Première année de Licence MIASHS

Introduction à la Statistique[1]

Julien GREPAT[2]

Contents

[2]Contact : julien.grepat@univ-grenoble-alpes.fr

Part I
Statistiques univariées

1 Statistiques descriptives

La statistique est une méthode scientifique qui consiste à organiser, analyser et interpréter des observations faites sur un ou plusieurs caractères (*poids, consommation,* $\cdots$) des individus (*chèvres, machines, pièces manufacturées,* $\cdots$) d'une population de taille N.

1.1 Vocabulaire

Le caractère ou variable statistique, noté x, est dit :

- discret si x prend une quantité dénombrable de valeurs $X = x_1, x_2, x_3, \cdots$ (*notes d'examen*),
- continu si x peut prendre toutes les valeurs d'un intervalle $[a, b]$ (*poids*),
- qualitatif si les valeurs de x sont une qualité (*couleur des cheveux*),
- quantitatif si les valeurs de x sont des nombres (*notes d'examen*).

L'ensemble des observations forme une série statistique.

1.2 Organisation

La démarche du statisticien est sensiblement la même quelque soit le type de variable. La première étape est d'établir le tableau statistique en triant les valeurs du caractère (dans l'ordre croissant lorsque c'est possible). Ensuite, il est important de représenter ces données. Le graphique permet souvent de dégager d'importantes informations et d'aiguiller la suite de l'étude. Il sera alors possible de faire l'analyse statistique qui pourra prendre différentes formes en fonction des attentes.

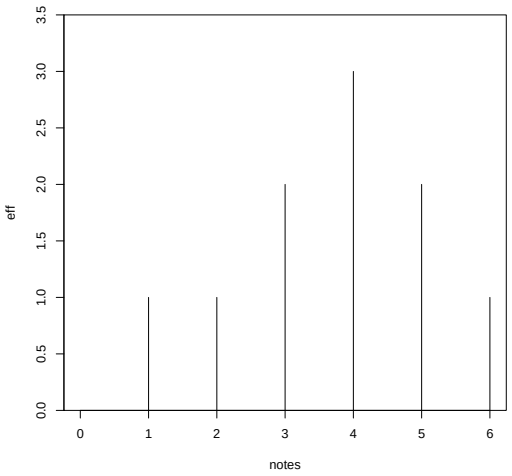
1.3 Variable discrète quantitative

Prenons l'exemple suivant d'une série de notes d'un examen noté sur dix points.

x_i	1	2	3	4	5	6
n_i	1	1	2	3	2	1
f_i	0.1	0.1	0.2	0.3	0.2	0.1

La série de note est donnée par la variable X et ses valeurs $x_1 = 1$, $x_2 = 2$, $\cdots$, appelées modalités de X, sont classées dans l'ordre croissant. Le nombre total d'individus est $N = 10$, l'effectif de la note x_i dans la population est n_i. Enfin, on a calculé la fréquence f_i de la note x_i qui représente la proportion d'individus ayant eu la note x_i dans la population. On a $f_i = n_i/N$.

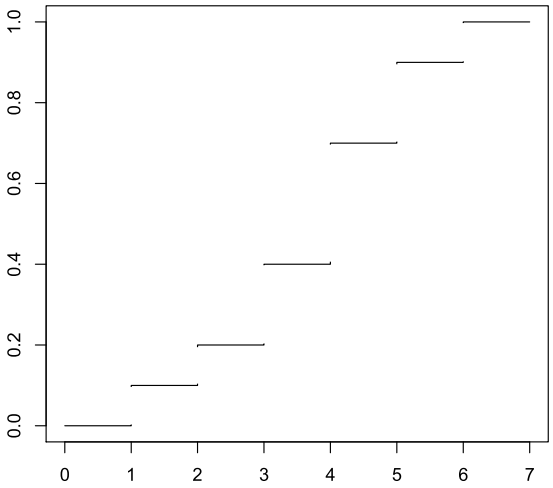
Il est à noter qu'il est totalement équivalent de parler en terme d'effectif n_i ou en terme de fréquences f_i.
On représentera graphiquement la série précédente par un diagramme en bâtons :



On pourra également tracer la fonction de répartition. Pour cela on ajoute la ligne des effectifs cumulés (*eff. c.*) et celle des fréquences cumulées (*f.c.*) au tableau précédent.

x_i	1	2	3	4	5	6
eff.c.	1	2	4	7	9	10
f.c.	0.1	0.2	0.4	0.7	0.9	1

Ces lignes permettent de répondre facilement à la question *combien d'élèves ont eu une note inférieure ou égale à* 4.
Le graphe des fréquences cumulées en fonction des valeurs de X est appelé fonction de répartition.

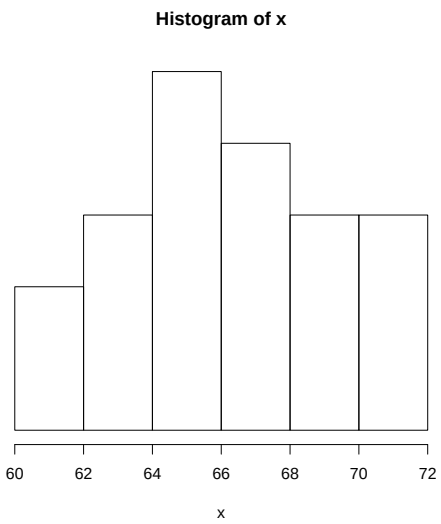


1.4 Variable continue quantitative

On a mesuré le poids de $N = 20$ personnes. Ici le caractère x, le poids, est une variable continue. On regroupe les mesures en classe et on obtient le tableau suivant :

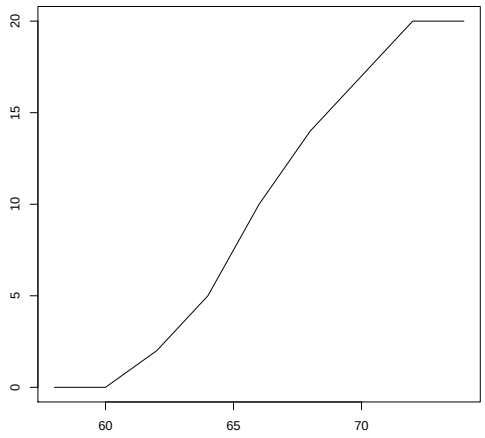
x_i	[60,62[	[62,64[	[64,66[	[66,68[	[68,70[	[70,72[
n_i	2	3	5	4	3	3

Il est également possible de s'exprimer en terme de fréquences. On représente ces données par un histogramme :



Remarque 1.1 *Sur l'histogramme, l'effectif se lit en terme d'aire des rectangles. Dans cet exemple (et c'est souvent le cas), les largeurs des classes sont les mêmes et donc les hauteurs des rectangles sont proportionnelles aux effectifs.*

Dans cette organisation en classe, on suppose qu'au sein d'une classe l'effectif est equiréparti, i.e. il y aura autant de personnes ayant un poids compris entre 60 et 61 kg qu'entre 61 et 62 kg. On a donc la fonction de répartition suivante :



1.5 Analyse

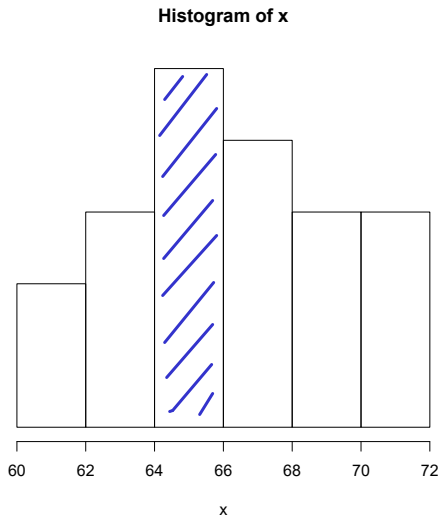
La représentation graphique nous donne une première idée de la série statistique. On peut ensuite calculer certaines caractéristiques pour avoir une analyse plus fine de la situation.

1.6 Caractéristiques de position

1.6.1 Le mode

Le mode, ou classe modale, d'une série statistique correspond intuitivement à la valeur du caractère (modalité.s) qui regroupe le plus grand effectif.

- Pour une variable discrète, c'est la modalité avec le plus grand effectif. Dans l'exemple 1.3, le mode est $x_4 = 4$ avec un effectif de 3.
- Pour une variable continue, c'est la classe $[a_i, a_{i+1}[$ dont la barre sur l'histogramme est la plus haute.



Autrement dit, la classe modale est la classe $[a_i; a_{i+1}[$ qui maximise la quantité

$$\frac{n_i}{l([a_i; a_{i+1}[)},$$

où l est la longueur de l'intervalle. Ici, la classe modale est $[64; 66[$.

1.6.2 Médiane, quartiles, quantiles

La médiane est une valeur du caractère qui sépare le classement en deux groupes de même taille. Le premier quartile (resp. le troisième) partage le classement dans les proportions 0.25 et 0.75 (resp. 0.75 et 0.25). On peut voir les quartiles comme les médianes des demi-groupes séparés par la médiane.

Plus généralement, le quantile d'ordre α, noté q_α est une valeur du caractère qui sépare le classement dans les proportions α et $(1 - \alpha)$.

Proposition 1.2 *La médiane est un antécédant de 0.5 par la fonction de répartition. Le quantile d'ordre α est un antécédant par la fonction de répartition de α.*

Dans l'exemple 1.3, l'effectif est de 10. On cherche donc à faire deux groupes de 5. Le milieu du classement est donc entre le cinquième et le sixième élève. Logiquement, on va prendre le milieu des notes des deux élèves concernés : $(4 + 4)/2 = 4$.

1.6.3 Moyenne

La moyenne est la valeur du caractère qu'auraient tous les individus de la population s'ils étaient identiques. Il s'agit du point d'équilibre (centre de gravité) du diagramme en batons ou de l'histogramme. En pratique, il s'agit de la moyenne des modalités pondérée par l'effectif :

$$m(x) = \bar{x} = \sum_i x_i \frac{n_i}{N}.$$

Dans l'exemple 1.3, la moyenne vaut 3.7.

Remarque 1.3 *Dans le cas d'une série continue regroupée en classes, il convient de remplacer les valeurs x_i du caractère par c_i, les valeurs du centre de la classe.*

Proposition 1.4 (Linéarité de la moyenne) *Soit a, b, c des réels, x, y des séries statistiques. On a*

$$m(ax + by + c) = am(x) + bm(y) + c.$$

1.7 Caractéristiques de dispersion

Il arrive que deux séries aient les mêmes caractéristiques de position mais soient totalement différentes. En effet, ces derniers indicateurs (hors quantiles) ne s'intéressent pas aux différences au sein de la population et donnent le stéréotype d'une population où tous les individus sont identiques. Pour aller plus loin dans l'étude statistique il nous faut donc introduire les notions suivantes concernant la dispersion des données.

1.7.1 Étendue

Il s'agit de l'écart entre la plus grande et la plus petite valeur prise par la série : $max\{x_i\} - min\{x_i\}$.

1.7.2 Variance et écart-type

La variance et l'écart–type sont des mesures des écarts à la moyenne : $(x_i - \bar{x})$. La variance pondère ces écarts au carré :

$$Var(x) = \sum_i (x_i - \bar{x})^2 \frac{n_i}{N}.$$

La variance est donc très influencée par les valeurs éloignées de la moyenne.

Proposition 1.5 *En pratique, on pourra calculer*

$$Var(x) = m(x^2) - (\bar{x})^2,$$

où

$$m(x^2) = \sum_i x_i^2 \frac{n_i}{N}.$$

Le résultat se vérifie en développant et en factorisant dans la formule initiale de la variance.

On définit l'écart-type σ par la racine carrée de la variance : $\sigma = \sqrt{Var(X)}$. Ainsi, l'écart-type renvoie à la distance pondérée entre les valeurs de la série et la moyenne. Il en découle la propriété suivante.

Proposition 1.6 *Soit a, b des réels. On a*

$$\sigma_{ax+b} = |a|\sigma_x, \qquad Var(ax + b) = a^2 Var(x).$$

1.8 Boîte de distribution

La boîte de distribution (ou boxplot) est une représentation graphique synthétique de la distribution des données. Elle résume quelques caractéristiques de position et de dispersion du caractère étudié (médiane, quartiles, minimum et maximum). Ce diagramme est utilisé essentiellement pour comparer un même caractère dans des populations différentes, ou une évolution au cours du temps.

(i) Tracer un rectangle qui s'étend du premier quartile au troisième.

(ii) Séparer ce rectangle en deux à la hauteur de la médiane. On obtient alors une boîte.

(iii) On complète ce rectangles par deux segments. Pour cela, on calcule

$$a = q_{0.25} - 1.5IQ \quad \text{et} \quad b = q_{0.75} + 1.5IQ,$$

avec la distance inter-quartile

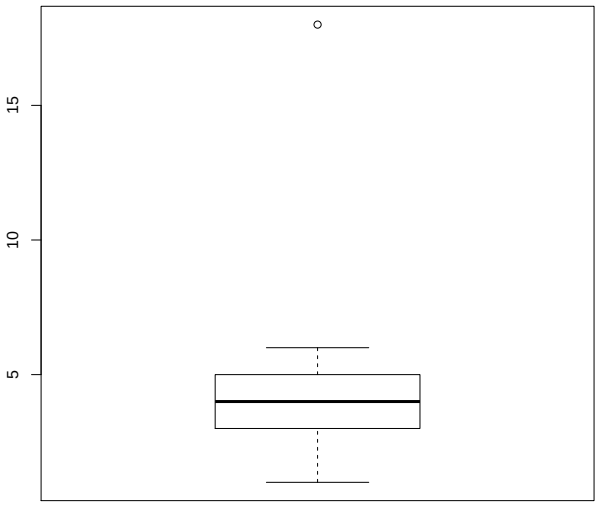
$$IQ = q_{0.75} - q_{0.25}.$$

On repère les valeurs :

$$x_a = \min\{x_i : x_i \geq a\} \quad \text{et} \quad x_b = \max\{x_i : x_i \leq b\}.$$

Ces valeurs sont appelées *valeurs adjacentes*. On relie ces valeurs aux cotés de la boîte.

(iv) Les valeurs qui ne sont pas comprises entre les valeurs adjacentes sont représentées par des points et sont appelées *valeurs extrêmes*.



Dans l'exemple 1.3 et en ajoutant un 18 (valeur extrême), on obtient la boîte de distribution suivante.

Part II

Distributions statistiques bivariées

Dans ce chapitre, on s'intéresse aux relations entre deux variables notées X et Y. Supposons que l'on observe ces deux variables sur n unités statistiques. À chaque individu i, on peut associer un couple d'observations (x_i, y_i). Chaque variable peut être quantitative ou qualitative. Dans cette première partie, on s'intéresse à la présentation des données et nous proposons un indicateur d'indépendance.

Plus tard, on pourra s'intéresser à la modélisation d'une relation linéaire entre les deux caractères étudiés.

2 Distributions

2.1 Distributions conjointe, marginale, conditionnelle

Notons $m_1^X, \cdots, m_J^X$ les J modalités de X et $m_1^Y, \cdots, m_K^Y$ les K modalités de Y. Si l'une des deux variables (ou les deux) est quantitative continue, les m_j^X ou les m_k^Y sont des classes modales. Introduisons les quantités suivantes :

- n_{jk} est le nombre de fois où le couple (X, Y) prend la modalité (m_j^X, m_k^Y),
- $n_{\bullet k}$ est le nombre de fois où la variable Y prend la valeur m_k^Y,
- $n_{j\bullet}$ est le nombre de fois où la variable X prend la valeur m_j^X.

On a

$$\sum_{j=1}^{J} n_{jk} = n_{\bullet k} \quad \text{et} \quad \sum_{k=1}^{K} n_{jk} = n_{j\bullet}$$

$$\sum_{k=1}^{K}\sum_{j=1}^{J} n_{jk} = \sum_{j=1}^{J} n_{j\bullet} = \sum_{k=1}^{K} n_{\bullet k} = n$$

Les données peuvent être représentées dans un tableau à double entrée appelé **Tableau de contingence**.

	m_1^Y	$\ldots$	m_k^Y	$\ldots$	m_K^Y	total
m_1^X	n_{11}	$\ldots$	n_{1k}	$\ldots$	n_{1K}	$n_{1\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_j^X	n_{j1}	$\ldots$	n_{jk}	$\ldots$	n_{jK}	$n_{j\bullet}$
$\vdots$	$\vdots$	$\ldots$	$\vdots$	$\ldots$	$\vdots$	$\vdots$
m_J^X	n_{J1}	$\ldots$	n_{Jk}	$\ldots$	n_{JK}	$n_{J\bullet}$
	$n_{\bullet 1}$	$\ldots$	$n_{\bullet k}$	$\ldots$	$n_{\bullet K}$	n

Le **tableau des fréquences** s'obtient en divisant les effectifs par le nombre d'unités statistiques n (effectif total). Comme précédemment on obtient

$$f_{jk} = \frac{n_{jk}}{n}, \quad f_{\bullet k} = \frac{n_{\bullet k}}{n} \quad f_{j\bullet} = \frac{n_{j\bullet}}{n}$$

	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
m_1^X	f_{11}	$\dots$	f_{1k}	$\dots$	f_{1K}	$f_{1\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_j^X	f_{j1}	$\dots$	f_{jk}	$\dots$	f_{jK}	$f_{j\bullet}$
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_J^X	f_{J1}	$\dots$	f_{Jk}	$\dots$	f_{JK}	$f_{J\bullet}$
	$f_{\bullet 1}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet K}$	1

2.2 Distributions marginales

A partir du tableau de contingence, on peut retrouver la distribution de chacune des variables séparément :

Modalité de Y	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
Fréquence empirique	$f_{\bullet 1}$	$\dots$	$f_{\bullet k}$	$\dots$	$f_{\bullet K}$	1

Modalité de X	m_1^X	$\dots$	m_j^X	$\dots$	m_J^X	total
Fréquence empirique	$f_{1\bullet}$	$\dots$	$f_{j\bullet}$	$\dots$	$f_{J\bullet}$	1

Les distributions de X et de Y sont appelées distributions marginales. Sur chaque variable, on peut calculer les indicateurs habituels (moyenne, variance, écart type si la variable est quantitative...). Ces paramètres sont qualifiés d'indicateurs marginaux.

2.3 Distributions conditionnelles

La ligne j du tableau de contingence représente la répartition sur les modalités (ou classes modales) $(m_1^Y, \cdots, m_K^Y)$ des individus pour lesquels le caractère X vaut m_j^X. Si on divise les lignes ou les colonnes par leur somme, on obtient les distributions empiriques constituées des fréquences conditionnelles. Pour $j = 1, \cdots, J$ et $k = 1, \cdots, K$ notons :

$$f_{k|j} = \frac{n_{jk}}{n_{j\bullet}} = \frac{f_{jk}}{f_{j\bullet}}.$$

La fréquence $f_{k|j}$ peut se lire fréquence de la modalité m_k^Y sachant que X prend la modalité m_j^X.

On peut alors construire le tableau des profils ligne :

	m_1^Y	$\dots$	m_k^Y	$\dots$	m_K^Y	total
m_1^X	$f_{1\|1}$	$\cdots$	$f_{k\|1}$	$\cdots$	$f_{K\|1}$	1
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_j^X	$f_{1\|j}$	$\cdots$	$f_{k\|j}$	$\cdots$	$f_{K\|j}$	1
$\vdots$	$\vdots$	$\dots$	$\vdots$	$\dots$	$\vdots$	$\vdots$
m_J^X	$f_{1\|J}$	$\cdots$	$f_{k\|J}$	$\cdots$	$f_{K\|J}$	1

Les profils colonnes sont les fréquences en colonne i.e. :

$$f_{j|k} = \frac{n_{jk}}{n_{\bullet k}} = \frac{f_{jk}}{f_{\bullet k}}.$$

Sur ces distributions conditionnelles, on peut calculer nos principaux indicateurs (moyenne, variance, écart type si la variable est quantitative...) conditionnels, c'est-à-dire sous a condition $X = m_j^X$ ou $Y = m_k^Y$.

3 Quantification de la dépendance

3.1 Statistique du χ^2

En présence de deux variables, l'un des enjeux principaux est d'étudier (c'est à dire quantifier voire expliquer) la dépendance entre les deux caractères.

Si on était dans le cadre des probabilités, ce qui n'est pas le cas, alors deux caractères sont indépendants si la valeur de l'un n'a aucune influence sur la distribution de l'autre. Si tel était le cas, alors les distributions conditionnelles seraient toutes semblables à la distribution marginale. Pour tout (j, k), on devrait avoir

$$f_{j|k} = f_{j\bullet} \qquad \text{et} \qquad f_{k|j} = f_{\bullet k}.$$

Ainsi, on aurait :

$$f_{kj} = f_{j|k} f_{\bullet k} = f_{j\bullet} f_{\bullet k}.$$

D'où, si les deux variables étaient indépendantes, on aurait

$$n_{kj} = \frac{n_{\bullet j} n_{k\bullet}}{n}.$$

En statistiques, on ne peut que "quantifier la distance à l'indépendance" par la statistique du χ^2,

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}}.$$

On peut remarquer que

$$D_{\chi^2} = n \left(\sum_{j=1}^{J} \sum_{k=1}^{K} \frac{n_{jk}^2}{n_{j\bullet} n_{\bullet k}} - 1 \right),$$

ou de façon équivalente

$$D_{\chi^2} = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j\bullet} n_{\bullet k}}{n}\right)^2}{\frac{n_{j\bullet} n_{\bullet k}}{n}},$$

où J et K sont le nombre de modalités de chacune des deux variables considérées.

Le cas d'indépendance probabiliste serait alors équivalent à $D_{\chi^2} = 0$. Puisque nous ne sommes pas dans le cadre des probabilités et que nous laissons aux série statistiques une certaine fluctuation due à l'onservation, nous accepterons l'indépendance, même si D_{χ^2} n'est pas tout à fait nulle. Nous proposons, dans ce qui suit, des règles simples pour décider de quand conclure à une dépendance et de quelle intensité.

3.2 Exemple

À partir de 200 dossiers d'une agence immobilière, on recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

(i) On donne le tableau des fréquences.
Pour calculer les fréquences, on divise chaque effectif par l'effectif total (ici 200) :

	Célibataire	En couple	Total
Dossier accepté	0.17	0.29	0.46
Dossier refusé	0.33	0.21	0.54
Total	0.5	0.5	1

(ii) On calcule la statistique du Chi-deux.
La statistique du Chi-deux est donnée par :

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}}.$$

Ici on a donc :

$$\begin{aligned} D_{\chi^2} =& 200\Bigg(\frac{(0.17 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.29 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.33 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \\ & + \frac{(0.21 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \Bigg) \\ =& 200\Big(0.016 + 0.016 + 0.013 + 0.013 \Big) \\ =& 11.6 \end{aligned}$$

3.3 Mesure d'association

Nous proposons trois mesures d'association, le ϕ, le C et le V de Cramér issues de la distance du χ^2.

3.3.1 Coefficients ϕ et C

Les coefficients ϕ et C découlent de la statistique du χ^2 par les formules

$$C = \sqrt{\frac{D_{\chi^2}}{D_{\chi^2} + n}}, \qquad \phi = \sqrt{\frac{D_{\chi^2}}{n}}.$$

En réalité ces deux coefficients sont une variante l'un de l'autre. L'avantage de C est qu'il est compris entre 0 et 1, alors que ce n'est pas le cas pour le ϕ. Plus ces indicateurs sont proche de zéro, plus il y a indépendance entre les deux variables X et Y étudiées.

3.3.2 V de Cramér

Comme pour le coefficient ϕ, plus le V de Cramér est proche de zéro, plus il y a indépendance entre les deux variables X et Y étudiées. Il vaut 1 en cas de complète dépendance.

Le coefficient V de Cramér nécessite l'utilisation de la statistique du χ^2 via la formule

$$V = \sqrt{\frac{D_\chi^2}{n \times min\{l-1; c-1\}}},$$

où n est l'effectif total de la population, c est le nombre de colonnes (nombre de modalités de Y) et l le nombre de lignes (modalités de X).

3.3.3 Interprétation

L'interprétation des coefficients ϕ et V est empirique et dépend du domaine d'application (sciences économiques, sciences humaines, médecine...). On peut considérer le tableau suivant pour l'interprétation (tout en vérifiant les valeurs frontières d'usage dans chaque domaine).

Valeur du V de Cramér	Intensité de la relation entre les variables
inférieur à 0,10	relation nulle ou très faible
entre 0,10 et 0,20	relation faible
entre 0,20 et 0,30	relation moyenne
au dessus de 0,30	relation forte

Il reste que ces mesures d'association sont difficiles à interpréter. Cependant quand elle sont correctement standardisées, elles peuvent être utilisées pour comparer les forces de la liaison de différents tableaux.

- Par exemple, il se peut que des tableaux fassent intervenir les mêmes variables avec les mêmes modalités mais que les observations proviennent de régions différentes.
- Il peut être opportun de se demander si la liaison ne serait pas plus forte dans une région que dans une autre.

3.4 Exemple (suite)

En reprenant notre exemple de dossiers acceptés ou refusés, on rappelle l'effectif total $n = 200$ et le nombre de modalités de chaque variable : 2. On se rappelle aussi de la statistique $D_{\chi^2} = 11,6$. On en déduit que le V de Cramér vaut :

$$V = \sqrt{\frac{11,6}{200 \times min\{2-1; 2-1\}}} = \sqrt{\frac{11,6}{200}} = 0.24$$

La relation entre les variables est moyenne.

Part III

Régressions linéaires

4 Statistiques à deux variables (séries doubles)

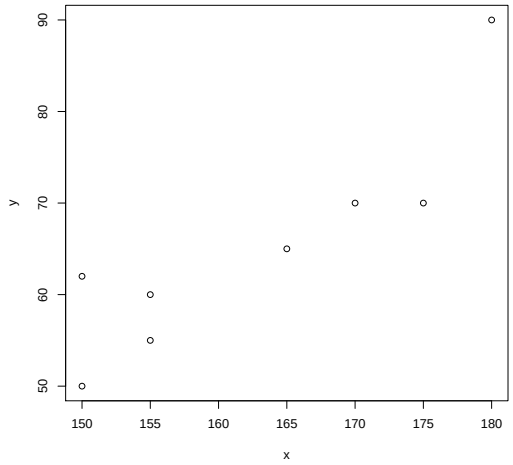
Dans certaines situations, on est amené à étudier deux caractères distincts d'une même population. On peut par exemple considérer la taille (x) et le poids (y) d'un ensemble d'individus. L'objectif principal de l'étude est de déterminer l'éventuel lien entre les deux variables x et y.

4.1 Nuage de points

On relève le couple (taille, poids) de 8 individus. On résume les données dans le tableau suivant.

taille	x	150	155	155	150	165	175	170	180
poids	y	50	55	60	62	65	70	70	90

Definition 4.1 *Soit une population de N individus. Le graphe des N points (x_i, y_i) est appelé nuage de points de la série.*



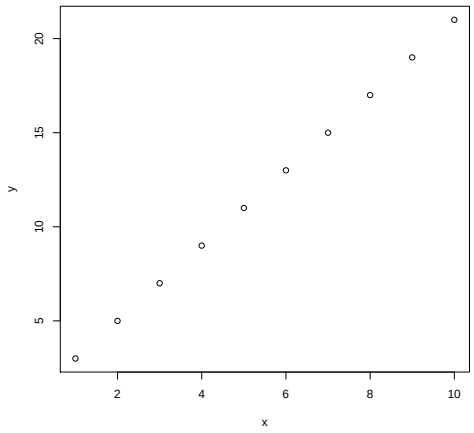
Definition 4.2 *Le point ayant pour coordonnées les moyennes $(\bar{x}, \bar{y})$ est appelé le point moyen.*

Il s'agit du centre de gravité du nuage. On rencontrera parfois cette dénomination. Dans notre exemple, le point moyen est $(65.2, 162.5)$.

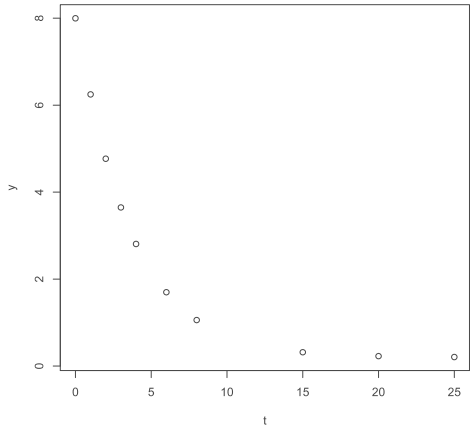
4.2 Forme du nuage de points

D'une manière générale, trois cas peuvent se présenter en ce qui concerne le profil du nuage.

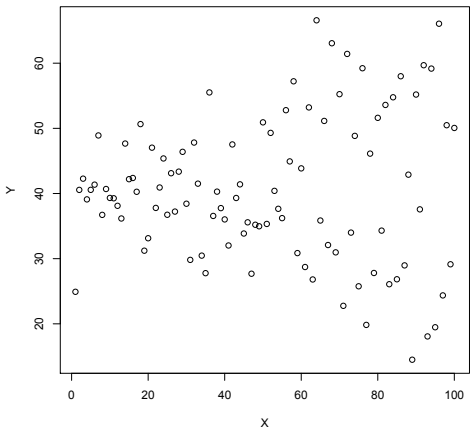
(i) forme allongée et rectiligne : les points sont plus ou moins alignés



(ii) forme allongée mais non rectiligne : les points ne sont pas alignés mais ont un profil ordonné



(iii) forme quelconque



5 Ajustement affine (droite de régression linéaire)

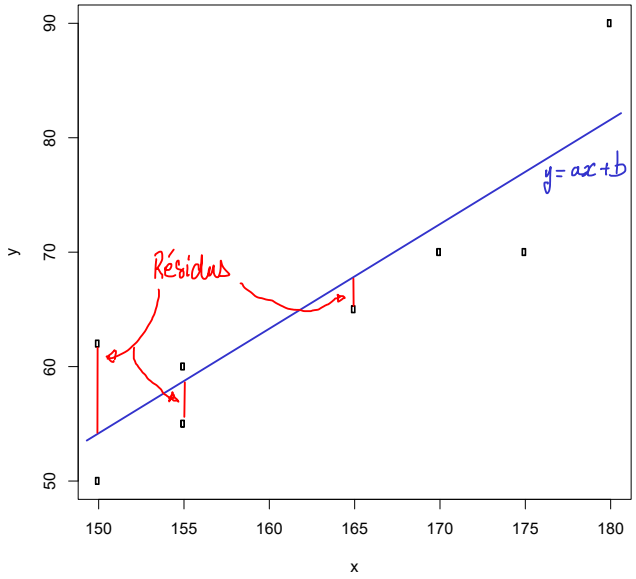
On s'intéresse plus particièrement au premier cas 4.2.1. Procéder à un ajustement affine revient à chercher une droite D d'équation

$$y = ax + b$$

qui passe au plus proche des points du nuage de points. Cette droite nous servira donc d'approximation. Bien évidemment, suivant la méthode utilisée pour la construire, on peut obtenir différentes droites. La méthode la plus utilisée car donnant la meilleure approximation est la méthode des moindres carrés.

5.1 La méthode des moindres carrés

L'idée de cette méthode est de chercher la droite$y = ax + b$ qui minimise la somme des carrés des écarts verticaux entre la droite et les points du nuage, les *résidus*.



La droite ainsi obtenue est unique. Cette droite s'appelle la droite de régression linéaire de y en x par la méthode des moindres carrés. On note

$$Cov(x, y) = \sum (x_i - \bar{x})(y_i - \bar{y})/N.$$

Cette quantité est nommée covariance de x et y. En pratique, on pourra calculer cette quantité par la formule

$$Cov(x, y) = m(xy) - \bar{x}\bar{y},$$

où

$$m(x, y) = \frac{1}{N} \sum_{i=1}^{N} x_i y_i.$$

Si la quatité σ_x est une distance entre les valeurs de x est $\bar{x}$, on peut considérer la covariance comme un produit scalaire entre les variables x et y. Ainsi, si la covariance est proche de 0, on peut penser que les variables ont une dynamique qui n'ont rien de commun (penser à l'orthogonalité), c'est à dire le nuage 4.2.3.

On a

$$\begin{aligned} a &= cov(x,y)/\sigma_x^2, \\ b &= \bar{y} - a\bar{x}. \end{aligned}$$

En pratique, on détermine les coefficients de la droite $D : y = ax + b$ à l'aide d'un tableur ou de `R`.

5.2 Coefficient de corrélation linéaire

Notons que la méthode des moindres carrés peut être utilisée pour n'importe quelle série double. On peut tout à fait obtenir une droite de régression dans le cas 4.2.3. Pour s'assurer de façon objective (et non purement visuelle) que l'ajustement est valide, on considère un autre paramètre de la série : le coefficient de corrélation r :

$$r = \frac{cov(x,y)}{\sigma_x \sigma_y}$$

Proposition 5.1 *On a les propriétés suivantes :*

(i) on a toujours $-1 \leq r \leq 1$ *;*

(ii) le coefficient directeur de la droite de régression et le coefficient de corrélation sont de même signe ;

(iii) le degré de corrélation est d'autant plus fort que r *est proche de* 1 *ou* -1*.*

C'est l'assertion 3.*iii* qui nous permet de dire si la droite de régression est proche des points. En pratique, une régression linéaire est légitime si $r > 0.9$ ou si $r < -0.9$.

5.3 Coefficient de détermination R^2

La variance est une bonne mesure de l'hétérogénéité d'une série (contrairement à la moyenne qui considère tous les individus comme semblables). La variance de la série x se décompose comme la variance expliquée par la droite de régression plus celle de l'erreur (résidus) :

$$Var(y) = Var(ax + b) + Var(e).$$

Le coefficient de détermination est le rapport de variance de y expliquée par la régression :

$$R^2 = \frac{Var(ax + b)}{Var(y)}. \tag{5.1}$$

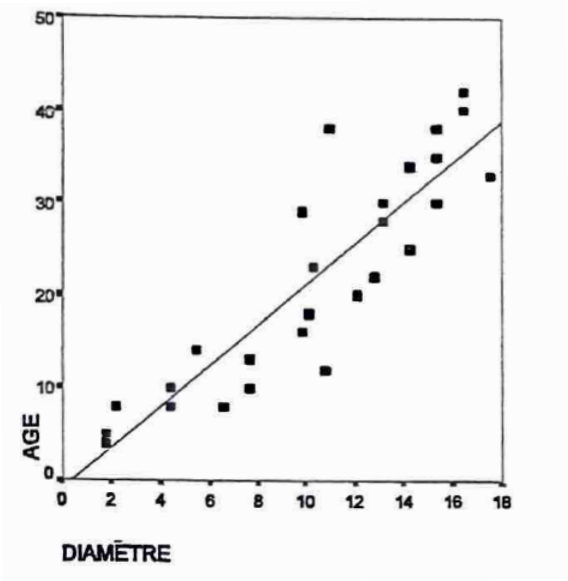
Il se trouve que R^2 est le carré du coefficient de corrélation linéaire r_{xy}.

5.4 Exemple

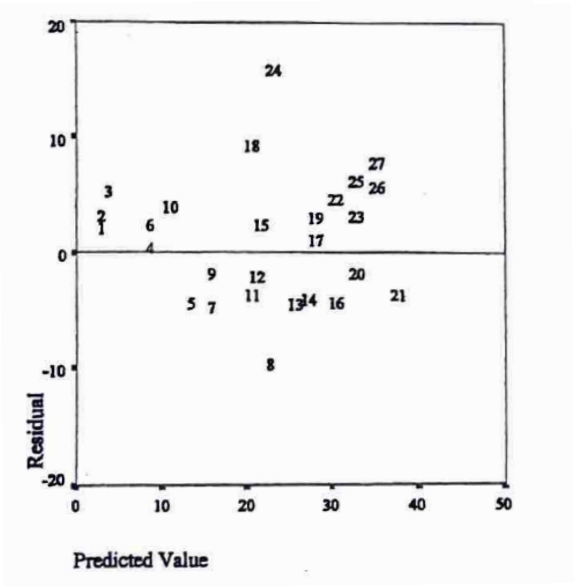
On sait que l'on mesure l'âge d'un arbre en comptant les anneaux sur une section transversale du tronc, mais cela nécessite de l'avoir abattu auparavant. Peut-on connaître l'âge à partir de la mesure de sa circonférence ?

Afin de répondre à cette question, on a effectué les mesures sur un échantillon de 27 arbres de la même espèce. À partir de ces données, on a effectué une régression de l'âge en fonction du diamètre. Les résultats ont été traités à l'aide du logiciel SPSS.

Nuage de points



Résidus



Résumé de la régression

	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	2905,549	1	2905,55	93,44	,000
Résidu	777,414	25	31,097		
Total	3682,963	26			

	Coefficients non standardisés		Coefficients standardisés		
	B	Erreur standard	Bêta	t	Signification
(constante)	-,974	2,604		-,374	,711
DIAMETRE	2,206	,228	,888	9,67	,000

En bref, on considère la régression linéaire d'équation

$$AGE = 2,206 \times DIAMETRE - 0.974.$$

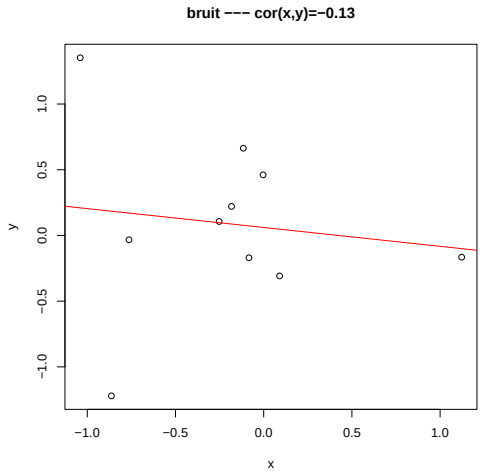
Le coefficient $R = 0,888$ est un peu faible pour établir assurer une forte liaison linéaire. Les points seront éloignés de la droite de régression. On se contentera d'une tendance. En regardant le graphe des résidus, il semble que l'arbre 24 soit une valeur extrême, qu'il faudrait peut-être ignorer.

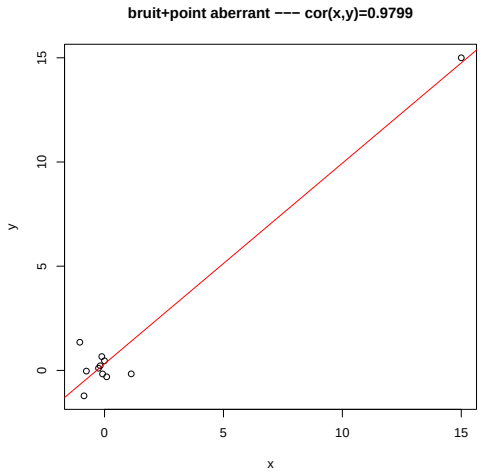
6 Discussions

6.1 Manipulation du coefficient r

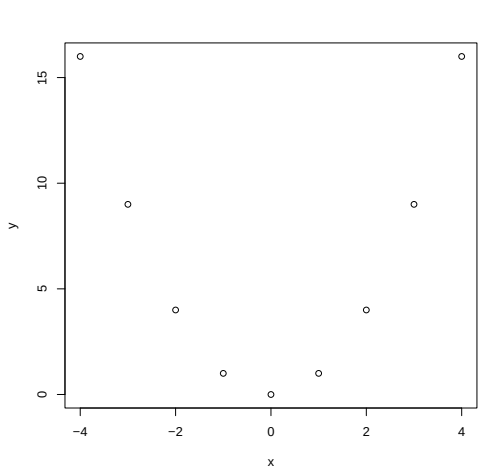
Le coefficient de corrélation linéaire r ou le coefficient de détermination R^2 mesurent le caractère rectiligne du nuage. Il ne suffira pas à décrire, seul, la pertinence d'un ajustement.

Il faudra, en particulier, faire attention aux points aberrants.





Notons également qu'ils passent pleinement à coté d'une forme en cloche. Bien qu'il y ait une liaison entre les variables, la nature de cette liaison n'est pas linéaire.



6.2 Régression $x = my + p$

Le choix de représenter y en fonction de x est bien souvent arbitraire. Lorsque le caractère y dépend du caractère x par un lien de cause à effet clair (concentration y d'un composant lors d'une réaction chimique en fonction du temps x), on utilisera bien entendu la régression $y = ax + b$. Dans le cas d'une interdépendance (chiffre d'affaire x et budget publicité y), le choix de la régression $y = ax+b$ ou $x = my+p$ se pose. En fait, les deux régressions sont tout à fait valides (ainsi que toute droite située entre les deux).

Dans les faits, puisque nous choisissons d'effectuer des régressions linéaires uniquement dans le cas où le coefficient de corrélation r est proche de 1 ou -1, ces deux droites seront très proches et amènent aux mêmes conclusions.

6.3 Changement de variable

Revenons au cas (2). La relation entre les deux caractères n'est pas linéaire. Une possibilité est de se ramener à une relation linéaire par un changement de variable deviné d'après la forme du nuage de points. On considère alors la régression

$$y = af(x) + b,$$

ou

$$y = f(ax + b),$$

avec f la fonction issue du changement variable et a et b les coefficients de la régression linéaire des moindres carrés. Le choix de la fonction f sera validé par la qualité de l'ajustement obtenu et donc par la valeur du coefficient de corrélation linéaire du nouveau problème.

Les courbes de type exponentielle, logarithme ou logistique sont bien adaptées à ce procédé. On évitera cependant les régressions polynômiales du type

$$y = a + bx + cx^2 + dx^3 + \cdots$$

qui ne permettent pas la prévision en dehors de l'intervalle d'étude.

6.4 Régressions linéaire multiple

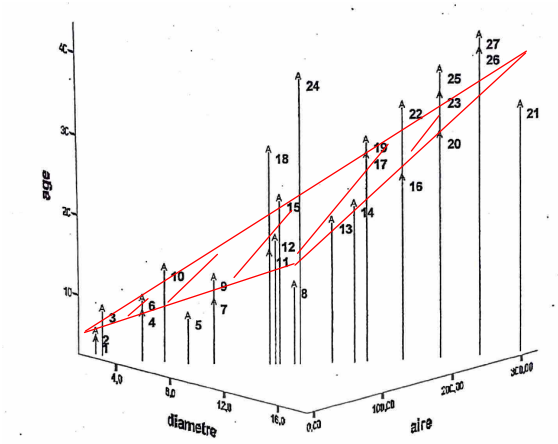
On peut souhaiter expliquer une variable réponse y par plusieurs variables explicatives $x_1, \cdots x_d$ en utilisant une équation cartésienne :

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d,$$

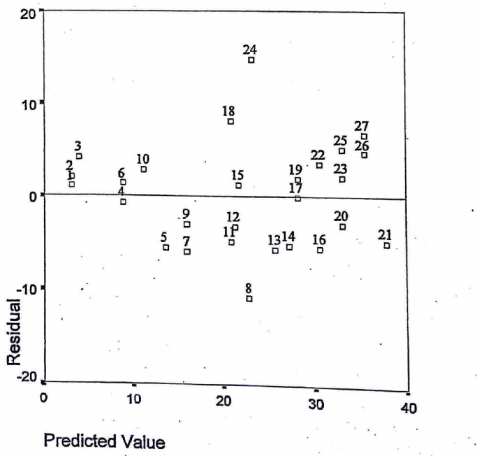
où les coefficients $\beta_0, \beta_1, \cdots, \beta_d$ sont les coefficients de la régression. On pourra généraliser directement le coefficient de détermination comme la proportion de variance de y expliquée par le modèle. On ne rentrera pas plus dans les détails des calculs qui deviennent très abstraits.

Reprenons notre exemple d'âge des arbres, et intéressons nous au diamètre et à la surface de la section (les arbres ne sont pas des cylindres parfaits, le diamètre n'explique pas totalement la surface). On obtient les graphes suivants

Nuage de points



Résidus



Résumé de la régression

	Coefficients non standardisés			
	B	Erreur standard	R	R-deux
(constante)	2,720	4,317		
DIAMETRE	1,161	1,001		
AIRE	,055	,051	,894[a]	,799

On considère donc la régression linéaire d'équation

$$AGE = 1,161 \times DIAMETRE + 0.055 \times AIRE + 2,72.$$

Le coefficient $R^2 = 0,799$ est un peu faible pour établir assurer une forte liaison linéaire. Les points seront éloignés de la droite de régression. On se contentera d'une tendance. En regardant le graphe des résidus, on fait la même remarque que dans la première régression, il semble que l'arbre 24 soit une valeur extrême, qu'il faudrait peut-être ignorer.

Dans ce problème, il semble que la connaissance de la surface de la section n'apporte pas grand chose.

Part IV

Notions sur les tests statistiques

Les tests statistiques forment une théorie assez complexe (qui nous occupera largement dans les prochaines années) dans le domaine des statistiques mais dont la formulation simple lui confère une grande popularité dans l'ensemble des domaines scientifiques. En effet, si nous reprenons la question de l'indépendance de deux variables, que nous avons traité en parlant de la distance du χ^2 (le nom s'éclairera dans les années à venir), nous pouvons procéder par la démarche d'un test statistique (test d'hypothèse) et poser la question "Est-ce que mes variables sont indépendantes ?". Le test, que nous allons détailler nous donnera la réponse par "oui" ou "non", avec une probabilté de se tromper vu la quantité de données.

Malgré sa popularité, son caractère catégorique et son apparente puissance, la théorie des tests n'est pas sans défaut. Nous aurons, dans la suite de la formation, le temps d'entrevoir certaines de ses limites.

7 Généralités sur les tests et test du χ^2

Détaillons le test du χ^2 en reprenant l'exemple du cours de la section distance du χ^2.

7.1 L'Hypothèse nulle

7.1.1 Définition

L'hypothèse nulle $\mathcal{H}_0$ est la proposition que l'on va chercher à rejeter sur notre jeu de donnée. On la formulera en termes simples.

7.1.2 Test du χ^2

Pour le test du χ^2, l'hypothèse nulle est $\mathcal{H}_0 =$"les deux variables sont indépendantes".

7.1.3 Exemple

Reprenons l'exemple du paragraphe 3. On recense les réponses positives et négatives selon la situation maritale du demandeur (célibataire ou en couple). On obtient les résultats suivants :

	Célibataire	En couple
Dossier accepté	34	58
Dossier refusé	66	42

On donne le tableau des fréquences.
Pour calculer les fréquences, on divise chaque effectif par l'effectif total (ici 200) :

	Célibataire	En couple	Total
Dossier accepté	0.17	0.29	0.46
Dossier refusé	0.33	0.21	0.54
Total	0.5	0.5	1

L'hypothèse nulle est $\mathcal{H}_0 =$"les variables *situation maritale* et *réponse* sont indépendantes".

7.2 La variable du test (Statistique du test)

7.2.1 Définition

De l'hypothèse nulle $\mathcal{H}_0$, on construit une quantité mathématique, une variable aussi appelée statistique du test. Elle doit nous aider à quantifier, caractériser mathématiquement la notion de l'hypothèse nulle.

7.2.2 Test du χ^2

Pour le test du χ^2, on va quantifier la distance à l'indépendance par la distance du χ^2, notée D_{χ^2}, définie dans le paragraphe 3.

7.2.3 Exemple

Dans notre exemple, on avait calculé la statistique du Chi-deux grâce à la formule

$$D_{\chi^2} = n \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{jk} - f_{j\bullet} f_{\bullet k})^2}{f_{j\bullet} f_{\bullet k}}.$$

On avait donc :

$$\begin{aligned} D_{\chi^2} =& 200 \Bigg(\frac{(0.17 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.29 - 0.46 \times 0.5)^2}{0.46 \times 0.5} + \frac{(0.33 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \\ & + \frac{(0.21 - 0.54 \times 0.5)^2}{0.54 \times 0.5} \Bigg) \\ =& 200 \Big(0.016 + 0.016 + 0.013 + 0.013 \Big) \\ =& 11.6 \end{aligned}$$

7.3 Le seuil α

7.3.1 Définition

On notera α la probabilité de faire l'erreur de prendre la décision de rejeter $\mathcal{H}_0$ alors qu'elle est vraie. Cette probabilité d'erreur, qu'on ne peut pas négliger, est appelée seuil. On prendra en général pour α les valeurs 0,01 ou 0,05.

7.3.2 Exemple

On va prendre, comme très souvent, un seuil $\alpha = 5\% = 0.05$.

7.4 La zone de rejet et l'interprétation

7.4.1 Définition

Pour un seuil donné α, on peut définir, en fonction du test, une zone de rejet pour un test *bilatéral* :

$$W_\alpha =]-\infty, -w_\alpha] \cup [w_\alpha, +\infty[,$$

ou, pour un test *unilatéral* :

$$W_\alpha = [w_\alpha, +\infty[.$$

Si la statistique du test est dans cette zone, on rejettera l'hypothèse nulle $\mathcal{H}_0$ au seuil α. Sinon, on ne rejettera pas $\mathcal{H}_0$.

Les quantités w_α sont issues de quantiles théoriques de lois de probabilités. Nous le verrons dans les prochaines unités de statistiques. Elle prennent en compte le seuil α mais aussi l'effectif total et d'autres paramètres, comme le nombre de modalités...

7.4.2 Test du χ^2

Rappelons que J et K sont le nombre de modalités de chacune des deux variables considérées. Au seuil $\alpha\%$ (le plus souvent $\alpha = 5\% = 0,05$), il faut comparer D_{χ^2} au quantile d'ordre $1-\alpha\%$ à savoir $q_{1-\alpha}$ ($q_{0,95}$ le plus souvent) d'une loi du χ^2_d (loi du *Khi2 à d degrés de libertés*), où

$$d = (J-1)(K-1)$$

est le degré de liberté de la loi (c'est à dire le paramètre de la loi du χ^2).

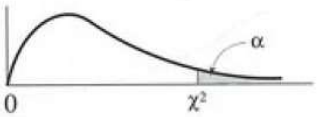
Il est à noter que les variables seront dépendantes, i.e. on rejettera $\mathcal{H}_0$, si D_{χ^2} est trop grande. L'interprétation est la suivante.

- Si $D_{\chi^2} \geq q_{1-\alpha}$, au seuil α, on rejette $\mathcal{H}_0$. On conclut que les deux variables sont dépendantes.
- Sinon, on ne rejette pas $\mathcal{H}_0$, on conclut qu'elles sont indépendantes.

7.4.3 Exemple

On compare la statistique du test 11,6 à la valeur de la table du Chi-deux à $(2-1)(2-1) = 1$ degré de liberté (2 modalités pour chaque variable).

Table χ^2 : points de pourcentage supérieurs de la distribution χ^2



dl	.995	.990	.975	.950	.900	.750	.500	.250	.100	.050	.025	.010	.005
1	0.00	0.00	0.00	0.00	0.02	0.10	0.45	1.32	2.71	3.84	5.02	6.63	7.88
2	0.01	0.02	0.05	0.10	0.21	0.58	1.39	2.77	4.61	5.99	7.38	9.21	10.60
3	0.07	0.11	0.22	0.35	0.58	1.21	2.37	4.11	6.25	7.82	9.35	11.35	12.84
4	0.21	0.30	0.48	0.71	1.06	1.92	3.36	5.39	7.78	9.49	11.14	13.28	14.86
5	0.41	0.55	0.83	1.15	1.61	2.67	4.35	6.63	9.24	11.07	12.83	15.09	16.75
6	0.68	0.87	1.24	1.64	2.20	3.45	5.35	7.84	10.64	12.59	14.45	16.81	18.55
7	0.99	1.24	1.69	2.17	2.83	4.25	6.35	9.04	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	5.07	7.34	10.22	13.36	15.51	17.54	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	5.90	8.34	11.39	14.68	16.92	19.02	21.66	23.59
10	2.15	2.56	3.25	3.94	4.87	6.74	9.34	12.55	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	7.58	10.34	13.70	17.28	19.68	21.92	24.72	26.75
12	3.07	3.57	4.40	5.23	6.30	8.44	11.34	14.85	18.55	21.03	23.34	26.21	28.30
13	3.56	4.11	5.01	5.89	7.04	9.30	12.34	15.98	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	10.17	13.34	17.12	21.06	23.69	26.12	29.14	31.31
15	4.60	5.23	6.26	7.26	8.55	11.04	14.34	18.25	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	11.91	15.34	19.37	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	12.79	16.34	20.49	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	13.68	17.34	21.60	25.99	28.87	31.53	34.81	37.15

On trouve 3.84.
On a donc $D_{\chi^2} > q_{0.95}$ et on conclue que les variables sont dépendantes : la situation maritale influence l'acceptation ou le refus du dossier.

7.5 La p-value

7.5.1 Définition

Les logiciels de statistiques (type R, Excel ...) calculent la $p-$value (ou valeur p). Il s'agit du seuil p minimal auquel on peut rejeter $\mathcal{H}_0$ vu la valeur de la statistique. C'est donc la probabilité de faire une erreur en rejetant $\mathcal{H}_0$. Ainsi, on retiendra qu'on rejettera l'hypothèse nulle si $p \leq \alpha/100$ (le plus souvent si $p \leq 0,05$.)

7.5.2 Test du χ^2

On retiendra qu'on rejettera l'hypothèse d'indépendance si $p \leq \alpha/100$ (le plus souvent si $p \leq 0,05$.)

7.5.3 Exemple

En utilisant `R`, on obtient

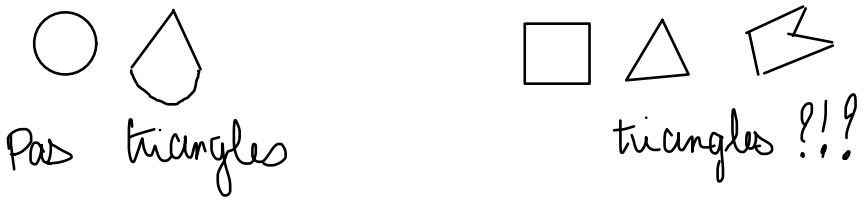
```
        Pearson's Chi-squared test

data:  contingences
X-squared = 11.594, df = 1, p-value = 0.0006616
```

On observe que la p-value est inférieure à 5%, on rejette donc l'indépendance significativement. On conclut que les variables sont dépendantes.

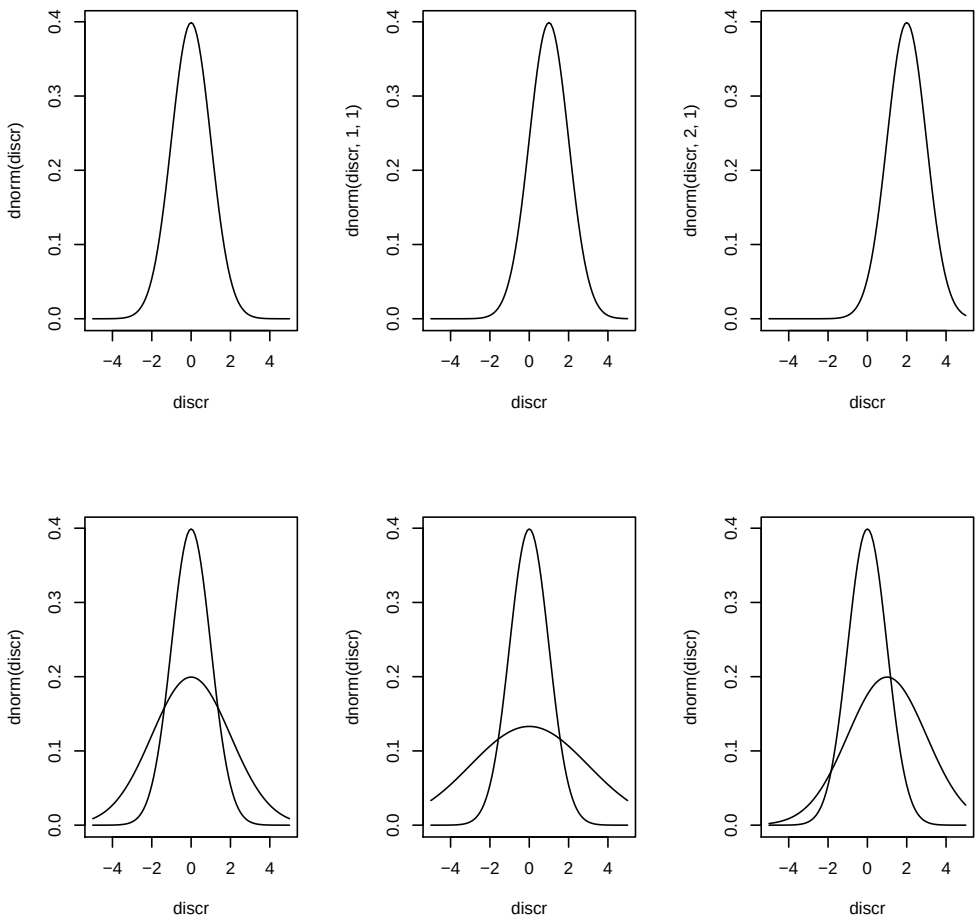
7.6 Discussions sur le test du χ^2

- Pour les tableaux de contingences 2×2 ou les effectifs trop faibles, `R` applique la correction de Yates. Cela peut légèrement modifier la valeur de la statistique du χ^2 par rapport à celle que l'on pourrait calculer à la main. Le plus souvent, cela n'a pas d'impact sur le résultat du test.
- En pratique, on évite d'utiliser le test du χ^2 si un effectif non nul du tableau est inférieur ou égal à 5.
- Ne pas rejeter $\mathcal{H}_0$ ne signifie pas, pour autant, qu'on l'accepte avec certitude. Cela signifie juste que rien de s'y oppose. Par analogie avec tout raisonnement mathématique,si on teste le fait d'être un triangle en supposant que ses bords sont des segments alors :



8 Les tests de normalité

Les lois normales (ou de Gauss) sont des lois de probabilités aux propriétés remarquables. Vous aurez l'occasion, pendant vos études, de voir leur omniprésence dans les mathématiques et les statistiques. Ces lois ont une répartition en cloche



On peut se demander si une série statistique a une répartition (histogramme ou diagramme en bâtons) en forme de cloche. On dira que la série est normalement distribuée. Pour valider ou infirmer ce caractère normal d'une distribution, on pourra un test de normalité.

8.1 Noms des différents tests de normalité

Il existe différents tests de normalité. Certains testeront l'adéquation des fonctions de répartitions quand d'autres s'intéresseront à d'autres propriétés de la loi normale. Ce sont les tests de Lilliefors, d'Anderson–Darling, d'Agostino, de Jarque–Bera, ou encore, et sans doute le plus populaire, le test de Shapiro–Wilk.

8.2 Hypothèse nulle

Tous ces tests de normalité testent l'hypothèse nulle $\mathcal{H}_0 =$"la distribution est normale".

8.3 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et il est improbable d'obtenir de telles données en supposant qu'elles soient normalement distribuées.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Rien ne s'oppose au fait que la série soit normale (Pour autant, rien ne l'assure non plus !).

8.4 Exemples

Simulons des observations normales et appliquons le test de Shapiro–Wilk.

```
> x=rnorm(50,11,2)
> x
 [1] 10.214512  9.081032  6.620966  9.362181  8.986913 11.673092 13.729038
 [8]  9.677063 11.311007 12.167166 15.143072 10.128579  9.741033  9.510930
[15] 11.741705 11.539940  9.512374 13.226112 11.564665 13.886239  7.412722
[22] 11.397585 11.934305 12.439496  7.633424 11.339367 11.585285  5.934555
[29]  9.705902 10.930072  7.706895 11.698624 13.606455 12.948415 12.573321
[36] 10.926673  7.010748  9.134820 10.628741 10.815647  8.714866  7.589055
[43] 12.306357  9.384695 10.401019 13.733044 11.899843  9.970912 14.845173
[50] 14.696913
> shapiro.test(x)

        Shapiro-Wilk normality test

data:  x
W = 0.98553, p-value = 0.7942
```

La $p-$value est supérieure à 5%, on ne rejette pas l'hypothèse nulle. On peut supposer que les observations sont normales.

Simulons des observations uniformes et appliquons le test de Shapiro–Wilk.

```
> y=runif(50,-5,15)
> y
 [1] -4.0879249 -3.1266726 -0.7170668 -2.2280176 -1.7824772 14.2489004
 [7]  2.1059444 -3.3355710 -2.7841404  1.3467159  1.0099143  7.3365456
[13]  7.8388273 -0.3247354  1.2960432  8.6903540 10.2471285  1.0189554
[19]  3.9594622  7.2452501 12.5912116  3.7068249  1.8631227  8.8284095
[25]  0.7586813  0.9331361 12.0389597 -2.1943549  1.0193574  9.9336526
[31] -3.5596714  8.6926976  9.7736992  1.4294532 -4.8870104 11.9398889
[37] 10.3700827  3.0211665  0.5607261 -2.7961990 -2.4244256  3.3730912
[43]  0.4950279  9.9933305  3.8206514  1.2576399  7.2472824 -1.5863573
[49]  3.6153996  4.0258651
> shapiro.test(y)

        Shapiro-Wilk normality test

data:  y
W = 0.94087, p-value = 0.01459
```

La $p-$value est inférieure à 5%, on rejette l'hypothèse nulle. On peut affirmer que les observations ne sont pas normales.

9 Tests sur régression linéaire

La question qu'on se pose est de savoir si la variable réponse est expliquée par les variables explicatives dans leur globalité, ou par telle ou telle variable explicative. Cela se traduit, mathématiquement, par la

non nullité des coefficients de la régression. En effet, si le coefficient d'une des variable explicative est nulle ou presque nul, cette variable explicative fait peu varier la régression linéaire, elle n'influence donc pas la variable réponse. Plaçons-nous dans le cadre de la régression linéaire multiple.

9.1 Les tests t de Student

Les tests de Student testent la nullité de chaque coefficient de la régression linéaire. Ainsi, on saura quelles variables explicatives ont un effet sur la variable expliquée.

9.1.1 Hypothèse nulle

L'hypothèse nulle de chaque test est $\mathcal{H}_0$ ="La variable x_i n'a pas d'effet sur la variable réponse"="$\beta_i = 0$".

9.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que la variable x_i a un effet sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. La variable x_i n'a pas d'effet sur la variable réponse.

9.3 Les tests F de Fisher - ANOVA

Le test de Fisher teste l'effet de l'ensemble des variables explicatives sur la variable réponse. Ainsi, on saura si la variable réponse est expliquée par les variables explicatives. On appelle cela une ANalyse de la (Of) VAriance.

9.3.1 Hypothèse nulle

L'hypothèse nulle du test est $\mathcal{H}_0$ ="Les variables x_i n'ont pas d'effet, dans leur globalité, sur la variable réponse"="la variance de l'erreur est très forte face à la variance expliquée par le modèle".

9.3.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que les variable x_i ont un effet global sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Les variables x_i n'ont pas d'effet sur la variable réponse.

9.4 Exemple : le cas de la régression linéaire simple

Reprenons l'exemple sur la régression de l'âge de l'arbre sur le diamètre. On avait le résumé de la régression suivant.

	Somme des carrés	ddl	Carré moyen	F	Signification
Régression	2905,549	1	2905,55	93,44	,000
Résidu	777,414	25	31,097		
Total	3682,963	26			

	Coefficients non standardisés		Coefficients standardisés		
	B	Erreur standard	Bêta	t	Signification
(constante)	-,974	2,604		-,374	,711
DIAMETRE	2,206	,228	,888	9,67	,000

Très logiquement, puisqu'il n'y a qu'une variable explicative, l'effet de toutes les variables explicatives est équivalent à l'effet de chaque (l'unique) variable explicative. Par conséquent, il n'est pas surprenant d'observer que $\sqrt{F} = t_{diametre}$. C'est à dire que, dans le cas de la régression linéaire simple, le test de Student et de Fisher sont totalement équivalents.

Dans notre exemple, nous avons, tant pour F que pour t, des $p-$values inférieures à 5%. Nous en concluons que le diamètre explique significativement l'âge des arbres.

9.5 Exemple : le cas de la régression linéaire multiple

On cherche à modéliser la relation entre poids des bébés à naissance et l'âge, le poids et le statut tabagique de la mère durant la grossesse. (Exemple fictif) On pose

- y = poids de naissance en grammes (bwt),
- x_1 = âge de la mère (age),
- x_2 = poids de la mère en kilos (weight),
- x_3 = statut tabagique de la mère pendant la grossesse (smoke) codé par un score à une échelle de 1 à 20.

```
> modele=lm(bwt~age+weight+smoke)
> summary(modele)

Call:
lm(formula = bwt ~ age + weight + smoke)

Residuals:
    Min      1Q  Median      3Q     Max
-385.81  -65.83   -0.70   68.17  290.66

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 3005.65519   22.99506  130.71   <2e-16 ***
age            0.02645    0.53148    0.05     0.96
weight         8.44845    0.30499   27.70   <2e-16 ***
smoke        -26.53764    1.82009  -14.58   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 97.58 on 996 degrees of freedom
Multiple R-squared:  0.4919,    Adjusted R-squared:  0.4904
F-statistic: 321.4 on 3 and 996 DF,  p-value: < 2.2e-16
```

En regardant la $p-$value du test F de Fisher, puisqu'elle est inférieure à 5%, on en déduit que les variables $(age, weight, smoke)$ ont globalement un effet sur la variable bwt. Dans le détail, en observant les $p-$values des test t de Student, on remarque que l'age n'a pas d'effet significatif sur bwt ($p \geq 0.05$) mais que les variables $weight$ et $smoke$ ont un effet significatif sur bwt ($p \sim 0$).

10 ANOVA à un facteur

Nous avons utilisé l'ANOVA dans l'étude du modèle linéaire

$$Y = aX + b + \epsilon.$$

Nous avions utilisé les test de Student et de Fisher afin de vérifier la non nullité de a, ce qui entrainerait l'absence d'effet de X sur Y, par l'étude des moyennes ou des variances.

La variable explicative X était alors quantitative. Il n'est cependant pas rare de rencontrer une variable explicative qualitative ξ. Le passage par une régression linéaire n'a plus de sens dès que la multiplication $a\xi$ n'en a plus. Prenons par exemple la variable ξ à deux modalités :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",

où plus :

- ξ_1 : "placébo",
- ξ_2 : "traitement expérimental",
- ξ_3 : "traitement expérimental à forte dose".

La variable ξ s'appelle le facteur. On pourra chercher à expliquer une variable réponse X, par exemple le taux d'une hormone. Pour chaque valeur ξ_i, on obtient un échantillon indépendant X_i. Dans le premier cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$

,

ou dans le deuxième cas,

$\xi = \xi_1$: "placébo"	$X_{1,1}, X_{1,2}, \cdots X_{1,n_1}$
$\xi = \xi_2$: "traitement expérimental"	$X_{2,1}, X_{2,2}, \cdots X_{2,n_2}$
$\xi = \xi_3$: "traitement expérimental à forte dose"	$X_{3,1}, X_{3,2}, \cdots X_{3,n_3}$

.

On considère le modèle

$$X_i = EX_i + \epsilon_i.$$

La question est de savoir si les $\mu_i = EX_i$ sont identiques (ξ n'as pas d'effet sur X) ou différents selon les valeurs ξ_i. Dans ce cas, ξ influence X.

10.1 Hypothèse nulle

L'hypothèse nulle du test d'ANOVA est $\mathcal{H}_0$ ="Le facteur ξ n'a pas d'effet sur la variable réponse"="la variance de l'erreur est très forte face à la variance expliquée par le modèle".

10.2 Décision

- Si la $p-$value est inférieure au un niveau α choisi (en général 0.05), alors on rejette l'hypothèse nulle et on considère que le facteur ξ a un effet sur la variable réponse.
- Si la $p-$value est supérieure au niveau α choisi (en général 0.05), alors on ne doit pas rejeter l'hypothèse nulle. Le facteur ξ n'a pas d'effet sur la variable réponse.

10.3 Exemple

On peut utiliser trois routes pour rentrer du travail. on a relevé les durées de trajet sur chacune des routes.

$\xi = \xi_1$: "route 1"	12 13 15 11 14 16
$\xi = \xi_2$: "route 2"	8 9 17 17 9 10
$\xi = \xi_3$: "route 3"	10 11 12 11 13 9

.

On demande la table d'ANOVA.

```
> anova(lm(temps~xi))
Analysis of Variance Table

Response: temps
          Df  Sum Sq Mean Sq F value Pr(>F)
xi         2  20.111 10.0556  1.3135 0.2981
Residuals 15 114.833  7.6556
```

On observe que la $p-$value est supérieure à 5%. On ne rejette donc pas le non effet du facteur. On en conclut qu'il n'y a pas de route à privilégier de manière significative, vu l'étude qui a été faite.