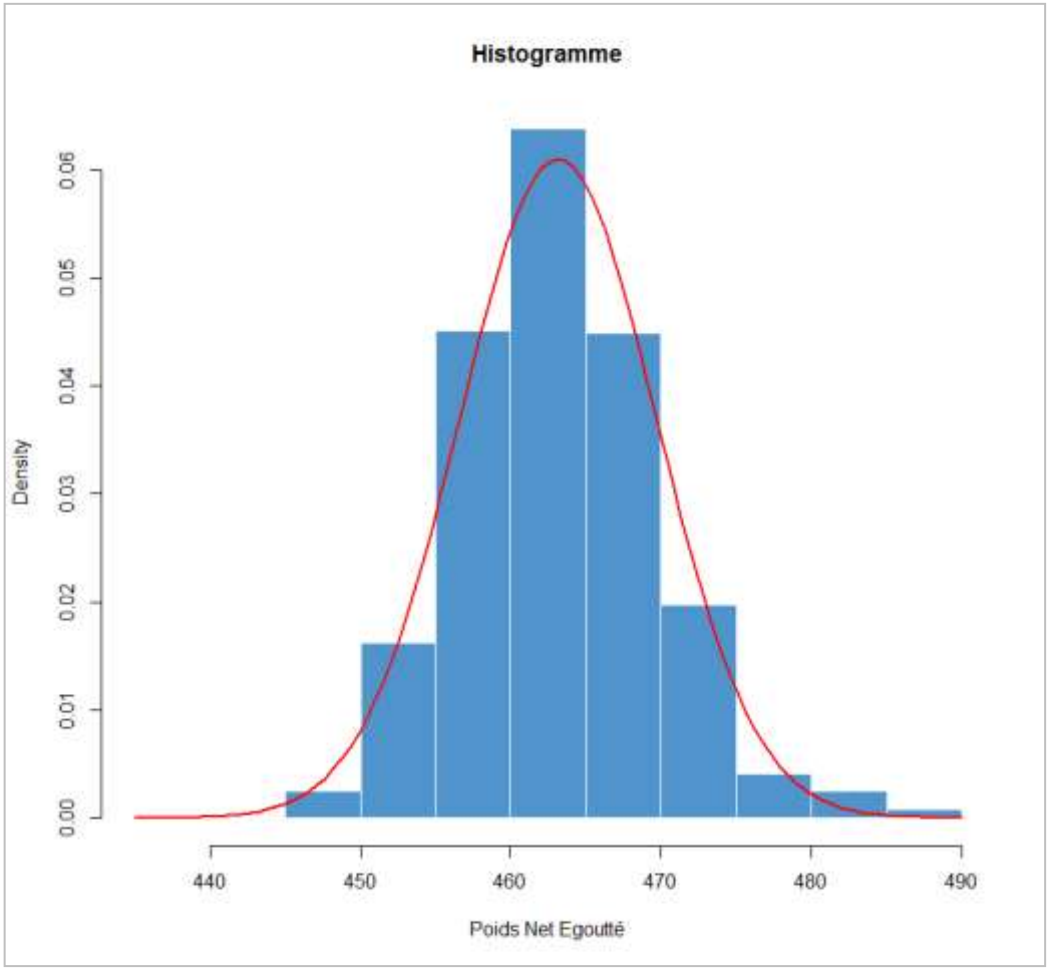
```
hist(vache$TB,
probability=TRUE, col="steelblue3",
border="white",
ylim=c(0,0.08),
breaks=30,
main="Histogramme",
xlab="Taux butyreux")

curve(dnorm(x, mean=mean(vache$TB),
sd=sd(vache$TB)), add=T,
col="red", lwd=2)
```



1. Notions de base – La loi normale

Exemple 2.

Poids net égoutté
de 2042 boîtes de conserve



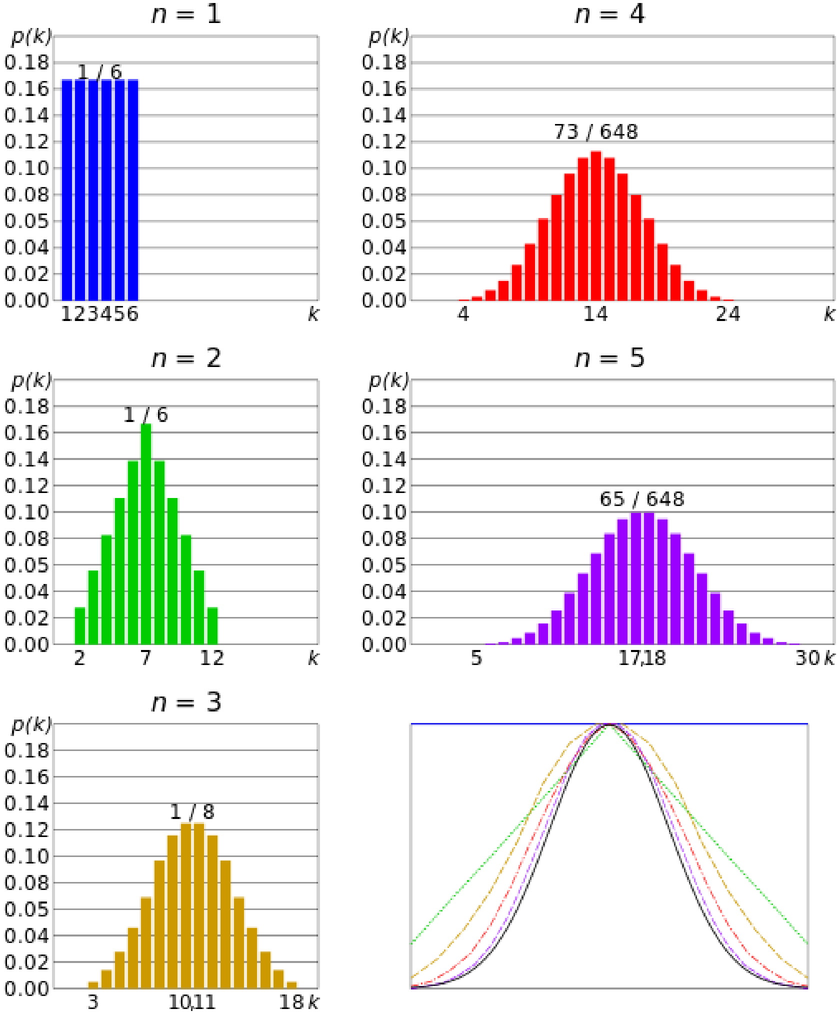
Autres utilisations ou applications

- Balistique (portée et direction)
- Erreur de mesure en astronomie
- Modélisation du Quotient Intellectuel
- Taille humaine (pour une classe d'âge donnée)
- Courbe de croissance (carnets de santé)
- Un caractère mesurable dans une population peut être modélisé à l'aide d'une loi normale s'il est codé génétiquement par de nombreux allèles ou si le caractère dépend d'un grand nombre d'effets environnementaux
- Accroissement du prix d'une denrée en Bourse (log – *N*)

1. Notions de base – La loi normale

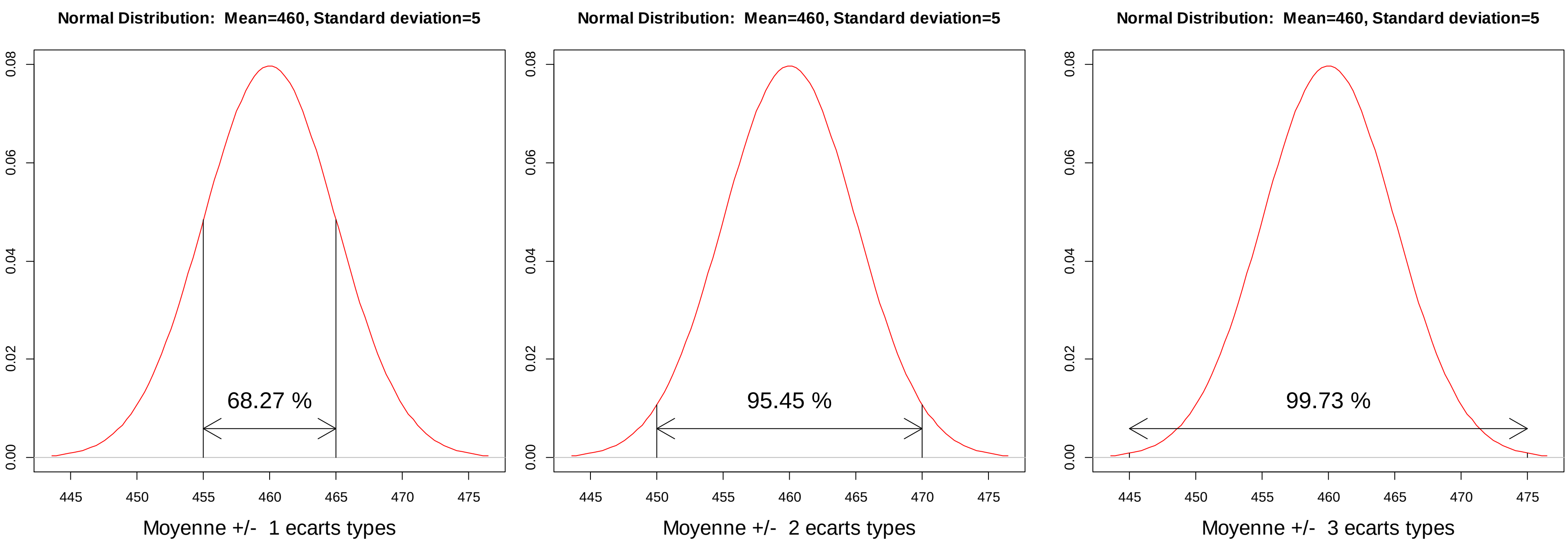
La loi de LG comme loi limite

Loi de la somme de *n* dés
(*source : Wikipedia*)



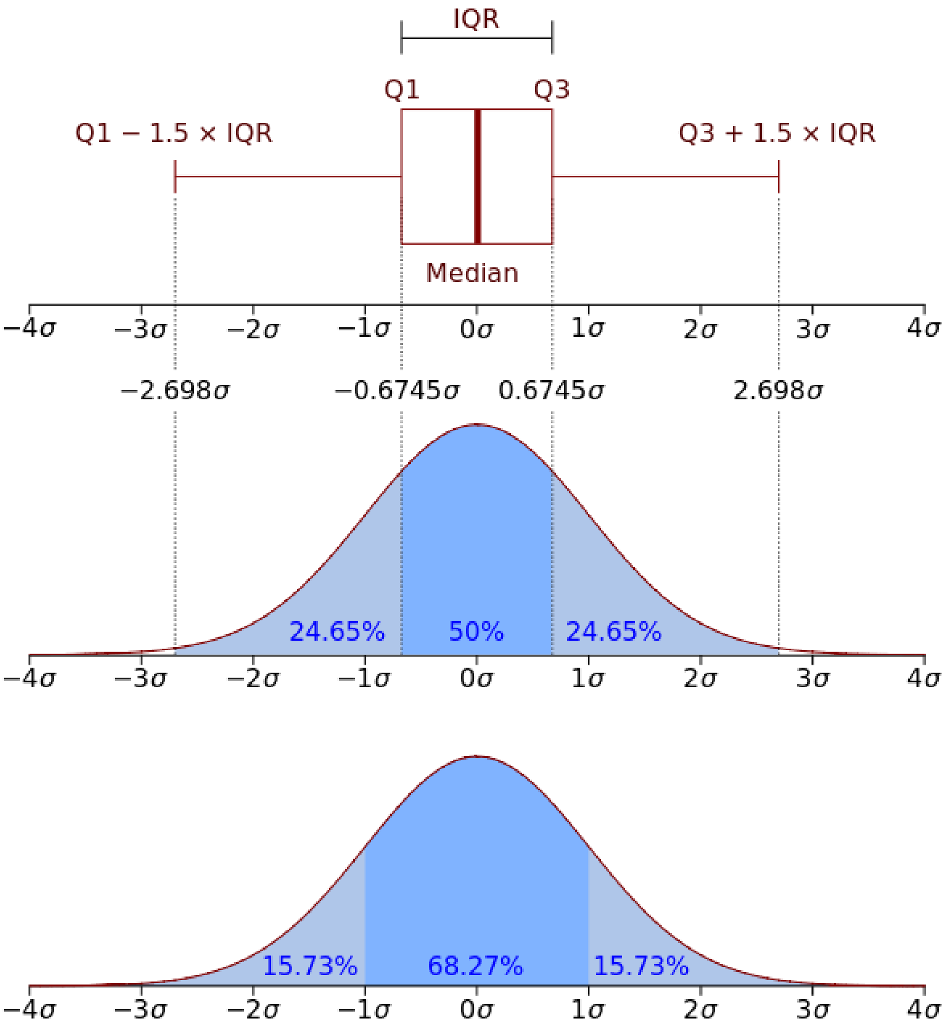
1. Notions de base – La loi normale

Quelques valeurs remarquables




```
s=3
plot(.x, dnorm(.x, mean=460, sd=5), xlab=paste("Moyenne +/- ",s,"ecarts types"), ylab="",
  main=paste("Normal Distribution:  Mean=460, Standard deviation=5"),type="l", cex.lab=1.5, col="red")
abline(h=0, col="gray")
segments(460-s*5,0,460-s*5,dnorm(460-s*5, mean=460, sd=5))
segments(460+s*5,0,460+s*5,dnorm(460+s*5, mean=460, sd=5))
arrows(460-s*5, 0.006, 460+s*5, 0.006, code=3)
text(460,0.012,paste(round(100*(2*pnorm(c(s), mean=0, sd=1, lower.tail=TRUE)-1),2),"%"),cex=1.7)
```

Boîte à moustache et loi normale



source : Wikipedia

La loi normale centrée réduite

$$X \sim \mathcal{N}(\mu, \sigma) \quad \Leftrightarrow \quad Z = \frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1)$$