0.1 Introduction

Lorsque les observations portent sur deux caractères, et lorsqu'elles sont nombreuses pour qu'on les cite une à une, on les présente sous la forme d'un tableau à double entrée. On définit alors la distribution conjointe, les distributions marginales et les distributions conditionnelles. L'étude de la distribution de deux variables se poursuit par celle de leur liaison.

L'étude de la liaison entre les variables observées, appelée : "étude des corrélations", dépend de leur nature. Dans l'ètude des corrèlation, il y a trois cas d'ètude possible : deux variables quantitatives, une variable quantitative et une variable qualitative, deux variables qualitatives.

Lorsque le domaine de variation d'une variable quantitative a été découpé en classes et que les observations sont présentées dans un tableau à double entrée, alors cette variable peut être traitée comme une variable qualitative et dans ce cas, on a plusieurs méthodes pour l'étude de la liaison.

Dans cette section, on s'intéresse a l'étude simultanée de deux variables X et Y, étudiées sur le même échantillon, toujours noté Ω. L'objectif essentiel des méthodes présentées est de mettre en évidence une éventuelle variation simultanée des deux variables, que nous appellerons alors liaison. Dans certains cas, cette liaison peut être considérée a priori comme causale, une variable X expliquant l'autre Y ; dans d'autres, ce n'est pas le cas, et les deux variables jouent des rôles symétriques. Dans la pratique, il conviendra de bien différencier les deux situations.

Les données sont présentées de la façon suivante : on dispose de deux séries **X** et **Y** représentant l'observation des variables X et Y sur les mêmes n individus : on a une série bidimensionnelle (X,Y)de taille n :

Individu	X	Y
1	x_1	y_1
2	x_2	y_2
...	...	...
i	x_i	y_i
...	...	...
n	x_n	y_n

où x_i est la valeur de X et y_i celle de Y pour l'individu n° i de la série.

0.1.1 Deux variables quantitatives

Pour étudier la liaison entre deux variables quantitatives (discrètes), on commence par faire un graphique du type nuage de points (scatterplot). La forme générale de ce graphique indique s'll existe ou non une liaison entre les deux variables. Il s'agit d'un graphique très commode pour représenter les observations simultanées de deux variables quantitatives.

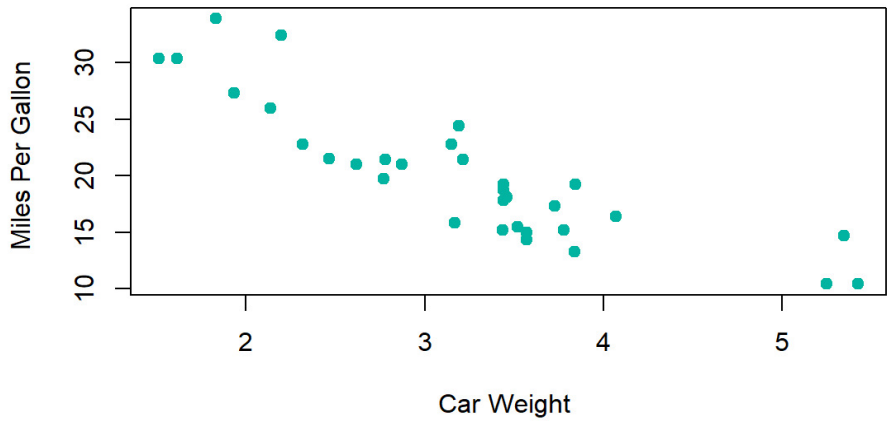

Figure 1 : Exemple de nuage de points représentant la consommation de carburant en fonction du poids de voitures. L'exemple extrait du dataset mtcars dans R.

Le choix des échelles a retenir pour réaliser un nuage de points peut s'avérer délicat. D'une façon générale, on distinguera le cas de variables homogènes (représentant la même grandeur et exprimées dans la même unité) de celui des variables hétérogenes. Dans le premier cas, on choisira la même échelle sur les deux axes (qui seront donc orthonormés) ; dans le second cas, il est recommandé soit de représenter les variables centrées et réduites sur des axes orthonormés, soit de choisir des échelles telles que ce soit sensiblement ces variables là que l'on représente (c'est en genéral cette seconde solution qu'utilisent, de façon automatique, les logiciels statistiques).

Coefficient de corrélation linéaire

On appelle coefficient de corrélation linéaire de Xet de Y la valeur définie par

$$\rho = \rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Ou $Cov(X,Y)$ est la covariance de X et de Y la valeur si elle existe

$$Cov(X,Y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$$

On peut montrer que $-1 \leq \rho(X,Y) \leq 1$.

Interprétation de ρ

- Le coefficient de corrélation est une mesure du degré de linéarité entre X et Y.
- Les valeurs de ρ proches de 1ou -1 indiquent une linéarité quasiment rigoureuse entre X et Y.
- Les valeurs de ρ proche de 0 indiquent une absence de toute relation linéaire.
- Lorsque $\rho(X,Y)$ est positif, Y a tendance à augmenter si X en fait autant.
- Lorsque $\rho(X,Y) \leq 0$,Y a tendance à diminuer si X augmente.
- Si $\rho(X,Y) = 0$, on dit que ces deux statistiques sont non corrélées.

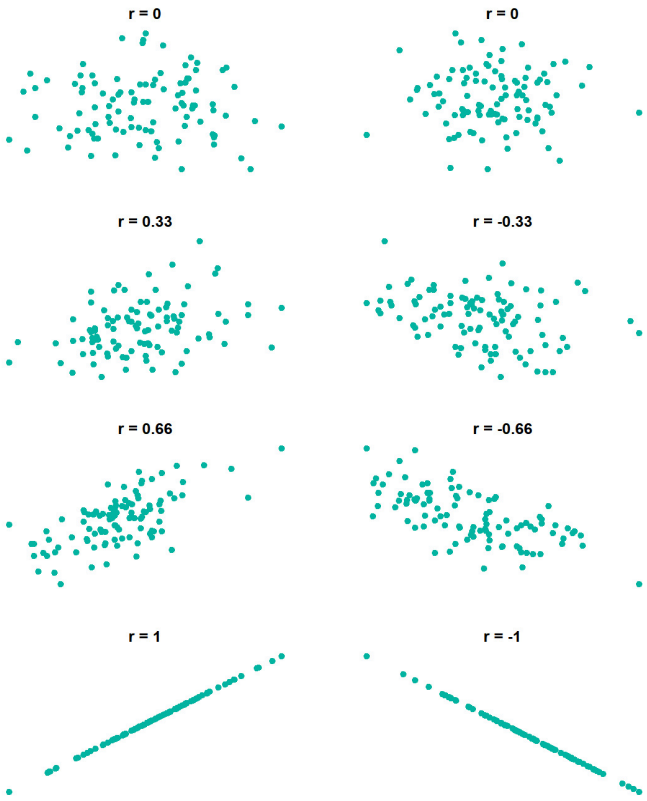

Figure 2 : Illustration de l'effet de la variation of the effect of varying the strength and direction of a correlation

La corrélation mesure l'association, pas la causalité. Dans certains cas, une liaison peut être considérée a priori comme causale, une variable expliquant l'autre. Dans d'autres, ce n'est pas le cas et les deux variables jouent alors des rôles symétriques.

Sur ce lien on trouve des exemples de variables associées linéairement mais non liées de manière causale.

D'autre part, une corrélation nulle ne signifie pas que deux variables sont indépendantes. Voici quelques exemples :

Figure 3 : Exemples où la corrélation est presque nulle mais les variables ne sont pas indépendantes.

Une variable quantitative et une qualitative

On dispose d'une variable qualitative X à p modalités $m_1, ..., m_p$ et une variable quantitative Y. On a alors p sous-populations déterminées par les p modalités de X.

L'étude de la liaison entre X et Y consiste en l'étude des différences entre ces sous-populations : il y aura absence de lien si on ne distingue pas de différence notoire dans les caractéristiques de ces différentes sous-populations.

Une façon commode de représenter les données dans le cas de l'étude simultanée d'une variable quantitative et d'une variable qualitative consiste à réaliser des boîtes à moustaches parallèles. Les boîtes à moustaches permettent de comparer facilement des groupes d'individus, par exemple ici les garçons et les filles parmi 237 étudiants :

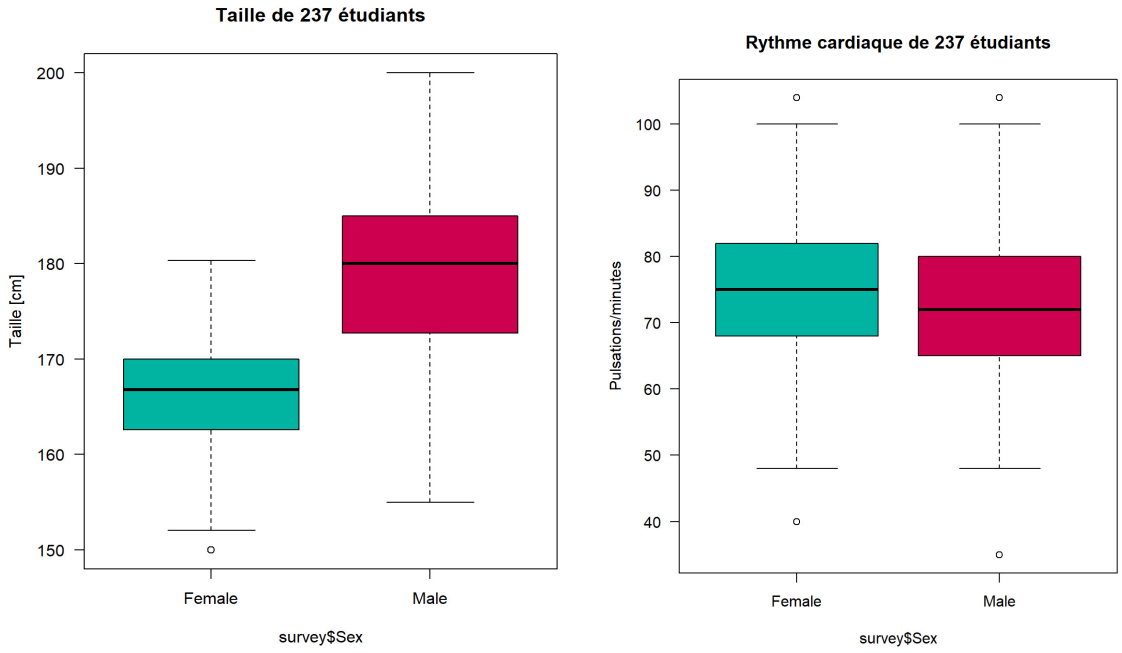

Figure 4 : Boîtes à moustaches par sexe

Deux variables qualitatives

On considère dans ce paragraphe deux variables qualitatives observées simultanément sur n individus. On suppose que la première, notée X, possède r modalités notées $x_1, ..., x_l, ..., x_r$,et que la seconde, notée Y, possède c modalités notées $y_1, ..., y_h, ..., y_c$.

Ces données sont présentées dans un tableau à double entrée, appelé tables de contingence, dans lequel on dispose les modalités de X en lignes et celles de Y en colonnes. Ce tableau est donc de dimension $r \times c$ et a pour élément générique le nombre n_{lh} d'observations conjointes des modalités x_l de X et y_h de Y ; les quantités n_{lh} sont appelées les effectifs conjoints.

Une table de contingence se présente donc sous la forme suivante :

	y_1	...	y_h	...	y_c	sommes
x_1	n_{11}		n_{1h}		n_{1c}	n_{11}
...						
x_l	n_{l1}		n_{lh}		n_{lc}	$n_{l.}$
...						
x_r	n_{r1}		n_{rh}		n_{rc}	$n_{r.}$
sommes	$n_{.1}$		$n_{.h}$		$n_{.c}$	n

On n'oubliera pas les différents modes d'études de la liaison de deux variables selon leur nature

Nature des variables et présentation des données
1- X et Y quantitatives :n couples (x_i, y_i), ou tableau de contingence

Étude de la liaison entre deux variables X et Y

1- Calcul du coefficient de corrélation linéaire :

$r = \frac{cov(x,y)}{s_x s_y}$ avec $-1 \leq r \leq 1$

Calcul et représentation graphique des deux droites des moindres carrés :

$y - \overline{y} = r\frac{s_y}{s_x}(x - \overline{x})$ $\quad$ $y - \overline{y} = \frac{1}{r}\frac{s_y}{s_x}(x - \overline{x})$

Elles se coupent au point moyen $(\overline{x} - \overline{y})$

Nature des variables et présentation des données
2- Y quantitative et X qualitative à k modalités (ou quantitative avec k classes de valeurs) Pour chaque modalité x_i de X, on dispose de : $n_{i.}$ = nbre de valeurs de Y associées à $\{X = x_i\}$ moyenne conditionnelle $\overline{y_i}$ pour $\{X = x_i\}$

Étude de la liaison entre deux variables X et Y

2- Calcul du rapport de corrélation de Y en x :

Si X est une variable quantitative classée, graphique de la courbe de régression de Y en x qui joint les points $(x_i, \overline{y_i})$

Nature des variables et présentation des données
3- X et Y quantitatives classées : tableau de contingence

Étude de la liaison entre deux variables X et Y
3- Calcul des rapports de corrélation de Y en x et de X en y : $\eta^2_{Y/X}$ et $\eta^2_{X/Y}$ Graphiques de la courbe de régression de Y en x qui joint les points $(x_i, \overline{y_i})$ les valeurs x_i étant ordonnées, et de la courbe de régression de X en y qui joint les points $\left(\overline{x_j}, y_j\right)$ les valeurs y_j étant ordonnées.

Nature des variables et présentation des données
4- X qualitative, Y qualitative : tableau de contingence

Étude de la liaison entre deux variables X et Y
4- Calcul du khi-deux : $\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - n^*_{ij}\right)^2}{n^*_{ij}} = n \sum_{i,j} \frac{\left(f_{ij} - f^*_{ij}\right)^2}{f^*_{ij}}$

Pour plus de détails sur la statistique bidimentionnelle voir le livre : Introduction à la méthode statistique manuel et exercices corrigés (https ://www.pdfdrive.com/introduction-a-la-methode-statistique-manuel-et-exercices-corriges-e186575810.html)