Master Eseec
Statistique pour l'expertise

Christian Lavergne

Université Paul Valéry - Montpellier 3
http://moodle-miap.univ-montp3.fr
http://www.univ-montp3.fr/miap/ens

L'ONG "e-enfance" mène une enquête sur l'utilisation d'Internet chez les adolescents (jeunes de 13 à 18 ans) par un sondage sur 1200 personnes. Le sondage met en avant les situations à risque suivantes :

- 53 % des jeunes déclarent avoir été exposé à des images choquantes.
- 29 % avoir reçu des propositions sexuelles, 48% avoir reçu des propositions (autres que sexuelles) de rendez-vous avec un inconnu, 23% n'avoir reçu aucune proposition.

Il est aussi demandé lors du sondage l'âge, le temps de connection quotidien moyen et le nombre d'enfants vivant sous le même toit.

Les premiers indices de la Statistique :

1. Indices de localisation , Indices de dispersion
2. Les quantiles
3. Intervalles basés sur les rangs et représentation graphique

Indices : définitions préliminaires

À la vue des observations d'une variable quantitative, on peut s'intéresser à résumer l'information de la distribution par des indices de :

- localisation : c'est une valeur qui reflète un endroit spécifique de la distribution.

2 indices de localisation naturels :

  - le minimun (min) : la valeur minimale observée.
  - le maximun (max) : la valeur maximale observée.

- dispersion : c'est une valeur qui renseigne sur la dispersion de la distribution indépendamment de sa localisation.

L'indice de dispersion naturel est :

  - l'étendue de la distribution : la valeur = **(max - min)**.

Le mode d'une distribution

- Cas d'une variable discrète :

  C'est l'observable le plus fréquement observé.

- Cas d'une variable continue :

On parlera plutôt de classe modale : la classe la plus souvent représentée.
C'est la classe du maximun de l'histogramme.
On définira alors *le mode comme le centre de la classe modale*.

Le mode est un indice de localisation

Indices basés sur les rangs : les quantiles

On peut aussi proposer des indices qui partagent la population en partie(s) égale(s) ; **les individus étant au préalable rangés par ordre croissant de la variable.**

- La médiane ("*Med*") : c'est la valeur observable qui partage en **deux** effectifs égaux la population rangée par ordre croissant de la variable.

▸ Exemples ▸ Caractérisation ▸ Propriétés ▸ Exercices

- Les 2 quartiles (Q_1 et Q_3) :

les 3 valeurs Q_1, *Med*, Q_3 partagent en **quatre** effectifs égaux la population rangée par ordre croissant de la variable.

Les 3 valeurs (observables) Q_1, *Med*, Q_3 sont les indices de localisation fondamentaux basés sur les rangs.

Exemples :

Quelle est la médiane des séries statistiques suivantes ?

- 03, 05, 06, 08, 10, 12, 14
- 4, 18, 12, 9, 7, 22, 10, 3, 6, 17, 14
- 1, 3, 2, 2, 3, 1, 4, 0, 2, 1, 3, 0, 2, 0, 1, 1, 3, 1, 3, 2, 2, 5, 1, 3, 5
- 1, 1, 2, 1, 2, 3, 3, 2, 1, 1, 2, 1, 3, 2, 0, 3, 3, 0, 2
- 5, 9, 19, 21, 24, 18, 43, 25, 26, 19

Caractérisation de la médiane - cas de données brutes :

1. Lorsque le nombre d'observations est impair, n = 2 k + 1, alors **Med** est la $(k+1)^e$ observation de la série triée.
2. Lorsque le nombre d'observations est pair, n = 2 k, alors **Med** est le milieu de la k^e et la $(k+1)^e$ observation de la série triée.

Propriété :
La médiane "*Med*" doit vérifier les 2 propriétés P_1 et P_2 :

1. au moins 1 individu sur 2 de la population à une valeur **inférieure ou égale** à *Med* :

$$P_1 \quad : \quad freq(observations \leq Med) \geq \frac{1}{2} \quad \text{ou} \quad F(Med) \geq \frac{1}{2}$$

2. au moins 1 individu sur 2 de la population à une valeur **supérieure ou égale** à *Med*.

$$P_2 \quad : \quad freq(observations \geq Med) \geq \frac{1}{2} \quad \text{ou} \quad G(Med) \geq \frac{1}{2}$$

Aspect géométrique. La médiane est la valeur pour laquelle le graphe de la fonction de répartition F franchit le palier 0.5.

Un tableau de répartition d'une variable discrète :

Nbre d'enfants dans une famille d'étudiants

N^b d'enfants	**N^b d'étudiants**
1	7
2	99
3	47
4	12
5	6
≥ 6	9
Total	180

Donner le mode, la médiane et les 2 quartiles de cette distribution ?

Un tableau de répartition d'une variable continue : l'information étant non exhaustive on ne calcule donc qu'une approximation de la médiane.

Longueur	Eff.	Fréq.	Cumul
de 30 à 34	6	4.00	4.00
de 34 à 36	6	4.00	8.00
de 36 à 38	20	13.33	21.33
de 38 à 40	30	20.00	41.33
de 40 à 42	37	24.67	66.00
de 42 à 44	23	15.33	81.33
de 44 à 46	20	13.33	94.67
de 46 à 50	8	5.33	100.00
Total	**150**	**100**	

Donner la classe modale et la classe médiane de cette distribution ?
Donner une valeur au mode, à la médiane ?

Autres indices basés sur les rangs

- Les déciles (D_1 à D_9) :
  les 9 valeurs D_1, D_2, D_3, D_4, Med, D_6, D_7, D_8, D_9, partagent en **dix** effectifs égaux la population rangée par ordre croissant de la variable.
- Les centiles (c_1 à c_{99})
- Les quantiles ($q_{\alpha\%}$) peuvent être associé à toute proportion
  Ex : $q_{2.5\%}$. Il y a 2.5% de la population qui a une valeur observée de la variable $\leq q_{2.5\%}$.

Tous ces indices sont des indices de localisation.

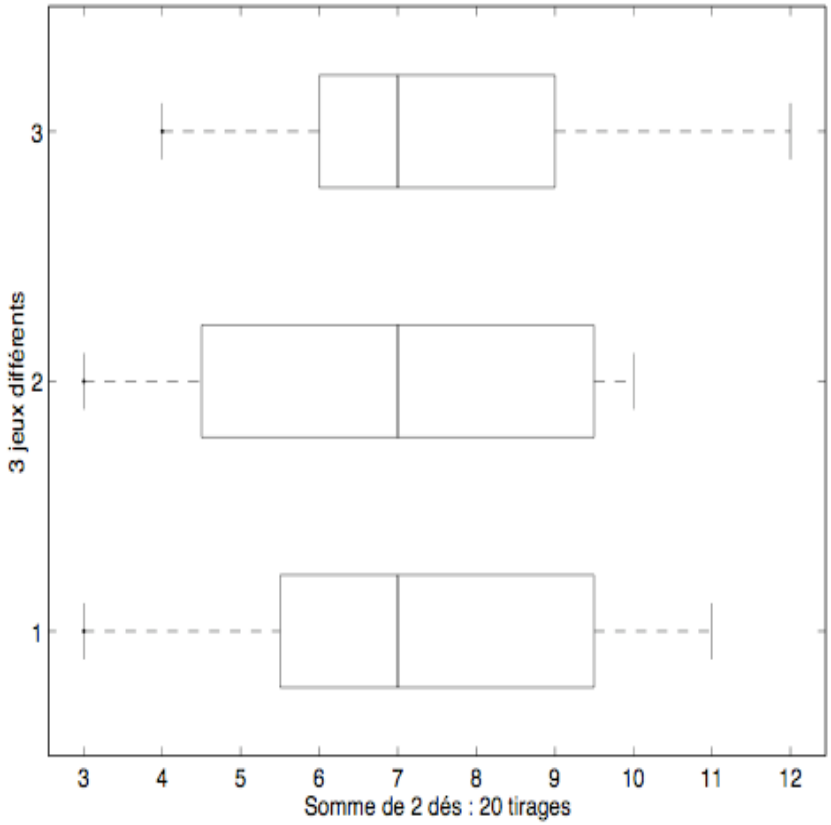
Intervalles basés sur les rangs et représentation graphique

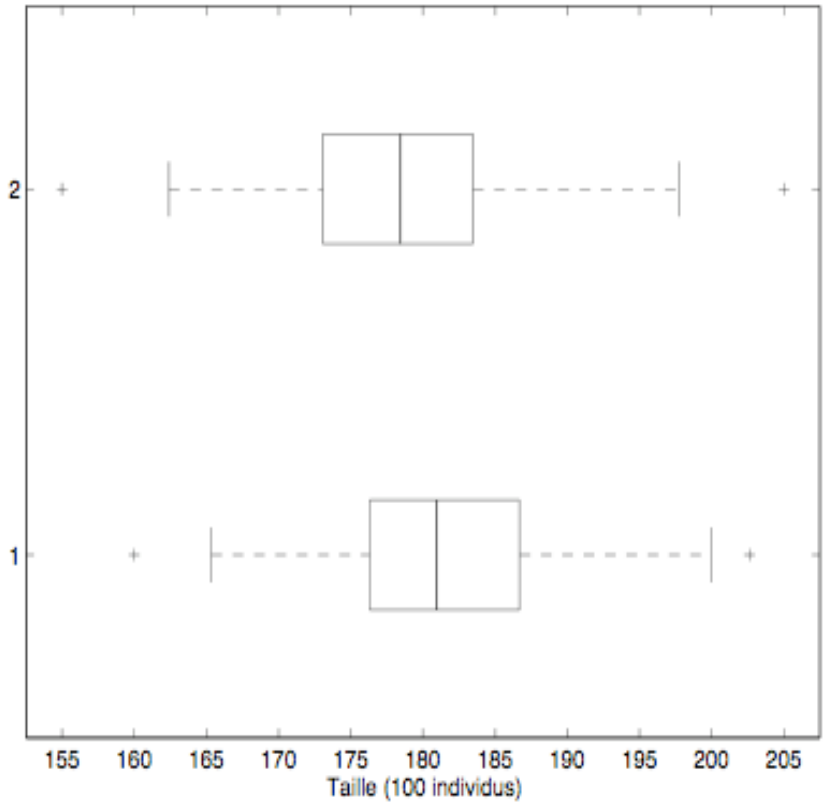
À l'aide des indices précédents, on peut construire autant d'intervalles centrés autour de la médiane (intervalles centrés au sens des quantiles) : ce sont autant d'indices de dispersion.

- l'intervalle **inter-quartile** : $Q_3 - Q_1$.
  C'est un intervalle à 50% centré autour de la médiane.
- l'intervalle à 80% centré autour de la médiane : $D_9 - D_1$
- l'intervalle à 60% centré autour de la médiane : $D_8 - D_2$
- l'intervalle à 95% centré autour de la médiane : $q_{97.5\%} - q_{2.5\%}$

Box-plot

Il existe enfin une représentation graphique fondée sur la médiane et les quartiles (“boîte à pattes", box and whiskers) :





Distance, dispersion, indice de localisation centrale

1. Définition intuitive d'un indice de localisation centrale
2. Distance et dispersion
3. Définition d'un indice de localisation centrale Moyenne et variance

Définition intuitive d'un indice de localisation centrale

Un indice de localisation centrale est :

- *définition 1* : une valeur "résumé" qui se situe **le mieux possible au milieu** des données
- *définition 2* : une valeur "résumé" qui est **proche de tous** les individus à la fois

On a donc besoin des 2 notions :

1. pour le mot "proche" on a besoin de définir une notion de distance
2. pour l'expression "proche de tous" on a besoin de définir une notion de dispersion.

Distance, dispersion

Soit une variable X observée sur n individus. On dispose donc des observations $x_1, x_2, \ldots, x_n$.
Soit a une valeur réelle quelconque. On note :

- $d(x_i, a)$ la distance entre la valeur de X pour i et la valeur a avec comme distance naturelle :
  1. la distance au sens des valeurs absolues : $d(x_i, a) = |x_i - a|$
  2. la distance au sens des carrés : $d(x_i, a) = (x_i - a)^2$
- $\text{Disp}(a) = \sum_{i=1}^{n} d(x_i, a)$ la dispersion totale des observations de la variable X autour de a associée à la distance d.

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $\lvert x_i - a \rvert$	3	6	7	13	16
pour $a = 2$, les $(x_i - a)^2$	9	36	49	169	256
pour $a = 8$, les $\lvert x_i - a \rvert$	3	0	1	7	10
pour $a = 8$, les $(x_i - a)^2$	9	0	1	49	100
pour $a = 12$, les $\lvert x_i - a \rvert$	7	4	3	3	6
pour $a = 12$, les $(x_i - a)^2$	49	16	9	9	36

	$a = 2$	$a = 8$	$a = 12$
$\text{Disp}^{(\lvert\rvert)}(a) = \sum_{i=1}^{5} \lvert x_i - a \rvert$	45	21	23
$\text{Disp}^{(2)}(a) = \sum_{i=1}^{5} (x_i - a)^2$	519	159	119

La valeur $a = 8$ est plus proche des données que $a = 2$ et $a = 12$ pour la dispersion $\text{Disp}^{(\lvert\rvert)}$

La valeur $a = 12$ est plus proche des données que $a = 2$ et $a = 8$ pour la dispersion $\text{Disp}^{(2)}$

Le cas d'une variable discrète présentée sous la forme du tableau de distribution.

n_1 individus prennent la valeur v_1
n_2 individus prennent la valeur v_2
$\vdots$
n_K individus prennent la valeur v_K

l'expression de la dispersion peut aussi s'écrire :

$$\text{Disp}(a) = \sum_{k=1}^{K} n_k d(v_k, a)$$

donc

$$\text{Disp}^{(||)}(a) = \sum_{k=1}^{K} n_k |v_k - a| \qquad \text{Disp}^{(2)}(a) = \sum_{k=1}^{K} n_k (v_k - a)^2$$

▸ Nbre d'enfants dans une famille d'étudiants : cas de la $\text{Disp}^{(\|)}$

							$\text{Disp}^{(\|)}$
	v_k	1	2	3	4	5	6	
	n_k	7	99	47	12	9	6	
$a=1$:	$\|v_k - a\|$	0	1	2	3	4	5	
$a=1$:	$n_k \times \|v_k - a\|$	0	99	94	36	36	30	295
$a=2$:	$\|v_k - a\|$	1	0	1	2	3	4	
$a=2$:	$n_k \times \|v_k - a\|$	7	0	47	24	27	24	**129**
$a=3$:	$\|v_k - a\|$	2	1	0	1	2	3	
$a=3$:	$n_k \times \|v_k - a\|$	14	99	0	12	18	18	161

La valeur $a=2$ est plus proche des données que $a=1$ et $a=3$ pour la dispersion $\text{Disp}^{(\|)}$

▸ Nbre d'enfants dans une famille d'étudiants : cas de la $\text{Disp}^{(2)}$

							$\text{Disp}^{(2)}$
	v_k	1	2	3	4	5	6	
	n_k	7	99	47	12	9	6	
$a=1$:	$(v_k-a)^2$	0	1	4	9	16	25	
$a=1$:	$n_k \times (v_k-a)^2$	0	99	188	108	144	150	689
$a=2$:	$(v_k-a)^2$	1	0	1	4	9	16	
$a=2$:	$n_k \times (v_k-a)^2$	7	0	47	48	81	96	279
$a=3$:	$(v_k-a)^2$	4	1	0	1	4	9	
$a=3$:	$n_k \times (v_k-a)^2$	28	99	0	12	36	54	**229**

La valeur $a=3$ est plus proche des données que $a=2$ et $a=1$ pour la dispersion $\text{Disp}^{(2)}$

Définition d'un indice de localisation centrale

Un indice de localisation centrale sera la valeur a :

- qui rend la dispersion totale minimale,
- donc telle que $\text{Disp}(a) = \sum_{i=1}^{n} d(x_i, a)$ est minimale

C'est donc un indice totalement dépendant de la distance choisie.

L'exemple pour la dispersion $\text{Disp}^{(|\,|)}$

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $\lvert x_i - a \rvert$	3	6	7	13	16
pour $a = 8$, les $\lvert x_i - a \rvert$	3	0	1	7	10
pour $a = 9$, les $\lvert x_i - a \rvert$	4	1	0	6	9
pour $a = 10$, les $\lvert x_i - a \rvert$	5	2	1	5	8
pour $a = 12$, les $\lvert x_i - a \rvert$	7	4	3	3	6

	$a = 2$	$a = 8$	$a = 9$	$a = 10$	$a = 12$
$\text{Disp}^{(|\,|)}(a)$	45	21	20	21	23

$$\text{Disp}^{(|\,|)}(9) = 20\,;\ \text{Disp}^{(|\,|)}(9+x) = 20 + |x| \ (\text{si } |x| \leq 1)\,;$$

$a = 9$ est **la valeur plus proche** des données pour la dispersion $\text{Disp}^{(|\,|)}$

- Si la distance choisie est : $d(x_i, a) = |x_i - a|$ alors le minimum de la dispersion est obtenu pour a = Med (la médiane) des observations.

  ▸ Nbre d'enfants dans une famille d'étudiants

- La dispersion totale des observations autour de la médiane est :

$$\text{Disp}^{(||)}(\text{Med}) = \sum_{i=1}^{n} |x_i - \text{Med}|.$$

  C'est l'écart absolu à la médiane.

- En divisant l'expression précédente par le nombre d'observations :

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \text{Med}|.$$

  on parle d'écart absolu moyen à la médiane.

L'exemple pour la dispersion $\text{Disp}^{(2)}$

pour la variable X , les x_i	5	8	9	15	18
pour $a = 2$, les $(x_i - a)^2$	9	36	49	169	256
pour $a = 8$, les $(x_i - a)^2$	9	0	1	49	100
pour $a = 10$, les $(x_i - a)^2$	25	4	1	25	64
pour $a = 11$, les $(x_i - a)^2$	36	9	4	16	49
pour $a = 12$, les $(x_i - a)^2$	49	16	9	9	36

	$a = 2$	$a = 8$	$a = 10$	$a = 11$	$a = 12$
$\text{Disp}^{(2)}(a)$	519	159	119	114	119

$$\text{Disp}^{(2)}(11) = 114\,;\ \text{Disp}^{(2)}(11 + x) = 114 + 5 \times x^2\,;$$

$a = 11$ est **la valeur plus proche** des données pour la dispersion $\text{Disp}^{(2)}$

- Si la distance choisie est : $d(x_i, a) = (x_i - a)^2$ alors le minimum de la dispersion est obtenu pour $a = \frac{1}{n}\sum_{i=1}^{n} x_i$ (la moyenne) noté $\overline{x}$.
- La dispersion **totale** des observations autour de la moyenne est :

$$\text{Disp}^{(2)}(\overline{x}) = \sum_{i=1}^{n}(x_i - \overline{x})^2.$$

- La dispersion **moyenne** des observations autour de la moyenne s'appelle la **variance** :

$$\text{Var}(x) = \frac{1}{n}\text{Disp}^{(2)}(\overline{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

On peut aussi vérifier que : $\text{Var}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2$

- la racine carrée de la variance est appelé **l'écart-type** :

$$\sigma(x) = \sqrt{\text{Var}(x)}.$$

La variance et l'écart-type sont des indices de dispersion ; mais contrairement à la variance, l'écart-type est un indice "compréhensible" puisque de même unité que les données.

- Pour une distribution symétrique et unimodale alors l'intervalle de dispersion centré autour de la moyenne :

$$\text{ID}(x) = [\overline{x} - 2 * \sigma(x), \overline{x} + 2 * \sigma(x)]$$

contiendra approximativement 95% de la population.

Remarque : **il est incohérent d'associer à la moyenne une mesure de dispersion qu'elle ne minimise pas. Autrement dit, il est incohérent d'associer à la moyenne une distance autre que le carré.**

Le cas d'une variable discrète présentée sous la forme du tableau de distribution.

n_1 individus prennent la valeur v_1
n_2 individus prennent la valeur v_2
$\vdots$
n_K individus prennent la valeur v_K

l'expression de la moyenne peut aussi s'écrire :

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k a_k$$

et celle de la variance :

$$\text{Var}(X) = \frac{1}{n} \sum_{k=1}^{K} n_k (v_k - \bar{x})^2$$

▸ Nbre d'enfants dans une famille d'étudiants : cas de la $\text{Disp}^{(2)}$

$\bar{x} = \frac{1}{180}(7v_1 + 99v_2 + 47v_3 + 12v_4 + 9v_5 + 6v_6) = 2.64.$

							$\text{Disp}^{(2)}$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	
$(v_k-2)^2$	1	0	1	4	9	16	
$n_k(v_k-2)^2$	7	0	47	48	81	96	279
$(v_k-3)^2$	4	1	0	1	4	9	
$n_k(v_k-3)^2$	28	99	0	12	36	54	229
$(v_k-\bar{x})^2$	2.7	0.41	0.13	1.9	5.6	11.3	
$n_k(v_k-\bar{x})^2$	18.8	40.6	6.1	22.2	50.1	67.7	**205.5**

La moyenne $a = 2.64$ est la valeur la plus proche des données pour la dispersion $\text{Disp}^{(2)}$

La dispersion vaut 205.5, la variance vaut 1.142, l'écart-type vaut 1.07.

Quelques remarques :

- La moyenne a toujours une précision plus "fine" que les observations. *Dans le cas d'une variable discrète elle ne sera donc jamais (ou presque) une valeur entière ; ce n'est donc presque jamais un observable.*
- La médiane est de la même précision que les observations. *Dans le cas d'une variable discrète elle sera donc (ou presque toujours) une valeur entière ; ce sera donc presque toujours un observable.*
- On a en général l'ordre suivant :

$$\text{Mode} \leq \text{Médiane} \leq \text{Moyenne}$$

  ou inversement.
- L'égalité de ces 3 indices traduit la symétrie de la distribution.

le cas d'une variable continue présentée sous la forme du tableau de distribution. On dispose de l'information non exhaustive suivante :

n_1 *individus prennent une valeur* $[b_0, b_1[$
n_2 *individus prennent une valeur dans l'intervalle* $[b_1, b_2[$
$\vdots$
n_K *individus prennent une valeur* $[b_{K-1}, b_K[$

on calculera une approximation de la moyenne par la même expression que précédemment, ici il y a K classes :

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{K} n_k c_k$$

en choisissant pour chaque c_k le milieu de la classe $[b_{k-1}, b_k[$:

c_2 *sera le milieu de l'intervalle* $[b_1, b_2[$,

c_3 *sera le milieu de la classe* $[b_2, b_3[$,

Exemple

Longueur	Eff.	c_k	Eff*c_k	Fréq.	Cumul
de 30 à 34	6	32	192	4.00	4.00
de 34 à 36	6	35	210	4.00	8.00
de 36 à 38	20	37	740	13.33	21.33
de 38 à 40	30	39	1170	20.00	41.33
de 40 à 42	37	41	1517	24.67	66.00
de 42 à 44	23	43	989	15.33	81.33
de 44 à 46	20	45	900	13.33	94.67
de 46 à 50	8	48	384	5.33	100.00
Total	**150**		**6102**	**100**	

$\overline{x} \approx$ **40.68**

la moyenne $\approx$ la médiane $\approx$ **40.7**

Le mode est aussi dans l'intervalle [40,42] :

la distribution est symétrique

Données centrées réduites, Indices

1. Données centrées réduites
2. Indices
3. Le coefficient de corrélation linéaire

Données centrées, réduites

Il existe 2 transformations qui sont classiquement appliquées à des observations d'une même variable. Ce sont les opérations de :

1. centrage : on retranche alors la moyenne à toutes les valeurs ;

$$y_i = x_i - \overline{x}$$

C'est une opération déjà rencontrée dans le calcul de la dispersion.

2. centrage et reduction : on divise toutes les valeurs centrées par l'écart type :

$$z_i = \frac{x_i - \overline{x}}{\sigma_X}$$

Les observations ainsi obtenues sont dites :

centrées (de moyenne nulle) et réduites (de variance 1).

On a donc :

- pour la variable centrée

$$\overline{y} = 0 \quad \sigma_y = \sigma_x$$

- pour la variable centrée, réduite

$$\overline{z} = 0 \quad \sigma_Z = 1$$

Retour sur le petit exemple

						$\sum$
les x_i	5	8	9	15	18	**55**

$$\bar{x} = 11$$

						$\sum$
les $(x_i - 11)$	-6	-3	-2	4	7	**0**
les $(x_i - 11)^2$	36	9	4	16	49	**114**

$$\sigma^2 = 22.8 \text{ et } \sigma = 4.775$$

						$\sum$
les $\frac{(x_i - 11)}{\sigma}$	-1.26	-0.63	-0.42	0.84	1.47	
les $\frac{(x_i - 11)^2}{\sigma^2}$	1.58	0.39	0.18	0.70	2.15	**5**

▸ Nbre d'enfants dans une famille d'étudiants :

$\bar{x} = 2.64$, $\sigma_x^2 = 1.142$ et $\sigma_x = 1.07$

							$\sum$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	**180**
$(v_k - \bar{x})$	-1.64	-0.64	0.36	1.36	2.36	3.36	
$n_k(v_k - \bar{x})$	-11.47	-63.25	16.97	16.33	21.25	20.17	**0**
$\frac{(v_k - \bar{x})}{\sigma_x}$	-1.532	-0.597	0.337	1.272	2.207	3.141	
$n_k \frac{(v_k - \bar{x})^2}{\sigma_x^2}$	16.46	35.39	5.37	19.47	43.94	59.35	**180**

2 nouveaux indices

- le skewness empirique

$$sk_X = \frac{1}{n}\sum_i^n z_i^3 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_X})^3$$

le skewness empirique "mesure" la symétrie d'une distribution,
le skewness empirique d'une distribution symétrique est proche de 0.

- le kurtosis empirique

$$k_X = \frac{1}{n}\sum_i^n z_i^4 = \frac{1}{n}\sum_i^n (\frac{x_i - \overline{x}}{\sigma_X})^4$$

le kurtosis empirique "mesure" les excès d'une distribution,
le kurtosis empirique d'une distribution symétrique *"Classique"* est proche de 3.

▸ Nbre d'enfants dans une famille d'étudiants :

$\bar{x} = 2.64$, $\sigma_x^2 = 1.142$ et $\sigma_x = 1.07$

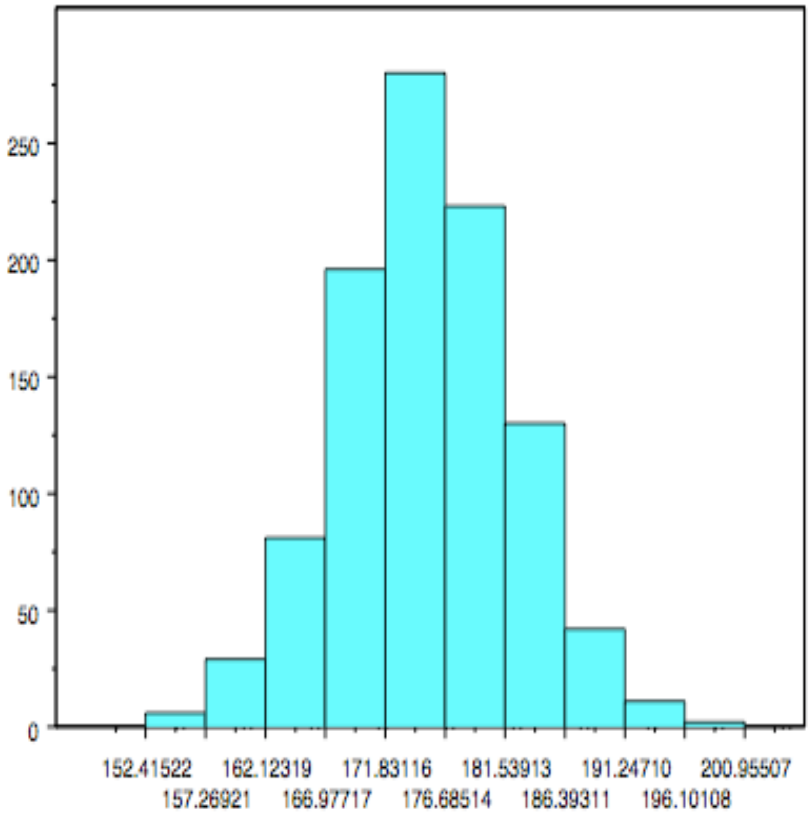
							$\sum$
v_k	1	2	3	4	5	6	
n_k	7	99	47	12	9	6	**180**
$(v_k - \bar{x})$	-1.64	-0.64	0.36	1.36	2.36	3.36	
$\frac{(v_k - \bar{x})}{\sigma_x}$	-1.53	-0.60	0.34	1.27	2.21	3.14	
$n_k(\frac{(v_k - \bar{x})}{\sigma_x})^3$	-25.2	-21.1	1.8	24.7	96.7	186	**263**
$n_k(\frac{(v_k - \bar{x})}{\sigma_x})^4$	38.5	12.6	0.6	31.4	213.4	584.2	**881**

Ainsi :

skewness $= 263/180 = 1.46$ kurtosis $= 881/180 = 4.89$

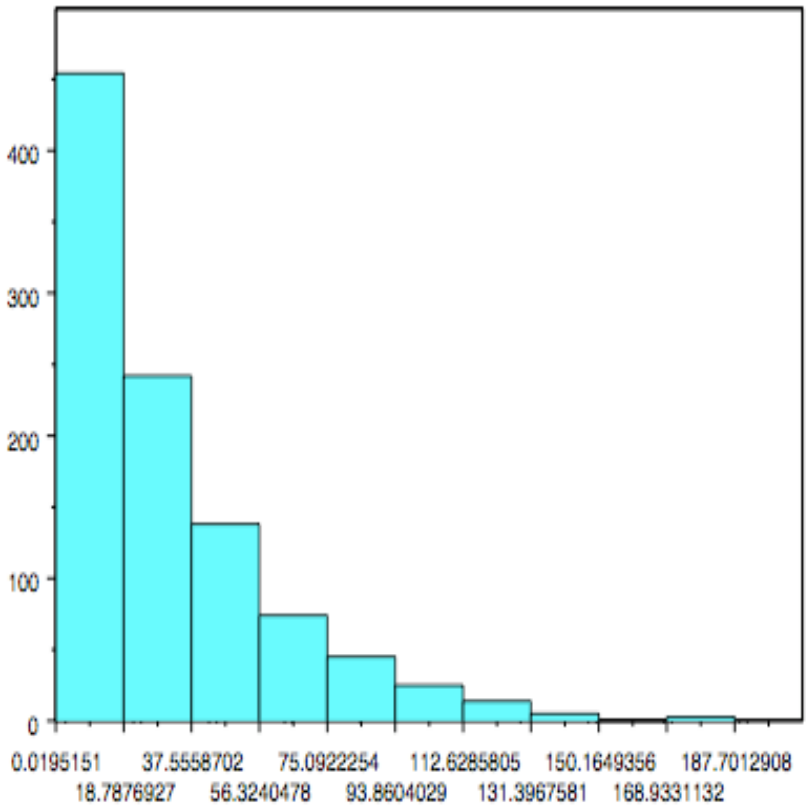
Deux variables continues (1000 observations)

la taille d'une population masculine



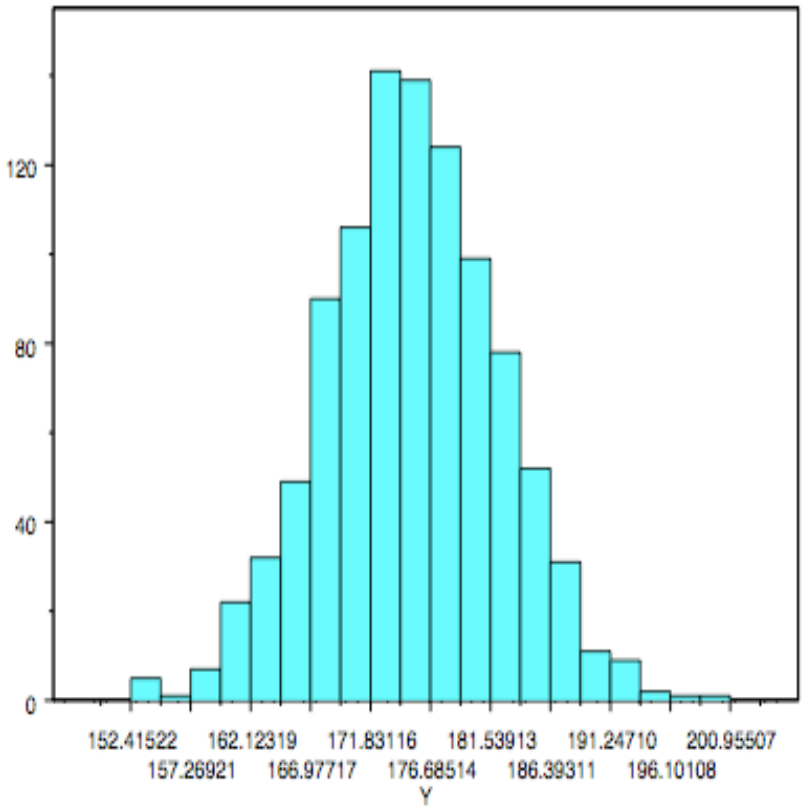
moyenne :	174.9
variance :	51.2
écart-type :	7.16
skewness :	0.047
kurtosis :	2.95

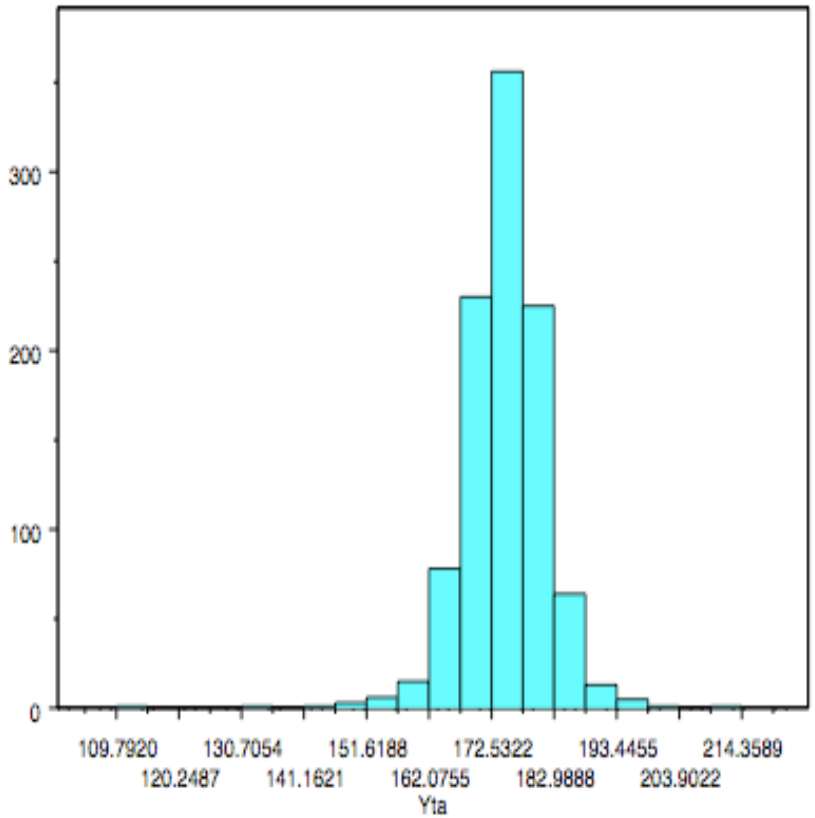
les précipitations pluvieuses



30.3
856.6
29.3
1.85
7.88

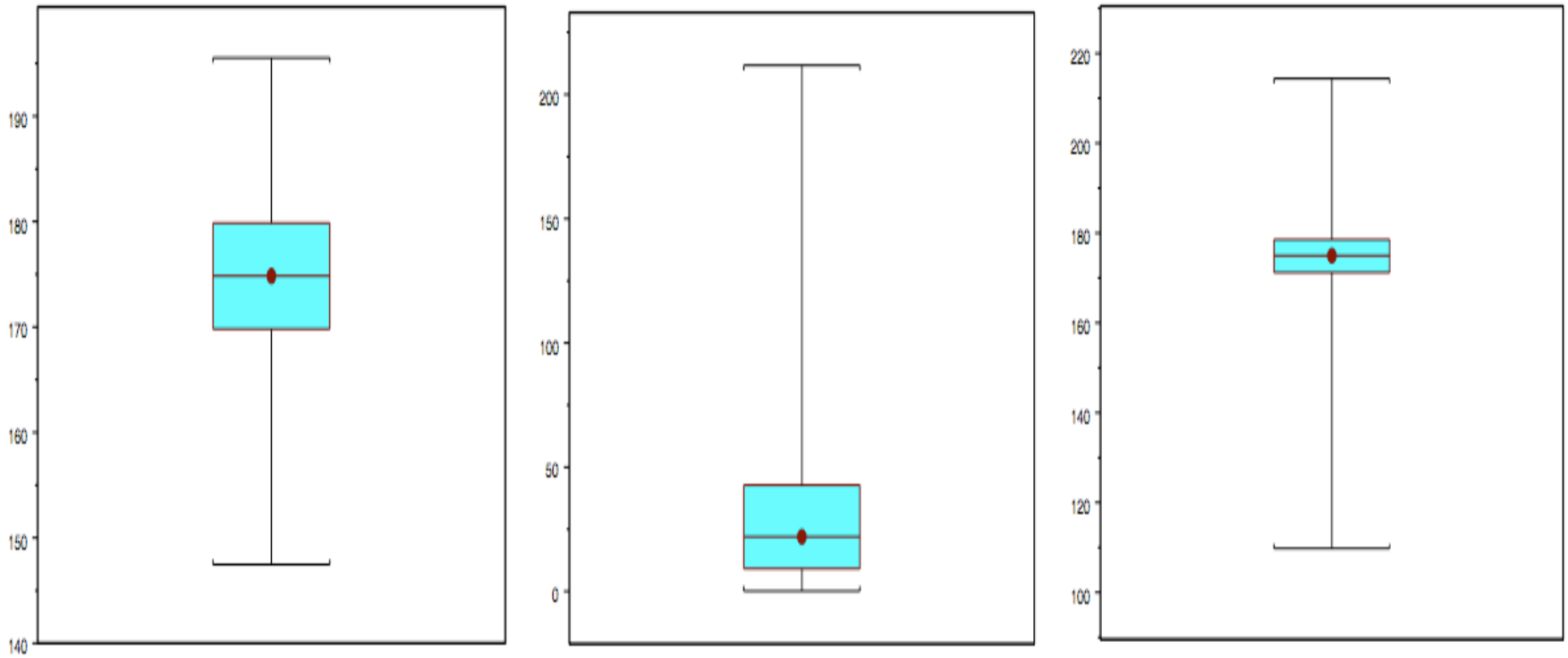
Histogrammes de 2 variables symétriques avec 20 classes (1000 observations)





moyenne :	174.9	174.7
variance :	51.2	47.5
écart-type :	7.16	6.89
skewness :	0.047	-0.94
kurtosis :	2.95	13.5

Box plot



Min :	147.46	0.091	109.79
1st Qu. :	169.82	9.109	171.20
Median :	174.85	21.97	174.98
3rd Qu. :	179.85	42.74	178.56
Max :	195.54	211.71	214.36

Le coefficient de corrélation linéaire

En présence de 2 variables X et Y mesurées sur les mêmes individus, un coefficient qui cherche à mesurer la liaison "linéaire" entre les 2 variables est le coefficient de corrélation linéaire empirique :

$$\rho(X,Y) = \rho_{XY} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})}{\sigma_X} \frac{(y_i - \bar{y})}{\sigma_Y}$$

qui possède la propriété suivante :

$$-1 \leq \rho_{XY} \leq 1.$$

Plus ce coefficient se rapproche de 1, plus les variables sont corrélées positivement, c'est-à-dire qu'elles varient dans le même sens. Plus il se rapproche de -1, plus elles varient en sens opposé. S'il se rapproche de 0, leurs variations ne sont pas liées linéairement.

Exemple 1 : Deux séries de notes observées sur 12 individus

X	14	13	17	15	14	15	16	12	14	12	13	13
Y	13	11	16	15	12	13	15	10	14	12	13	12
$\bar{X} = 14$												
$\bar{Y} = 13$												
$X - \bar{X}$	0	-1	3	1	0	1	2	-2	0	-2	-1	-1
$Y - \bar{Y}$	0	-2	3	2	-1	0	2	-3	1	-1	0	-1
Produits	0	2	9	2	0	0	4	6	0	2	0	1

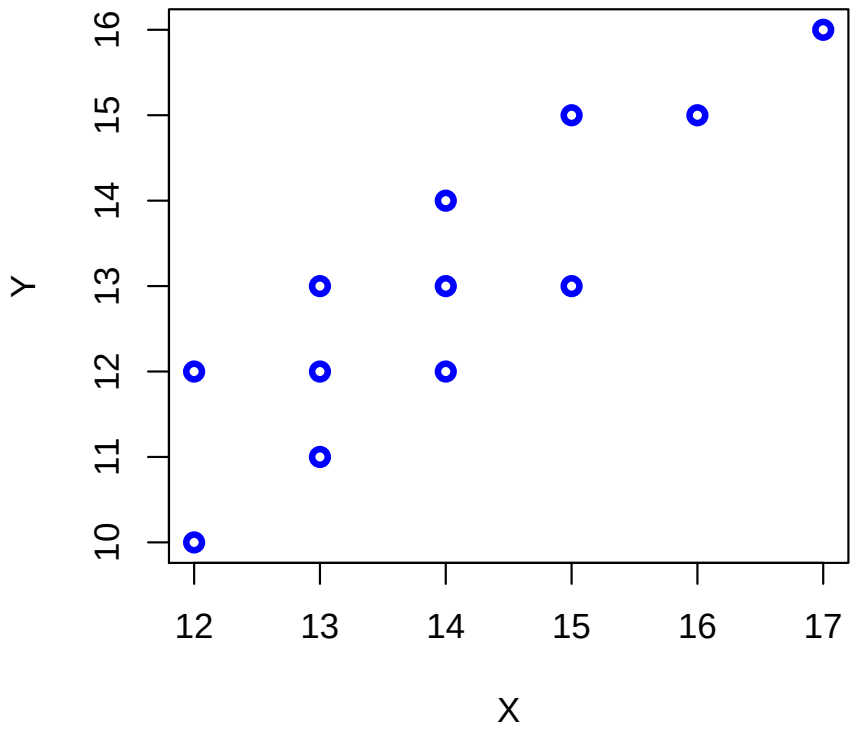
$$\sum_{i=1}^{12}(x_i - \bar{x})^2 = 26, \text{ donc } \sigma_X^2 = \frac{26}{12} = 2.167 \text{ et } \sigma_X = 1.472$$

$$\sum_{i=1}^{12}(y_i - \bar{y})^2 = 34 \text{ , donc } \sigma_Y^2 = \frac{34}{12} = 2.833 \text{ et } \sigma_Y = 1.683$$

$$\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y}) = 26$$

Le coefficient de corrélation linéaire est donc :

$$\rho = \frac{1}{n}\frac{\sum_{i=1}^{12}(x_i - \bar{x})(y_i - \bar{y})}{\sigma_X \times \sigma_Y} = \frac{26}{12 \times \sigma_X \times \sigma_Y} = 0.875$$



Exemple 2 : “fichier notes"

	Math	Phys	Chim	Angl	Fran	Hist
1	15	17	18	9	8	10
2	6	7	5	10	7	5
3	7	4	4	13	15	19
4	18	19	19	18	14	16
5	8	12	10	10	11	9
6	15	14	19	12	6	8
7	6	10	5	19	13	16
8	14	16	12	17	11	15
9	8	7	8	9	10	10
10	7	9	7	7	5	5
11	9	10	11	12	14	13
12	14	18	15	6	7	5
13	5	7	9	18	16	16
14	6	11	10	9	5	8
15	14	16	18	16	11	10
16	16	12	12	14	11	14
17	14	16	15	8	7	10
18	9	8	13	12	9	10
19	6	4	7	15	14	17
20	8	4	3	8	9	8
21	12	15	13	15	12	13
22	7	4	3	17	15	13
23	5	8	7	9	9	9
24	16	14	17	7	9	6
25	7	11	12	11	10	9

Variable	Moyenne	Écart-type	Minimum	Maximum
Mathém	10.08	4.17	5.00	18.00
Physique	10.92	4.69	4.00	19.00
Chimie	10.88	5.07	3.00	19.00
Anglais	12.04	3.94	6.00	19.00
Français	10.32	3.22	5.00	16.00
Histoire	10.96	4.03	5.00	19.00

	Math.	Phys.	Chim.	Angl.	Fran.	Hist.
Mathém	1.00	0.82	0.83	-0.00	-0.15	-0.05
Physique	0.82	1.00	0.87	-0.04	-0.29	-0.18
Chimie	0.83	0.87	1.00	-0.05	-0.25	-0.17
Anglais	-0.00	-0.04	-0.05	1.00	0.76	0.80
Français	-0.15	-0.29	-0.25	0.76	1.00	0.85
Histoire	-0.05	-0.18	-0.17	0.80	0.85	1.00

Ce tableau s'appelle le tableau ou matrice des corrélations.

!! Attention

Un coefficient de corrélation ne traduit pas nécessairement une relation de cause à effet :

"Une bonne note en math n'implique pas une bonne note en chimie."

Autre exemple : la corrélation entre le revenu et le débit de carte bancaire est fortement positif. Il existe ici une relation évidente : plus le revenu est élevé plus le débit de carte bancaire va augmenter et pas le contraire!!!

!! La relation n'est pas contenue dans les données.

En présence de 2 variables

1. Test du khi2 d'adéquation
2. Analyse de la variance

Un test d'adéquation : le test du χ^2

Exemple 1 : "lancer de dés"

Une expérience consiste à lancer deux dés, et à relever la somme des chiffres lus. On fait l'expérience $n = 1000$ fois, et on obtient :

S	2	3	4	5	6	7	8	9	10	11	12
n_k	32	56	81	115	142	160	143	105	89	53	24

Exemple 2 : "Les familles de 8 enfants"

On a observé, en étudiant 53680 familles de 8 enfants, les résultats suivants (k désigne le nombre de garçons et n_k le nombre de familles ayant k garçons) :

k	0	1	2	3	4	5	6	7	8
n_k	215	1485	5331	10649	14959	11929	6678	2092	342

Exemple 3 : "10 000 premières décimales du nombre π"
La répartition de ces décimales est donnée dans le tableau suivant :

décimale	0	1	2	3	4	5	6	7	8	9
effectifs	968	1026	1021	974	1012	1046	1021	970	948	1014

Se répartissent-elles de manière uniforme ?

Exemple 4 : "Sondage sur le niveau d'acceptation d'un nouveau système"

Appréciation / Âge des sondés	Très difficile	Assez difficile	Peu/pas difficile
de 18 à 29 ans	81	138	132
de 30 à 40 ans	126	131	94
50 ans et plus	203	78	69

Distance entre 2 tableaux : le tableau des observations $[n_k]$ et le tableau sous l'hypothèse d'un modèle [Eff théo$_k$]

$$\chi^2 = \sum_k \frac{(n_k - \text{Eff théo}_k)^2}{\text{Eff théo}_k}$$

Cette distance sera donc d'autant plus grande que le tableau des observations sera loin du tableau sous l'hypothèse du modèle

Pourquoi diviser par Eff théo$_k$?

Théo	1000	100	10
Obs	1010	110	20
écart	négligeable	faible	important
$\chi^2_{cellule}$	1/10	1	10

Quand doit-on considérer cette distance du χ^2 comme grande ?

C'est un problème décisionnel : donc problème de test statistique.
Pour répondre à cette question, on s'aidera d'une table appelée "Table du χ^2" qui donne pour chaque degré de liberté et différente "probabilité d'erreur" une valeur limite.
On décidera de la dépendance entre 2 caractères lorsque la distance du χ^2 sera supérieure à une valeur limite.

Table du χ^2

ν : nombre de degrés de liberté

$$P(\chi^2_\nu < l_\nu) = p$$

Exemple : $P(\chi^2_4 < 11.1433) = 0.975$

p	0.8000	0.9000	**0.9500**	0.9750	0.9900	0.9950
ν						
1	1.6424	2.7055	3.8415	5.0239	6.6349	7.8794
2	3.2189	4.6052	5.9915	7.3778	9.2103	10.5966
3	4.6416	6.2514	7.8147	9.3484	11.3449	12.8382
4	5.9886	7.7794	9.4877	**11.1433**	13.2767	14.8603
5	7.2893	9.2364	11.0705	12.8325	15.0863	16.7496
6	8.5581	10.6446	12.5916	14.4494	16.8119	18.5476
7	9.8032	12.0170	14.0671	16.0128	18.4753	20.2777
8	11.0301	13.3616	15.5073	17.5345	20.0902	21.9550
9	12.2421	14.6837	16.9190	19.0228	21.6660	23.5894
10	13.4420	15.9872	18.3070	20.4832	23.2093	25.1882
11	14.6314	17.2750	19.6751	21.9200	24.7250	26.7568
12	15.8120	18.5493	21.0261	23.3367	26.2170	28.2995
13	16.9848	19.8119	22.3620	24.7356	27.6882	29.8195
14	18.1508	21.0641	23.6848	26.1189	29.1412	31.3193
90	101.0537	107.5650	113.1453	118.1359	124.1163	128.2989
100	111.6667	118.4980	124.3421	129.5612	135.8067	140.1695

Analyse de la variance ou Anova

1 - Introduction

Objectif : Étudier l'effet d'une ou plusieurs variables qualitatives sur une variable quantitative

Le cas d'une variable qualitatitive : on observe sur des individus à la fois une variable quantitative et une une variable qualitative. On cherhche alors à "savoir" si les différentes modalités de la variable qualitative influencent la variable quantitative.

Exemple : On considère 6 échantillons de patients correspondant à des localisations différentes. Pour chaque patient, on observe une donnée clinique :

1	2	3	4	5	6
1602	1472	1548	1435	1493	1585
1615	1477	1555	1438	1498	1592
1624	1485	1559	1448	1509	1598
1631	1493	1563	1449	1516	1604
	1496	1575	1454	1521	1609
	1504		1458	1523	1612
	1510		1467		
			1475		

Information disponible pour chaque patient :

- Y : donnée clinique
- X : code de la localisation

Question : peut-on considérer que la localisation a une influence sur la donnée clinique des patients ?

... ou encore : la variable X a-t-elle une influence sur la variable Y ?

... ou encore : la modélisation de l'espérance μ de Y doit-elle être différente selon les modalités de X ou non ?

Vocabulaire

On appelle facteur (ou facteur explicatif, cause contrôlée) la variable qualitative X qui sert à expliquer Y.
On parle de niveaux d'un facteur (ou traitements) pour désigner les différentes modalités de cette variable.
Lorsqu'on étudie l'effet de plusieurs facteurs sur Y, on peut regarder leurs effets cumulés mais aussi l'effet de leur interaction, i.e. le croisement de deux modalités a une influence particulière sur Y.
On appelle unité expérimentale le sujet que l'on soumet à un traitement et sur lequel on mesure Y.
De façon générale ici, on suppose que l'on ne soumet pas une unité expérimentale à plusieurs traitements, autrement dit on suppose qu'il n'y a pas de répétition de la mesure de Y. Il y a donc autant d'observations que d'unités expérimentales.

ANOVA : pourquoi ?

- Analyse de la variance :
  - variations inter-groupes : écart entre les moyennes des groupes, dispersion des moyennes autour de la moyenne globale.
  - variations intra-groupes : écart entre les données à l'intérieur des groupes, dispersion des données autour de leur moyenne de groupe.

- ANOVA
  - ↗ technique de comparaison de moyennes
  - ↘ décider de l'égalité des moyennes

Approche intuitive sur un exemple

Un étudiant a mesuré le temps de parcours pour ce rendre à la fac selon trois types de trajet.

T1	17.5	20.0	18.0	17.0	16.5
T2	15.1	16.0	13.0	12.0	14.5
T3	10.0	13.0	10.0	11.0	12.0

Structure générale des données :

On a le tableau des données suivant :

facteur	niveau 1	...	niveau i	...	niveau I
observations	y_{11}	...	y_{i1}	...	y_{I1}
indépendantes	y_{12}	...	y_{i2}	...	y_{I2}
de la variable	$\vdots$		$\vdots$		$\vdots$
quantitative	y_{1n_1}	...	y_{in_i}	...	y_{In_I}

- y_{ik} : k^e observation du niveau i
- n_i : nombre d'observations du niveau i

- $n = \sum_{i=1}^{I} n_i$: nombre total d'observations
- $\bar{y}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$: moyenne des observations pour le niveau i
- $\bar{y}_{..} = \frac{1}{n} \sum_{i=1}^{I} \sum_{k=1}^{n_i} y_{ik}$: moyenne globale

!!! Attention
La moyenne des observations $\bar{y}_{..}$ n'est pas égale à la moyenne des moyennes par niveau du facteur $\bar{y}_{i.}$

Quelques hypothèses naturelles

Les Y_{ik} sont les résultats aléatoires de l'expérience étudiée et on suppose que leur "valeur espérée" est un paramètre noté μ_i qui ne dépend que du niveau du facteur contrôlé.

- μ_i est appelé l'effet fixe du niveau i du facteur contrôlé. C'est un paramètre inconnu : l'espérance de l'observation Y_{ik}
- les Y_{ik} sont aussi supposés être indépendants (donc associés à des sujets distincts.

Dans cette modélisation, il y a donc I paramètres inconnus liés à l'espérance : un paramètre pour chacun des niveaux du facteur.

Équation d'analyse de la variance :

$$\sum_{ik}(y_{ik} - \bar{y}_{..})^2 = \sum_{ik}(y_{ik} - \bar{y}_{i.})^2 + \sum_{i} n_i(\bar{y}_{i.} - \bar{y}_{..})^2$$

dispersion totale = dispersion **INTRA** + dispersion **INTER**

$$SS_T = SS_R + SS_F$$

$$\uparrow \quad \uparrow \quad \uparrow$$

Sum of Square Total = Sum of Square Résiduelle + Sum of Square due au facteur contrôlé

Degrés de liberté associés aux SS

$$\text{ddl}(SS_T) = \text{ddl}(SS_R) + \text{ddl}(SS_F)$$

$$n - 1 = n - I + I - 1$$

Données :

1	2	3	4	5	6	
1602	1472	1548	1435	1493	1585	
1615	1477	1555	1438	1498	1592	
1624	1485	1559	1448	1509	1598	
1631	1493	1563	1449	1516	1604	
	1496	1575	1454	1521	1609	
	1504		1458	1523	1612	
	1510		1467			
			1475			
4	7	5	8	6	6	n_i
1618	1491	1560	1453	1510	1600	$\bar{y}_{i.}$
470	1152	404	1296	760	534	$\sum_k (y_{ik} - \bar{y}_{i.})^2$

Analyse de la variance sur l'exemple des patients :

i	1	2	3	4	5	6
$\bar{y}_{i.}$	1618	1491	1560	1453	1510	1600

$$\bar{y}_{..} = \frac{1}{n}\sum_{ik} y_{ik} = \frac{1}{n}\sum_{i} n_i \bar{y}_{i.} = 1528 \qquad \text{ici } n = 36$$

Tableau d'analyse de la variance

Source de dispersion	Somme des carrés	ddl
INTER	125145 $\sum_i n_i(\bar{y}_{i.} - \bar{y}_{..})^2$	5 $I-1$
INTRA	4616 $\sum_{ik}(y_{ik} - \bar{y}_{i.})^2$	30 $n-I$
TOTALE	129761 $\sum_{ik}(y_{ik} - \bar{y}_{..})^2$	35 $n-1$

Test de l'égalité des espérances :

• l'hypothèse H_0 : absence d'effet du facteur contrôlé ; égalité des espérances $\mu_1 = \mu_2 \cdots = \mu_I$
• l'hypothèse alternative H_1 : effet significatif du facteur contrôlé ; différence des espérances $\mu_i,\ i = 1, \ldots, I$

$$\frac{\text{INTER}}{\text{INTRA}} * (n - I) = \frac{\sum_{ik}(y_{ik} - \bar{y}_{..})^2 - \sum_{ik}(y_{ik} - \bar{y}_{i.})^2}{(\sum_{ik}(y_{ik} - \bar{y}_{i.})^2)/(n - I)}$$

suit sous H_0 une loi de χ^2 à $(I - 1)$ degrés de liberté
dès que le nombre d'observations est grand

Re-écriture du paramètre μ_i et interprétation

$$\mu_i = \mu + \alpha_i \quad i = 1, \ldots, I \tag{2}$$

avec $I + 1$ paramètres pour l'espérance dont seulement I sont libres et identifiables $\Longrightarrow$ il y a sur-paramétrisation.
Différentes contraintes peuvent alors être envisagées :

- $\sum_i \alpha_i = 0$
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \frac{1}{I}\sum_{i=1}^{I} \mu_i = \bar{\mu}_. \qquad \text{et} \qquad \alpha_i = \mu_i - \bar{\mu}_.$$

  Le paramètre μ représente alors l'effet moyen général
  Le paramètre α_i représente alors l'effet différentiel du niveau i à la "moyenne"

- $\alpha_1 = 0$, par défaut dans de nombreux logiciels de statistiques.
  Lien avec la paramétrisation de l'équation (1) :

$$\mu = \mu_1 \qquad \text{et} \qquad \alpha_i = \mu_i - \mu_1$$

  Le paramètre μ représente alors l'effet du niveau 1 du facteur
  Le paramètre α_i représente alors l'effet différentiel du niveau i à l'effet du niveau 1
  Ici le traitement 1 sert de référence mais on peut prendre l'un quelconque des I traitements comme référence.

Estimation des paramètres :

Par moindres carrés : $\min_{\mu_i} \sum_{i=1}^{I} \sum_{k=1}^{n_i} (y_{ik} - \mu_i)^2$

- $\hat{\mu}_i = \bar{y}_{i.} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik}$
- contraste "sum" :

$$\hat{\mu} = \hat{\bar{\mu}}_{.} = \frac{1}{I} \sum_{i=1}^{I} \hat{\mu}_i = \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{i.} \qquad \text{et} \qquad \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} = \bar{y}_{i.} - \frac{1}{I} \sum_{i=1}^{I} \bar{y}_{i.}$$

Remarque : si $\forall i \in \{1, \ldots, I\} \quad n_i = K$ alors $\hat{\mu} = \bar{y}_{..}$ et $\hat{\alpha}_i = \bar{y}_{i.} - \bar{y}_{..}$

- contraste "treatment" :

$$\hat{\mu} = \hat{\mu}_1 = \bar{y}_{1.} \qquad \text{et} \qquad \hat{\alpha}_i = \hat{\mu}_i - \hat{\mu}_1 = \bar{y}_{i.} - \bar{y}_{1.}$$