Chapitre 3

Liaison entre deux variables

3.1 Liaison linéaire entre deux variables quantitatives

On considère un couple de variables (X, Y). On dispose d'observations de ce couple de variables sur un échantillon de taille n : pour chaque individu on connaît le couple d'observations (x_i, y_i).

3.1.1 Covariance

Définition 1. *On définit la* ***covariance*** *de X et Y par :*

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x})(y_i - \bar{y})\right].$$

L'unité dans laquelle est exprimée la covariance est le produit des unités de X et de Y.

Remarque 1. *Lien avec la variance : $Cov(X,X) = Var(X)$.*

Remarque 2. *Formule pratique :*

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

Exemple : on s'intéresse à la liaison entre la taille T et la pointure P dans une population de 20 individus. On connaît l'ensemble des observations du couple : $\{(t_i, p_i), 1 \leq i \leq 20\}$. A partir de ces observations, on a calculé les quantités suivantes :
$\sum_{i=1}^{20} t_i = 34.91$, $\sum_{i=1}^{20} p_i = 832$, $\sum_{i=1}^{20} t_i^2 = 61.10$, $\sum_{i=1}^{20} p_i^2 = 34774$, $\sum_{i=1}^{20} t_i p_i = 1454.91$.
Caculer la covariance entre la taille et la pointure.

Propriété 1. *Changement d'échelle : soient a, b, c, d des constantes réelles. On a*

$$Cov(aX+b, cY+d) = acCov(X,Y).$$

Proposition 1. *Expression de la variance d'une somme de variables :*

$$Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y).$$

Proposition 2. *Inégalité de Cauchy-Schwarz :*

$$|Cov(X,Y)| \leq \sigma_X \sigma_Y.$$

Exemple : Dans l'exemple précédent du couple (T,P), calculer les écart-types des deux variables et vérifier l'inégalité.

Preuve de la proposition 2 : Pour tout réel a, on peut développer grâce à la Proposition 1 la quantité $Var(X+aY) \geq 0$:

$$\begin{aligned} Var(X+aY) &= Var(X) + Var(aY) + 2Cov(X,aY) \\ &= VarX + a^2 Var(Y) + 2aCov(X,Y) \text{ par la Propriété 1} \qquad (3.1) \\ &\geq 0. \end{aligned}$$

Le polynôme du second degré en a étant de signe constant, son discriminant est négatif ou nul :

$$4\left(Cov(X,Y)\right)^2 - 4Var(X)Var(Y) \leq 0,$$

d'où l'inégalité recherchée.

Remarquons au passage que le cas d'égalité se produit lorsque le discriminant de l'équation (3.1) est nul. Dans ce cas, l'équation admet une racine double :

$$\begin{aligned} a &= -\frac{2Cov(X,Y)}{2Var(Y)} = -\frac{Cov(X,Y)}{Var(Y)}. \\ &= \begin{cases} -\frac{\sigma_X}{\sigma_Y} & \text{si } Cov(X,Y) = +\sigma_X\sigma_Y \\ \frac{\sigma_X}{\sigma_Y} & \text{si } Cov(X,Y) = -\sigma_X\sigma_Y \end{cases} \end{aligned}$$

Dans le premier cas, cela signifie que $X - \frac{\sigma_X}{\sigma_Y}Y$ a une variance nulle, donc est une constante, d'où

$$X = \frac{\sigma_X}{\sigma_Y}Y + constante.$$

Dans le second cas,

$$X = -\frac{\sigma_X}{\sigma_Y}Y + constante.$$

Ces deux cas sont les seuls cas d'égalité dans la Proposition 2. Ils correspondent au fait que les variables X et Y s'obtiennent l'une à partir de l'autre par une application affine.

3.1.2 Coefficient de corrélation

Définition 2. *Le coefficient de corrélation $r(X,Y)$ est défini par :*

$$r(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

C'est un coefficient sans unité. Sa valeur absolue est invariante par translation et changement d'échelle des variables : pour toutes constantes réelles $a \neq 0$, b, $c \neq 0$ et d,

$$r(aX+b, cY+d) = \frac{ac}{|ac|} r(X,Y).$$

Exemple : reprendre l'exemple précédent du couple (T,P) et calculer le coefficient de corrélation entre les deux variables.

Propriété 2. *Il découle de la Proposition 2 que*

$$-1 \leq r(X,Y) \leq 1.$$

De plus, les cas d'égalité sont les suivants :

- $r(X,Y) = 1$ *si et seulement si les deux variables satisfont une relation affine du type* $Y = aX + b$ *avec* $a > 0$.
- $r(X,Y) = -1$ *si et seulement si les deux variables satisfont une relation affine du type* $Y = aX + b$ *avec* $a < 0$.

Lorsque le nuage des points (x_i, y_i) est exactement situé sur une droite (cas idéal), on est dans la situation où $r(X,Y) = \pm 1$. Lorsque $r(X,Y)$ est proche de ± 1 (pour fixer les idées : $|r(X,Y)| \geq 0.8$), alors il y a une liaison linéaire importante entre X et Y. Lorsqu'au contraire $r(X,Y)$ est proche de 0, alors il n'existe pas de relation linéaire entre X et Y. Attention, il peut y avoir quand même un autre type de liaison entre X et Y.

3.1.3 Régression linéaire

On suppose à présent que les observations du couple de variables (X,Y) satisfont une relation de la forme suivante :

$$y_i = ax_i + b + \epsilon_i, \ i = 1, \ldots, n, \tag{3.2}$$

où a et b sont des coefficients réels. Il s'agit ici de la régression **de** Y **sur** X. Le terme ϵ_i désigne un *bruit*, c'est à dire une perturbation supposée *petite*. Dans ce cours, on ne cherchera pas à donner un sens précis à la mesure de ce bruit.

Disposant des observations $(x_i, y_i)_{i=1}^n$ du couple (X,Y), on cherche à trouver les coefficients a et b qui permettent le mieux d'ajuster les données à une relation du type (3.2), au sens du critère des moindres carrés. On cherche

$$\min_{a,b} \sum_{i=1}^{n} (y_i - b - ax_i)^2. \tag{3.3}$$

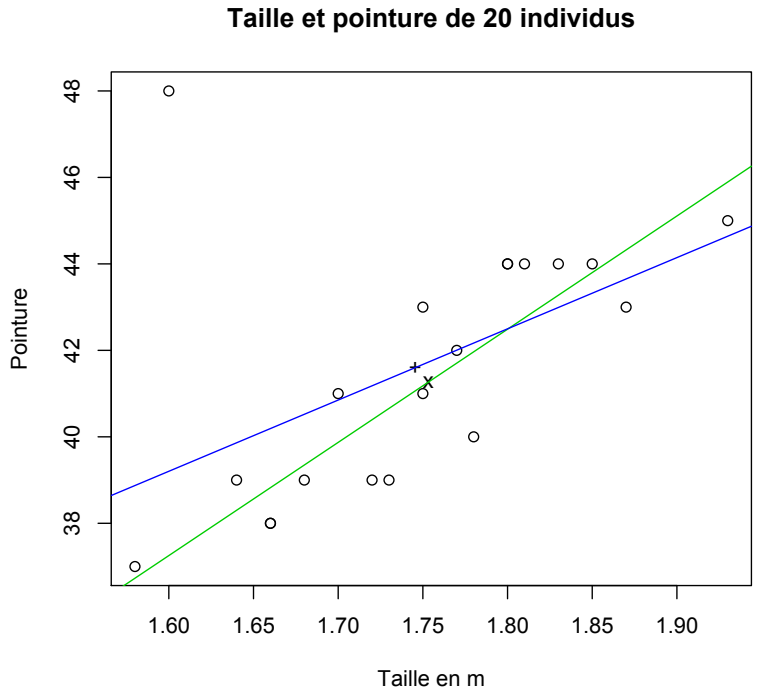
La solution, qui s'obtient en annulant les dérivées partielles de la fonction de (a, b) qui est minimisée dans (3.3), est :

$$\begin{aligned} \hat{a} &= \frac{Cov(X,Y)}{Var(X)}, \\ \hat{b} &= \bar{y} - \hat{a}\bar{x}, \end{aligned}$$

où $\bar{x}$ et $\bar{y}$ désignent les moyennes respectives de X et Y. La droite des moindres carrés est la droite d'équation : $y = \hat{a}x + \hat{b}$. On peut remarquer qu'elle passe toujours par le barycentre $(\bar{x}, \bar{y})$ du nuage de points. Sa pente peut aussi s'écrire à l'aide du coefficient de corrélation : $\hat{a} = r(X,Y)\frac{\sigma_Y}{\sigma_X}$.

Exemple : droite des moindres carrés de régression de la pointure sur la taille.
Reprendre les données relatives au couple (T, P) des exemples précédents et calculer l'équation de la droite des moindres carrés de régression de la pointure sur la taille.

Le graphique suivant représente le nuage de points du couple (T, P). Deux droites ont été ajoutées sur le graphique. Laquelle correspond à la droite de régression de la pointure sur la taille ? On verra en cours comment a été obtenue la deuxième droite.



Prédiction.
Pour une valeur x_0 de la variable X qui ne fait pas partie des observations, on peut faire une prédiction de la valeur correspondante de Y en calculant l'ordonnée du point d'abscisse x_0 sur la droite des moindres carrés :

$$y_0 = \hat{a}x_0 + \hat{b}.$$

Exemple : prédire la pointure d'un individu qui mesure 1.71 m et retrouver la valeur obtenue par une méthode graphique.

3.1.4 Régression linéaire après transformation d'une variable

On suppose que les observations $(x_i, y_i)_{i=1}^n$ satisfont une relation du type

$$y_i = af(x_i) + b + \epsilon_i,$$

pour une certaine fonction f donnée et des bruits ϵ_i. On peut estimer les coefficients de la droite de régression de Y sur $f(X)$ par la méthode décrite auparavant. Des exemples seront vus en TD.

3.2 Liaison entre deux variables qualitatives

On observe une série statistique $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ composée de n couples d'observations d'un couple de variables qualitatives (X, Y). On suppose que X a I modalités notées $C_1, \ldots, C_I$ et Y a J modalités notées $D_1, \ldots, D_J$. Pour $1 \leq i \leq I$ et $1 \leq j \leq J$, on note n_{ij} l'effectif des couples d'observations égaux à (C_i, D_j).

3.2.1 Table de contingence

Dans la table de contingence, on regroupe les effectifs n_{ij}. On peut compléter la table de contingence en ajoutant les totaux en lignes et en colonnes.
On note $n_{i.} = n_{i1} + \cdots + n_{iJ} = \sum_{j=1}^{J} n_{ij}$ le total sur la ligne i de la table de contingence, $n_{.j} = n_{1j} + \cdots + n_{IJ} = \sum_{i=1}^{I} n_{ij}$ le total sur la colonne j de la table de contingence.

Y / X	D_1	D_2	$\ldots$	D_J	Total
C_1	n_{11}	n_{12}	$\ldots$	n_{1J}	$n_{1.}$
C_2	n_{21}	n_{22}	$\ldots$	n_{2J}	$n_{2.}$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
C_I	n_{I1}	n_{I2}	$\ldots$	n_{IJ}	$n_{I.}$
$Total$	$n_{.1}$	$n_{.2}$	$\ldots$	$n_{.J}$	n

Exemple : Les données suivantes reprennent l'exemple donné dans le polycopié d'Yves Tillé (Université de Neuchâtel). Voir également les codes **R** proposés dans ce cours. On a relevé la couleur des yeux et le sexe de 200 individus. Les données sont rapportées dans la table de contingence suivante :

Couleur yeux / Sexe	Bleu	Vert	Marron	Total
Homme	10	50	20	80
Femme	20	60	40	120
Total	30	110	60	200

Pour simplifier, on notera dans la suite, si besoin, X la variable « sexe » et Y la variable « couleur des yeux ».

3.2.2 Distribution marginale

La distribution marginale de la variable X est la donnée des **effectifs marginaux** $n_{1.}, \ldots, n_{I.}$. C'est la distribution de la variable X. On peut la présenter dans un tableau et calculer les fréquences marginales ($f_{i.} = n_{i.}/n$), qui sont les proportions associées à chaque modalité de la variable X. On peut calculer de même la distribution marginale de la variable Y.

Distribution marginale de X :

X	C_1	$\ldots$	C_I	Total
Effectif	$n_{1.}$	$\ldots$	$n_{I.}$	n
Proportion	$f_{1.} = n_{1.}/n$	$\ldots$	$f_{I.} = n_{I.}/n$	1

Distribution marginale de Y :

Y	D_1	...	D_I	Total
Effectif	$n_{.1}$	...	$n_{.J}$	n
Proportion	$f_{.1} = n_{.1}/n$	...	$f_{.J} = n_{.J}/n$	1

Exemple : Couleur des yeux des 200 individus.

3.2.3 Distribution conditionnelle

a) Profils-lignes

La distribution conditionnelle de Y sachant la modalité C_i de X est la distribution dont les proportions sont données dans le tableau suivant :

$Y_{\|X=C_i}$	D_1	...	D_I	Total
Proportion	$n_{i1}/n_{i.}$	...	$n_{iJ}/n_{i.}$	1

Une telle distribution est appelée **profil-ligne**. L'ensemble des profils-lignes peut être présenté dans un tableau :

$Y_{\|X}$ / X	D_1	D_2	...	D_J	Total
C_1	$n_{11}/n_{1.}$	$n_{12}/n_{1.}$	...	$n_{1J}/n_{1.}$	1
C_2	$n_{21}/n_{2.}$	$n_{22}/n_{2.}$	...	$n_{2J}/n_{2.}$	1
...	...	...	...	...	...
C_I	$n_{I1}/n_{I.}$	$n_{I2}/n_{I.}$	...	$n_{IJ}/n_{I.}$	1

Exemple : Distribution conditionnelle de la variable « Couleur des yeux » sachant la modalité « femme » ou : distribution de la couleur des yeux des femmes.

b) Profils-colonnes

De même, l'ensemble des distributions conditionnelles de X sachant les modalités de Y est l'ensemble des profils-colonnes, que l'on peut présenter dans le tableau suivant :

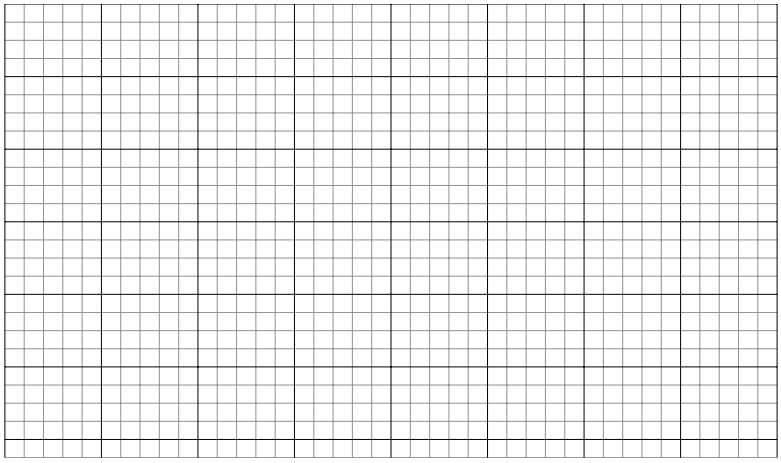
Y / $X_{\|Y}$	D_1	D_2	...	D_J
C_1	$n_{11}/n_{.1}$	$n_{12}/n_{.2}$	...	$n_{1J}/n_{.J}$
C_2	$n_{21}/n_{.1}$	$n_{22}/n_{.2}$	...	$n_{2J}/n_{.J}$
...	...	...	...	...
C_I	$n_{I1}/n_{.1}$	$n_{I2}/n_{.2}$	...	$n_{IJ}/n_{.J}$
$Total$	1	1	...	1

Exemple : Ensemble des profils-colonnes du couple de variables « sexe » et « couleur des yeux ».

3.2.4 Représentation graphique

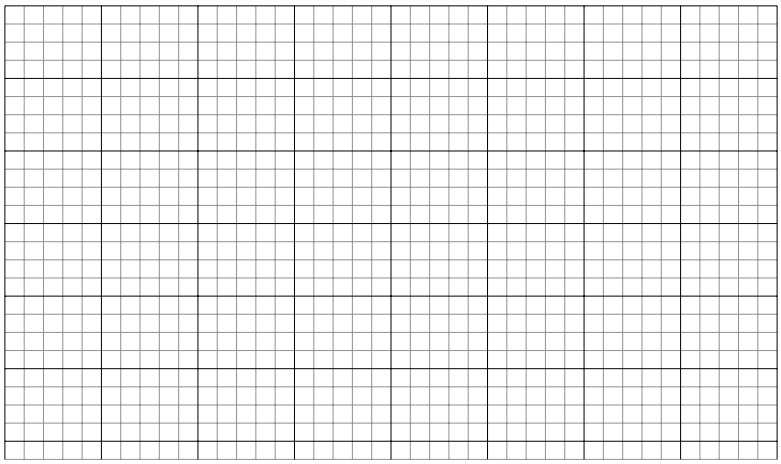
a) Distribution jointe

Exemple : Diagramme en barres de la distribution jointe des variables « Sexe » et « Couleur des yeux ».



b) Distribution conditionnelle

Exemple : Diagramme en barres de la distribution de la variable « Couleur des yeux » sachant la variable « Sexe ».



3.2.5 Mesure de la liaison entre deux variables qualitatives

a) Comparaison qualitative des profils-lignes ou des profils-colonnes

Il y a indépendance stricte entre X et Y lorsque tous les profils-lignes sont identiques. Ils sont dans ce cas tous identiques à la distribution marginale de Y.

De la même manière, l'indépendance a lieu lorsque tous les profils-colonnes sont égaux à la distribution marginale de X.

Ceci implique : pour tous i, j,

$$n_{ij} = \frac{n_{i.} n_{.j}}{n}. \tag{3.4}$$

Réciproquement, si (3.4) a lieu, alors il y a indépendance stricte entre X et Y.

Preuve :

b) La distance du $\chi 2$ pour mesurer l'écart à l'indépendance

Dans la pratique, cette indépendance stricte ne s'observe jamais sur un échantillon. On peut être plus ou moins éloigné de cette situation parfaite. La distance du χ^2 d'écart à l'indépendance permet de mesurer le degré de dépendance entre X et Y. Elle se base sur la comparaison entre n_{ij} et $\frac{n_{i.} n_{.j}}{n}$.

Définition 3. *La **distance du** $\chi 2$ observée sur la série statistique $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ est définie par*

$$\chi^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}} \right).$$

Exemple : Distance du χ^2 pour mesurer l'écart à l'indépendance entre les variables « Sexe » et « Couleur des yeux ».

Propriétés :

- la grandeur $\chi^2 = 0$ si et seulement si il y a indépendance stricte entre X et Y.
- la grandeur χ^2 est d'autant plus élevée que la liaison est forte : il existe alors des cellules (i, j) avec un écart important $n_{ij} - \frac{n_{i.}n_{.j}}{n}$.
- l'inégalité suivante est toujours vérifiée :

$$\frac{\chi^2}{n} \leq \min\{I - 1, J - 1\}.$$

Définition 4. *On appelle **contribution au** χ^2 du couple de modalités (C_i, D_j) de (X, Y) la quantité* $\frac{\left(n_{ij} - \frac{n_{i.}n_{.j}}{n}\right)^2}{\frac{n_{i.}n_{.j}}{n}}$.

Plus la contribution est forte, plus la liaison entre les modalités C_i et D_j est importante.

Définition 5. *L'association entre les modalités C_i et D_j est dite **positive** si $n_{ij} - \frac{n_{i.}n_{.j}}{n} > 0$. Elle est dite **négative** si $n_{ij} - \frac{n_{i.}n_{.j}}{n} < 0$.*

Exemple : qualifier la liaison entre la modalité « Femme » de la variable « Sexe » et la modalité « Marron » de la variable « Couleur des yeux ».

Définition 6. *Le coefficient C de Cramer est défini par :*

$$C = \sqrt{\frac{\chi^2}{n. \min\{I - 1, J - 1\}}}.$$

Propriétés :

- $0 \leq C \leq 1$
- $C = 0$ lorsqu'il y a indépendance. De petites valeurs de C signifient que la liaison entre X et Y est très faible. Des valeurs proches de 1 signifient qu'il y a une liaison forte entre X et Y.
- Ce coefficient, qui varie entre 0 et 1, permet de comparer la liaison entre plusieurs couples de variables.

Exemple : Calcul du C de Cramer pour mesurer l'écart à l'indépendance entre les variables « Sexe » et « Couleur des yeux ».

3.3 Liaison entre une variable qualitative et une variable quantitative

On observe des couples $\{(x_i, y_i), 1 \leq i \leq n\}$ d'observations du couple de variables (X, Y) avec :

- X qualitative à I modalités : $C_1, \ldots, C_I$
- Y quantitative, discrète ou continue, avec données brutes ou regroupées en classes.

Exemple : On connaît le sexe et la taille en mètres de 20 individus. Parmi ces 20 individus, il y a 10 hommes et 10 femmes. La taille moyenne chez les hommes est de 1.80 m et chez les femmes, de 1.69 m. La variance de la taille est égale à 0.0082. On connaît également la variance de la taille chez les hommes, qui est égale à 0.0055. Chez les femmes, elle est égale à 0.0047.

3.3.1 Classement des données et distributions marginales

La distribution marginale de X est la distribution associée à la série statistique $(x_1, \ldots, x_n)$ (variable qualitative). La distribution marginale de Y est la distribution associée à la série statistique $(y_1, \ldots, y_n)$ (variable quantitative). On note $\bar{y}$ la moyenne marginale de la variable Y et σ_Y^2 sa variance marginale.

On note $n_{1.}, \ldots, n_{I.}$ les effectifs marginaux de la variable X. C'est-à-dire : $n_{1.}$ est l'effectif des observations pour lesquelles X prend la modalité C_1, etc...

On peut regrouper les couples d'observations (x_i, y_i) qui comportent la même modalité x_i. Après regroupement, on obtient la nouvelle énumération :

$$\begin{aligned}
(x_{11}, y_{11}), (x_{12}, y_{12}), \ldots, (x_{1n_{1.}}, y_{1n_{1.}}) &= (C_1, y_{11}), (C_1, y_{12}), \ldots, (C_1, y_{1n_{1.}}) \\
(x_{21}, y_{21}), (x_{22}, y_{22}), \ldots, (x_{2n_{2.}}, y_{2n_{2.}}) &= (C_2, y_{21}), (C_2, y_{22}), \ldots, (C_2, y_{2n_{2.}}) \\
&\ldots \\
(x_{I1}, y_{I1}), (x_{I2}, y_{I2}), \ldots, (x_{In_{I.}}, y_{In_{I.}}) &= (C_I, y_{I1}), (C_I, y_{I2}), \ldots, (C_I, y_{In_{I.}})
\end{aligned}$$

3.3.2 Distribution conditionnnelle

Pour $1 \leq i \leq I$, la distribution conditionnelle de Y sachant la modalité C_i de X est la distribution de la série statistique $(y_{i1}, y_{i2}, \ldots, y_{in_{i.}})$.

Cette distribution possède une moyenne, encore appelée moyenne conditionnelle de Y sachant $X = C_i$, notée $\bar{y}_i$:

$$\bar{y}_i = \frac{1}{n_{i.}} \sum_{j=1}^{n_{i.}} y_{ij}.$$

De même, la variance conditionnelle σ_i^2 de Y sachant $X = C_i$ est :

$$\sigma_i^2 = \frac{1}{n_{i.}} \sum_{j=1}^{n_{i.}} (y_{ij} - \bar{y}_i)^2 = \frac{1}{n_{i.}} \sum_{j=1}^{n_{i.}} y_{ij}^2 - (\bar{y}_i)^2.$$

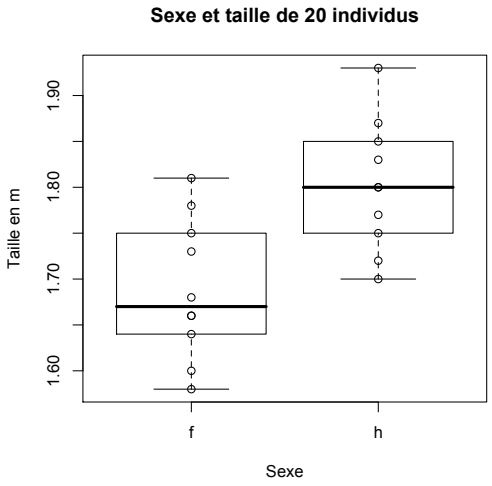
Exemple :

3.3.3 Représentations graphiques

On peut représenter, pour chaque modalité de X en abscisse, les données brutes en les plaçant sur l'axe des ordonnées.

Pour une représentation plus synthétique des données, on peut représenter les différentes boîtes à moustaches des distributions conditionnelles de Y sachant chaque modalité de X, sur un même graphique.

Exemple :



3.3.4 Rapport de corrélation

Définition 7. *On appelle* ***variance intercatégories*** *la quantité :*

$$\frac{1}{n}\sum_{i=1}^{I} n_{i.}(\bar{y}_i - \bar{y})^2.$$

Cette quantité est d'autant plus grande que Y a un comportement différent sur les différentes classes définies par les modalités de X. Inversement, elle est petite si Y a sensiblement la même moyenne sur chaque classe de modalité de X.

Définition 8. *On appelle* ***variance intracatégories*** *la quantité :*

$$\frac{1}{n}\sum_{i=1}^{I} n_{i.}\sigma_i^2.$$

Cette quantité est d'autant plus petite que les valeurs prises par Y sont homogènes à l'intérieur des différentes classes définies par les modalités de X.

Proposition 3. *La variance de Y est la somme des variances intercatégories et intracatégories :*

$$\sigma_Y^2 = \frac{1}{n}\sum_{i=1}^{I} n_{i.}(\bar{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{I} n_{i.}\sigma_i^2.$$

Preuve :

Définition 9. *On appelle* ***rapport de corrélation*** *le rapport "variance inter/variance totale" :*

$$e = \frac{\frac{1}{n}\sum_{i=1}^{I} n_{i.}(\bar{y}_i - \bar{y})^2}{\sigma_Y^2}.$$

Propriétés :
- $0 \leq e \leq 1$.
- si e est proche de 0 : alors $\bar{y}_1 \simeq \bar{y}_2 \simeq \cdots \simeq \bar{y}_I \simeq \bar{y}$. On dit qu'il y a absence de dépendance en moyenne entre X et Y.
- si e est proche de 1 : alors la variance intra est proche de 0 et il y a une grande homogénéité dans les valeurs de Y à l'intérieur de chaque classe de modalité de X. Il y a donc une forte liaison entre X et Y.

Exemple :

3.4 Cas d'une variable quantitative regroupée en classes

On peut remplacer dans les définitions précédentes les calculs de moyenne et variance de Y à partir de données brutes par les calculs habituels dans le cas de données regroupées en classes. Cela permet de définir un rapport de corrélation (qui sera, bien sûr, approché).

Par ailleurs, regrouper la variable quantitative continue Y en classes permet également d'étudier la liaison entre X et Y comme celle d'un couple de variables qualitatives.