Institut Supérieur du Sport et d'Education Physique, le KEF

Formation à distance

LFEP

Statistique

Semestre 6

Guelmami Noeman

Plan

Intro : généralités

I La statistique dans le cursus de l'étudiant

C'est une méthode de raisonnement permettant d'interpréter des données (en s'appuyant sur les probabilités) en tenant compte de la variabilité.

II Introduction à la statistique

II 1 Pile ou face ?

Tirer à pile ou face. La probabilité de gagner est de 50 / 50 pour les deux joueurs A et B.

- Pièce = P = 1
- Pièce face = Pf = 0.5 Pièce pile = Pl = 0.5 $Pf = \frac{1}{2}$.

La pièce tombe la première fois sur face. On la lance à nouveau. Face encore.

$Pf = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$

Elle tombe 4 fois sur face. Il faudrait lancer la pièce un certain nombre de fois pour atteindre une probabilité Pf = 0.5 et Pp = 0.5

Estimation de la proportion de pile ou face joués.

II 2 Combien y a t'il de gauchers à l'université ?

Estimation de la proportion d'étudiants gauchers à entretenir et à relever pour avoir une idée du pourcentage de gaucher à Orsay.

Rang	D	G	
1	10	0	
2	9	1	
3	8	2	
4	10	0	
5	10	0	
6	10	0	
Total	57	3	D + G = 60

Tableau 1

D + G = nombre d'élèves total d'élèves par rang. Moyenne générale (connue) = 1 gaucher pour 10 (= 6 / 60). Mais ici, il y en a moins d'1 pour 10.

Pourquoi le chiffre 60 a t'il été choisit ? Pour que cet échantillon soit **représentatif**, il faut prendre au moins 60 sujets. En deçà de cette limite, on ne pourra prendre en compte les **fluctuations d'échantillonnage**. Pour les maîtriser, il faut prendre la plus grande population possible.

En général, les méthodes les plus fréquemment utilisées pour construire un échantillonnage représentatif sont :

1. **L'échantillon au hasard** : dans ce cas, on tire << n >> sujets (ou classe de sujets) au hasard dans la population parente.
2. **L'échantillon stratifié** : on retient toutes les classes de sujets composant la population parente (CSP, sexe, rural / urbain)
3. **L'échantillon stratifié pondéré** : Les strates sont retenues proportionnellement à leur représentation à l'intérieur de la population parente. (ex. des gauchers, on prend un échantillon d'élèves dans les STAPS, les maths, la physique... en fonction du nombre total d'étudiants dans chacune de ces disciplines (car les étudiants dans les différents UFR sont plus ou moins nombreux)

chances de se tromper, cependant, on constate parfois que des statistiques sans pourcentage ou marge d'erreur ne veulent rien dire.

Tests statistiques : la corrélation.

II 3 La fourchette ?

III Statistiques et activités physiques et sportives : quelques exemples

III a L'évaluation : les tests de valeur physique

TEST de Cooper (il varie avec les époques)

III b L'évaluation : les barèmes

Il existe les barèmes généraux, mais on peut les adapter à la population à laquelle on a affaire.

III c L'évaluation : l'observation des élèves

Qualitatif : Appréciation sur les élèves.

Quantitatif : On fait le point en début d'apprentissage et l'on donne une note ou lettre ayant une **fonction diagnostique**. En cours d'apprentissage, on fait une autre évaluation; celle-ci a une **fonction formative**. Et en fin d'apprentissage, la dernière évaluation aura une **fonction sommative**.

Ces évaluations ont pour but de ne pas plaquer des connaissances toutes faites et pas nécessairement faites pour les élèves, mais plutôt d'adapter la connaissance à leur niveau.

III.d Relation entre performance : la corrélation

Trouver des relations simples entre un grand nombre d'élèves dans différents types de spécialités.

Objectif :On veut travailler une qualité physique particulière (ex. : la force explosive)

L'expression de cette qualité physique se retrouve dans plusieurs activités physiques et celles-ci ont beaucoup de relations entre elles au niveau de la variable "force explosive". Ou bien, on veut travailler un ensemble de qualités physiques (force, souplesse, endurance, etc.) et pour cela, on prend des activités physiques très variées.

	Hauteur (en m)	Poids	Zh	Zp	Zh x Zp
1	1	5	- 0.56	- 1.57	0.88
2	1.10	5.50	- 0.56	- 0.26	0.14
3	1.10	6	- 0.56	1.05	0.59
4	0.90	5.50	- 1.7	- 0.26	0.44
5	1.15	6	1.1	1.05	1.16
S	0.09	0.38			
Moyenne	1.05	0.90 5.6		3.22 S ?	

Tableau 2

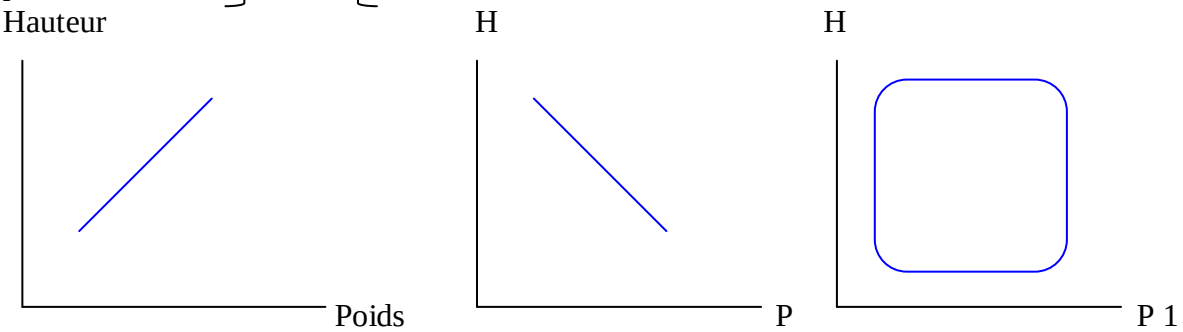
S est l'écart type.

La **corrélation (r)** entre les deux séries de performances est à 0.66, donc r = 0.66 : c'est la valeur de ce test. Cette corrélation est positive, elle signifie que plus je saute haut dans l'exercice du saut en hauteur, plus je lance loin dans l'exercice du lancer. Si la corrélation avait été négative, cela aurait signifié que plus je saute haut, moins je lance loin. La corrélation parfaite est égale à 1. La corrélation à 0.66 est assez forte.

Pour calculer la corrélation, c'est à dire la note centrée réduite, (Z), on utilise la formule suivante : z hauteur = zh et z poids = zp.

$$z = \frac{x-M}{S} \qquad z = \frac{3x\ 3y}{N}$$

Il existe trois types de corrélations : la corrélation positive, négative et nulle, et elles plafonnent entre] - 1 ; 1 [



- Corrélation positive = 1 - Corrélation négative = -1 - Corrélation illisible = 0

Fig. 1

Ex. pour tracer la droite de la corrélation. On reporte les performances dans un tableau, puis sur deux axes.

Tableau 3

Elèves	Saut haut = h	Saut long = l
A	1.5	5
B	1.7	7
C	1.4	4
D	1.8	8

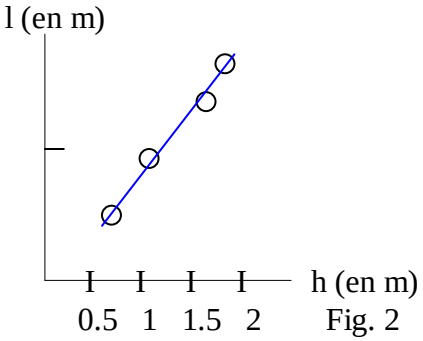


Fig. 2

Ici, la droite est croissante et positive, elle fait un angle à environ 45° avec les abscisses, elle montre donc une forte corrélation (proche de 1).

Ex. de courbe en cloche renversée de l'écart type :

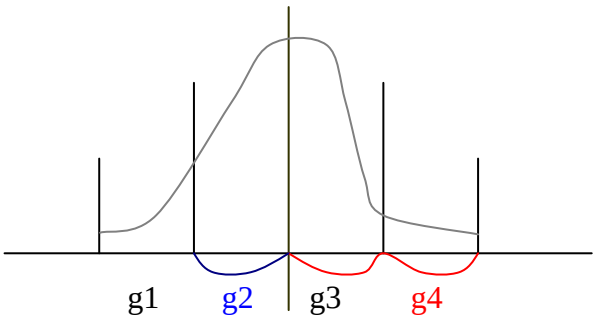


Fig. 3

Le groupe 4 (g4) est à écarts types (S). Le groupe 2 n'est qu'à un S.

La forme de cloche renversée est la **distribution normale** à partir de laquelle on peut faire parler les statistiques.

IV Concepts élémentaires en Statistiques

IV a Les différents types de recherches

Il y a deux types de recherches :

La **recherche corrélationnelle** :
Les variables ne sont pas **influencées** (en tout cas, volontairement) puisqu'il est seulement question de les observer.

La **recherche expérimentale** :
Ici, les variables sont manipulées et c'est l'effet que cette manipulation produit sur d'autres variables qui est examiné.
Ex. : On manipule le "climat perçu" :
Deux groupes. L'un travaille sans pression, dans la tranquillité. L'autre groupe est soumis au stress, on travaille en présence d'une caméra, on le note sans cesse... A partir de cela, on examine comment évoluent les notes des deux groupes.

IV b Les différents types de statistiques

Il y en a deux types :

Les **statistiques descriptives** :
Elle consiste à décrire les variables telles que la médiane, la moyenne, l'écart type, la variance, le mode... Elles ont une représentation graphique (un histogramme)

Les **statistiques comparatives ou statistiques inférentielles** :
On essaie d'interpréter un résultât. On passe donc par l'utilisation de tests statistiques. (le CHI2 ou χ 2. Le << t >> de Student – l'analyse de variances – les corrélations – la régression – l'analyse en composantes principales – l'analyse factorielle des correspondances, etc...

IV c Variables indépendantes et variables dépendantes

Elles sont ce qu'on appelle des mesures.

La **variable dépendante** (V.D) est celle qui l'on enregistre et que l'on mesure. Elle dépend de l'autre variable.

La **variable indépendante** (V.I) est celle que l'on manipule. (?)
Si on cherche à prouver que les longs cheveux se retrouvent plus souvent chez les femmes. Pour cela, on prend deux variables : la longueur des cheveux (V.D), et le sexe (V.I). On constate que du fait d'être une femme ou un homme, on a les cheveux plus longs.
Erreur : Du fait d'avoir les cheveux longs (V.I), on est une femme ou un homme (V.D)
Il s'agit de diminuer les variables indépendantes pour avoir un maximum de variables dépendantes, c'est à dire contrôler au maximum son expérience.

IV d Variables continues et variables discontinues

Variables continues (V.C) : c'est une variable telle qu'entre deux variables quelconques, il est

Variables discontinues ou discrètes (V.d) : C'est une variable non continue qui varie non pas de façon progressive, mais en effectuant des sauts sur un ensemble lui- même discrêt, en passant d'une valeur ponctuelle à une autre valeur ponctuelle arrêtée. On ne peut intercaler de valeur intermédiaire.

V - Les niveaux de mesure

Les variables diffèrent par la qualité de leur mesure, c'est à dire par la quantité d'informations qu'elles délivrent. Chaque degré de mesure est déterminé par la quantité d'informations qu'elles nous donnent.

V A - Les variables quantitatives ou nominales

Classement ou échelles de mesures :

Les **variables nominales** (V.N) de type qualitatif (sexe, CSP,...)

Les **variables ordinales ou de rang** (V.O) permettent de donner un ordre, de hiérarchiser les items mesurés (comme le classement d'une course par ex. : 1° 2° ...)

Les **variables d'intervalles** (V.i) permettent non seulement de donner un ordre mais de quantifier et de comparer l'ampleur de différence entre les items mesurés (en général, toutes les échelles de mesures) Ex. : de (1 à 3) (4 à 6) (7 à 9) ...

Les **variables de ratio** ou **échelles de rapport** sont généralement considérées comme des V.i.: elles en possèdent les mêmes qualités, mais en plus, possèdent un zéro absolu (température, temps, espace...).
N.B : Les procédures d'analyse statistiques ne distinguent pas les deux derniers types d'intervalles.

S'il est fait de façon quelconque, on se retrouve devant un certain nombre de résultats, mais il faut y mettre de l'ordre pour analyser, interpréter ces données.
On les ordonnent généralement du plus petit au plus grand. Pour les nombres répétitifs, on donne le nombre d'apparitions = l'**effectif** de chaque caractéristique de la variable observée.
Quand on trace la **médiane** des effectifs, on indique la séparation de l'effectif en deux, on sépare l'effectif avec 50% de chaque côté.
Exemple :

Performance de l'élève en mètres	P (m)	1.05	1.10
Note	n	12	13.5
Effectif (Nb d'élèves ayant la même note)	e	7	3
Effectif cumulé	e.c	7	10

Tableau 4

Chapitre I Variables quantitatives
Le recueil des données et leur représentation

Pour décrire les séries statistiques, il faut observer trois étapes :

1 – Rendre compte de la **forme de la série** (cloche, asymétrique...)

[1] - Langouet & Porlier. Mesure et statistiques en milieu éducatif. Ed° ESF
- ex. pour les variables continues : performance d'un concours de saut en hauteur : Parlebas & Cyffers. 1992, p. 4-5

2 – Les **valeurs centrales** de cette série statistique sont **le mode, la médiane et la moyenne.**

3 – Les **indices de dispersion** de ce mode sont **l'étendue, la variance et l'écart type**.

II Les représentations (couramment utilisées.)

■ Série 1

II 1 Diagramme en bâton

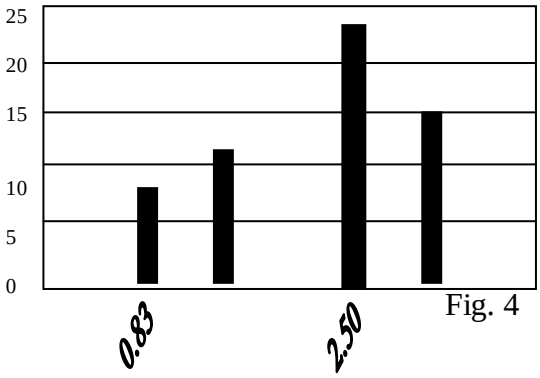


Fig. 4

Performance (en m)	0.83	1.60	2.50	3.20
Effectif	8	12	24	15
Effectif cumulé	8	20	44	59

Tableau 5

Cette représentation graphique du tableau 5 (que l'on peut exécuter avec Excel) donne la figure suivante (Fig. 4). On l'utilise dans le cas des **variables discrètes** (discontinues)
La série est représentée de manière discontinue (entre chaque bâton, il y a des espaces).

II 2 L'histogramme

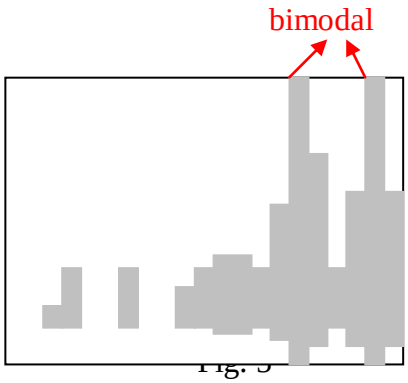
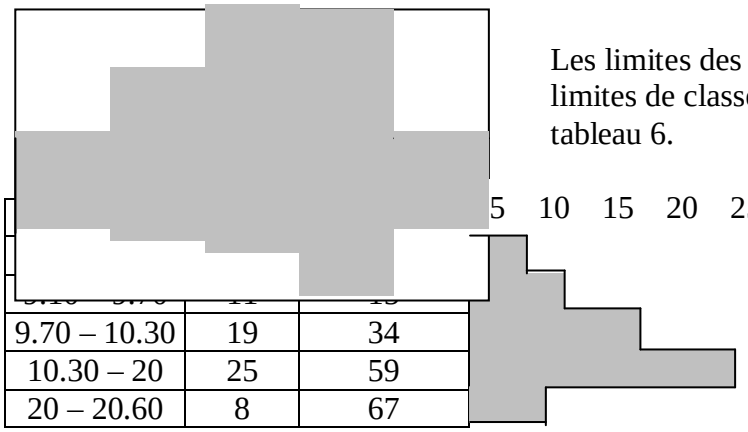


Fig. 5

Performance sur 60 m plat

sec dix°	.0	.3	.5	.9	total
8					0
9	1			4	5
10	6	4	4		10
11	2	10	3	0	15
12	0	7	5	6	18
13	0	0	1	4	5

Tableau 6

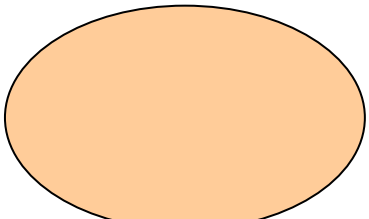


Les limites des rectangles sont les limites de classes trouvées dans le tableau 6.

9.70 – 10.30	19	34
10.30 – 20	25	59
20 – 20.60	8	67

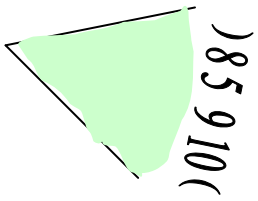
Tableau 7

Camembert en diagramme circulaire.



Le problème est que l'on ne peut voir si la distribution est symétrique. On ne peut pas voir la moyenne non plus.

Fig. 7



II 3 Les effectifs cumulés

Ce type de représentation des effectifs cumulés a une utilité lors de plusieurs calculs (ex. : calcul de la médiane)
Le quantilage = "Découpage en tranches" de la variable. (100 tranches = centilage)
Ex. : La variable est la distribution des tailles suivantes :

Tailles	158	159	160	161	162	163	164	165	166	167	168	169
Effectif	1	0	0	1	0	1	1	2	0	0	0	1
Effectif cumulé	1	1	1	2	2	3	4	6	6	6	6	7
Tailles	170	171	172	173	174	175	176	177	178	179	180	181
Effectif	0	2	2	3	0	3	1	2	2	2	1	0
Effectif cumulé	7	9	11	14	14	17	18	20	22	24	25	25
Tailles	182	183	184	185	186	187	188	189	190	191	192	
Effectif	1	2	0	0	1	0	0	0	0	0	1	
Effectif cumulé	26	28	28	28	29	29	29	29	29	29	30	

Tableau 8

Tailles inf. ou égales à	162	167	172	177	182	187	189
Effectif cumulé	2	6	11	20	26	29	30

Tableau 9

On effectue un rangement par classes de cette variable :

Valeurs extrêmes	Valeurs centrales	effectif	Effectifs cumulés	fréquence	Fréquences cumulées
158-162	160	2	2	0.1	0.1
163-167	165	4	6	0.1	0.2
168-172	170	5	11	0.2	0.4
173-177	175	9	20	0.3	0.7
178-182	180	6	26	0.2	0.9
183-187	185	3	29	0.1	1
188-192	190	1	30	0.03	1

	N = 30		Total = 1.00	

Tableau 10

fréquence FQ = effectif / effectif total

La fréquence est le pourcentage. Ex. : 0.07 = 7 % des sujets ont une taille inférieure à 162.
Cela donne la représentation suivante :

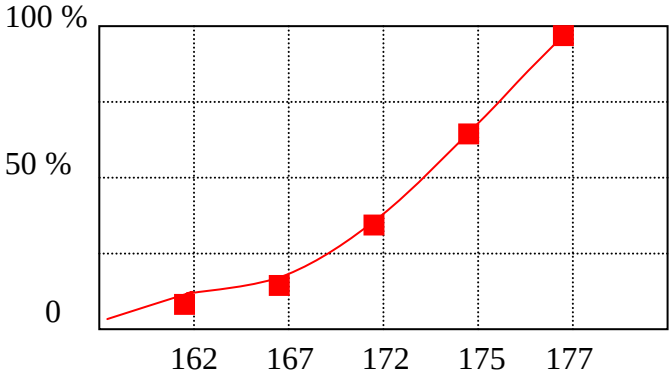


Fig. 8

II 3.1 Les effectifs cumulés croissants

La représentation en cloche est caractéristique d'une distribution normale.

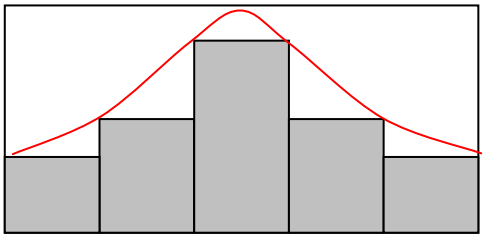
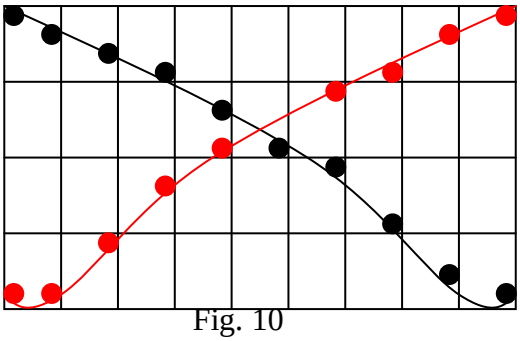


Fig. 9

II 3.2 Les effectifs cumulés décroissants

On inverse le tableau. On trace les deux courbes sur un même diagramme, à l'intersection, les valeurs qui séparent l'effectif à 50 / 50, cela nous donne les effectifs cumulés croissants et décroissants. On a à présent la possibilité de calculer la médiane.



Fréquence cumulées croissante

Fréquence cumulées décroissante

Fig. 10

III Les valeurs centrales ou d'opposition

III 1 La moyenne

Parler de moyenne, c'est sous-entendre que l'on parle de moyenne arithmétique.
La moyenne représentée par M est un indice de tendance centrale qui, pour être normale, situe sur l'axe des abscisses la valeur pour laquelle il y a autant de valeurs de la variable qui lui sont supérieures d'un côté, et autant de valeurs de la variable qui lui sont inférieures.
Moyenne des tailles : 170 car 11 sujets < 170
19 sujets > 170
Valeur de la variable = 10,30

$$M = \frac{\overset{\text{Sigma}}{\overline{\quad}} \Sigma\, ni.xi}{N}$$

ni = nombre de sujet pour une des valeurs de la variable.
xi = 175 (taille)
N = effectif (nombre total de sujets)

III 1.1 Calcul de la moyenne après groupement par classe

(Cf. tableau 8)
Calcul : 158 + 161 + 163 + 164 + 165 + 165...etc. = 5222 / 30 (nombre d'élèves)
= 174.07
□ ni . xi = (158 x 1) + (159 x 0) + (160 x 0) + (161 x 1) + (162 x 0) + (163 x 1) ...
C'est la moyenne de notre échantillon. Mais elle n'est qu'une estimation de la moyenne de la population parente. C'est pourquoi on note la moyenne M = m ou x. Face a un grand échantillon, on procède à un regroupement par classes. On attribue à la variable la valeur centrale de la classe. On prend les valeurs extrêmes.

Taille en cm Valeurs extrêmes	Valeur centrale xi	Effectif ni	Ni . xi
158-162	160	2	320
163-167	165	4	660
168-172	170	5	850
173-177	175	9	1575
178-182	180	6	1080
183-187	185	3	555
188-192	190	1	190
		□ ni = 30	□ ni . xi = 5420

Tableau 11

III 2 La médiane

La médiane = C'est la valeur de la variable telle que 50 % des effectifs lui sont supérieurs et 50 % inférieurs. (On divise son groupe en deux : j'ai 50 sujets, j'en ai 25 au-dessus, 25 en dessous)

III 2.A - Cas d'une variable discrète

Il n'y a généralement pas de calcul à faire. On classe les observations dans l'ordre croissant. La **médiane** de la variable correspond à l'élément qui se trouve au milieu. C'est elle qui me permet de déterminer le rang médian. $\frac{23}{2} = 11.5$ (le rang que prend ma variable dans les effectifs entre la 11ème et la 12ème classe = 11.5)

Nombre de frères et soeurs	effectif	Effectif cumulé
0	4	4
1 (Me = 1)	10	**14** (rang Me = 11.5)
2	5	19
3	4	23

Tableau 12

Si N est pair, dans ce cas, la médiane (notée Me) correspond à 50 % des effectifs. Elle occupe le rang :

$$\frac{N + 1}{2}$$

III 2.B - Cas d'une variable continue

III 2.B.1 - Sur les données brutes

Si N est pair, Me occupe le rang : $\frac{N + 1}{2}$

Si N est impaire, Me occupe le rang : $\frac{N}{2}$

22 observations, $\frac{23}{2} = 11.5$. Rang médian = entre 11ème et 12ème. Donc la taille médiane est entre 1.58 et 1.59 m (Me = 1.585 m. Si N est impair, la médiane prend la valeur de l'observation qui se situe au milieu de la variable : c'est la valeur centrale.

1.47	1.48	1.51	1.52	1.52	1.54	1.54	1.55
1.58	1.58	1.58	1.59	1.59	1.59	1.60	1.63
1.63	1.64	1.64	1.65	1.67	1.69		

Tableau 13

III 2.2.2 Sur les données classées

Exemple : distribution avec 2692 élèves. On procède à un regroupement par tailles.

Taille	effectif	Effectif cumulé	fréquence	Fréquence en %
1.45 à 1.50	514	514	0.2	19.1
1.50 à 1.55	1221	1735	0.5	45.4
1.60 à 1.65	259	1994	0.1	9.6
1.65 à 1.70	698	2692	0.3	25.9
Total	2692		1	100

Tableau 14

La médiane correspond au rang $\frac{N+1}{2} = \frac{2692+1}{2} = 1346.5$

La médiane est située entre la 1346 ° observation et la 1347 ° observation, c'est à dire que nous avons affaire à une classe dont les élèves mesurent entre 1m50 et 1m55.

1.50	514
Me	1346.5
1.55	1735

C'est une interpolation linéaire.

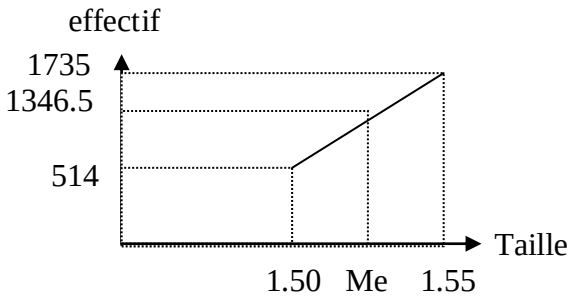


Fig. 11

Par le coefficient directeur de la droite :
1.50 ↔ 514
Me ↔ 1346.5
1.55 ↔ 1735

$\frac{Me - 1.50}{1.55 - 1.50} = \frac{1346.5 - 514}{1735 - 514}$ ⇨ Me = 1.534 m

III 3 Le mode

- **Le mode** = c'est la valeur de la variable la plus fréquente. Il y a plusieurs modes (Bimodal = deux valeurs de la variables ont le même mode. (ex. Fig. 5), plurimodal...)

- **La médiane** = Valeur de la variable qui partage une série statistique ordonnée en deux sous populations de même effectif.

- **La moyenne** = quotient de la somme des valeurs d'une série statistique par le nombre de ses valeurs.

- **Le mode** = Valeur de la variable qui correspond à l'effectif le plus élevé.

IV Les indices de dispersion

IV 1 L'étendue

C'est la différence entre l'observation la plus élevée et l'observation la plus faible de la série statistique. (ex. : 1.69 m – 1.47 m = 0.22
1.70 m – 1.45 m = 0.25. Mais ce n'est pas très précis ni très parlant)

IV 2 La variance

On évoque la dispersion des valeurs (observations) de la variable autour de la moyenne. Elle peut être plus ou moins grande.

Exemple : 2 profs notent des copies. Les profs A et B.

Copies / Profs	a	b	c	d	e	Etendue
A	09	11	06	13	16	10
B	08	10	03	15	19	16

Tableau 15

Données classées dans l'ordre : A = 6 9 11 13 16 : L'étendue entre les deux est de 10

B = 3 8 10 15 16 19 : L'étendue est de 16

MA = $\frac{\sum xi}{N} = \frac{55}{5} = 11$ MB = $\frac{\sum xi}{N} = \frac{55}{5} = 11$

Cela permet de construire un outil qui rend compte de chaque écart avec la moyenne.

Variance (V) = $\frac{\Sigma ni\,(xi - M)^2}{N} = \frac{\mathbf{\Sigma\, ni\,.\, xi^2}}{\mathbf{N}} \mathbf{- M^2}$

Où <<xi>> représente les valeurs que peut prendre la variable, <<ni>> le nombre de classes, <<M>> la moyenne, <<N>> le nombre d'observations (effectif)

$\sum (x-M)^2$ → A = 58 ; → B = 154

$\sum (x-M)^2$ = indice de dispersion

Le prof B a une notation plus dispersée que le prof A.

IV 2.1 Calcul pratique d'une variance

Variance du prof VA : 58 / 5 = 11.60 SA = V11.6 = 3.4
Variance du prof VB : 154 / 5 = 30.80 SB = V30.8 = 5.5

L'écart type est la racine carrée de la variance. $S = \sqrt{V}$

Loi normale : Mode = M = Me

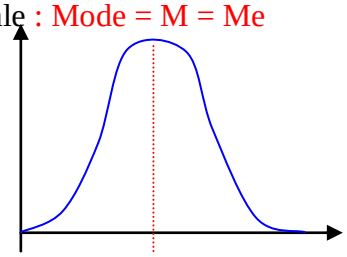


Fig. 12

IV 2.1 Calcul pratique de la variance

Copies / Profs	a	b	c	d	e	Etendue
A	09	11	06	13	16	10
B	08	10	03	15	19	16

Tableau 16

On schématise l'indice de dispersion du tableau 16:

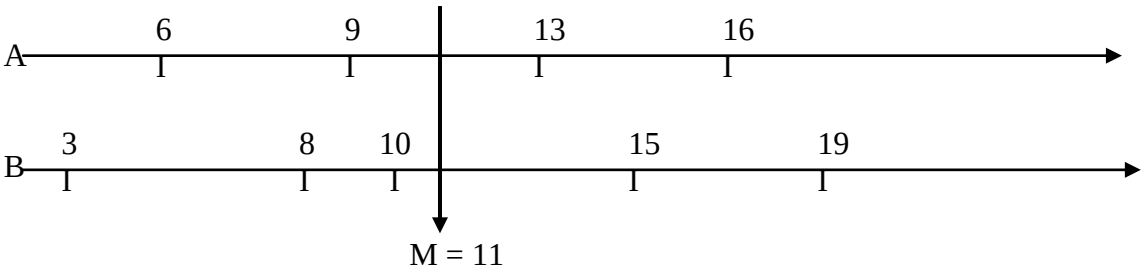


Fig. 13

L'indice de dispersion est égal à o car il y a autant de valeurs de chaque côté et à égales distances. Donc, ça n'indique pas beaucoup de dispersion, mais plutôt une symétrie. Si l'indice de dispersion était positif, cela voudrait dire qu'il y aurait plus de valeurs du côté positif, et s'il était négatif, cela voudrait dire le contraire.
Qu'y a t'il au dessus et au dessous de la moyenne ?

Calcul de l'indice de dispersion :

$\Sigma(x-M)^2$

Ex. A = $(6 – 11)^2 + (9 – 11)^2 + ... = 58$
B = $(3 – 11)^2 + (8 – 11)^2 + .. = 154$
Le chiffre est supérieur en B, il a donc plus de dispersion que chez le A.

Mais cette formule est valable si on calcule valeur par valeur. Pour plus de données, il faut procéder à des regroupements. Dans ce cas, le meilleur estimateur de la variable devient la formule :

$$V = \frac{\Sigma ni\,(xi – M)^2}{N – 1}$$

<<N – 1>> est le degré de liberté.

Plus le nombre de sujets est grand, plus l'estimation est précise.

IV 2.2 La notion de degré de liberté

C'est la contrainte imposée à notre test statistique.

C'est le nombre de valeurs qu'il faut connaître pour deviner (à coup sûr) la suite de la distribution.

Ex. des notes du prof B. Si je connais seulement la somme (55) de ces variables, ou bien la somme et la première variable, je ne peux deviner la suite. Mais si je connais les 4 premières valeurs et le total, je devine à coup sûr la $5^{ème}$ valeur. Donc la $5^{ème}$ valeur dépend des 4 premières valeurs prises par la variable. Donc la contrainte que j'impose à ma variable N – 1 (5 – 1 = 4) est le degré de liberté.

IV.3 L'écart type

Données groupées. $S = V\frac{\Sigma ni\,(xi – M)^2}{N – 1} = V\frac{\Sigma ni.xi^2 - \frac{(\Sigma ni.xi)^2}{N}}{N - 1}$

V La forme de distribution

Il faut ensuite caractériser la forme de la distribution.

Forme de gauss : Fig. 14

Formes binomiales

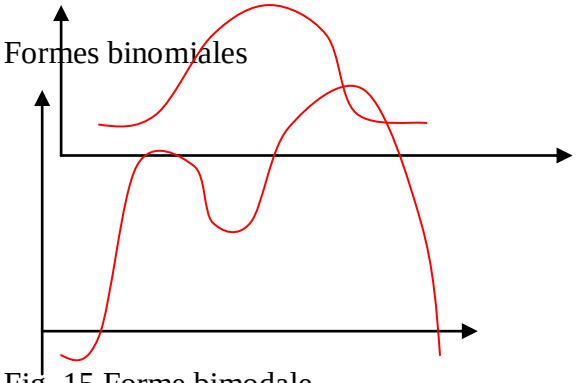
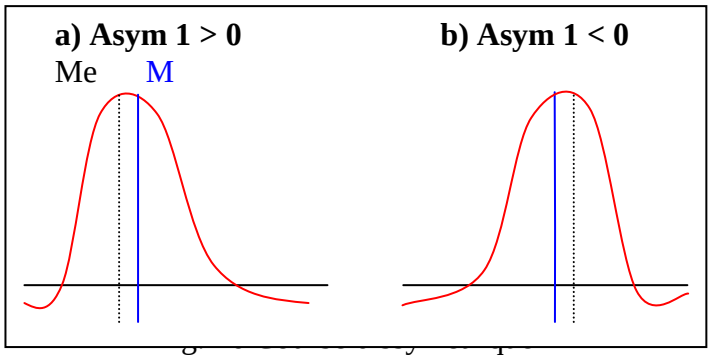


Fig. 15 Forme bimodale



"Asym. 1 > 0" signifie que Me < M. La classe b a de meilleurs résultats que la classe a car l'asymétrie de sa courbe est négative (Me > M). Si cet indice est positif, la moyenne est au dessus de la médiane.

Indice d'asymétrie

$$Ass1 = \frac{M - Me}{S}$$

V.1 Pourquoi la <<LOI NORMALE>> est importante ?

Elle est importante car elle permet de faire une bonne estimation de la fonction qui relie la valeur que prend la variable et le test statistique utilisé. De nombreux tests statistiques ont une forme ou des fonctions dérivées de cette loi normale. Elle (courbe en cloche ou de gauss) est caractérisée par deux indices :

1 / l'indice de tendance centrale = M
2 / L'écart type = S
Dans une loi normale, peu importe la valeur de la variable.

Loi normale centrée réduite :

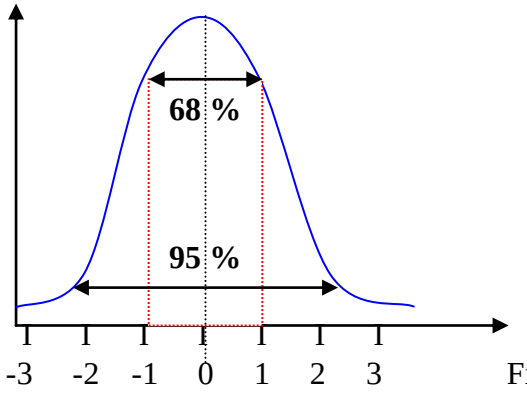


Fig. 17

68 % (de la population), c'est à dire un écart type. Tandis que 95 %, cela correspond à 2 S.
On retrouve souvent les valeurs suivantes liées à ces pourcentages dans les tests statistiques :
50 % : plus ou moins 0.67 S
68 % : plus ou moins 1 S
95 % : plus ou moins 1.96 S
99 % : plus ou moins 2.58 S, c'est à dire qu'on a 1 % de chances de se tromper.
99.8 % : plus ou moins 3.09 S

Quand on cherche un seuil significatif, on tente de se rapprocher des trois dernières valeurs.

Soit un tableau de test de détente verticale sur 205 sujets :

perf. en cm	eff	eff cum	perf. en cm	eff	eff cum	perf. en cm	eff	eff cum
32	1	1	53	4	63	74	2	191
33	0	1	54	4	67	75	3	194
38	11	12	...	...	...	...	...	...
41	29	41						
47	...	...						
Etc.								

Tableau 17

Si la distribution se distribuait normalement, elle ressemblerait à la figure 14, donc on pourrait dire : au delà de telle valeur, il y a tel pourcentage de la population. En deçà, entre telle et telle valeur, il y a tant... S'il était à un écart type de la moyenne, on pourrait dire que la performance de tel est bonne... On peut dresser des tables de correspondances connues d'après la loi normale (cf. table de Faverge)

VI Le quantilage

C'est un découpage de la population en tranches ou classes égales, elles ont toutes le même effectif et sont appelées les **inter quantiles**.

Noté : q
Ex. : on a 9 valeurs, on fait 10 classes :

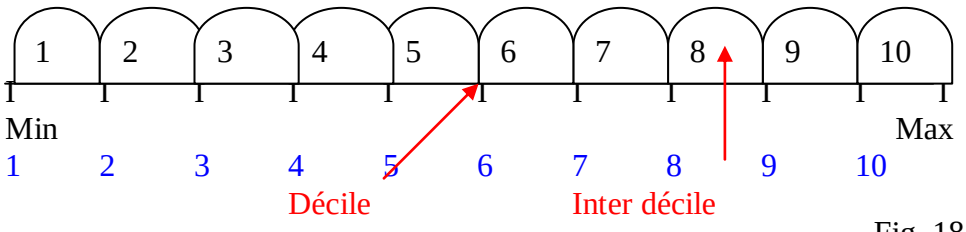


Fig. 18

Ex. pour 4 classes :

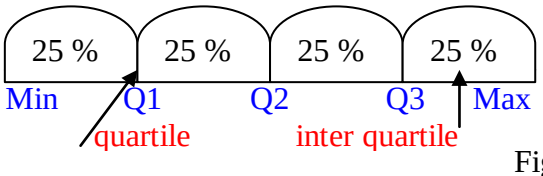


Fig. 19

Q2 = Me car il départage la variable en deux parties égales.

Il en est de même pour 99 centiles + les valeurs extrêmes des variables et les inter centiles.

A présent, comment déterminer sur la distribution de N observations le rang k qui correspond à un quantile dépassé par un rapport d'observation ? Pour déterminer la valeur que prend le décile, il existe trois règles :

1/ Appeler k le nombre entier qui est immédiatement supérieur à $\frac{qN + 1}{100}$

$$k = \frac{qN + 1}{100}$$

k est le rang que prend la variable suivante.

Ex. : N = 66. On veut déterminer le 3ème quartile.

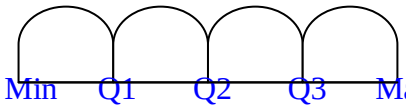


Fig. 20

= 75 % de la population

Dans ce cas, k = 50 (car c'est la 50ème observation) et parce que l'on calcule :

$$qN = \frac{75 \times 66}{100} = 49.5$$

On a 66 observations qui vont de 1 à 66. On les découpe en 4 tranches.

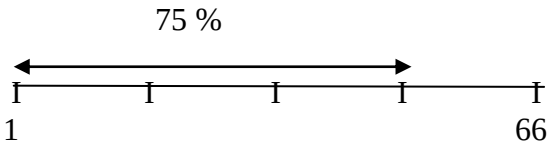


Fig. 21

Quelle est la valeur du 3ème quantile qui fait que j'ai 75 % de la population à ce niveau ? Il faut faire une règle de trois pour déterminer à quel rang va se trouver la valeur du quantile.
75 % de 66 = 49.50

Exemple : N = 320 et k = 33 1er décile ? $N = \frac{10 \times 320}{100} = 32$

2/ Lorsqu'une séparation tombe dans une classe où il y a plusieurs observations (cf. tableau 17) on dépasse la cotation si on les prend toutes. Mais si on ne les prend pas toutes, on accepte d'être en dessous du cotât. C'est celui fait le quantilage, qui décide de les prendre toutes ou non.

On veut répartir les sujets en 20 classes, donc on répartit les sujets tous les 5 % ($\frac{205}{20}$ = 10.25) c'est à dire tous les 10.25 sujets. On choisira de prendre 11 (?) sujets à un moment, et 12 (?) le moment suivant (10 et 11). C'est une côte mal taillée mais répartie approximativement.

- 1er quantile (application du point 1°) = 5 x 205 / 100 = 10.25. On choisit de prendre 11 sujets. On voit sur le tableau 17, dans la colonne des effectifs cumulés (11ème sujet) que 2 d'entre eux ont fait 38 cm.

- $2^{ème}$ quantile : 10 x 205 / 100 = 20.5 (21) ...etc

Proposition d'un barème à partir d'une technique de quantilage :

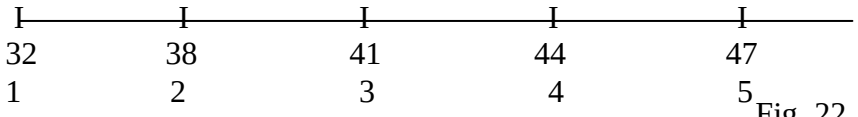


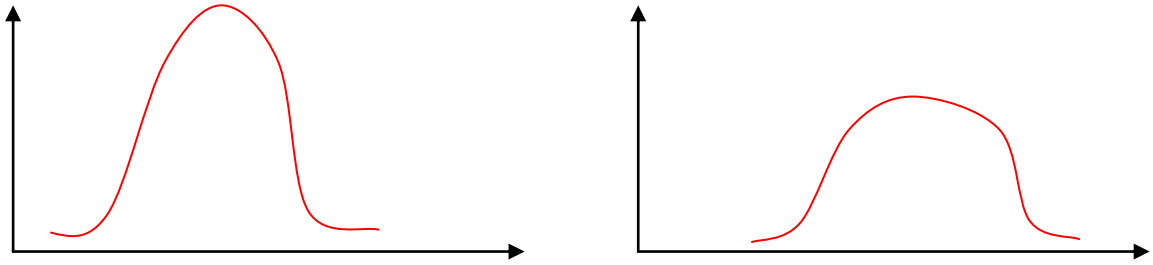
Fig. 22

Barre	De : (>)	A : (S)	Fréquence	Fréquence cum	Pourcentage
0		32			
1	32	38			
2	38	41			
3	41	44			
4	44	47			

Tableau : 18

VI . 2 Calcul par interpolation linéaire

Correction des exercices
1 / L'écart type sert à caractériser la distribution



Forme en pique — Forme en coupe

Fig. 23

Valeurs : 4 / 5 / **6** / 7 / 8
moyenne

$$\text{L'écart type} = V\frac{\square\ (xi - M)^2}{N} = V\frac{(4-6)^2 + (5-6)^2 + (6-6)^2 + (7-8)^2 + (8-6)^2}{5}$$

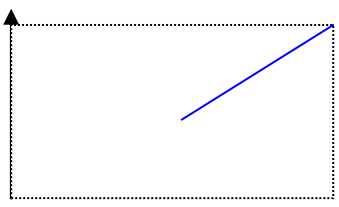
$$= V\frac{4 + 1 + 0 + 1 + 4}{5} = V\frac{10}{5} = V\,2 = 1{,}41$$

2 / Quand on veut décrire une variable, on se demande :
Quelle est la nature de la variable ? (maximale, ordinale, d'intervalle ?)
Pour une série quantitative, on donne :

- La moyenne, le mode, la médiane. Ce sont les indices de tendance centrale.
- L'étendue, la variance (pas forcément nécessaires) et l'écart type.
- On regarde si elle est symétrique.

3 / Voir l'exemple dans le cours
4 / Quand on se trouve face à un calcul par interpolation linéaire :
(exemple : déjà vu pour la moyenne)

rang + z



Nous avions cherché
sur une droite la valeur

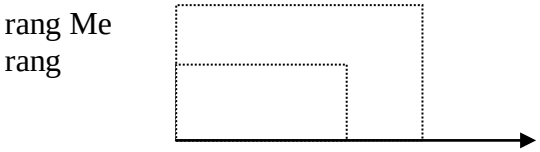


Fig. 24

On reprend le même raisonnement en essayant d'abord de dresser un tableau des effectifs :

xi	3.40 3.60	3.60 3.80	3.80 4	4 4.20	4.20 4.40	4.40 4.60	4.60 4.80	4.80 5	5 5.20
ni	3	10	12	15	25	23	17	8	4
eff cum	3	13	**25**	**40**	65	88	106	114	117

Tableau : 19

On divise la variable en 4 classes :

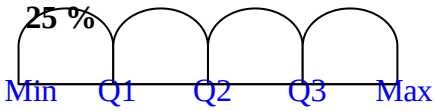


Fig. 25

On cherche le rang de l'observation qui correspond à la valeur de la variable.

$$\frac{25}{100} \times 117 = 29.25$$

rang Q1 = 30ème observation. Celle-ci est entre 4 m et 4.20 m.

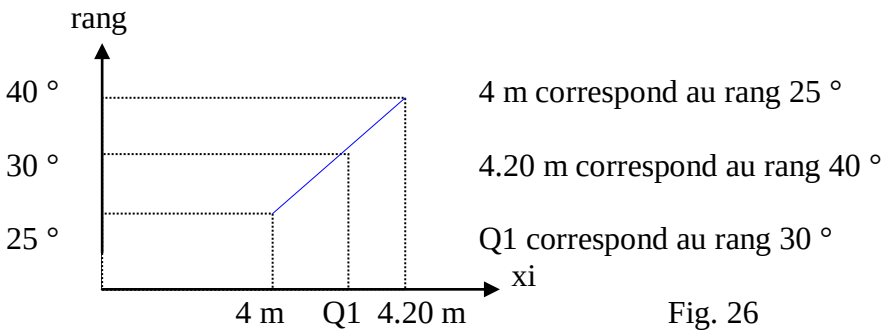


Fig. 26

$\frac{Q1 - 4.00}{4.20 - 4.00} = \frac{30 - 25}{40 - 25}$ On fait correspondre les égalités par la droite

$\frac{Q1 - 4.00}{0.20} = \frac{1}{3}$ $\frac{Q1}{0.20} = \frac{1}{3} + 4.00$ $Q1 = (\frac{1}{3} \times 0.20) + 4.00$

Q1 = 4.07 m On arrive à la valeur du premier écart type.

On peut attribuer les notes ainsi :

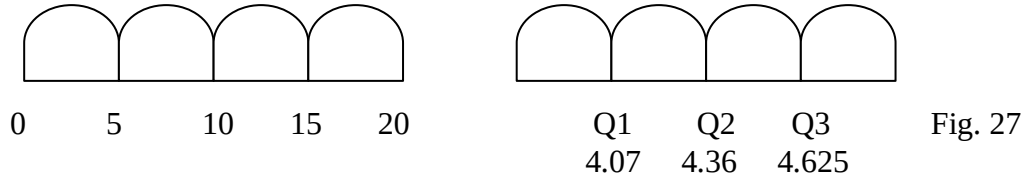


Fig. 27

VII – Les notes centrées réduites
VII 1-A Un exercice pratique

Ces valeurs sont très utilisées car elles sont normalisées. On transforme n'importe quelle valeur de n'importe quelle distribution dont on connaît la moyenne.
exemple :

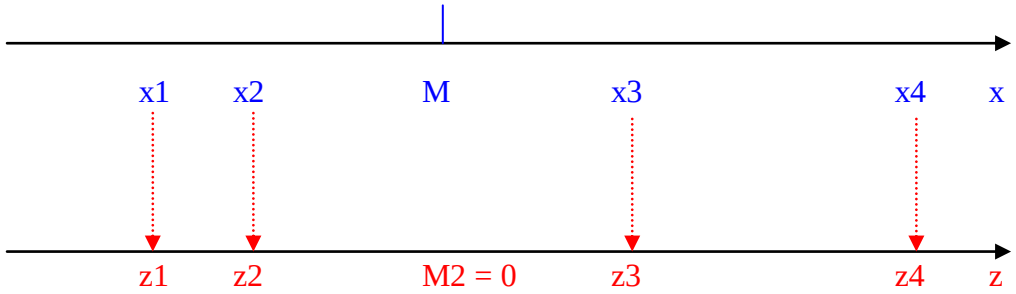


Fig. 28

Les notes x sont transformées terme à terme.
La moyenne des "z" est toujours égale à 0, et l'écart type égal à 1
/ z / = valeur absolue de z. Même si c'est négatif, je n'en tient pas compte.
La distribution x a un écart type. Pour la transformer en distribution z, j'applique la formule :

$$/ z / = \frac{x1 - M /}{S}$$

La moyenne des z = 0 et l'écart type de la distribution des z = 1
La distribution des x n'est pas forcément normale. On la transforme en distribution qui est toujours normale. On sait alors qu'à tel écart type, on a tant de pourcentage de la population.

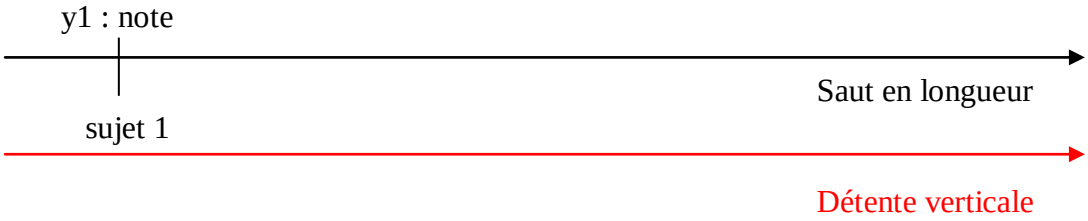


Fig. 29

Attention : on ne peut comparer un temps à une longueur car ce n'est pas la même unité. C'est pourquoi il faut transformer x1 en z1.

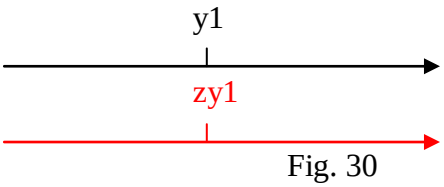


Fig. 30

Exemple pratique : on a relevé la performance d'un grand nombre d'élèves pour 3 épreuves d'athlétisme : le poids, le saut en hauteur et le 60 m plat. Nous disposons des indices suivant les 3 distributions.

	P	H	60 m
Moyenne	6.20 m	1.15 m	9.02 sec

Ecart type	0.40 m	0.05 m	9.03 sec

Tableau 20

2 élèves a et b ont obtenu les performances suivantes :

	P	H	60 m
a	6.50 m	1.10 m	9.4 sec
b	5.20 m	1.25 m	8.7 sec

Tableau 21

Performances tout à fait théoriques, choisies pour réaliser cet exercice.

- 1° Dans quelle épreuve l'élève a est-il le meilleur (en comparant avec les performances des autres élèves) ?
- 2° Quelle performance au poids faudrait-il que l'élève a réalise pour égaler la meilleure performance de b ?

On dresse un tableau avec les notes centrées réduites :

	Poids	z poids	hauteur	z haut	60 m	z 60 m
a	6.50 m	0.75	1.10 m	- 1	9.45	-0.66
b	5.20 m	- 2.5	1.25 m	+ 2	8.75	+ 1.66

Tableau : 22

Pour le z poids du premier : $\frac{6.50\text{ m} - 6.20\text{ m}}{0.40} = 0.75\text{ m}$

Poids du second sujet : $z = \frac{5.20\text{ m} - 6.20\text{ m}}{0.40\text{ m}} = -2.5\text{ m}$

Si on fait le même calcul sur les trois performances, on s'aperçoit que comparativement, le sujet a a sa meilleure performance en poids, sa plus mauvaise en hauteur et sa performance intermédiaire au 60 m. Le sujet b est meilleur en hauteur, mauvais en poids, intermédiaire en 60 m.

1° : le poids
Plus on lance loin, meilleure est la note. Mais plus le temps est grand, plus la note est mauvaise. Il faut donc inverser quand il s'agit de temps. On fait donc :

$$\textbf{z temps} = \frac{\textbf{M} - \textbf{x}}{\textbf{S}}$$

2° L'élève b est meilleur au saut en hauteur. Pour za poids = zb hauteur, on calcule :

$\frac{\text{xa poids} - \text{M poids}}{\text{S poids}} = \frac{1.25\text{ m} - 1.15}{0.05}$ $\qquad$ $\frac{\text{xa poids} - 6.20}{0.40} = \frac{1.25\text{ m} - 1.15}{0.05}$

$\frac{\text{xa poids} - 6.20}{0.40} = 2$ $\qquad$ xa poids = (2 x 0.40) + 6.20 = 7

La performance au poids de a doit être égale à 7 m

Autre exemple (correction de la question 4)
Echantillon a Ma = 4.40 m S = 0.40 m
Echantillon b Mb = 4.55 m S = 0.42 m

Combien doit sauter l'élève A pour être au même classement que l'élève B qui saute 4.70 m ?

z de la performance = 4.70 m $\frac{xa - 4.40}{0.40} = \frac{4.70 - 4.55}{0.42}$

xa = 4.54 m
Pour être au même classement, il faudrait qu'il saute 4.54 m.

Tous les tests statistiques sont basés sur les valeurs qui viennent du z ou qui en sont dérivées, c'est à dire :

$$/ z / = / \frac{x1 - M}{S} /$$

VIII – Compléments

VIII 1 – Moyennes

Jusque là, on a utilisé la moyenne arithmétique, mais il en existe d'autres...

VIII 1 A - Moyenne Harmonique

Elle sert par exemple dans le calcul d'une vitesse moyenne.
Exemple : un coureur fait un 1500 m, les 3 derniers tours (derniers 1200 m), il parcourt 400 m à vitesse moyenne, le 2ème tour à vitesse moyenne V2 et le 3ème = V3. Pour calculer sa vitesse moyenne, on utilise la moyenne harmonique notée "h". (cf. Parlebas et Cyffers...)

$$\frac{1}{h} = \frac{1}{n}\left(\frac{1}{x1} + \frac{1}{x2} + ... + \frac{1}{xn}\right)$$

$$h = \frac{n}{\sum_{i=1}^{n} \frac{1}{xi}}.$$

VIII 1 B – Moyenne géométrique

Sert à calculer une **accélération moyenne**.
Exemple : sur 100 m, le coureur part à la vitesse V1, il accélère. Entre le temps T1 et le temps T2, son accélération moyenne est :

$$g = \sqrt{x1 \times x2 \times x3 ...} = \sqrt[n]{\prod_{i=1}^{n} xi}$$

Moyenne notée "g"
"**n**" : Enième = autant que l'on a de données.

VIII 2 - Indice d'asymétrie

Quand on caractérise une courbe, on essaie de voir si elle est symétrique par rapport à ses indices.

$$\text{Indice 1} = \frac{Q3 - Q2}{Q2 - Q1}$$

Exemple :

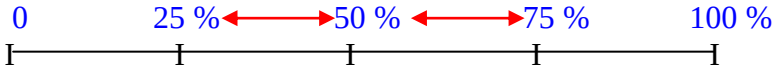


Fig. 31

On compare Q3 – Q2 à Q2 – Q1. On va comparer cet indice à la valeur 1.

On sait qu'il y a plus de valeurs de la variable entre Q3 et Q2 si c'est supérieur à 1 :

> 1 Q3 – Q2 > Q2 – Q 1

Si c'est inférieur à 1, il y aura plus de valeurs de la variables Q 2 et Q1 :

< 1 Q3 – Q 2 < Q2 – Q1

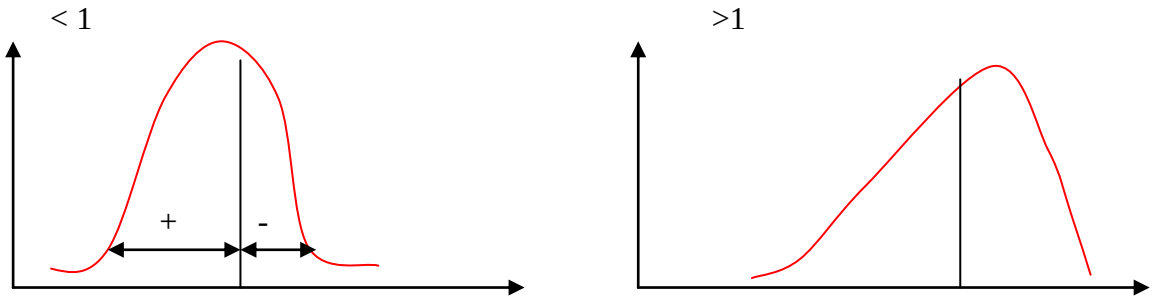


Fig. 32

$$\text{Indice 2} \qquad \text{Asym} \rightarrow 0$$
$$\text{Asym 1} = \frac{M - Me}{S}$$

Id. fig. 16

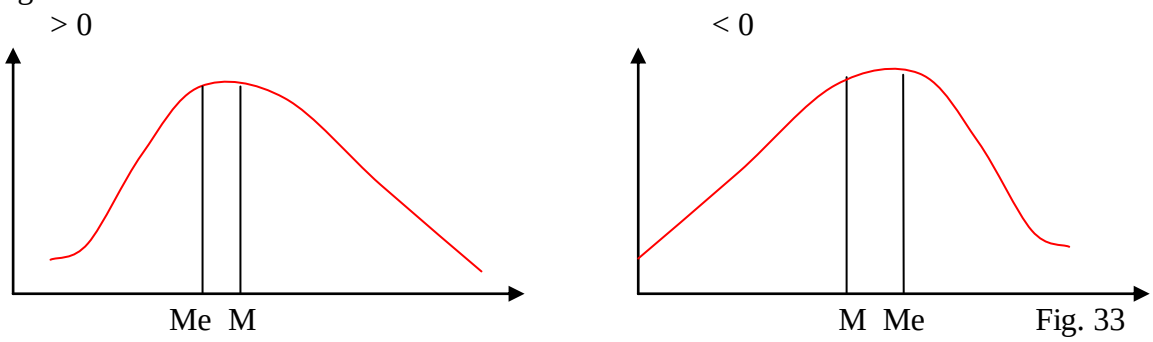


Fig. 33

$$\text{Indice 3} \qquad \text{Asym 2} = \frac{\Sigma z}{N}$$

S'il est **égal** à 0, la distribution est **normale**. S'il est **supérieur à 0**, il y a plus de valeurs de la variables au dessus qu'en dessous de la moyenne. S'il est **inférieur à 0**, c'est le contraire, c à d qu'il y a **plus de valeurs négatives.**

Chapitre II : Aperçu de la Loi Normale

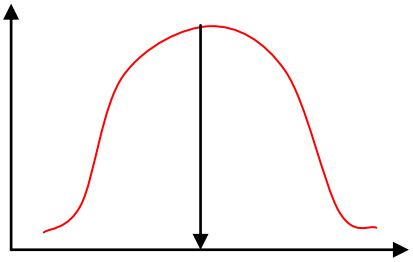
C'est grâce à la Loi Normale qu'on peut connaître le pourcentage de population représenté à l'intérieur ou de chaque côté de la variable et que l'on peut lire un test statistique.

(Statistique inférentielle : elle permet de dire si les choses analysées sont en lien ou si elles sont dues au hasard)

II Pourquoi la Loi Normale est-elle importante ?

C'est une loi de référence. Par rapport à cette référence, on peut qualifier la nature des observations.

Un mode

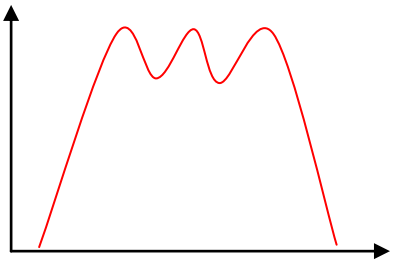


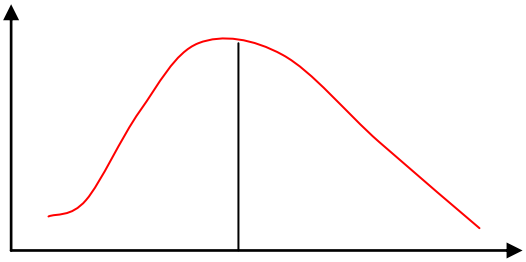
Moyenne

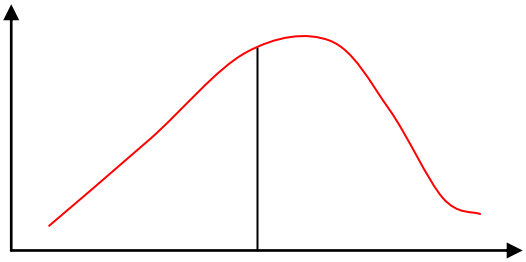
Fig. 34

Elle est symétrique par rapport à la moyenne et elle a un mode. On peut qualifier toutes les distributions par rapport à cette norme. On peut aussi qualifier la forme de la courbe qui qualifie la distribution. On peut encore dire si une performance est dans les 30 % les meilleurs, ou savoir dans quelle tranche de performance on risque de se retrouver...

mode 1 2 et 3







Courbes asymétriques.

Fig. 35

Le niveau de significativité veut dire qu'on se donne des chances de se tromper. C'est le niveau de confiance accordé au test (plus ou moins relatif) Si l'on peut relier la valeur du test au niveau de significativité, c'est parce que sur une Loi Normale, en connaissant l'équation de la courbe, on calcule un "z" et à ce "z", on fait correspondre un f (z).

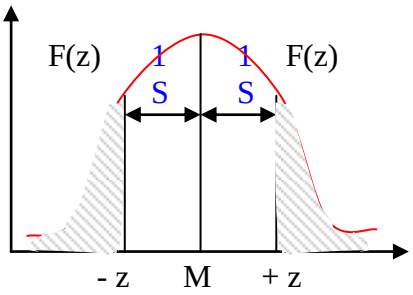
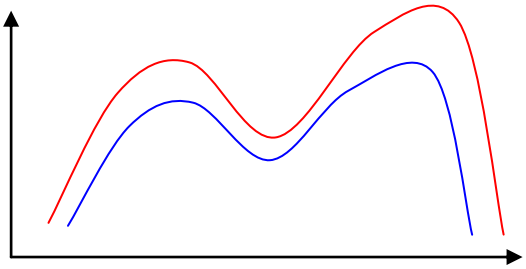


Fig. 36

On a une valeur et cette valeur est la probabilité d'avoir une valeur en dehors de l'intervalle représentée par ⌋ – z ; + z ⌊ . Donc, connaissant la courbe et la fonction, à chaque z, je vais pouvoir calculer une valeur qui correspond à une probabilité, et cette probabilité est d'être hors de la zone hachurée.

P = 0.05 = .05 = 5 % de chances de se tromper.
P = 0.01 = .01 = 1 % ...
On a vu que l'une des caractéristiques de la Loi Normale était d'avoir dans un écart type + ou – 50 % de la population.


Distribution bi modales Fig. 37

50 % de la population ± 0.67 écart type
68 % de la population ± 1 écart type
95 % de la population ± 1.96 écart type
99 % de la population ± 2.58 écart type
99.8 % de la population ± 3.09 écart type

Conclusion : pour avoir 5 % des valeurs supérieures à , il faut que – z soit égal à – 1.96 et que z soit égal à 1.96

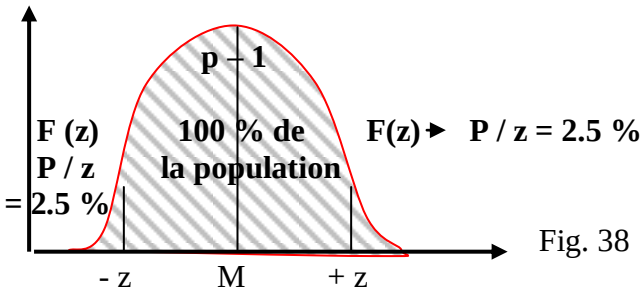


Fig. 38

Pour avoir une performance meilleure de 5 % par rapport aux autres, il faut qu'elle se trouve en dehors de la zone hachurée.

III Tous les tests statistiques sont-ils distribués normalement ?

La majorité des tests statistiques sont dérivés de la Loi Normale. Donc, pour appliquer des tests, il faut d'abord mesurer si la distribution est normale. Il faut aussi avoir un nombre de données suffisant pour être sûr que la loi est normale. A partir du moment où j'ai suffisamment d'informations, je n'ai pas besoin de faire cette vérification.

La limite du nombre d'observations nécessaire est : N > 30
N > 40
N > 60 Il faut prendre au moins 40 observations, si possible 60.

III 1 - Les principales propriétés de la Loi Normale

Deux indices définissent la Loi Normale. Sa moyenne qui est toujours égale à 0 et son écart type qui est toujours égal à 1 (dans le cas d'une Loi Normale "centrée réduite")

Courbe en cloche :
M = 0 ; S = 1.

Une distribution peut avoir la même moyenne mais un écart type différent.
Exemple sur la notation de trois profs : le premier note entre 6 et 12, le deuxième entre 9 et 14 et le dernier entre 12 et 18.

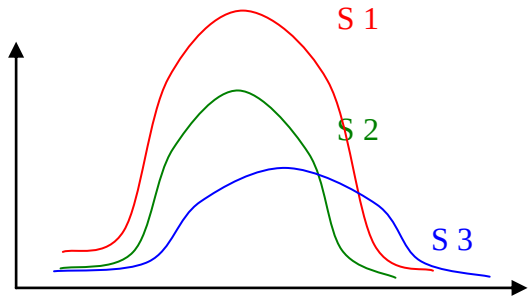


Fig. 39

Propriété dont découlent toutes les autres :
Pour combinaison linéaire de plusieurs variables normales indépendantes est une nouvelle variable normale.

?
?
?

1) Si x est une variable qui se distribue normalement (M, S) et si a et b sont deux constantes, la variable y = ax + b se distribue aussi normalement (M = aM + b et S = aS). Concrêtement, cela permet de placer la moyenne où bon nous semble dans un barême en mettant a = 1 et b = 1 ou 2 ou toute valeur qui semble justifiée.
2) Si x est une variable qui se distribue normalement (M, S) alors on peut définir une variable "z" par la transformation suivante

de la moyenne $z = \frac{x - M}{S}$ et de l'écart type

$S_1 = 1$

$M_z = 0$

Cette nouvelle variable est appelée variable normale réduite

C'est grâce à ces propriétés que l'on a pu dresser une table de correspondances et que l'on a pu dire qu'entre telle valeur S1 et telle valeur S2, on retrouvait tant de % de la population. Cela permet de déterminer le nombre d'éléments par classe (= proportion d'éléments)

La distribution peut être divisée en plusieurs classes (5, 7, 9, 11...) Les pourcentages vont correspondre de manière symbolique de la 2ème à la 2ème classe, de la 3ème à la 3ème classe, etc... Si une distribution ne correspond pas à ces pourcentages, elle n'est pas normale.

classes	1	2	3	4	5	6	7	8	9	10	11
11	3.6 %	4.5 %	7.8 %	11.6 %	14.6 %	15.8 %	14.6 %	11.6 %	7.8 %	4.5 %	3.6 %
9	4 %	6.5 %	12.1 %	17.5 %	19.8 %	17.5 %	12.1 %	6.5 %	4 %		
7	4.8 %	11.1 %	21.2 %	25 %	21.2 %	11.1 %	11.1 %				
5	6.7 %	28.2 %	34.2 %	28.2 %	6.7 %						

Tableau : 23

IV La Loi Normale et la significativité des tests statistiques

Quel est réellement le rapport avec le niveau "P" (niveau de significativité des tests statistiques) et qu'est ce que ça veut dire ?
Un test statistique (inférentiel) est quelque chose qui s'intéresse à la relation qui existe ou non entre deux variables. On note deux caractéristiques du test statistique :

1) Son intensité
2) Sa fiabilité

Lorsqu'on cherche avec un test statistique à mesurer la relation entre deux variables, on calcule ce test et on indique deux choses (intensité et fiabilité). La valeur est la force, l'intensité de la relation. Plus il est grand, plus l'intensité est forte.

Intensité : c'est la "force" avec laquelle deux variables sont liées. Plus l'intensité de la relation entre deux variables est élevée, plus on peut prévoir la valeur de l'une en fonction de la valeur de l'autre.

La fiabilité : C'est la probabilité de trouver une relation similaire à celle mise en évidence si l'expérience était à nouveau menée sur d'autres échantillons issus de la même population (et plus encore à la population totale dont est issu l'échantillon).

Rappel : L'usage veut que l'on considère trois niveaux de significativité d'un résultat :
à P < .05 : les résultats sont statistiquement significatifs (avec une probabilité d'erreur de 5 % qui n'est pas négligeable)
à P < .01 : les résultats sont statistiquement significatifs
à P < .001 : les résultats sont statistiquement **très** significatifs.

Test : la corrélation
On veut donner la force de relation entre la variable x et la variable y. S'il n'y a aucune relation.

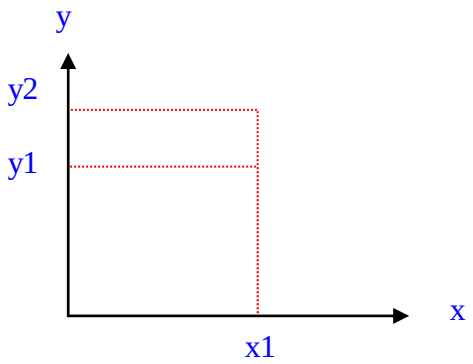


Fig. 40

Je ne peux dire, à partir de la valeur x1, la valeur de y1. Corrélation = 0.
Mais si les points s'organisent sur une droite quasiment parfaite (ils s'alignent) je pourrais, par la force de la corrélation, deviner à coup sûr y à partir de x. Corrélation parfaite = 1.

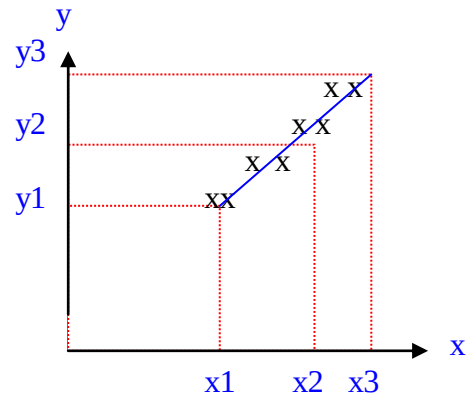


Fig. 41

Corr. = 1

La corrélation oscille entre –1 et + 1, c'est une corrélation très forte mais elle est inversée. Corrélation = -1.

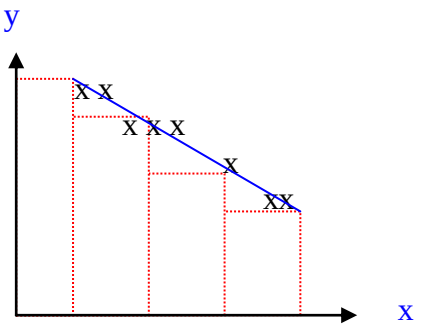


Fig. 42

Corr. = - 1

Deuxièmement, il faut indiquer le niveau de significativité du test, cad sa fiabilité. Pour cela, on prend un échantillon. Il doit avoir la même corrélation que tout le test.

Trois niveaux de significativité d'un résultat :
quelque chose existe, mais j'ai 5 % de chances de me tromper.
..... 1 % de chances .....
..... 1 chance sur 1000.....

Premier test : le Chi 2 ou □ ² .
Probabilité entre deux variables = .01 (1 chance sur 100)
Plus □ ² est petit (cad proche de 0) moins la relation est forte.

V Trois applications pédagogiques

1ère application pédagogique : Un enseignant d'EPS fait un test de cloche pied sans élan et note cette épreuve de 0 à 20 points.

1.25	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65	1.70	1.75	1.80	1.85
1	1	1	1	1	1	1	1	1	2	2	2	2
1.90	1.95	2.05	2.10	2.15	2.20							
2	2	3	4	1	1							

Tableau : 24

Tableau des distributions :

Performance	1.20	1.25	1.30	1.35	1.40	1.45	1.50	1.55	1.60	1.65	1.70
notes	0	1	2	3	4	5	6	7	8	9	10
Effectif	0	1	1	1	1	1	1	1	1	1	2
Effectif cumulé	0	1	2	3	4	5	6	7	8	9	11
Performance	1.75	7.80	1.85	1.90	1.95	2.05	2.10	2.15	2.20	2.25	
notes	11	12	13	14	15	16	17	18	19	20	
Effectif	2	2	2	2	2	3	4	1	1	0	
Effectif cumulé	13	15	17	19	21	24	28	29	30	30	

Tableau : 25

Pour dresser un barème pour toutes les autres classes de l'établissement, il faut que je normalise la distribution suivante qui ne correspond qu'à une classe.
1 - Je réunis les sujets tous les 3 points de 1 à 20 :

Classes	Val. extrêmes	Val. centrales	Effectif	E. cumulé
1	0-2	1	2	2
2	3-5	4	3	5
3	6-8	7	3	8
4	9-11	10	5	13
5	12-14	13	6	19
6	15-17	16	9	28
7	18-20	19	2	30

Tableau : 26

Mais je m'aperçois que ce n'est pas toujours normal.
On a pris une distribution normale en 7 classes (de façon aléatoire). On regarde sur le tableau de la distribution normale en 7 classes (cf polycopié)

On finit par trouver une distribution normale. Fig. 43

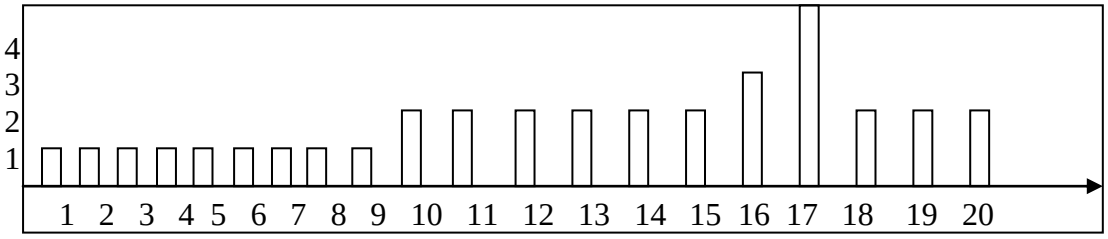
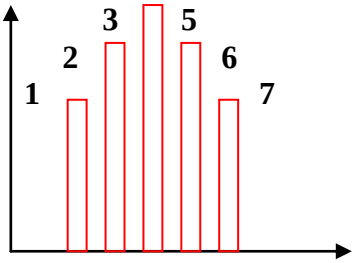


Fig. 44

On veut trouver une distribution normale.



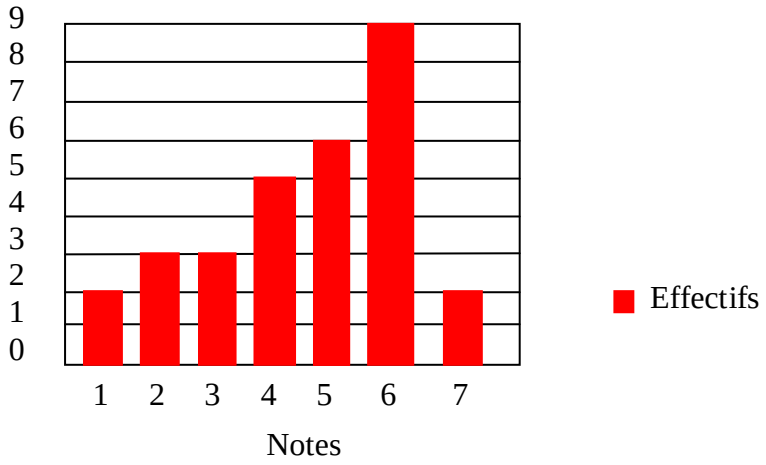


Fig. 45

Dans le cas d'une loi normale à 7 classes, on sait que le recoupement est à :

1	2	3	4	5	6	7
4.8 %	11.1 %	21.2 %	25.8 %	21.2 %	11.1 %	4.8 %

Tableau : 27

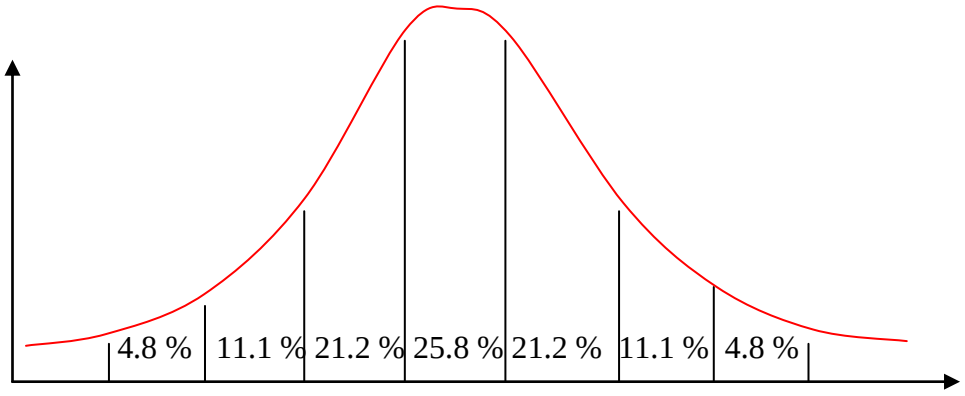


Fig. 46

J'ai 30 sujets. Je calcule combien font 4.8 % de 30 sujets, puis 11.1 % etc...

Rappel : le quantilage consiste en une opposition qui utilise le découpage de la fonction de répartition en tranches le plus généralement équidistantes. Le nombre de tranches permettra sa dénomination.

V 1 - Tout doit-il être normalisé ?
V 2 - Comment constituer des groupes de niveau à deux inconnues ?

Table 1
Exemple : Si on prend / z / = 1.50, on lit .134. Cela correspond à la partie hachurée des deux côtés de la courbe et est égale à 13.4 %. 13.4 / 2 = 6.7 %

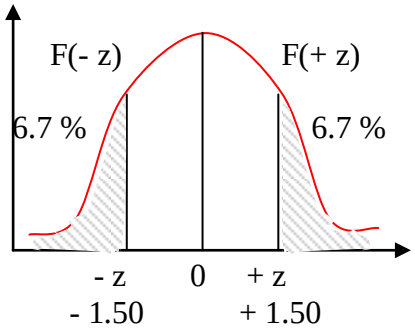


Fig. 47

Pour / z / = .44
Extrait de la table 1

z	.00	.01	.02	.03	.04
.00	1.000				
.10					
.20					
.30					
.40					.660

Tableau : 28

/ z / = .44 ➧ .660 = 66 %

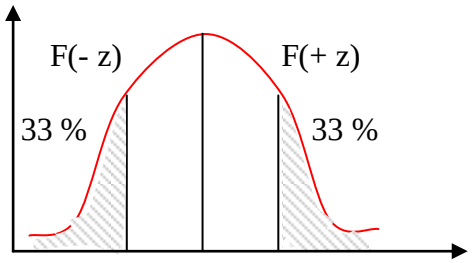


Fig. 48

Si on cherche à quelle valeur de z correspond 10 % dans l'air hachuré (cad 5 % de chaque côté), on trouve 1.65. / z / = 1.65.

Exercice : Un prof d'EPS a relevé un grand nombre de performances au triple saut. Il en a relevé suffisamment pour qu'il y ait normalité.
M = 7.50 m
S = 1.2 m

Combien d'élèves (en pourcentage) dans cette distribution saute plus de 9 m ou moins de 6.30 m ? (> 9 m ou < 6.3 m)

$/ z / = \frac{x1 - M}{S}$ x1 = 9 x2 = 6.3 m

$/ z / = \frac{9 - 7.50}{1.2} = \frac{1}{1.2} = / - 1 / = 1$

$/ z / = 1.25 \longrightarrow .211 = \frac{21.1\ \%}{2} = 10.55$

J'ai donc 10.55 % des élèves au-dessus de 9 m.

$/ z / = \frac{6.3 - 7.5}{1.2} = \frac{1}{1.2} = / - 1 / = 1$

$z = 1 \rightarrow .317 = 31.7$ % (des effectifs sont dans l'air hachuré), donc $\frac{31.7\ \%}{2} = 15.85$ % des élèves sont en dessous de 6.30 m.

On veut constituer des groupes de niveau à partir des performances et à partir de n'importe quelle classe et quelque soit les hasards de son échantillonnage :
Groupe A : Les bons élèves. La performance doit être dans les 30 % les meilleures.
Groupe B : les moyens. Ils doivent se trouver dans les 40 % suivants.
Groupe C : groupe faible. Ils se situeront au niveau des 30 derniers pour cents.

- Quelle performance permet de savoir dans quel groupe on se trouve ?

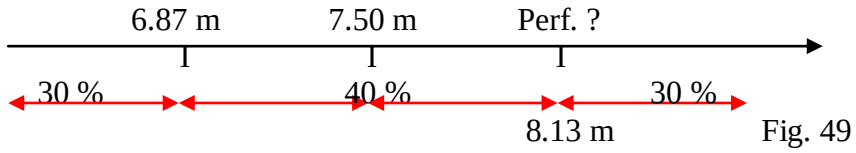


Fig. 49

On commence par les groupes A et C, c'est le plus simple à calculer :
Table 1 : 30 % + 30 % = 60 %. On cherche donc .600 dans la table :
.603 → / z / = .52
.296 → / z / = .53
} On prend la valeur intermédiaire : / z / = .525

$/ z / = \frac{x - M}{S}$

$x = (/ z / \times S) + M$

(+ z x 1.2) + 7.5 = 8.13 m
(- z x 1.2) + 7.5 = 6.87

Autre problème :
On a 200 étudiants. Au javelot, ils établissent ces performances :

M = 36.10 m
S = 6.35 m
La distribution des performances est considérée comme approximativement normale.

- Combien d'étudiants auront une performance entre 40 et 45 m ?

/ z 40/ = $\frac{40 - 36.1}{6.35} = \frac{3.9}{6.35} = 0.61$ → .542 = 54.2 % → 54.2 / 2 = 27.1 % x 2

/ z45 / = $\frac{45 - 36.1}{6.35} = \frac{8.9}{6.35} = 1.4$ → .162 = 16.2 % → 16.2 / 2 = 8.1 x 2

/ z45 / → 100 % - 8.1 % = 91.9 %
/ z40 / → 100 % - 27.1 % = 72.9 %

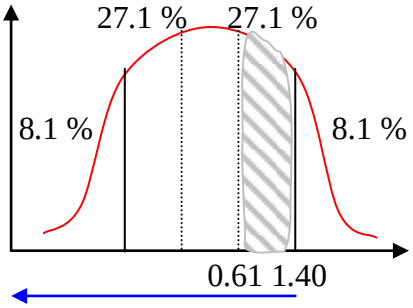
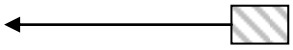


Fig. 50

91.9 % d'étudiants sont inférieurs à 45 m



72.9 % sont inférieurs à 40 m.

19 % de 200 étudiants = 38 étudiants.

V 3 Comparer deux moyennes et pourquoi ?

Cela permet d'expliquer comment fonctionnent les test statistiques.
Comment comparer deux moyennes : On considère que la distribution suivant est normale. On a deux populations :

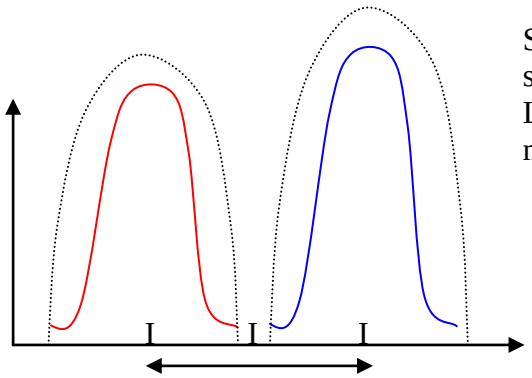


Fig. 51

Si les courbes des moyennes de ces populations sont tracées ainsi, on constate une différence (a) Les deux populations ne font pas partie de la. même population parente.

Autre exemple

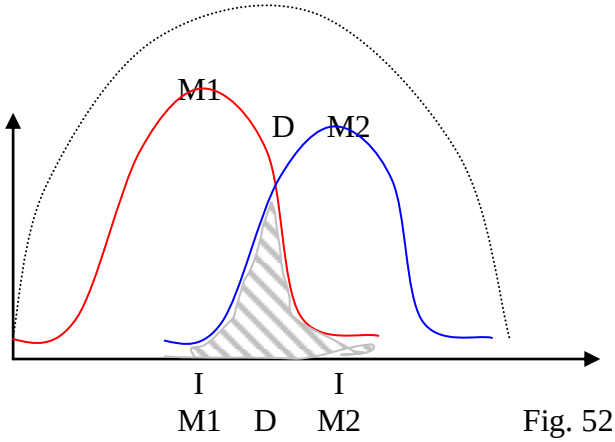


Fig. 52

Les deux populations suivantes ont des courbes beaucoup plus proches.
Il y a une différence mais ce la vient du tirage aléatoire du test.
On considère donc que ces deux échantillons ont la même population parente car ils ont à peu près la même moyenne.

Le "t" de student.

Une petite différence sur un échantillon n'est pas significative. Mais la même différence trouvée chez des milliers de sujets le devient. Donc, on pondère la distribution suivante avec la formule :

$$/ t / = \frac{/ Ma - Mb / = - 0}{V \frac{Sa^2}{Na} + \frac{Sb^2}{Nb}}$$

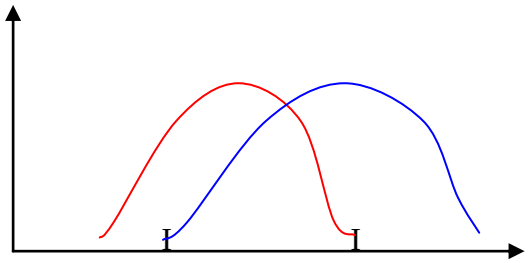


Fig. 513

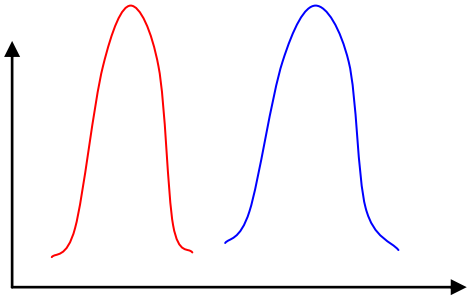


Fig. 54

La moyenne est la même. Mais la dispersion (S) est plus importante dans la courbe fig.

Exemple : On a deux distributions, les notes en APS des filles et celles des garçons.

Mf = 9	Sf = 4	Nf = 70
Mg = 10.5	Sg = 3	Ng = 80

Application de la formule : $/ t / = \frac{/9 - 10.5 /}{V \frac{4^2}{70} + \frac{3^2}{80}} = 2.23$

La particularité de ce test vient du fait qu'en ayant un nombre de sujets > 60, le / t / de student se confond avec les / z / :

N > 60 / t / = / z /
ou / t / ↔ / z / (N > 60)

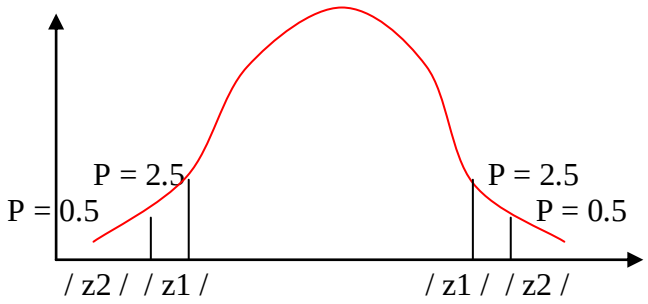


Fig. 55

Proba = P = 0.5 (5 % de chances de se tromper)
Combien j'ai de chances que la valeur " t " soit à l'extérieur des seuils que l'on a définit ?
Si on lit sur la loi normale :
A 0.5 % on trouve 1.96 et à 1 %, on trouve 2.58. Et la valeur du test statistique est / z / = 2.23. Si le / z / calculé est supérieur au / z / lu, alors il existe une différence.

/ z / cal > / z / lu ➔ il y a une différence
/ z / cal < / z / lu ➔ il n'y a pas de différence

Dans le cas de la figure , / z / cal > / z / lu car 2.23 > 1.96 : il y a une différence. Mais 2.58 > 2.23 : il n'y a pas de différence. Ce n'est pas sûr... On peut vérifier si cette différence existe toujours en prenant davantage de personnes dans l'échantillon.
Tous les tests statistiques sont basés sur l'étendue de la loi normale et ces valeurs vont quasiment systématiquement se retrouver.

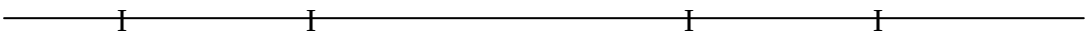


Fig. 56

Calculer le pourcentage de chances qu'une valeur a de se retrouver à l'extérieur ou à l'intérieur

V 4 Autres exercices (cf. poly)

$$1.035 = \frac{(16 - M)}{S}$$

$$0.255 = \frac{9 - M}{S}$$

Ici, j'ai deux inconnues : il faut les éliminer.

$$1.035 - 0.255 = \frac{(16 - M)}{S} - \frac{9 - M}{S}$$

$$1.290 = \frac{(16 - M)}{S} - \frac{9 - M}{S} \qquad 1.290 = \frac{7}{S} \qquad S = \frac{7}{1.290} = 5.43$$

	Disque	Perche	T S
M	19 m	2.20 m	9 m
S	4 m	0.30 m	1 m

Tableau : 29

	Disque	Perche	T S
Sujet a	22 m	2.35 m	7 m
Sujet b	18 m	2.20 m	10.50 m

Tableau : 30

Dans quelle épreuve le sujet a est-il le moins bon ? Est-il meilleur en disque ou en perche ?

- On compare les performances de disque et de perche :

za disq $= \frac{22 - 19}{4} = 0.75$

za perch $= \frac{2.35 - 2.20}{0.30} = 0.50$

Le sujet a est meilleur en disque car za dis > za perch

za TS $= \frac{7-9}{1} = -2$

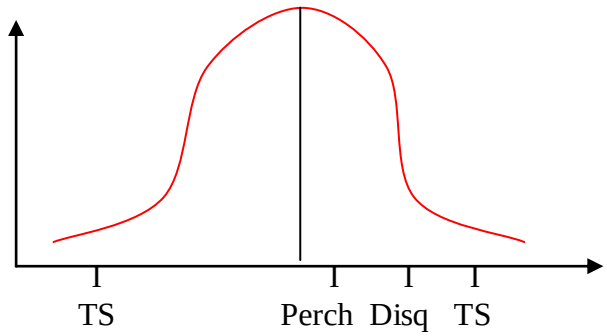


Fig. 57

- Est ce que a à la perche est meilleur, comparativement aux élèves de sa classe, que le sujet b au TS ?

zb TS $= \frac{10.50 - 9}{1} = 1.5$ (za perch = 0.50)

- Le sujet b est meilleur au TS que le sujet a n'est bon à la perche.

Chapitre III : Intro à la statistique inférentielle

Loi du x^2

Variable nominale quantitative (continue, discrète...)

Question : Quel type d'information m'apporte la variable ?
Cela permet de déterminer le test statistique à utiliser.

Univarié (ou **tri à plat**) : statistique descriptive (socio)
Bivarié : point de vue de deux variables que l'on étudie en même temps. Il s'agit de savoir ce que ces variables peuvent avoir en commun ou de différent.
Multivarié : le χ^2 n'admet pas qu'il y ait deux variables nominales.

Il y a trois niveaux d'analyse :

Pour répondre à ces questions, on utilise, pour les variables nominales :
- **le x^2**.

Pour les variables nominales et ordinales, on utilise :
- **le "t" de student ou l'analyse de la variance** (exemple : quand on a deux catégories (bleu / jaune...) : on applique le "t" de student et pour plusieurs catégories (enfants / ados / adultes) on applique l'analyse de la variance)

Et pour deux variables d'intervalle (comme deux mesures (tailles, poids) ou deux performances...) on applique :
- **le test de corrélation ou les régressions** (liées à la corrélation R)

La plupart des tests présentés supposent que la distribution soit normale. Ce sont des **tests paramétriques**. Quand on a plus de 60 sujets, il n'est pas obligatoire de vérifier que la distribution est normale. On a souvent plusieurs hypothèses expérimentales (H1, H2, H3...). Le statisticien part toujours d'une hypothèse 0 (hypothèse nulle) cad qu'il calcule toujours de la même façon...
Si l'hypothèse nulle (le test) est supérieure à l'hypothèse, l'hypothèse nulle est rejetée.
Si l'hypothèse nulle est inférieure à l'hypothèse, on garde l'hypothèse nulle.

Etude de la normalité de la variable : on a une variable se distribuant d'une certaine manière (un peu décalée vers la droite)

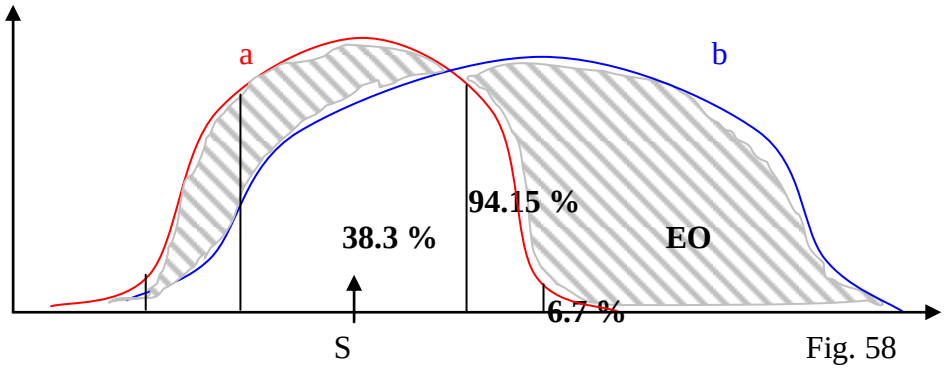


Fig. 58

On a une différence plus ou moins importante.
La différence est très grande, cela veut dire que la variable b ne se distribue pas normalement.
La différence est petite, on ne peut pas dire qu'il y ait une réelle différence avec une distribution telle qu'elle serait si elle était normale.

On distribue la variable b en 5 classes. On sait que dans une loi normale, pour 5 classes, on a un pourcentage fixe. Ce sont les **effectifs théoriques**. On applique le x^2 : une des valeurs de ce test est inférieure à 5.

~~5~~ < x < 10

Exemple : on observe une variable. On a 300 sujets. On en trouve 17 dans le premier intervalle (6.7 % de 300 = 17) Il existera une différence entre les effectifs observés et les effectifs théoriques car les effectifs observés ne sont pas totalement normaux. Plus la différence est grande, plus le x^2 est grand.

On sait que x^2 cal > x^2 lu → on rejette l'hypothèse nulle.
x^2 cal < x^2 lu → on accepte l'hypothèse nulle.
Table du x^2 (degrés de liberté : DDL)
Dans les deux cas, on voit que x^2 calculé est supérieur à x^2 lu. Il n'y a pas de différence entre la distribution et une distribution normale, donc cette variable se distribue normalement.